Faculty of Engineering
Tel Aviv University

# Random Signals and Noises
# 0512.3632

Lecturer: Professor Uri Erez, 2014, First semester
Edited by Eran Avidor, January 14, 2014
Translated by Tamir Zchut

February 28, 2021

# Contents

Professor Uri Erez, uri@eng.tau.ac.il

# Introduction/Goal

Random processes are everywhere. Even in “deterministic” systems there are some variables that are not fully determined – there is some randomness in the system. We will model the unknown variables as the random part of the system.

**Example 1 – “A Little Randomness”** Consider the instantaneous voltage in an electric socket at the moment of sampling. Clearly, we will get a sample of the sine wave, but we can’t fully determine the wave because we don’t know its phase. In other words, the voltage as a function of time can be expressed as

$$V(t) = A\sin(2\pi t + \phi), \phi \sim U([0, 2\pi]),$$

where $\phi \sim U([0, 2\pi])$ means that $\phi$ is uniformly distributed over the interval $[0, 2\pi]$, which we assume to be the case as the sampling time is arbitrarily chosen so there is no “preference” for any phase.

Actually, $V(t)$ is random for every time $t$, as the phase $\phi$ is unknown. We would like to know what is the probability to get a value $V(t)$ for a given sampling time instance $t$, say $t = 0$. More precisely, we can ask what is probability density function $f_{V(0)}(v)$? In particular, we as will be developed in the problem sets, we can observe that see that the probability to get values in the vicinity of $V(0) = A$ is larger than the probability to get values in the vicinity of $V(0) = 0$.

**Example 2 – “Very Random”** Two speakers are in a room. Speaker A (child, for example) and speaker B (adult). We call the speaking signals $X_1(t)$ and $X_2(t)$, respectively. Assume that the speakers are stationary, so we can model the system (the room, reflections, etc.) as a linear system. The transfer function of each signal is $h_1(t)$, $h_2(t)$ (if the speakers were in an open space, the transfer function were time invariant). The speaking signals can be described as: $Y(t) = h_1(t)X_1(t) + h_2(t)X_2(t)$. Assume we want to listen to speaker A. both signals $X_1(t)$,$X_2(t)$ contain information: one is our signal (here it’s $X_1$) and the other is noise. Our goal is to filter the noise so we can get a better estimation of the signal $X_1(t)$. The human brain does it amazingly easily and intuitively. In order to plan such a system, we want to know some initial information about the signals. We know they are speaking signals of a child and an adult; hence the signals are spectrally different, that is the signals are in different frequency domains (the voice of a child is higher than the voice of an adult). From this initial information we can plan the needed filter (in our case, simple HPF) and estimate the information signal. In fact, the system (filter) is just a function:

$$Y(t) \rightarrow [g(t)] \rightarrow \tilde{X}(t) = g(t) \star Y(t).$$

We want to know the statistical relation between the input and the output of the system. That is, how the function alters the statistics of the input signal and its properties. In general, we want to model the signal/random process, analyze in a certain domain (for example, the frequency domain) and then build an estimator according to the needed quality. During the course we will learn about the next input types:

- Single random variable
- Random vector
- Random (stochastic) process in discrete time
- Random (stochastic) process in continuous time.

In order to get intuition about the different states we assume a random signal $X(t)$, from which we obtain all four mathematical objects:

- Single random variable will be the value of the signal at certain time: $X(t_0)$
- Random vector will be the finite set of samples of the signal: $\underline{X} = (X(t_0), X(t_1), \ldots, X(t_n))$
- Random process in discrete time corresponds to the countably “infinite vector”: $\{X_i\}_{i=1}^{\infty} = \{X(n), \forall n \in \mathbb{N}\}$ obtained from sampling the random signal $X(t)$.
- Random process in continuous time is $X(t)$ itself, where the time index is $t \in \mathbb{R}$.

# Part A: Random Variables and Operations

## Probability and random variables

### Review on probability fundamentals

**Probability Space (definition):** Probability space is defined by 3 parameters: $\Omega, F, \Pr$

- $\Omega$ – sample space: The set of all possible outcomes of the experiment. The sample space can be finite (for example, rolling a dice) or infinite (for example, choosing a rational number)
- $F$ – set of events:

  An event is a sub-set of the sample space $\Omega$. This means

  that every event is actually a “possible scenario”. The set of event contains only (and every) question that can be asked about the experiment. Beside the fact that $F$ is contained in $\Omega$, $F$ must apply the next terms: The set of events $F$ is closed under the countable union, countable intersection and complement. That must happen in order for each scenario to be defined. The objects $\Omega$, $\emptyset$ is in $F$ (that is $\Omega$,$\emptyset \in F$)
- Pr – Probability measure Probability measure is a function from the set of event to the closed interval $[0, t]$. This function defines the probability to get the event. $\Pr : F \to [0, 1]$.

The set $\Omega, F, \Pr$ must apply the terms:

1. $\forall A \in F : \Pr(A) \geq 0$
2. $\Pr(\Omega) = 1$, i.e. the probability of the certain event is 1.
3. If $\{A_i\}$ is the group of mutually exclusive pairs of events (i.e. $A_i \cap A_j = \emptyset, \forall i \neq j$) then
$$\Pr(\cup A_i) = \sum \Pr(A_i).$$

From these requirements we get the next properties:

1. $\forall A \in F : P(A) \leq 1$
   Proof:
   $1 = \Pr(\Omega) = P(A \cup \bar{A}) = \Pr(A) + \Pr(\bar{A}) \overset{1}{\to} P(A) \leq 1 \; \forall A \in F$
2. $\Pr(\emptyset) = 0$

3. $\Pr(\bar{A}) = 1 - \Pr(A)$

**Independent events (definition):** Events $A, B \in F$ are statistically independent $\iff \Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$

**Example 1:** Look at the experiment of rolling a single dice.

- Sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$
- The set of events will be all the subsets of $\Omega$, i.e. the power set of $\Omega$ $F = \{\emptyset, \{1\}, \{1, 2\}, \ldots, \{1, 2, 3, 4, 5, 6\}\}, |F| = 2^{|\Omega|} = 2^6 = 64$
- The probability measure sets a number for each of the 64 events.

In fact, it is enough to define Pr only on the elementary events 1,2,...,6, and from these values we can calculate the probability measure for each event we will need using the properties of the probability space. $\Pr(1) = \Pr(2) = \cdots = \Pr(6) = \frac{1}{6}$
And for example:
$\Pr(\textit{even result}) = \Pr(\{2, 4, 6\}) = \Pr\{\{2\} \cup \{4\} \cup \{6\}\} = \Pr\{\{2\}\} + \Pr\{\{4\}\} + \Pr\{\{6\}\} = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$
Note:
On the same sample space (and even on the same specific experiment) we can define different sets of events. for example, we can define this set of events for the experiment "even/not even": $F_2 = \{\emptyset, \{1, 3, 5\}, \{2, 4, 5\}, \{1, 2, 3, 4, 5, 6\}\}$ But in fact there is no reason to define such set of events, as I is contained in the complete set of events. So while the sample space is finite (or even countable) we would rather work with the power set of the sample space and the probability function to define on elementary events only.

**Example 2:** Uniformly picking a point in the closed interval $[0, 1]$. This is the "closest" experiment to a dice rolling with a continuous sample space.

- The sample space is $\Omega = [0, 1]$
- The set of event will be all the subgroups of $\Omega$, that is all the inner interval inside $[0, 1]$ It is important to notice that in $F$ there are non-continuous intervals. In fact, in the set of event there are all the intervals with the from $[a, b] \subseteq [0, 1]$ and union/intersection/completion of

[1] $\Pr(A), \Pr(\bar{A}) \geq 0$

every possible combination. If we want to get to a single point, we will make an infinite series of cutting so that in every step we will make the interval smaller, but still stay with an interval and not a point (in the limit we get infinitesimal interval which will represent a point with a certain “accuracy”). The reason for that is that a point is not a subset of the sampling set, hence we cant define Pr for a point (the chance to pick a point is 0, and if we define the set of events as the set of all the possible points, we would get that the probability of each event is 0).

- The probability measure Pr will be defined for basic event, in this case continuous intervals (not necessarily closed, since the probability measure of a point is 0) and $\Pr([a,b]) = \frac{b-a}{1-0}$, $[a,b] \subseteq [0,1]$

**Review on the basics of single random variable**

We would like to “get rid of” the abstract language of probability spaces and $\{\Omega, F, P\}$ and discuss issues using a more engineering-related language. This “language” is abstract because the elements of $\Omega$ are abstract (for each experiment the set is different, and the elements of the set has no mathematical meaning – for example, in coin flipping experiment $\Omega = \{Heads, Tail\}$ and in dice rolling experiment $\Omega = \{1,2,3,4,5,6\}$). Hence, in order to “get rid of” this abstract language we would like to map the elements of $\Omega$ to a uniform language – the language of real numbers.

**Random variable (definition):** Random (stochastic) variable $X$ is a mapping (function) that maps each experiment outcome $\omega \in \Omega$ a real number $X(\omega) \in \mathbb{R}$, so that $\{\omega : X(\omega) \leq x\} \in F, \forall x \in \mathbb{R}$ . that is, for each $x \in \mathbb{R}$ the set $\{\omega : X(\omega) \leq x\}$ is an event(belongs to the set of events)
Notes:

1. Random variable will always be capital letter.
2. The mapping can be arbitrary. But in every use in engineering, the mapping will be natural and reasonable.
3. The requirement above is needed in order to the define the CDF function (which will be learned later).

**Example:** For a dice rolling experiment, $\Omega = \{1,2,3,4,5,6\}$ and we can define random variable

$$X(\omega) = \begin{cases} 1, \omega = 1, 3, 5 \\ 2, \omega = 2, 4, 6 \end{cases}$$

i.e. random variables with values 1 (for odd result) and 2 (for even result).

### Cumulative Distribution Function (CDF) and Probability Density Function (PDF)

We exchanged the abstract results of the experiments with numerical values by mapping. Now we would like to develop tools to calculate the probability to get certain values of the random variable. i.e., we would like to define functions that we could use to answer every probability-related question on a certain experiment (for example $\Pr(X \in S)$)

### Cumulative Distribution Function (CDF):

The CDF is represented using the letter $F$, with the name of the random variable as an index, and an argument that represents the requested value. Its value represents the probability to get a value that is not bigger (smaller or equal) than the argument. for a random variable $X$ we define the CDF as:
$F_X(x) \triangleq \Pr(\{\omega : X(\omega) \le x\}) \overset{2}{=} \Pr(X \le x)$

### Properties of the CDF:

1. $0 \le F_X(x) \le 1$
2. $\lim\limits_{x \to \infty} F_X(x) = 1$
3. $\lim\limits_{x \to -\infty} F_X(x) = 0$
4. time continuity: $F_X(x_0) = \lim\limits_{x \to x_0} F_X(x) = F_X(x_0^+)$
5. non decreasing function: $\forall x_2 \ge x_1 : F_X(x_2) \ge F_X(x_1)$
6. $\Pr(a \le X \le b) = F_X(b) - F_X(a)$, because: $\Pr(a \le X \le b) = \Pr(-\infty \le X \le b) - \Pr(\infty \le X \le a) = F_X(b) - F_X(a)$

[2] the calculation of probability is done on events, but we will write them by using the values of the random variables to keep things simple.

Notes:

1. $\Pr(X \le a) = F_X(a)$ (from 3 and 6, or from the definition)
2. The probability of a single point is the limit $\Pr(X = a) = F_X(a^+) - F_X(a^-) \overset{3}{=} F_X(a) - F_X(a^-)$
3. property 4 is an arbitrary property which happens because we define the function as also equal, not just smaller. If we remove the equality, we would get continuity from the left.

**Examples:**

- Randomly choosing uniformly in $[0,1]$ we define the mapping $X(\omega) = \omega$ (where $\omega \in \Omega = [0,1]$) and get:
  The function is not differentiable in every point but is continuous everywhere.
- Rolling a dice. We define the "Natural" mapping, and get:

$$F_X(x) = \sum_i \Pr(X = x_i) u(x - x_i) = \frac{1}{6}\sum_{i=1}^{6} u(x - i)$$

Distinguish:

- We define random variable as discrete if $F_X(x)$ is composed of only steps (discrete "jumps").
- We define random variable as continuous if $F_X(x)$ is continuous for every $x \in \mathbb{R}$
- We define random variable as mixed if $F_X(x)$ is composed from steps and continuous parts

we can write the CDF for each random variable by the formula:
$F_X(x) = \alpha F_X^D(x) + (1-\alpha)F_X^C(x)$, $0 \le \alpha \le 1$ Where $F_X^D(x)$ is the discrete CDF and $F_X^C(x)$ the continuous CDF.
Note:
For a general (neither purely discrete nor purely continuous) random variable, points that have a non-zero probability are called atoms. These are the points with a "jump" in the CDF, and we say they have probability mass.

[3] from continuity from the right – property 4

**Probability Density Function (PDF):** The PDF is represented using the letter $f$, where the name of the random variable is the index, and the needed value is the argument. Its value represents the probability density around a specific point (the argument).
For a random variable $X$, the PDF is defined as:

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx$$

Properties:

1. $$\int_{-\infty}^{\infty} f_X(x)dx = 1$$
2. $\forall x, f_X(x) \geq 0$
3. $$\Pr(a \leq X \leq b) = F_X(b) - F_X(a) \overset{4}{=} \int_{a}^{b} f_X(x)dx$$

Notes:

1. for a $F_X(x)$ differentiable for every x (stronger than continuity) then $f_X(x) = \frac{d}{dx}F_X(x)$
2. $f_X(x)$ is not unique. This happens because the values at a specific discrete points don't change the value of the integral.
   these point exist at the points where the function $F_X(x)$ is not differentiable, and we can set $f_X(x)$ any value in these points.

**Examples:**

- Randomly choose an number uniformly in $[0, 1]$
  at the point $x = 0^-, 1^+$ we can put any finite value we want.
- Rolling a dice:

$$f_X(x) = \frac{d}{dx}F_X(x) = \frac{1}{6}\sum_{i=1}^{6}\delta(x-i)$$

[4] from the definition of the PDF

**Summary** Now we can answer questions like $\Pr(X \in S)$, where $S \in \mathbb{R}$, using the function $f_X(x), F_X(x)$.
We will now show a simple example of a continuous random variable (Gaussian random variable, which we will see in the future)
$S = \{[a,b],[c,d]\}$
$\Pr(x \in S) = \Pr(x \in \{[a,b],[c,d]\}) = \Pr(x \in [a,b] \cup [c,d]) = \Pr(x \in [a,b]) + \Pr(x \in [c,d]) \overset{5}{=} [F_X(b) - F_X(a)] + [F_X(d) - F_X(c)]$

### Functions of random variables

A random variable $X$ is transformed using the function $g(\cdot)$.
Knowing $F_X(x)$ and $f_X(x)$ (meaning we the CDF and PDF are known) and the characterizations of $g(\cdot)$, we would like to know the statistics of the output of the transformation $Y = g(X)$, meaning $f_Y(y), F_Y(y)$.
We get a composite function:
$Y = g(X) = g(X(\omega)) \to Y(\omega) = g(X(\omega))$
A composite function of a random variable is still a random variable since it is still a mapping $\Omega \to \mathbb{R}$.
We would like to find a solution for $\Pr(Y \in S), S \in \mathbb{R}$.
$\Pr(Y \in S) = \Pr(g(X) \in S) \overset{6}{=} \Pr(X \in g^{-1}(S))$
Conclusion:
The CDF of Y can be calculated using the CDF of X: $F_Y(y) \triangleq \Pr(Y \in (-\infty, y]) = \Pr(g(X) \in (-\infty, y]) = \Pr(X \in g^{-1}((-\infty, y]))$

**Example:** For the next system, we would like to know the CDF and PDF of the output. when developing a solution, we will notice the importance of the sign of $a$

$$\Theta \sim (0, 2\pi] \to a\Theta + b, PDF/CDF =?$$

$$\begin{aligned} F_Y(y) &= \Pr(g(\Theta) \le y) = \Pr(a\Theta + b \le y) = \Pr(a\Theta \le y - b) \\ &= \begin{cases} \Pr(\Theta \le \frac{y-b}{a}), a > 0 \\ \Pr(\Theta \ge \frac{y-b}{a}), a < 0 \end{cases} \\ &= \begin{cases} F_\Theta(\frac{y-b}{a}), a > 0 \\ 1 - F_\Theta(\frac{y-b}{a}), a < 0 \end{cases} \end{aligned}$$

[5] Because $\Pr(a \le X \le b) = F_X(b) - F_X(a)$ and because the random variable is continuous we can say: $\Pr(a \le X \le B) = \Pr(a \le X \le b)$

[6] from definition $g^{-1}(S) = \{x : g(x) \in S\}$ - i.e. all the x-es that g maps to S.
g is not necessarily invertible

$$\Rightarrow F_Y(y) = \begin{cases} F_\Theta(\frac{y-b}{a}), a > 0 \\ 1 - F_\Theta(\frac{y-b}{a}), a < 0 \end{cases} \xrightarrow{\frac{d}{dy}} f_Y(y) = \begin{cases} \frac{1}{a} f_\Theta(\frac{y-b}{a}), a > 0 \\ -\frac{1}{a} f_\Theta(\frac{y-b}{a}), a < 0 \end{cases}$$

**Moments,Characteristic functions and Moments-generating functions**

Since the CDF and PDF of the function is not always completely known and we know only part of the statistical parameters, we need to define a partial statistical characterization. from the partial characterization we get a bound instead of an exact probability.
In addition, we can test the robustness (the ability of the system to fight changes) of the system by using a partial statistics (for example, dependency of a system using up to 2nd order moments or assuming 2nd order moment to get a linear approximation).

**n-th order moment of a random variable:** The n-th order moment of a random variable is:

$$m_n(X) = m_n \triangleq \mathbb{E}[X^n] \overset{7}{=} \int_{-\infty}^{\infty} x^n \cdot f_X(x)dx$$

$m_n$ represents the “center of mass of the distribution” of the random variable $X$.

**Central n-th order moment of a random variable:** The central n-th order moment of a random variable is:

$$\mu_n(X) = \mu_n = m_n(X - \eta_X) = \mathbb{E}[(X - \eta_X)^n] \overset{8}{=} \int_{-\infty}^{\infty} (x - \eta_X)^n \cdot f_X(x)dx$$

$\mu_n$ represents the “center of mass of the distribution” of the random variable $(X - \eta_X)^n$.

**Expected value, or Mean:** The mean of a random variable is the 1st order moment of the random variable.
The mean of a random variable is defined as:

[7] Assuming the integral exists.

[8] See previous footnote

$$\eta_X = m_1(x) = \mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)dx$$

For a discrete random variable we can use:

$$\mathbb{E}[X] = \sum_i x_i \cdot \Pr(x_i)$$

The mean represents the “center of mass of the distribution”.
Note:
The mean is a linear function, because of the properties of the integral.
Properties:

1. if $Y = g(X)$ then:

$$\mathbb{E}[Y] \triangleq \int_{-\infty}^{\infty} y \cdot f_Y(Y)dy = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)dx$$

2. Linearity: $\mathbb{E}\left[aX+b\right] = a\mathbb{E}\left[X\right] + b$
Proof:

$$\begin{aligned}\mathbb{E}\left[aX+b\right] &= \int_{-\infty}^{\infty} (ax+b) \cdot f_X\left(x\right)dx \\ &= \int_{-\infty}^{\infty} ax \cdot f_X\left(x\right)dx + \int_{-\infty}^{\infty} b \cdot f_X\left(x\right)dx \\ &= a\int_{-\infty}^{\infty} x \cdot f_X\left(x\right)dx + b\int_{-\infty}^{\infty} \cdot f_X\left(x\right)dx = a \cdot \mathbb{E}\left[X\right] + b \cdot 1 \\ &= a\mathbb{E}\left[X\right] + b\end{aligned}$$

**Variance:** The variance of a random variable is the 2nd order central moment of th random variable.
It is calculated by:

$$\begin{aligned}\mathrm{Var}\left(X\right) = \sigma_x^2 = \mu_2\left(x\right) &= \mathbb{E}\left[\left(X-\eta_x\right)^2\right] \\ &= \mathbb{E}\left[X^2 - 2X\eta_X + \eta_X^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2\eta_X\mathbb{E}\left[X\right] + \eta_X^2 \\ &= \mathbb{E}\left[X^2\right] - 2\eta_X \cdot \eta_X + \eta_X^2 \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}^2\left[X\right] \\ &= m_2 - \left(m_1\right)^2\end{aligned}$$

Note:
Defining all the moments up to 2nd order of a random variable, is the same as knowing the mean and variance of the random variable.
Explanation:
The moments up to 2nd order are:

$$m_0 = 1, m_1 = \eta_X, m2 = \mathbb{E}\left[X^2\right]$$
$$\mu_0 = 1, \mu_1 = 0, \mu_2 = \sigma_X^2 = var\left(X\right) = m_2 - \left(m_1\right)^2$$

Out of then,3 are trivial ($m_0, \mu_0, \mu_1$ are trivial and can be calculated regardless the statistics of the random variable) and 3 more that are defined by the other 2.

**Finite Order Statistics:** k-th order statistic means knowing all the moment up to (and include) the k-th order, i.e. knowing all he moments $\{m_i\}_{i=0}^k$ (or knowing $\{\{\eta_i\}_{i=0}^k, m_1 = \eta\}$), that is knowing all the central moments up to the k-th order and the 1st order moment - the mean.)
Ignoring the trivial parameters we defined earlier,k-th order statistics means:

$$\begin{cases}\{m_i\}_{i=0}^k \\ \{\{\mu_i\}_{i=0}^k, m_1 = \eta\}\end{cases} = \begin{cases} m_0, m_1, m_2, ..., m_k \\ \mu_0, \mu_1, \mu_2, ..., \mu_k, m_1\end{cases} \Longleftrightarrow$$
$$\begin{cases} m_1, m_2, ..., m_k \\ \mu_2, ..., \mu_k, m_1\end{cases} = \begin{cases}\{m_i\}_{i=1}^k \\ \{\{\mu_i\}_{i=2}^k, m_1 = \eta\}\end{cases}$$

Note:
Mathematically,knowing one set is equal to knowing the other set (but engineering-related there is a difference - for example, for 2nd order,$X^2$ represents power, while $\left(X - \eta_X\right)^2$ represents a squared deviation from the center)
Proof:
This direction: ⇒
Assume we know $\{\{\mu_i\}_{i=0}^k, m_1 = \eta\}$, we would like to use them express $m_k$:

$$m_k = \mathbb{E}\left[X^k\right] = \mathbb{E}\left[\left(\left(Xx - \eta_X\right) + \eta_X\right)^k\right]$$
$$\overset{9}{=} \mathbb{E}\left[\sum_{i=0}^{k} \binom{k}{l} \left(X - \eta_X\right)^l \left(\eta_X\right)^{k-l}\right] = \sum_{i=0}^{k} \binom{k}{l} \mu_l \left(\eta_X\right)^{k-l}$$

The other direction $\Leftarrow$:
Assume we know $\{m_i\}_{i=1}^k$. We would like to express $\mu_k$:

$$\begin{aligned}\mu_k &= \mathbb{E}\left[(X-\eta_X)^k\right] \overset{10}{=} \mathbb{E}\left[\sum_{i=0}^{k}\binom{k}{l}(X)^l(-\eta_X)^{k-l}\right]\\ &\overset{11}{=} \sum_{i=0}^{k}\binom{k}{l} m_l(-m_1)^{k-l}\end{aligned}$$

As noted, partial statistics will not give us precise probability, but a bound.

**Chebyshev's Inequality** Chebyshev's inequality describes the probability to be at a certain distance from the mean.

$$Pr(|X-\eta_X|\geq a)\leq\frac{\sigma_X^2}{a^2}, a\geq 0$$

By choosing $a=k\sigma_X$, i.e. choosing $a$ as a multiplication of the square root of the variance, we ask what is the probability that $X$ is at a distance of $k$ steps of $\sigma_X$ away from the mean. In that case, we get:

$$Pr(|X-\eta_X|\geq k\sigma_X)\leq\frac{1}{k^2}$$

**Markov's Inequality** Markov's inequality describes the probability of a non-negative random variable to go over a certain bound $b$:

$$Pr(Y\geq b)\leq\frac{\eta_Y}{b}$$

### Proof of Markov's Inequality

$$\begin{aligned}Pr(Y\geq b) &= \int_b^\infty f_Y(y)\,dy = \int_b^\infty 1\cdot f_Y(y)\,dy\\ &\overset{12}{\leq} \int_b^\infty \frac{y}{b}f_Y(y)\,dy \overset{13}{\leq} \int_0^\infty \frac{y}{b}\cdot f_Y(y)\,dy\\ &= \frac{1}{b}\int_0^\infty y\cdot f_Y(y)\,dy \overset{14}{=} \frac{1}{b}\int_{-\infty}^\infty y\cdot f_Y(y)\,dy = \frac{1}{b}\mathbb{E}[Y] = \frac{\eta_Y}{b}\end{aligned}$$

[9] According to Newton's Binomial theorem

[10] See previous footnote.

[11] because $\eta_X = m_1$

[12] $y \geq b$

[13] $\frac{y}{b} \to 0$, positive integrand

[14] Y is non-negative random variable, hence $f(y) = 0 \forall y \in (-\infty, 0)$

**Proof of Chebyshev's Inequality** For a random variable $X$ we define $Y = (X - \eta_X)^2$. Y is a non-negative random variable.

$$Pr\left(|X - \eta_X| \geq a\right) = Pr\left(|X - \eta_X|^2 \geq a^2\right) = Pr\left(Y \geq a^2\right) \overset{15}{\leq} \frac{\eta_Y}{a^2} = \frac{\sigma_X^2}{a^2}$$

Note:
Knowing the moments for each order(i.e., knowing the infinite set of moments) is equal to the full statistics.

**Characteristic Function** Characteristic Function (CF) of a random variable $X$ is the Fourier transform of the PDF of $X$:

$$\Phi_X(\omega) = \int_{-\infty}^{\infty} e^{j\omega x} f_X(x)\, dx = \mathbb{E}\left[e^{j\omega x}\right]$$

Notes:

- We notice that this is actually the function of a random variable $Y(X) = e^{j\omega X}$, this function creates a composite random variable.
- Because the CF is created by Fourier transform, all the properties of Fourier transform hold.

Properties:

1. CF is an alternative description of a full statistical information, and so knowing the CF is equal to knowing CDF(and its PDF, which is not always defined) an vice versa. Hence, statistically, $\Phi_X(\omega) \iff F_X(x) \iff f_X(x)$

$$f_X(x) = \mathcal{F}^{-1}\left(\Phi_X(\omega)\right) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-j\omega x}\Phi_X(\omega)\, d\omega$$

2. $\Phi_X(0) = 1$, Proof:

$$\Phi_X(0) = \int_{-\infty}^{\infty} e^{0} f_X(x)\, dx = \int_{-\infty}^{\infty} f_X(x) = 1$$

3. $|\Phi_X(\omega)| \leq 1, \forall\omega$. Proof:

$$|\Phi_X(\omega)| = \left|\int_{-\infty}^{\infty} e^{j\omega x} f_X(x)\, dx\right| \leq \int_{-\infty}^{\infty} \left|e^{j\omega x}\right| f_X(x)\, dx \overset{16}{=} 1$$

[15] Markov's Inequality

4. The connection between CF and moments: $\frac{d^n\Phi_X(\omega)}{d\omega^n}|_{\omega=0} = j^n \cdot m_n$. Proof:

$$\begin{aligned}\frac{d^n\Phi_X(\omega)}{d\omega^n}|_{\omega=0} &= \left[\int_{-\infty}^{\infty}(jx)^n e^{j\omega x} f_X(x)\,dx\right]|_{\omega=0} \\ &= j^n\left[\int_{-\infty}^{\infty}(x)^n e^{j\omega x} f_X(x)\,dx\right]|_{\omega=0} = j^n \cdot m_n\end{aligned}$$

5. CF is a linear function (from the linearity of the integral and the properties of the exponent): $Y = a \cdot X + b \Rightarrow \Phi_Y(\omega) = e^{j\omega b} \cdot \Phi_X(a\omega)$. Proof:

$$\Phi_Y(\omega) = \mathbb{E}\left[e^{j\omega Y}\right] = \mathbb{E}\left[e^{j\omega(a\cdot X+b)}\right] = e^{j\omega b}\mathbb{E}\left[e^{j\omega(a\cdot X)}\right] = e^{j\omega b} \cdot \Phi_X(a\omega)$$

**Moment Generating Function** The MGF of a random variable $X$ is the Laplace transform of the PDF of $X$:

$$M_X(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x)\,dx = \mathbb{E}\left[e^{sX}\right]$$

Note:
CF is always absolutely continuous (Fourier transform) while the MGF can sometimes have problems of converging (Laplace transform). On the other side, the CF contains a complex coefficient $j$ that is not in the MGF.
Properties:

1. MGF is an alternative description of the full statistical information.
2. the connection between the MGF and moments is: $\frac{d^n M_X(s)}{ds^n}|_{s=0} = m_N$

**Chernoff Bound:** The Chernoff bound describes the probability of a random variable $X$ to go above a certain top bound $a$:

$$Pr(X \geq a) \leq e^{-sa} M_X(s)$$

Note:
On the right side of the bound we use the MGF. Knowing it is equal ro knowing the full statistic, hence why would we use the bound if we have the whole statistic?
The answer is that calculating the bound using the MGF is very simple and useful (the bound is very close to the statistics) to calculate the probability with low threshold to the i.i.d. random variable: $X = \sum_i a_i X_i, \{X_i\} i.i.d$

[16] $e^{j\omega x} = 1$

**i.i.d.** A set of random variables is independent identically distributed (i.i.d) if they are independent and with the same PDF.

### Proof of Chernoff Bound

$$Pr\left(X \geq a\right) \overset{17}{=} Pr\left(sX \geq sa\right) = Pr\left(e^{sX} \geq e^{sa}\right) \overset{18}{\leq} \frac{M_X\left(s\right)}{e^{sa}} = e^{-sa}M_X\left(s\right)$$

Note:
As said before, knowing moments of every order equal to knowing the entire statistics of the random variable. We assume that $\{m_i\}_{i=0}^{\infty}$ is known.
We develop the CF to a Maclaurin series:

$$\Phi_X\left(\omega\right) = \sum_{n=0}^{\infty} \frac{1}{n!}\frac{d^n\Phi_X\left(\omega\right)}{d\omega^n}|_{\omega=0}\omega^n = \sum_{n=0}^{\infty}\frac{1}{n!}j^n m_n \omega^n = \sum_{n=0}^{\infty}\frac{\left(j\omega\right)^n}{n!}m_n$$

And so, we described the MGF (which is equal to knowing the PDF) by a group of moments. This formula is valid only in case the CF is analytic.
A sufficient condition for analyticality is that the random variable $X$ is bounded, i.e. there is always an $A$ such that $|X| \leq A$, or $Pr\left(-A \leq X \leq A\right) = 1$

## Random Vectors

A random vector is a generalization of random variables. It is a set of several random variables(that may be dependent on each other), which are defined on the same probability space $X_1(\omega), X_2(\omega), ..., X_n(\omega)$.
Mathematically, random vector $\underline{X}$ is defined on the same probability space $\{\Omega, F, P\}$ as a random variable, but with a mapping of: $\underline{X} : \Omega \rightarrow \mathbb{R}^n$
We will demand it will map $\Omega$ so the set $\{\omega : X_1\left(\omega\right) \leq x_1, ..., X_n\left(\omega\right) \leq x_n\}$ is an event in F for each $\underline{x} \in \mathbb{R}^n$.
Usually we will right random vectors as a column vector:

$$\underline{X} = \begin{pmatrix} X_1\left(\omega\right) \\ . \\ . \\ . \\ X_n\left(\omega\right) \end{pmatrix}, \underline{X} : \Omega \rightarrow \mathbb{R}^n$$

Like in random variables, we will describe the statistical information of the random vector $\underline{X}$, by:

[17] $s > 0$

[18] Markov’s Inequality

- jCDF
- jPDF
- jMGF

Note: j fot joint.

## Joint Cumulative Distribution Function and Joint Probability Density Function

### Joint Cumulative Distribution Function-(jCDF)

The jCDF function is defined as:

$$\begin{aligned} F_X(x) &\triangleq Pr\left(\{\omega : X_1(\omega) \le x_1, ..., X_n(\omega) \le x_n\}\right) \\ &\overset{19}{=} Pr\left(X_1 \le x_1, ..., X_n \le x_n\right) = Pr\left(\underline{X} \le \underline{x}\right) \end{aligned}$$

Properties:

1. $0 \le F_{\underline{X}}(\underline{x}) \le 1, \forall \underline{x} \in \mathbb{R}^n$
2. $F_{\underline{X}}(\underline{x})$ is a monotonic non-decreasing function in every variable $x_i, i = 1, 2, .., n$.
3. $F_{\underline{X}}(\underline{x})$ is continuous drom the right, in every variable $x_i, i = 1, 2, .., n$.
4. $$\lim_{\underline{x} \to \text{"}\underline{\infty}\text{"}} F_{\underline{X}}(\underline{x}) = \lim_{x_1 \to \infty, x_2 \to \infty, ..., x_n \to \infty} F_{\underline{X}}(x_1, x_2, ..., x_n) = 1$$
   The reason is that at $\infty$ every variable is already mapped from $\Omega$.
5. $\lim_{X_i \leftarrow -\infty} F_{\underline{x}}(\underline{x}) = 0$. This happens because there is no source from $\Omega$ that will map $X_i$ to a smaller value then $-\infty$, because $\{\omega : X_i(\omega) \le -\infty\}$.
6. Calculating the CDF of a sub-vector of $\underline{X}$:
   $$F_{X_k}(x) = \lim_{x_i \to \infty, i \ne k} F_{\underline{X}}(x_1, x_2, ..., x_n)$$
   $$F_{X'}(x') = \lim_{x_i \to \infty, X_i \notin \underline{X'}} F_{\underline{X}}(x_1, x_2, ..., x_n)$$
   This is done by taking the not-included variables to $\infty$

[19] Calculation of probability should get events as inputs, but we will write the values of random variables values for simplicity.

Note:
The definition of jCDF is enough to answer any question from the form $Pr\left(\underline{X} \in S\right)$.
Explanation:
We will show for 2-dimensions - variables X,Y:
We first assume that $S$ is an area that can be divided into rectangles. For $S$ that can’t be simply divided into rectangles, we can divide it into infinitesimal rectangles. Now we want to answer the question, what is sthe probability to “be” in a single rectangle (as small as we want) and then we can sum the probabilities of all the rectangles.

$$\begin{aligned} Pr\left((X,Y) \in S\right) &= Pr\left(x \leq X \leq x_{\Delta}x, y \leq Y \leq y + \Delta y\right) \\ &= F_{XY}\left(x + \Delta x, y + \Delta y\right) - F_{XY}\left(x, y + \Delta y\right) \\ &\quad - F_{XY}\left(x_{\Delta}x, y\right) + F_{XY}\left(x, y\right) \end{aligned}$$

The reason for that is that $F_{XY}(a,b)$ is in fact the probability that $(X,Y)$ will be found in the rectangle below - left -to the point $(a,b)$, and that is why we add parts of rectangles.

### Joint Probability Density Function (jPDF)

For a random vector $\underline{X}$ the jPDF will be defined as:

$$F_{\underline{X}}\left(\underline{x}\right) = \int\int_{-\underline{\infty}}^{\underline{x}} \cdots \int f_{\underline{X}}\left(\underline{x}\right) \underline{dx} = \int_{-\infty}^{x_1}\int_{\infty}^{x_2} \cdots \int_{\infty}^{x_n} f_{\underline{X}}\left(\underline{x}\right) dx_1 dx_2 \ldots dx_n$$

Properties:

1. $\int_{\underline{X}\in\mathbb{R}} \cdots \int f_{\underline{X}}\left(\underline{x}\right) \underline{dx} = 1$
2. $\forall \underline{x} : f_{\underline{x}}\left(\underline{x}\right) \geq 0$
3. Calculating the PDF of a random variables or a sub-vector:

$$f_{X_k}\left(x\right) = \int_{-\infty}^{\infty} \underbrace{\cdots}_{\forall X \neq X_i} \int_{-\infty}^{\infty} f_{\underline{X}}\left(\underline{x}\right) \underline{dx}$$

$$f_{X'}\left(\underline{x}'\right) = \int_{-\infty}^{\infty} \underbrace{\cdots}_{\forall X_i \notin \underline{X}'} \int_{-\infty}^{\infty} f_{\underline{X}}\left(\underline{x}\right) \underline{dx}$$

We do so by integrating in the domain $(-\infty, \infty)$ for every random variable that doesn’t belong to our needed jPDF.

Proof for 2D:
$F_X(x) = F_{XY}(x,\infty)$

$$\Rightarrow f_X(x) = \frac{d}{dx}F_{XY}(x,\infty) = \frac{d}{dx}Pr(-\infty \le X \le x, -\infty \le Y\infty)$$
$$= \frac{d}{dx'}\int_{y'=-\infty}^{\infty}\int_{x'=-\infty}^{x} f_{XY}(x',y')\,dx'dy' = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy$$

Note:
For $F_{\underline{X}}(\underline{x})$ differentiable for each $\underline{x}$ (stronger then continuity) we can say: $f_{\underline{X}}(\underline{x})\frac{d^n}{dx_1\cdot\ldots\cdot dx_n}F_{\underline{X}}(\underline{x})$ We assume that the derivative exists. Explanation:
For 2D:

$$f_{XY}(x,y) = \lim_{\Delta x\to 0,\Delta y\to 0}\frac{1}{\Delta x\Delta y}Pr(x\le X\le x+\Delta x, y\le Y\le y+\Delta y) = \frac{d^2}{dxdy}F_{XY}(x,y)$$

Conclusion:

$$Pr(\underline{X}\in S)\int\underbrace{\ldots}_{\underline{x\in S}}\int f_{\underline{X}}(\underline{x})\,\underline{dx}$$

**Two Random Variables(Marginal and Conditional Distribution):**

First we talk about a random vector with size of 2, i.e. two random variable. The discussion about the 2 random variables holds for a general random vector, but we will look at 2D private case for educational purposes.

**Example** : For a dice rolling experiment we define 2 variables:

1. $X_1(\omega)$ even indicator
2. $X_1(\omega)$ is bigger than 1 indicator.

$$X_1(\omega) = \begin{cases}0, \omega = 1,3,5\\ 1, \omega = 2,4,6\end{cases}, X_2(\omega) = \begin{cases}0, \omega > 1\\ 1, \omega = 1\end{cases}$$

We can calculate the probabilities:

$$Pr(X_1,X_2) = \begin{cases}\frac{1}{6}, (X_1,X_2) = (0,0)\\ \frac{1}{3}, (X_1,X_2) = (0,1)\\ 0, (X_1,X_2) = (1,0)\\ \frac{1}{2}, (X_1,X_2) = (1,1)\end{cases}$$

We can see that there is a dependence between the variables because $Pr\left(X_1=\alpha, X_2=\beta\right) \neq Pr\left(X_1=\alpha\right) \cdot Pr\left(X_2=\beta\right)$
The jPDF and jCDF are:
We see that the jPDF is composed of an impulse train and the jCDF is composed of jumps. This happens because the random variables that the random vector is composed from, are discrete.

$$\begin{aligned} F_{X_1,X_2}\left(x_1,x_2\right) &= \sum_{X_1=0}^{1}\sum_{X_2=0}^{1} Pr\left(X_1,X_2\right)U\left(x_1-X_1\right)U\left(x_2-X_2\right) \\ &= \frac{1}{6}U\left(x_1-0\right)U\left(x_2-0\right)+\frac{1}{3}U\left(x_1-0\right)U\left(x_2-1\right) \\ &\quad+\frac{1}{2}U\left(x_1-1\right)U\left(x_2-1\right) \end{aligned}$$

$$\begin{aligned} \Rightarrow f_{x_1,x_2}\left(x_1,x_2\right) &= \frac{d^2}{dx_1dx_2}F_{X_1,X_2}\left(x_1,x_2\right)=\frac{d}{dx_2}\left[\frac{d}{dx_1}F_{X_1,X_2}\left(x_1,x_2\right)\right] \\ &= \frac{1}{6}\delta\left(x_1\right)\delta\left(x_2\right)+\frac{1}{3}\delta\left(x_1\right)\delta\left(x_2-1\right)+\frac{1}{2}\delta\left(x_1-1\right)\delta\left(x_2-1\right) \end{aligned}$$

**Conditional probability, reminder:** Given events $A, B$, and assuming that $\Pr(B) \neq 0$, so the probability of event $A$,denoting event $B$ is:

$$\Pr(A|B)=\frac{\Pr(A\cap B)}{\Pr(B)}$$

**Conditional Probability for two random variables:** For a random vector $(X,Y)$, consider two events $X\in S_X, Y\in S_Y$. We can assume that each of the domains $S_X, S_Y$ is a union of intervals.

$$\begin{aligned} \Pr\left(X\in S_X|Y\in S_Y\right) &= \frac{\Pr\left(X\in S_X\cap Y\in S_Y\right)}{\Pr\left(Y\in S_Y\right)} \\ &= \frac{\int_{x\in S_X}\int_{y\in S_Y} f_{XY}(x,y)\,dxdy}{\int_{y\in S_Y} f_Y(y)\,dy} \\ &= \frac{\int_{X\in S_X}\int_{y\in S_Y} f_{XY}(x,y)\,dxdy}{\int_{x=-\infty}^{\infty}\int_{y\in S_Y} f_{XY}(x,y)\,dxdy} \end{aligned}$$

**Pointwise conditioning:** Very often, we wish to condition a random variable: Since the probability of a point is zero, $\Pr(Y=y)=0$, it follows that we can't use the conditional probability formula, as we can't divide by zero.

To deal with this problem we define:

$$\begin{aligned}
\Pr\left(X \in S_X | Y = y\right) &= \lim_{\Delta y \to 0} \Pr\left(X \in S_X | Y \in (y, y + \Delta y)\right) \\
&\overset{20}{=} \lim_{\Delta y \to 0} \Pr\left(X \in S_X | Y \in S_Y\right) \\
&= \lim_{\Delta y \to 0} \frac{\Pr\left(X \in S_X \cap Y \in S_Y\right)}{\Pr\left(Y \in S_Y\right)} \\
&= \lim_{\Delta y \to 0} \frac{\int_{x \in S_X} \int_{y \in S_Y} f_{XY}(x,y) dx dy}{\int_{y \in S_Y} f_Y(y) dy} \cdot \frac{\frac{1}{\Delta y}}{\frac{1}{\Delta y}} \\
&= \lim_{\Delta y \to 0} \frac{\frac{1}{\Delta y} \int_{X \in S_X} \int_{v=y}^{y+\Delta y} f_{XY}(u,v) du dv}{\frac{1}{\Delta y} \int_{v=y}^{y+\Delta y} f_Y(v) dv} \\
&\overset{21}{=} \frac{\lim_{\Delta y \to 0} \left[\frac{1}{\Delta y} \int_{v=y}^{y+\Delta y} \left(\int_{X \in S_X} f_{XY}(u,v) du\right) dv\right]}{\lim_{\Delta y \to 0} \left[\frac{1}{\Delta y} \int_{v=y}^{y+\Delta y} f_Y(v) dv\right]} \\
&\overset{22}{=} \frac{\int_{x \in S_X} f_{XY}(x,y) dx}{f_Y(y)} = \int_{x \in S_x} \frac{f_{XY}(x,y) dx}{f_Y(y)} \\
&= \int_{x \in S_x} f_{X|Y}\left(x \mid y\right) dx
\end{aligned}$$

**Conditional distribution and conditional distribution functions:** We mat define the conditional distribution functions (CDF,PDF), i.e. we define functions that we can use to answer questions of the form

$$\Pr\left(X \in X_X \mid Y = y\right) = ?$$

Conditional CDF:

$$F_{X|Y}\left(x \mid y\right) \overset{23}{=} F_{X|Y}\left(x, y\right) \triangleq \Pr\left(X \in (-\infty, x) \mid Y = y\right) \overset{24}{=} \int_{x \in S_X} \frac{f_{XY}(x,y) dx}{f_Y(y) dy}$$

[20] $S_Y = (y, y + \Delta y)$

[21] Assuming the limit exists. Changing order of integration

[22] L'Hospital's rule. Continuity of the function in the integration boundaries.

Conditional PDF:

$$f_{X|Y}(x|y) = f_{X|Y}(x,y) \triangleq \frac{d}{dx} F_{X|Y}(x,y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$

Note - Bayes' Theorem:
By switching the roles of variables, we can get Bayes' theorem for a conditional PDF:

$$f_{XY}(x,y) = f_X(x) \cdot f_{Y|X}(y \mid x) = f_Y(y) \cdot f_{X|Y}(x \mid y)$$

$$f_{Y|X}(y \mid x) = \frac{f_Y(y) \cdot f_{X|Y}(x \mid y)}{f_X(x)} = \frac{f_Y(y) \cdot f_{X|Y}(x \mid y)}{\int_{-\infty}^{\infty} f_{XY}(x,y)dy} = \frac{f_Y(y) \cdot f_{X|Y}(x \mid y)}{\int_{-\infty}^{\infty} f_Y(y) f_{X|Y}(x \mid y)\, dy}$$

This law is defined for both continuous and discrete cases. For a discrete case, we use the PMF $P(\cdot)$ instead of the PDF $f(\cdot)$.

**Remark 1** *The conditional PDF/CDF satisfy all the properties of PDF/CDF functions. It is convenient to think of them as the PDF/CDF - of a variable "$X|Y = y$".*

**Conditional mean and conditional variance:** We can define a conditional mean, conditional variance and so on for every moment we want:

$$\eta_{X|Y=y} = \mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x,y)dx$$

$$\mathrm{Var}(X \mid Y = y) = \int_{-\infty}^{\infty} \left(x - \eta_{X|Y=y}\right)^2 \cdot f_{X|Y}(x,y)dx$$

**Statistical Independence:** We can say that $X, Y$ are independent random variables if:

$$\Pr(X \in S_X, Y \in S_Y) = \Pr(X \in S_X) \cdot \Pr(Y \in S_Y), \quad \forall S_X, S_Y \in \mathbb{R}$$

We note that the next statements are equivalent:

1. $(X, Y)$ is an independent random vector.
2. $F_{XY}(x,y) = F_X(x) \cdot F_Y(y)$, for all $x, y$.

[23] Both forms of the formula are equivalent because this is a 2D function, and we can change the way we write it for our convenience. The index is what is important in this formula.

[24] $S_X = (-\infty, x)$

3. $f_{XY}(x,y) = f_X(x) \cdot f_Y(y)$ for all $x, y$.

Proof:

1. $\rightarrow$ 2.

$$\begin{aligned} F_{XY}(x,y) &\triangleq \Pr\left(X \in (-\infty, x), Y \in (-\infty, y)\right) \\ &\overset{25}{=} \Pr\left(X \in (-\infty, x)\right) \cdot \Pr\left(Y \in (-\infty, y)\right) \\ &= F_X(x) F_Y(y) \end{aligned}$$

2. $\rightarrow$ 3.

$$\begin{aligned} f_{XY}(x,y) &= \frac{d^2}{dxdy} F_{XY}(x,y) \overset{26}{=} \frac{d^2}{dxdy}\left[F_X(x) \cdot F_Y(y)\right] \\ &= \frac{d}{dx} F_X(x) \cdot \frac{d}{dy} F_Y(y) \\ &= f_X(x) \cdot f_Y(y) \end{aligned}$$

3. $\rightarrow$ 1.

$$\begin{aligned} \Pr\left(X \in S_X, Y \in S_Y\right) &= \int_{x \in S_X} \int_{y \in S_Y} f_{XY}(x,y) dxdy \\ &\overset{27}{=} \int_{x \in S_X} \int_{y \in S_Y} f_X(x) \cdot f_Y(y) dxdy \\ &= \int_{x \in S_X} f_X(x) dx \int_{y \in S_Y} f_Y(y) dy \\ &= \Pr\left(X \in S_X\right) \cdot \Pr\left(Y \in S_Y\right) \end{aligned}$$

Generalization: We say that a random vector $\underline{X}$ has independent components if for any choice of sets $S_{X_i}$, we have:

$$\Pr\left(X_1 \in S_{X_1}, ..., X_n \in S_{X_n}\right) = \prod_{i=1}^{n} \Pr(X_i \in S_{X_i}).$$

Equivalently, the jCDF and the jPDF can be expressed as a product of their marginals.

[25] Statistically independent

[26] Property 2

[27] Property 3

[28] Property 4

[29] Property 4

**Remark 2** *The jPDF of a random vector $X$ determines the PDF of each of its entries (as well as the jPDF of every sub-vector) by integrating over $(-\infty, \infty)$ for every other random variable. The opposite is false, however, unless we know that the components are independent. In the latter case, e.g. in the case of a two-dimensional random vector: $X, Y$ independent $\rightarrow$ $f_{XY}(x, y) = f_X(x)f_Y(y)$.*

**Example: pairwise independence doesn't imply independence** We may construct a random vector $(X, Y, Z)$ with pairwise independent components, i.e., $(X, Y), (X, Z), (Y, Z)$ are independent, as follows. Let $X, Y$ be independent and identically distributed as

$$X = \begin{cases} 1, & w.p.\ 0.5 \\ 0, & w.p.\ 0.5 \end{cases}$$

and

$$Y = \begin{cases} 1, & w.p.\ 0.5 \\ 0, & w.p.\ 0.5 \end{cases} .$$

Define

$$Z = X \oplus Y = \begin{cases} 1, & w.p.\ 0.5 \quad \text{(corresponding to } \{0,1\}, \{1,0\}) \\ 0, & w.p.\ 0.5 \quad \text{(corresponding to } \{0,0\}, \{1,1\}) \end{cases} .$$

Notice that indeed all pairs of random variables are independent. But the random vector is *not* independent, because:

$$\Pr\left(X = 1, Y = 1, Z = 1\right) \neq \Pr(X = 1) \cdot \Pr(Y = 1) \cdot \Pr(Z = 1)$$

since

$$\Pr(X = 1, Y = 1, Z = 1) = 0$$

whereas

$$\Pr(X = 1) \cdot \Pr(Y = 1) \cdot \Pr(Z = 1) = \frac{1}{8}.$$

**Example:** Suppose we have a random vector $(X, Y)$ where it is known that the marginals satisfy $X, Y \sim U(-0.5, 0.5)$. That is, we know that

$$f_X(x) = \begin{cases} 1, & -0.5 < x < 0.5 \\ 0, & else \end{cases} ,$$

and also

$$f_Y(Y) = \begin{cases} 1, & -0.5 < y < 0.5 \\ 0, & else \end{cases}$$

.

We would like to understand the effect of the dependence between the two random variables by looking at three different scenarios:

Case 1: If we know that $X, Y$ are independent, then:

$$f_{XY}(x,y) = f_X(x) \cdot f_Y(y) = \begin{cases} 1, & \begin{cases} -0.5 < x < 0.5 \\ -0.5 < y < 0.5 \end{cases} \\ 0, & \text{else} \end{cases}$$

Case 2: If it is known that $X = Y$, then:[28]

$$f_{X|Y}\left(x \mid y\right) = \delta(x-y), \quad |y| < 0.5$$

Hence, we get:

$$f_{XY}(x,y) = f_Y(y) \cdot f_{X|Y}\left(x \mid y\right) = \begin{cases} 1 \cdot \delta(x-y), & \begin{cases} -0.5 < x < 0.5 \\ -0.5 < y < 0.5 \end{cases} \\ 0, & else \end{cases}$$

which is in fact a "wall of deltas" on the line $x = y$ in the domain that corresponds to the distribution of the random variable ($|x|, |y| < 0.5$).

Case 3: Let $Y = B \cdot X$ where $B = \pm 1$ with equal probability, $B, X$ being independent. That is,

$$B = \begin{cases} 1, & w.p.\ 0.5 \\ 0, & w.p.\ 0.5 \end{cases}$$

Further, it is known that $X \sim U([-0.5, 0.5])$, i.e., it is uniform over the interval $[-0.5, 0.5]$ . It follows that the random variable $-X$ has the same distribution as $X$. Therefore, no matter if $B = 1$ or $B = -1$, we have that $Y, X$ have the same distribution and are each uniformly distributed over the interval $[-0.5, 0.5]$. We therefore get:

$$f_{Y|X}\left(y \mid x\right) = \begin{cases} \frac{1}{2}\delta(y-x) + \frac{1}{2}\delta(y+x), & |x| < 0.5 \\ 0, & else \end{cases}$$

and consequently:

$$f_{XY}(x,y) = f_X(x) \cdot f_{Y|X}\left(y \mid x\right) = \begin{cases} 1 \cdot \left[\frac{1}{2}\delta(y-x) + \frac{1}{2}\delta(y+x)\right], & \begin{cases} -0.5 < x < 0.5 \\ -0.5 < y < 0.5 \end{cases} \\ 0, else \end{cases}$$

This is in fact are two "walls of deltas" on the lines $x = y$ and $x = -y$ in the domain $|x|, |y| < 0.5$

[30] Given $Y = y$, we know that $X = y$.

### Vector of random variables

Now we will generalize our discussion and talk about random vectors of any dimension. Nonetheless, illustrations will still be given in $\mathbb{R}^2$ .

**Function of a random vector** Suppose we apply the transformation $g(\cdot)$ to a random vector $\underline{X}$. The output of the transformation is a random vector $\underline{Y}$ which don't necessarily have the same size/dimension. Knowing the jCDF and jPDF of the input, and also the system $g(\cdot)$, we would like to know the statistics of the output $\underline{Y} = g(\underline{X})$, i.e., to find $f_{\underline{Y}}(\underline{y}), F_{\underline{Y}}(\underline{y})$. In fact, we get a composite function:

$$\underline{Y} = g(\underline{X}) = g(\underline{X}(\omega)) \Rightarrow \underline{Y}(\omega) = g(\underline{X}(\omega))$$

A composite function of a random vector is still a random vector because it is still a mapping $\Omega \to \mathbb{R}^k$. We would like to answer the question $\Pr(\underline{Y} \in S) =?$ for every $S \in \mathbb{R}^M$

**Example:** For the next system (where the joint distribution of the input is known, hence so is the distribution of each of the variables): $Z = X + Y$. We would like to find $F_Z(z)$:

$$\begin{aligned} F_Z(z) &= \Pr(Z \leq z) = \Pr\left(g(X,Y) \leq z\right) \\ &\overset{29}{=} \Pr\left((X,Y) \in g^{-1}\left((-\infty, z]\right)\right) \\ &= \iint\limits_{(X,Y)\in g^{-1}(-\infty,z]} f_{XY}(x,y)dxdy \end{aligned}$$

So far, the development was general for $g : \mathbb{R}^2 \to R$, in our case $g(X,Y) = X + Y$, so:

$$\begin{aligned} F_Z(z) &= \Pr(Z \leq z) = \Pr(X + Y \leq z) \\ &= \int_{x=-\infty}^{\infty} \left( \int_{y=-\infty}^{z-x} f_{XY}(x,y)dy \right) dx \end{aligned}$$

$$\Rightarrow f_Z(z) = \frac{d}{dz} F_Z(z) \overset{30}{=} \int_{x=-\infty}^{\infty} f_{XY}(x, z-x)dx$$

[31] The inverse function is: $g^{-1}\left((-\infty, z]\right) = \{(x,y) : g(x,y) \in (-\infty, z]\}$

If we add that $(X, Y)$ is an independent random vector, so $f_{XY}(x,y) = f_X(x) \cdot f_Y(y)$, we obtain:

$$f_Z(z) = \int_{x=-\infty}^{\infty} f_{XY}(x, z-x)dx = \int_{x=-\infty}^{\infty} f_X(x) \cdot f_Y(z-x)dx \overset{31}{=} f_X(x) * f_Y(y)$$

**Conclusion**: For a series of independent random variables $\{X_i\}_{i=1}^{N}$, we can define a random variable $Z = X_1 + ... + X_N = \sum_{i=1}^{N} X_i$. The PDF of this sum is givenby: $f_Z(z) = f_{X_1} * \cdots * f_{X_N}$.
Note that as $N \to \infty$ we will get a Gaussian, regardless of the PDF of $X$ as a consequence of the central limit theorem. For any distribution we start with, we eventually approach a Gaussian distribution.

For example, We can show this if we start with a uniform distribution, then: for a single random variable we get a rectangle, for two we get a triangle (convolution of 2 rectangles), and in the limit we approach a Gaussian.

**Mean of a random vector and mean of a function of a random vector:** Consider a random vector $(X, Y)$ with $F_{XY}(x,y)$ (or in the $n$-dimensional case: $\underline{X}$ with $f_{\underline{X}}(\underline{x})$). Also, we assume transformation $g(X, Y)$ (or, respectively,$g(\underline{X})$).

1. Smoothing Theorem/Law of total expectation
   We may calculate the mean of a random variable from a vector as:

$$\begin{aligned}\mathbb{E}[Y] &= \int_{-\infty}^{\infty} y \cdot f_Y(y)dy = \int_{-\infty}^{\infty} y \cdot \left(\int_{-\infty}^{\infty} f_{XY}(x,y)dx\right) dy \\ &= \int_{-\infty}^{\infty} y \cdot \left(\int_{-\infty}^{\infty} f_X(x) f_{Y|X}(y \mid x)\, dx\right) dy \\ &= \int_{-\infty}^{\infty} f_X(x) \left(\int_{-\infty}^{\infty} y \cdot f_{Y|X}(y \mid x)\, dx\right) dy \\ &= \int_{-\infty}^{\infty} f_X(x) \mathbb{E}\left[Y \mid X = x\right] dy = E_X\left[E_Y\left[Y \mid X\right]\right]\end{aligned}$$

   i.e. we first calculate the mean of $Y$ first by conditioning on $X$, and then we calculate the mean by all the possible values of $X$ (that is how in the inner mean we "got rid of" the random variable $Y$).

[32] Fundamental theorem of algebra: $\frac{d}{dt}\int_a^b f(t)dt = f(b)$

[33] Definition of convolution:
$f(x) * g(x) = \int_{t=-\infty}^{\infty} f(t-x) \cdot g(t)dt = \int_{t=-\infty}^{\infty} g(t-x) \cdot f(t)dt = g(x) * f(x)$

2. Smoothing Theorem/Law of total expectation - Mean of a function of a random vector
   We calculate the mean of a function of a random vector by:

$$\mathbb{E}\left[g\left(\underline{x}\right)\right] = \int \underset{\forall \underline{x}}{\cdots} \int g\left(\underline{x}\right) f_{\underline{X}}(\underline{x}) \underline{dx}$$

   For the 2-D case:

$$\begin{aligned}\mathbb{E}\left[g(X,Y)\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_X(x) f_{Y|X}\left(y \mid x\right) dxdy \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} g(x,y) f_{Y|X}\left(y \mid x\right) dy\right) f_X(x)dx \\ &= \int_{-\infty}^{\infty}\left[E_Y\left[g(X,Y)|X=x\right]\right] f_X(x)dx \\ &= E_X\left[E_Y\left[g(X,Y)|\right]\right] \\ &\overset{32}{=} E_Y\left[E_X\left[g(X,Y|)Y\right]\right].\end{aligned}$$

Notes:

1. A special case of the smoothing theorem corresponds to taking $g(x,y) = y$ or $g(x,y) = x$.

2. We can generalize these formulas to the $n$-dimensional case, for instance:

$$\mathbb{E}\left[g(X,Y,Z)\right] = \mathbb{E}_Y\left[\mathbb{E}_X\left[\mathbb{E}_Z\left[g(X,Y,Z) \mid X,Y\right]\right]\right].$$

**Joint moments of a random vector:** For a random vector $\underline{X}$ of dimension $N$, we define the joint moments:

$$m_{k_1,\ldots,k_N} = \mathbb{E}\left[X_1^{k_1}\cdot\ldots\cdot X_N^{k_N}\right] = \mathbb{E}\left[\prod_{i=1}^{N} X_i^{k_i}\right].$$

We define the order of the moment $k$ as: $k = \sum_{i=1}^{N} k_i$. Similarly, a joint central moment of order $k$ is defined as:

$$\mu_{k_1,\ldots,k_n} = \mathbb{E}\left[(X_1-\eta_{k_1})^{k_1}\cdot\ldots\cdot(X_N-\eta_{k_N})^{k_N}\right] = \mathbb{E}\left[\prod_{i=1}^{N}(X_i-\eta_{k_i})^{k_i}\right].$$

[34] From symmetry.

For a 2-D random vector $(X, Y)$:

$$m_{nk} = \mathbb{E}\left[X^n \cdot Y^k\right], \quad \mu_{nk} = \mathbb{E}\left[(X - \eta_X)^n \cdot (Y - \eta_Y)^k\right].$$

We may write down all the non-trivial moments up to order 2, which we refer to as *second order statistics*:

$$m_{1,0} = \eta_X, m_{0,1} = \eta_Y, m_{1,1} = \mathbb{E}[XY] = R_{XY}$$

$$\mu_{2,0} = \sigma_X^2 = \mathrm{Var}(X), \mu_{0,2} = \sigma_Y^2 = \mathrm{Var}(Y), \mu_{1,1} = \sigma_{XY} = \mathrm{Cov}(XY)$$

**Covariance:** The covariance of two random variables $X, Y$ is defined as:

$$\mathrm{Cov}(X, Y) = \sigma_{XY} = \mu_{1,1} = \mathbb{E}\left[(X - \eta_X) \cdot (Y - \eta_Y)\right] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

**Uncorrelatedness:** We say that two random variables $X, Y$ are uncorrelated if $\mathrm{Cov}(X, Y) = 0$. Uncorrelation is a weaker property than statistical independence.

Claim: If two random variables $X, Y$ are independent, then they are uncorrelated.
Proof:

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \overset{33}{=} \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0$$

Claim: If two random variables $X, Y$ are uncorrelated, then: $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$
Proof:
We recall that:

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \eta_X)^2\right] = \mathbb{E}[X^2] - \mathbb{E}^2[X] = \mathbb{E}[X^2] - \eta_X^2$$

Therefore:

$$\begin{aligned}
\mathrm{Var}(X + Y) &= \mathbb{E}\left[\left((X - \eta_X) + (Y - \eta_Y)\right)^2\right] \\
&\overset{34}{=} \mathbb{E}\left[(x - \eta_X)^2\right] + 2 \cdot \mathbb{E}\left[(X - \eta_X)(Y - \eta_Y)\right] + \mathbb{E}\left[(Y - \eta_Y)^2\right] \\
&= \mathrm{Var}(X) + 2 \cdot \mathrm{Cov}(X, Y) + \mathrm{Var}(Y) \\
&\overset{35}{=} \mathrm{Var}(X) + \mathrm{Var}(Y)
\end{aligned}$$

[35] Independence.

[36] Linearity of mean

[37] Because $X, Y$ is uncorrelated, $\mathrm{Cov}(X, Y) = 0$

**Orthogonality:** Two random variables $X, Y$ are orthogonal if $R_{XY} = \mathbb{E}[XY] = m_{1,1} = 0$.

**Remark 3** *We note that*

1. *Generally, independence don't necessarily mean orthogonality.*
2. *If a pair of random variables $X, Y$ are orthogonal, and in addition $\mathbb{E}[Y] = \eta_Y = 0$ or $\mathbb{E}[X] = \eta_X = 0$ (or both), then they are uncorrelated (we have $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0 = 0$).*
3. *If a pair of random variables $X, Y$ are uncorrelated, and in addition we know that $\mathbb{E}[X] = \eta_X = 0$ or $\mathbb{E}[Y] = \eta_Y = 0$ (or both), then they are orthogonal (we have $\mathbb{E}[XY] = \text{Cov}(X, Y) + \mathbb{E}[X]\mathbb{E}[Y] = 0 + 0 = 0$)*

Conclusion:
If $\mathbb{E}[X] = \eta_X = 0$ or $\mathbb{E}[Y] = \eta_Y = 0$, then orthogonality $\iff$ uncorrelation.

**Linear correlation coefficient (Pearson coefficient and correlation coefficient):** Given two random variables, we define

$$r(X, Y) = \frac{\mathbb{E}[X \cdot Y]}{\sqrt{\mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]}}$$

It will be shown in the sequel that $|r(X, Y)| = 1 \iff \exists\alpha : Y = \alpha X$, i.e. there exists a linear correlation that goes through the point (0,0) that describes the relationship between $X, Y$.

We further define

$$\rho(X, Y) = r(X - \eta_X, Y - \eta_Y) = \frac{\mathbb{E}[(X - \eta_X) \cdot (Y - \eta_Y)]}{\sqrt{\mathbb{E}[(X - \eta_X)^2] \cdot \mathbb{E}[(Y - \eta_Y)^2]}} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

It will be shown in the sequel that $|\rho(X, Y)| = 1 \iff \exists\alpha : (Y - \eta_Y) = \alpha(X - \eta_X)$, i.e there is a linear relationship between $X, Y$.

**Remark 4** *By definition, we have*

1. *A pair of random variables $X, Y$ are orthogonal $\iff$ $r(X, Y) = 0$ (because $\mathbb{E}[X \cdot Y] = 0$)*
2. *A pair of random variables $X, Y$ are uncorrelated $\iff$ $\rho(X, Y) = 0$ (because $\text{Cov}(X, Y) = 0$)*

Claim:
For every pair of random variables $X, Y$: $-1 \leq r(X,Y) \leq 1$ (namely $|r(X,Y)| \leq 1$)
Proof:
Look at the matrix:

$$\begin{bmatrix} \mathbb{E}[X^2] & \mathbb{E}[X \cdot Y] \\ \mathbb{E}[X \cdot Y] & \mathbb{E}[Y^2] \end{bmatrix}$$

This is a PSD matrix (Positive Semi Definite) as will be shown below. Hence, all of its eigen values $(\lambda_1, \lambda_2)$ are non-negative. therefore, the determinant of the matrix is non-negative. Hence:

$$\begin{vmatrix} \mathbb{E}[X^2] & \mathbb{E}[X \cdot Y] \\ \mathbb{E}[X \cdot Y] & \mathbb{E}[Y^2] \end{vmatrix} = \mathbb{E}[X^2] \cdot \mathbb{E}[Y^2] - \mathbb{E}^2[X \cdot Y] \geq 0$$

From this equation, we can deduce Cauchy-Schwarz Inequality:

$$\mathbb{E}^2[X \cdot Y] \leq \mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]$$

From the definition of Pearson coefficient we get:

$$r^2(X,Y) = \frac{\mathbb{E}^2[X \cdot Y]}{\mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]} \leq 1 \Rightarrow |r(X,Y)| \leq 1$$

Claim:
For every two random variables $X, Y$: $-1 \leq \rho(X,Y) \leq 1$, (namely, $|\rho(X,Y)| \leq 1$).
Proof:
We already proved $|r(X,Y)| \leq 1$ for every two random variables $X, Y$. Now we define a pair of new variables: $\begin{cases} X' = X - \eta_X \\ Y' = Y - \eta_Y \end{cases}$. And because $|r(X',Y')| \leq 1 \Rightarrow |\rho(X,Y)| \leq 1$ .

**Joint Characteristic Function (jCF):** The joint CF of a random vector $\underline{X}$ is the Fourier transform (with $-j$ relpaced by $j$) of the PDF of $\underline{X}$:

$$\Phi_{\underline{X}}\left(\underline{\omega}\right) = \int \underset{\underline{X} \in \mathbb{R}^N}{\cdots} \int e^{j\underline{\omega}^T\underline{x}} f_{\underline{X}}\left(\underline{x}\right) \underline{dx} = \mathbb{E}\left[e^{j\underline{\omega}^T\underline{x}}\right]$$

Notes:

- $\underline{\omega}, \underline{X}$ are of the same size, and the multiplication can be written as: $\underline{\omega}^T \cdot \underline{X} = \sum_{i=1}^{N} \omega_i \cdot X_i$.

- jCF is an alternate description of the statistical information about the random vector.
- The argument of the jCF will be written as a row vector or as a column vector, as needed.

Properties:

1. $\Phi_{\underline{X}}(\underline{0}) = 1$
2. $|\Phi_{\underline{X}}(\underline{\omega})| \leq 1$
3. $\forall x_i, \Phi_{X_i}$ can be calculated out of the jCF by putting $\omega_k = 0, k \neq i$:
$$\Phi_{X_i}(\omega_i) = \Phi_{\underline{X}}(0, ..., 0, \underset{i-th\ place}{\omega_i}, 0, ..., 0)$$
i.e. by inserting $\underline{\omega I}$ to the jCF, where $\underline{I}$ contains ones only in the relevant places.
4. Reversed Fourier transform:
$$f_{\underline{X}}(\underline{x}) = \frac{1}{(2\pi)^n} \int \underset{\underline{\omega} \in \mathbb{R}^n}{\cdots} \int e^{-j\underline{\omega}^T \underline{x}} \Phi_{\underline{X}}(\underline{\omega}) \underline{d\omega}$$
5. The relation between CF and moments is:
$$m_{k_1,...,k_n} = \frac{1}{j^{k_1+...+k_n}} \left[ \frac{d^{k_1+...+k_n}}{d\omega_1^{k_1} \cdot ... \cdot d\omega_n^{k_n}} \Phi_{\underline{X}}(\underline{\omega}) \right] \Bigg|_{\omega=0}$$
6. jCF is a linear function (from the linearity of integral and properties of exponent).
We will show the effect of a linear transformation on jCF.
Denote a random vector $\underline{X}$ with a jCF $\Phi_{\underline{X}}(\underline{\omega})$. Suppose we apply the linear transformation to obtain:
$$\underline{Y}_{Mx1} = \underline{\underline{A}}_{MxN} \cdot \underline{X}_{Nx1} + \underline{b}_{Mx1}.$$
Then,
$$\Phi_{\underline{Y}}(\underline{\omega}) = e^{j\underline{\omega}^T \underline{b}} \cdot \Phi_{\underline{X}}(\underline{\underline{A}}^T \underline{\omega})$$
Explanation:
$$\begin{aligned} \Phi_{\underline{Y}}(\underline{\omega}) &= \mathbb{E}\left[e^{j\underline{\omega}^T \underline{Y}}\right] = \mathbb{E}\left[e^{j\underline{\omega}^T (\underline{\underline{A}} \cdot \underline{X} + \underline{b})}\right] \overset{36}{=} e^{j\underline{\omega}^T \underline{b}} \cdot \mathbb{E}\left[e^{j\underline{\omega}^T \cdot \underline{\underline{Ax}}}\right] \\ &\overset{37}{=} e^{j\underline{\omega}^T \underline{b}} \cdot \mathbb{E}\left[e^{j(\underline{\omega}^T \cdot \underline{\underline{A}})^{T^T} \cdot \underline{X}}\right] \overset{38}{=} e^{j\underline{\omega}^T \underline{b}} \mathbb{E}\left[e^{j(\underline{\underline{A}}^T \cdot \underline{\omega})^T \cdot \underline{X}}\right] \\ &= e^{j\underline{\omega}^T \underline{b}} \cdot \Phi_{\underline{X}}\left(\underline{\underline{A}}^T \underline{\omega}\right) \end{aligned}$$

7. The random vector $(X, Y)$ is independent $\iff$ $\Phi_{XY}(\omega_1, \omega_2) = \Phi_X(\omega_1) \cdot \Phi_Y(\omega_2)$
   Proof of direction 1:

$$\begin{aligned}
\Phi_{XY}(\omega_1, \omega_2) &= \int\int e^{j(\omega_1,\omega_2)^T(x,y)} f_{XY}(x, y)dxdy \\
&\overset{39}{=} \int\int e^{j(\omega_1 x+\omega_2 y)} f_X(x)f_Y(y)dxdy \\
&= \int\int e^{j\omega_1 x}e^{j\omega_1 x}e^{j\omega_2 y} f_X(x)f_Y(y)dxdy \\
&= \int e^{j\omega_1 x} f_X(x)dx \int e^{j\omega_2 y} f_Y(y)dy \\
&= \Phi_X(\omega_1) \cdot \Phi_Y(\omega_2)
\end{aligned}$$

   Proof of direction 2: to prove independence - the same transition like the first proof.

Example of properties:

- Example for property 7:
  Denote $(X, Y)$ an independent random vector. Define two random variables: $V = h(Y)$, $U = g(X)$. We will show that $U, V$ are also independent.

$$\begin{aligned}
\Phi_{UV}(\omega_1, \omega_2) &= \mathbb{E}\left[e^{f(\omega_1,\omega_2)^T(U,V)}\right] \\
&= \int\int e^{f(\omega_1,\omega_2)^T(U,V)} f_{UV}(u, v)dudv \\
&= \int\int e^{f(\omega_1,\omega_2)^T(g(X),h(Y))} f_{XY}(x, y)dxdy \\
&\overset{40}{=} \int\int e^{f(\omega_1 g(x),\omega_2 h(y))} f_X(x)f_Y(y)dxdy \\
&= \int e^{j\omega_1 g(x)} f_X(x)dx \int e^{j\omega_2 h(y)} f_Y(y)dx \\
&= \Phi_U(\omega_1)\Phi_V(\omega_2)
\end{aligned}$$

[38] $\mathbb{E}[cX] = c\mathbb{E}[X], c = const.$

[39] Double transpose

[40] By inserting the transpose into the brackets, each of elements is being transposed and they switch order.

[41] Independence.

This means that (scalar) functions of independent random variables produce independent random variables.

- Example of property 6:
Denote $X, Y$ two independent random variables, nd $Z = X + Y$. We've seen that $f_Z = f_X * f_Y$. We will prove it again:
We write $Z$ as a linear system:
$$A = (1,1),\ b = \underline{0},\ Z = AX + b$$
$$\Phi_Z(\omega) = \Phi_{XY}(A^T \cdot \omega) = \Phi_{XY}\left((1,1)^T \cdot \omega\right) = \Phi_{XY}\left(\begin{pmatrix}\omega\\ \omega\end{pmatrix}\right) \overset{41}{=} \Phi_X(\omega) \cdot \Phi_Y(\omega)$$

**Second order statistics of a random vector:** We will see that in linear estimation and also in the characterization of a Gaussian random vector depend only on second order statistics.

For a random vector $\underline{X}$, we can express all moments of order two (the means of the random variables, their variances, and the covariances of every pair) by joint moments as we have seen:

$$m_{0,\dots,0,\underset{i-thplace}{1},0,\dots,0} = \mathbb{E}[X_i] = \eta_{X_i}$$
$$m_{0,\dots,0,\underset{i-thplace}{1},0,\dots,0,\underset{k-thplace}{1},0,\dots,0} = \mathbb{E}[X_i \cdot X_k]$$
$$\mu_{0,\dots,0,\underset{i-thplace}{2},0,\dots,0} = \mathbb{E}[(X_i - \eta_i)^2] = \sigma^2_{X_i} = \mathrm{Var}(X_i)$$
$$\mu_{0,\dots,0,\underset{i-thplace}{1},0,\dots,0,\underset{k-thplace}{1},0,\dots,0} = \mathbb{E}[(X_i - \eta_i) \cdot (X_k - \eta_k)]$$

Expressing the second order statistics in this manner is cumbersome. Therefore, it will be convenient to adopt matrix notation to succinctly represent the 2-nd order statistics of a random vector.

**The vector mean:** The mean of a random vector $\underline{X}$ represents all the moments of the 1-st order of the vector:

$$\eta_{\underline{X}} = \mathbb{E}[\underline{X}] = \begin{bmatrix}\mathbb{E}[X_1]\\ \cdot\\ \cdot\\ \cdot\\ \mathbb{E}[X_N]\end{bmatrix} = \begin{bmatrix}\eta_{X_1}\\ \cdot\\ \cdot\\ \cdot\\ \eta_{X_N}\end{bmatrix}$$

[42] See previous footnote.

[43] See previous footnote.

**The correlation matrix:** The correlation matrix of a random vector $\underline{X}$ represents all the moments of the 2-nd order of the vector:

$$R_{\underline{X},\underline{X}} = \mathbb{E}[\underline{X} \cdot \underline{X}^T] = \begin{bmatrix} \mathbb{E}[X_1^2] & \dots & \mathbb{E}[X_1 \cdot X_N] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[X_N \cdot X_1] & \dots & \mathbb{E}[X_N^2] \end{bmatrix}$$

We will notice that te multiplication $\underline{X} \cdot \underline{X}^T$ gives us a matrix of dimensions $N \times N$ and we calculate the mean of each element in the matrix separately.

**Remark 5** *Note that,*

- *This is a symmetric matrix, as* $\forall i, j : \; \mathbb{E}[X_i \cdot X_j] = \mathbb{E}[X_j \cdot X_i]$
- *We will “access” the elements in the matrix by:* $R_{\underline{X},\underline{X}}(i,j) = \mathbb{E}[X_i \cdot X_j]$
- $R_{\underline{X},\underline{X}}$ *is diagonal iff the elements of* $\underline{X}$ *are jointly orthogonal (*$\forall i \neq j, \; \mathbb{E}[X_i \cdot X_j] = 0$*).*

**The Covariance Matrix:** The covariance matrix of a random vector $\underline{X}$ represents all the centralized moments of the 2-nd order of the vector:

$$C_{\underline{X},\underline{X}} = R_{\underline{X}-\underline{\eta_X},\underline{X}-\underline{\eta_X}} = \mathbb{E}\left[(\underline{X} - \eta_{\underline{X}}) \cdot (\underline{X} - \eta_{\underline{X}})^T\right]$$

$$= \begin{bmatrix} \mathbb{E}\left[(X_1 - \eta_{X_1})^2\right] & \dots & \mathbb{E}\left[(X_1 - \eta_{X_1}) \cdot (X_N - \eta_{X_N})\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}\left[(X_N - \eta_{X_N}) \cdot (X_1 - \eta_{X_1})\right] & \dots & \mathbb{E}\left[(X_N - \eta_{X_N})^2\right] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{X_1}^2 & \dots & \mathrm{Cov}(X_1, X_N) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_N, X_1) & \dots & \sigma_{X_N}^2 \end{bmatrix}$$

**Remark 6** *Note that*

- *Of course this is also a symmetric matrix, and we will also “access” it using indexes.*
- $C_{\underline{X},\underline{X}} = R_{\underline{X},\underline{X}} - \eta_{\underline{X}} \cdot \eta_{\underline{X}}^T$
- $C_{\underline{X},\underline{X}}$ *is diagonal iff the elements of* $\underline{X}$ *are pairwise uncorrelated:* $\forall i \neq j : \; \mathrm{Cov}(X_i, X_j) = 0$

**Properties of the correlation and covariance matrices:**

1. $C_{\underline{X},\underline{X}}$, $R_{\underline{X},\underline{X}}$ are symmetric and therefore may be orthonormaly diagonolized.

2. $C_{\underline{X},\underline{X}}$, $R_{\underline{X},\underline{X}}$ are positive semi-definite matrix - we will prove this shortly.

**Symmetric matrix:** A matrix $\underline{\underline{A}}$ is a symmetric matrix of it is a square matrix and $\underline{\underline{A}}^T = \underline{\underline{A}}$.

**Property of symmetric matrices:** Any symmetric matrix $\underline{\underline{A}}$ may be orthonormaly diagonalized, i.e. there is a matrix $\underline{\underline{P}}$ such that:

- $\underline{\underline{P}} \cdot \underline{\underline{P}}^T = \underline{\underline{I}}$
- $\underline{\underline{P}}^T \cdot \underline{\underline{A}} \cdot \underline{\underline{P}} = \underline{\underline{\Omega}}$, where $\Omega$ is a diagonal matrix.

This matrix $\underline{\underline{P}}$ is composed of eigenvectors (forming its columns) of the matrix $\underline{\underline{A}}$, the matrix $\underline{\underline{\Lambda}}$ is a diagonal matrix with the eigenvalues of $\underline{\underline{A}}$ on the main diagonal (in the order of the eigenvectors in the matrix $\underline{\underline{P}}$).

$$\underline{\underline{P}} = \begin{bmatrix} \underline{v}_1 | & \dots & | \underline{v}_N \end{bmatrix}$$
$$\{\lambda_i\} : det(\underline{\underline{A}} - \lambda \underline{\underline{I}}) = 0,\ \forall i :\ ||\underline{v}_i|| = 1,\ \forall i \neq j : \underline{v}_i \cdot \underline{v}_j = 0$$
$$(\underline{\underline{A}} - \lambda \underline{\underline{I}}) \cdot \underline{v}_i = 0$$

**PSD matrix - positive semi-definite matrix:** A matrix $\underline{\underline{A}}$ is will be called PSD if $\forall \underline{v} \in \mathbb{R}^N : \underline{v}^T \cdot \underline{\underline{A}} \cdot \underline{v} \geq 0$.

**Remark 7** *We note that*

- *A PSD matrix $\underline{\underline{A}}$ is in particular symmetric and therefore diagonalizable.*
- *A PSD matrix $\underline{\underline{A}}$ has non-negative eigenvalues($\lambda_1, ..., \lambda_2 \geq 0$).*
- *If a symmetric matrix $\underline{\underline{A}}$ has only non-negative eigen values($\forall i :\ \lambda_i \geq 0$) then it is PSD.*

In other words, symmetric+non-negative eigenvalues is equivalent to PSD.

**Linearity of expectation:** Given a random matrix $\underline{\underline{X}}_{N\times M}$ and deterministic matrices $\underline{\underline{A}}_{N\times M}, \underline{\underline{B}}_{N\times M}, \underline{\underline{C}}_{N\times M}, \underline{\underline{D}}_{N\times M}$, we have:

$$\mathbb{E}[A\cdot X] = A\cdot\mathbb{E}[X],\ \mathbb{E}[X\cdot B] = \mathbb{E}[X]\cdot B,\ \mathbb{E}[X+C] = \mathbb{E}[X]+C$$

$$\Rightarrow \mathbb{E}[A\cdot X\cdot B + D] = A\cdot\mathbb{E}[X]\cdot B + D$$

We will prove one of the rules, we will prove for one element in the matrix.

$$\begin{aligned}
\left(\mathbb{E}[A\cdot X]\right)_{j,k} &= \mathbb{E}\left[(A\cdot X)_{j,k}\right] = \mathbb{E}\left[\sum_{i=1}^{N} a_{j,i}\cdot x_{i,k}\right] \\
&= \sum_{i=1}^{N}\mathbb{E}\left[a_{j,i}\cdot x_{i,k}\right] = \sum_{i=1}^{N} a_{j,i}\cdot\mathbb{E}\left[x_{i,k}\right] \\
&= \left(A\cdot\mathbb{E}[X]\right)_{j,k}
\end{aligned}$$

**Proof of PSD property of correlation and covariance matrix:** For a random vector $\underline{X}$, we know that $R_{\underline{X},\underline{X}}$ is symmetric. we would like to prove that

$$\forall \underline{b}\in\mathbb{R}^N:\ \underline{b}^T\cdot R_{\underline{X},\underline{X}}\cdot\underline{b}\geq 0$$

Define $Y = \underline{b}^T\cdot\underline{X}$. Then:

$$\begin{aligned}
0 &\overset{42}{\leq} \mathbb{E}[Y^2] = \mathbb{E}[Y\cdot Y] \overset{43}{=} \mathbb{E}[Y\cdot Y^T] = \mathbb{E}\left[\left(\underline{b}^T\cdot\underline{X}\right)\cdot\left(\underline{b}^T\cdot\underline{X}\right)^T\right] \\
&= \mathbb{E}\left[\left(\underline{b}^T\cdot\underline{X}\right)\cdot\left(\underline{X}^T\cdot\underline{b}\right)\right] = \mathbb{E}[\underline{b}^T\cdot\underline{X}\cdot\underline{X}^T\cdot\underline{b}] \\
&\overset{44}{=} \underline{b}^T\cdot\mathbb{E}[\underline{X}\cdot\underline{X}^T]\cdot\underline{b} = \underline{b}^T\cdot R_{\underline{X}\underline{X}}\cdot\underline{b}
\end{aligned}$$

$$\Rightarrow R_{\underline{X}\underline{X}}\ is\ \text{PSD}$$

We can prove in the same way that $C_{\underline{X}\underline{X}}$ is PSD by defining $Y = \underline{b}^T\cdot(\underline{X}-\eta_{\underline{X}})$. Alternatively, We can claim that the matrix $C_{\underline{X}\underline{X}}$ is a special case of $R_{\underline{X'}\underline{X'}}$ for $\underline{X'} = \underline{X}-\eta_{\underline{X}}$ and hence the properties of $R_{\underline{X}\underline{X}}$ carry over.

[44] $\forall Y\in\mathbb{R}: Y^2\geq 0$

[45] Adding transpose on a number:$\forall a\in\mathbb{R}: a^T = a$

[46] Properties of the mean of function multiplication.

**Cross-Correlation and Cross Covariance Matrices:** For a pair of random vectors $\underline{X}_{N\times 1}$, $\underline{Y}_{M\times 1}$ we define:

- Cross correlation matrix:

$$R_{\underline{X},\underline{Y}} = \mathbb{E}[\underline{X}\cdot\underline{Y}^T] = \begin{bmatrix} \mathbb{E}[X_1\cdot Y_1] & \dots & \mathbb{E}[X_1\cdot Y_M] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[X_N\cdot Y_1] & \dots & \mathbb{E}[X_N\cdot Y_M] \end{bmatrix}$$

- Cross covariance matrix:

$$C_{\underline{X},\underline{Y}} = R_{\underline{X}-\eta_{\underline{X}},\underline{Y}-\eta_{\underline{Y}}} = \mathbb{E}\left[(\underline{X}-\eta_{\underline{X}})\cdot(\underline{Y}-\eta_{\underline{Y}})^T\right] = \begin{bmatrix} \text{Cov}(X_1,Y_1) & \dots & \text{Cov}(X_1,Y_M) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_N,Y_1) & \dots & \text{Cov}(X_N,Y_M) \end{bmatrix}$$

Properties:

1.
$$R_{\underline{Y},\underline{X}} = \left(R_{\underline{X},\underline{Y}}\right)^T,\ C_{\underline{Y},\underline{X}} = \left(C_{\underline{X},\underline{Y}}\right)^T$$

2.
$$C_{\underline{X},\underline{Y}} = R_{\underline{X},\underline{Y}} - \eta_{\underline{X}}\cdot\eta_{\underline{Y}}^T$$

Notes:

- Matrices $C_{\underline{X},\underline{Y}}, R_{\underline{X},\underline{Y}}$ are not necessarily square (and therefore not necessarily symmetric), in fact their size is $[|\underline{X}|\times|\underline{Y}|]$
- This is in fact a generalization of $C_{\underline{X},\underline{X}}, R_{\underline{X},\underline{X}}$, because we can choose $\underline{X}=\underline{Y}$.
- For two random vectors $\underline{X}_{N\times 1}$, $\underline{Y}_{M\times 1}$, we can define a new random vector $\underline{Z}_{N+M\times 1}$ that is made of a concatenation of $\underline{X},\underline{Y}$.
  We can define the correlation (and covariance) matrix of Z, and it will include the next sub-matrices:

$$\underline{Z} = \begin{bmatrix}\underline{X}\\ \underline{Y}\end{bmatrix},\quad R_{\underline{Z},\underline{Z}} = \mathbb{E}[\underline{Z}\cdot\underline{Z}^T] = \mathbb{E}\left[\begin{bmatrix}\underline{X}\\ \underline{Y}\end{bmatrix}\cdot\begin{bmatrix}\underline{X} & \underline{Y}\end{bmatrix}\right] = \begin{bmatrix}R_{\underline{X},\underline{X}} & R_{\underline{X},\underline{Y}}\\ R_{\underline{Y},\underline{X}} & R_{\underline{Y},\underline{Y}}\end{bmatrix}$$

### Random Vectors in Linear Systems

**Output Statistics of a Vector Linear System:** Denote $\underline{X}_N$ a random vector, that goes through the described linear system (where $\underline{b}, \underline{\underline{A}}$ are deterministic:)

$$\underline{Y}_{M\times 1} = \underline{\underline{A}}_{M\times N} \cdot \underline{X}_{N\times 1} + \underline{b}_{M\times 1}$$

In these systems we will discuss only 2-nd order statistics(because the system is linear).
The input statistics are known, meaning we are given $C_{\underline{X},\underline{X}}$ (or $R_{\underline{X},\underline{X}}$) and the mean vector $\eta_{\underline{X}}$. We would like to know the output statistics and the cross-statistics between the input and the output (i.e. how we can express them by using the input statistics and the parameters of the system)

Claims:

1. The output mean vector is:

$$\eta_{\underline{Y}} = \mathbb{E}[\underline{Y}] = \mathbb{E}[\underline{\underline{A}} \cdot \underline{X} + \underline{b}] = \underline{\underline{A}} \cdot \eta_{\underline{X}} + \underline{b}$$

2. The output covariance matrix is:

$$C_{\underline{Y},\underline{Y}} = \underline{\underline{A}} \cdot C_{\underline{X},\underline{X}} \cdot \underline{\underline{A}}^T$$

Proof:

$$\begin{aligned} C_{\underline{Y},\underline{Y}} &= \mathbb{E}\left[(\underline{Y} - \eta_{\underline{Y}}) \cdot (\underline{Y} - \eta_{\underline{Y}})^T\right] \\ &= \mathbb{E}\left[\left[(\underline{\underline{A}} \cdot \underline{X} + \underline{b}) - (\underline{\underline{A}} \cdot \eta_{\underline{X}} + \underline{b})\right] \cdot \left[(\underline{\underline{A}} \cdot \underline{X} + \underline{b}) - (\underline{\underline{A}} \cdot \eta_{\underline{X}} + \underline{b})\right]^T\right] \\ &= \mathbb{E}\left[\left[\underline{\underline{A}} \cdot (\underline{X} - \eta_{\underline{X}})\right] \cdot \left[\underline{\underline{A}} \cdot (\underline{X} - \eta_{\underline{X}})\right]^T\right] \\ &= \mathbb{E}\left[\underline{\underline{A}} \cdot (\underline{X} - \eta_{\underline{X}}) \cdot (\underline{X} - \eta_{\underline{X}})^T \cdot \underline{\underline{A}}^T\right] \\ &= \underline{\underline{A}} \cdot C_{\underline{X},\underline{X}} \cdot \underline{\underline{A}}^T \end{aligned}$$

3. The output-input cross-covariance matrix is:

$$C_{\underline{YX}} = \underline{\underline{A}} \cdot C_{\underline{XX}}$$

Proof:

$$\begin{aligned}
C_{\underline{Y},\underline{X}} &= \mathbb{E}\left[(\underline{Y}-\eta_{\underline{Y}})\cdot(\underline{X}-\eta_{\underline{X}})^T\right] \\
&= \mathbb{E}\left[\left[(\underline{\underline{A}}\cdot\underline{X}+\underline{b})-(\underline{\underline{A}}\cdot\eta_{\underline{X}}+\underline{b})\right]\cdot\left[(\underline{X}-\eta_{\underline{X}})\right]^T\right] \\
&= \mathbb{E}\left[\left[\underline{\underline{A}}\cdot(\underline{X}-\eta_{\underline{X}})\right]\cdot\left[(\underline{X}-\eta_{\underline{X}})\right]^T\right] \\
&= \mathbb{E}\left[\underline{\underline{A}}\cdot(\underline{X}-\eta_{\underline{X}})\cdot(\underline{X}-\eta_{\underline{X}})^T\right] = \underline{\underline{A}}\cdot C_{\underline{X},\underline{X}}
\end{aligned}$$

And of course:

$$C_{\underline{XY}} = (C_{\underline{YX}})^T = (\underline{\underline{A}}\cdot C_{\underline{XX}})^T = C^T_{\underline{XX}}\cdot\underline{\underline{A}}^T = C_{\underline{XX}}\cdot\underline{\underline{A}}$$

Note:
In the same way we can calculate the correlation(and cross-correlation) matrices by adding/subtracting the mean vector:

$$R_{\underline{YY}} = \mathbb{E}[\underline{Y}\cdot\underline{Y}^T] = \mathbb{E}\left[(\underline{\underline{A}}\cdot\underline{X}+\underline{b})\cdot(\underline{\underline{A}}\cdot\underline{X}+\underline{b})^T\right] = \underline{\underline{A}}\cdot R_{\underline{XX}}\cdot\underline{\underline{A}}^T+\underline{\underline{A}}\eta_{\underline{X}}\cdot\underline{b}^T+\underline{b}\eta_{\underline{X}}\cdot\underline{\underline{A}}^T+\underline{bb}^T$$

$$R_{\underline{XY}} = \mathbb{E}[\underline{X}\cdot\underline{Y}^T] = \mathbb{E}\left[\underline{X}\cdot(\underline{\underline{A}}\cdot\underline{X}+\underline{b})^T\right] = \mathbb{E}\left[\underline{X}\cdot(\underline{X}^T\cdot\underline{\underline{A}}^T\underline{b}^T)\right] = R_{\underline{XX}}\cdot\underline{\underline{A}}^T+\eta_{\underline{X}}\cdot\underline{b}^T$$

$$R_{\underline{YX}} = \mathbb{E}[\underline{A}\cdot\underline{X}^T] = \mathbb{E}\left[(\underline{\underline{A}}\cdot\underline{X}+\underline{b})\cdot\underline{X}^T\right] = \underline{\underline{A}}\cdot R_{\underline{XX}}+\underline{b}\cdot\eta^T_{\underline{X}} = (R_{\underline{XY}})^T$$

### Gaussian Random Vector

Random vector $\underline{X}_{N\times 1}$ will be named Gaussian random vector if $\forall\underline{a}\in\mathbb{R}^N$, the random variable $Y=\underline{a}^T\cdot\underline{X}$ is a Gaussian random variable.
Namely, if for every linear combination of the vector elements ($Y=\underline{a}^T\cdot\underline{X}=\sum_{i=1}^N a_i\cdot X_i$) we will get a Gaussian random variable.
In fact, we demand that the inner product of the random vector withe every vector with the same size $\langle\underline{a},\underline{X}\rangle$, i.e. the projection of $\underline{X}$ on every vector $\underline{a}\in\mathbb{R}^N$, will be a Gaussian random variable.
For $\underline{X}$ Gaussian random variable we can say: $\underline{X}\sim N(\eta_{\underline{X}},C_{\underline{X},\underline{X}})$ (distributes according to the mean vector and the covariance matrix)

Conclusions:

1. All the elements of a Gaussian random vector are Gaussian random variables, namely for $\underline{X}_N$ Gaussian random vector, we get that $X_i$ is a Gaussian random variable for every $i=1,...,N$

2. A Gaussian random vector is closed under linear transformation, i.e. for $\underline{X}$ Gaussian random vector, $Y = \underline{\underline{A}} \cdot \underline{X} + \underline{b}$ is a Gaussian random vector.

3. jCF of Gaussian random vector $\underline{X} \sim N(\eta_{\underline{X}}, C_{\underline{X}\underline{X}})$ is given by:

$$\Phi_{\underline{X}}(\underline{\omega}) = e^{j\cdot\underline{\omega}^T\cdot\eta_{\underline{X}}-\frac{1}{2}\cdot\underline{\omega}^T C_{\underline{X}\underline{X}}\cdot\underline{\omega}}$$

4. jPDF of a Gaussian random vector $\underline{X} \sim N(\eta_{\underline{X}}, C_{\underline{X}\underline{X}})$ (if $C_{\underline{X}\underline{X}}$ is invertible, i.e. $|C_{\underline{X}\underline{X}}| \neq 0$):

$$f_{\underline{X}}(\underline{x}) = \frac{1}{\sqrt{(2\pi)^N |C_{\underline{X}\underline{X}}|}} \cdot e^{-\frac{1}{2}(\underline{x}-\eta_{\underline{x}})^T\cdot(C_{\underline{X}\underline{X}})^{-1}\cdot(\underline{x}-\eta_{\underline{x}})}$$

5. If $\underline{Z} = \begin{bmatrix} \underline{X} \\ \underline{Y} \end{bmatrix}$ is a Gaussian random vector, the random vector $\underline{Y}|\underline{X}$ is a Gaussian random vector:

$$\underline{Y}|\underline{X} \sim N(\eta_{\underline{Y}|\underline{X}}, C_{\underline{Y}(\underline{Y}|\underline{X})})$$

$$\eta_{\underline{Y}|\underline{X}=\underline{x}} = \eta_{\underline{X}\underline{Y}} + C_{\underline{Y}\underline{X}} \cdot (C_{\underline{X}\underline{X}})^{-1} \cdot (\underline{x} - \eta_{\underline{X}})$$
$$C_{\underline{Y}(\underline{Y}|\underline{X})} = C_{\underline{Y}\underline{Y}} - C_{\underline{Y}\underline{X}} \cdot (C_{\underline{X}\underline{X}})^{-1} \cdot C_{\underline{X}\underline{Y}}$$

Note:
If $\underline{X} \sim N(\eta_{\underline{X}}, C_{\underline{X}\underline{X}})$ a Gaussian random vector and we know that its elements are uncorrelated (namely $C_{\underline{X}\underline{X}}$ is a diagonal matrix) so $\underline{X}$ is a independent random vector.
Explanation:
We know that the covariance matrix is diagonal:

$$C_{\underline{X}\underline{X}} = \begin{bmatrix} \sigma^2_{X_1} & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \sigma^2_{X_N} \end{bmatrix}$$

And so, by using the jCF, we get:

$$\begin{aligned}
\Phi_{\underline{X}}(\underline{\omega}) &= e^{(j\cdot\underline{\omega}^T\cdot\eta_{\underline{X}}-\frac{1}{2}\cdot\underline{\omega}^T\cdot C_{\underline{X}\underline{X}}\cdot\underline{\omega})} \\
&= e^{(j\cdot\sum_{i=1}^{N}\omega_i\cdot\eta_{X_i}-\frac{1}{2}\cdot\underline{\omega}^T\underline{\underline{\Omega}}\cdot\underline{\omega})} \\
&= e^{(j\cdot\sum_{i=1}^{N}\omega_i\cdot\eta_{X_i}-\frac{1}{2}\cdot\sum_{i=1}^{N}\omega_i^2\cdot\sigma_{X_i}^2)} \\
&= e^{\sum_{i=1}^{N}(j\cdot\omega_i\cdot\eta_{X_i}-\frac{1}{2}\cdot\omega_i^2\cdot\sigma_{X_i}^2)} \\
&= \prod_{i=1}^{N} e^{(j\cdot\omega_i\cdot\eta_{X_i}-\frac{1}{2}\cdot\omega_i^2\cdot\sigma_{X_i}^2)} \\
&= \prod_{i=1}^{N} \Phi_{X_i}(\omega_i)
\end{aligned}$$

Conclusion:
If two general random vector $\underline{X}, \underline{Y}$ are independent they are also uncorrelated. If in addition the mean vector of at least one of them is 0 (namely $\eta_{\underline{X}}, \eta_{\underline{Y}} = 0$) than they are orthonormal. If we know that 2 uncorrelated random vectors are jointly Gaussian, then they are independent.
Note:
If we know that $X$ a Gaussian random variable, and $Y$ also a Gaussian random variable, and they are uncorrelated, it doesn't necessarily mean that they are independent.
To know information about every Gaussian element is not enough - the elements must be added into a Gaussian random vector.
i.e. if a random vector $\underline{X}$ is a vector that every one of its elements is Gaussian random variable, even uncorrelated, it doesn't necessarily mean that $\underline{X}$ is a Gaussian random vector.
But if $\underline{X}$ is a random vector that composed of independent, Gaussian random variables, then $\underline{X}$ is Gaussian random vector.

**Example:** We take 2 uncorrelated Gaussian random variables, and we will create from them a random vector that is not Gaussian.

$$X \sim N(0,1), B = \begin{cases} 1, \ w.p. \ 0.5 \\ -1, \ w.p. \ 0.5 \end{cases}, Y = B \cdot X$$

$X, B$ are independent we we get that $Y$ is a Gaussian random variable:

$$(Y|B=1) = X \sim N(0,1)$$

$$(Y|B=-1) = -X \sim N(0,1)$$

In addition, we can show that $X, Y$ ar uncorrelated:

$$\begin{aligned}\mathbb{E}[X\cdot Y] = E_B\left[\mathbb{E}[X\cdot Y|B]\right]\Pr(B=1)\cdot\mathbb{E}[X\cdot Y|B=1]\\ +\Pr(B=-1)\cdot\mathbb{E}[X\cdot Y|B=-1]\\ =\frac{1}{2}\cdot\mathbb{E}[X^2]+\frac{1}{2}\cdot\mathbb{E}[-X^2]=\frac{1}{2}\cdot\mathbb{E}[X^2]-\frac{1}{2}\cdot\mathbb{E}[X^2]=0\end{aligned}$$

$$\mathrm{Cov}(X,Y) = \mathbb{E}[X\cdot Y]-\mathbb{E}[X]\cdot\mathbb{E}[Y] = 0-0\cdot 0 = 0$$

They are clearly not independent, and they are not jointly Gaussian. We show the jPDF:

$$f_{XY}(x,y) = f_X(x)\cdot f_{Y|X}(y|X=x) = f_X(x)\cdot\left[\frac{1}{2}\cdot\delta(y-x)+\frac{1}{2}\cdot\delta(y+x)\right]$$

The conditional distribution is not a Gaussian distribution because it is not composed of single delta functions (we can look at delta as Gaussian random variable: $\delta(x_0)\sim N(x_0,0)$ i.e. a Gaussian random variable that exists only in one point)
We get a jPDF that composed of two “delta-mountains” according to the Gaussian distribution of $X$ (these “mountains” are on the lines $Y=\pm X$ and the height of the deltas are like the Gaussian) the elevation of this jPDF is not elliptical. In addition, if the jPDF was Gaussian, we could write it using Gaussian distributions.

**Whitening of a Gaussian Random Vector:** Denote Gaussian random vector $\underline{X}$ with known mean vector $\eta_{\underline{X}}$ and covariance matrix $C_{\underline{X}\underline{X}}$. We would like to linearly transform it so that the output will be with the same dimension with $\underline{0}$ mean vector and diagonal covariance matrix,namely uncorrelated elements (and because this is Gaussian random vector, they are also independent).
The process will go as follows:

1. We reduce from $\underline{X}$ its mean vector $\eta_{\underline{X}}$, namely we define: $\underline{X'}=\underline{X}-\eta_{\underline{X}}$. That way we get a Gaussian random vector with mean vector $\eta_{\underline{X'}}=\underline{0}$ and covariance matrix: $C_{\underline{X'}\underline{X'}}=C_{\underline{X}\underline{X}}$

2. Because the covariance matrix is PSD, we know that exists a diagonalizing orthonormal matrix $\underline{\underline{P}}$ and diagonal matrix $\underline{\underline{\Lambda}}$ such that $\underline{\underline{P}}^T\cdot C_{\underline{X}\underline{X}}\cdot\underline{\underline{P}}=\underline{\underline{\Lambda}}$ ($\underline{\underline{P}}$ is orthonormaly diagonalize $C_{\underline{X'}\underline{X'}}=C_{\underline{X}\underline{X}}$ and $\underline{\underline{\Lambda}}$ contains

the eigen values of $C_{\underline{X}\underline{X}}$)
Hence we apply on $\underline{X'}$ the transformation $\underline{\underline{P}}^T$ and denote the output as $\underline{Y}$: $C_{\underline{Y}\underline{Y}} = \underline{\underline{P}}^T \cdot C_{\underline{X}\underline{X}} \cdot \underline{\underline{P}} = \underline{\underline{\Lambda}}$. Because the mean vector of $\underline{X'}$ is 0, we get $\eta_{\underline{Y}} = 0$ (Properties of random vectors in a linear system).

If in addition we want the elements of the output vector will be i.i.d, we in fact would like to crate a diagonal covariance matrix with identical elements in the diagonal, for simplicity we would demand the covariance of output to be the unitary matrix: $\underline{\underline{I}}$ .
To do so we will apply on $\underline{Y}$ the transformation $(\underline{\underline{\Lambda}}^{0.5})^T = \underline{\underline{\Lambda}}^{-0.5}$ and denote the output as $\underline{Z}$. We notice we get:

$$C_{\underline{Z}\underline{Z}} = \underline{\underline{\Lambda}}^{-0.5} \cdot C_{\underline{Y}\underline{Y}} \cdot \underline{\underline{\Lambda}}^{0.5} = \underline{\underline{\Lambda}}^{-0.5} \cdot \underline{\underline{\Lambda}} \cdot \underline{\underline{\Lambda}}^{0.5} = \underline{\underline{I}}$$

**Creating/Coloring a Gaussian Random Vector:** We would like to “create” a Gaussian random vector $\underline{X} \sim N(\eta_{\underline{X}}, C_{\underline{X}\underline{X}})$. To do that we start with a Gaussian random vector $\underline{Z} \sim N(\underline{0}, \underline{\underline{I}})$ with the same dimension, i.e. Gaussian random vector i.i.d with a mean 0 and variance of 1. We will do the process in the opposite way ($\{\lambda_1, ..., \lambda_N\}$ are eigen values of $C_{\underline{X}\underline{X}}$)

**Geometric image of Creating/Coloring Gaussian Random Variable Process:** As remembered, the jPDF of a Gaussian random vector is denoted as:

$$f_{\underline{X}}(\underline{x}) = \frac{1}{\sqrt{(2\pi)^N |C_{\underline{X}\underline{X}}|}} \cdot e^{-\frac{1}{2}(\underline{x}-\eta_{\underline{x}})^T \cdot (C_{\underline{X}\underline{X}})^{-1} \cdot (\underline{x}-\eta_{\underline{x}})}$$

In order to visualize the steps of the process, we would like to observe the elevate of the jPDF in different stages( elevate of a certain constant is a group of vectors ($\{\underline{X} | f_{\underline{X}}(\underline{x}) = const.\}$). From the jPDF we can demand that only the argument of the exponent will stay constant, i.e. demand that $(\underline{x} - \eta_{\underline{X}})^T \cdot (C_{\underline{X}\underline{X}})^{-1} \cdot (\underline{x} - \eta_{\underline{X}}) = const.$

We will recognize that the demand is similar to $\underline{X}^T \cdot \underline{\underline{A}} \cdot \underline{X} = const.$ . This pattern is a square matrix pattern. For a square matrix pattern with a PSD matrix $\underline{\underline{A}}$ the solution of the equation is an ellipsoid with the dimensions of $\underline{\underline{A}}$.
For simplicity, we will look at the process for 2D, the ellipsoid is becoming a 2D ellipsoid. we will show elevate map for the different stages:

Explanation:

- For $\underline{Z} = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$ we discuss two i.i.d Gaussian random variables, with distribution of $Z_1, Z_2 \sim N(0,1)$
  The demands of elevation comes from the jPDF is $Z_1^2 + Z_2^2 = const.$, namely a circle around the origin.

- For $\underline{Y} = \underline{\underline{\Lambda}}^{0.5} \cdot \underline{Z} = \begin{bmatrix} \sqrt{\lambda_1} \cdot Z_1 \\ \sqrt{\lambda_2} \cdot Z_2 \end{bmatrix}$ we will get the elevation equation $\frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} = cont.$, meaning an vertical ellipsoid (with axes (0,1),(1,0)) around the origin (in the drawing, we assumed $\lambda_2 \geq \lambda_1$)

- For $\underline{X'} = \underline{\underline{P}} \cdot \underline{Y} = [\underline{v_1}, \underline{v_2}] \cdot \underline{Y}$ (where $\underline{v_1}, \underline{v_2}$ are the eigen vectors of $C_{\underline{XX}}$) we get rotating mapping (we know that the eigen vectors create a new unitary basis) where $T\left(\underline{y} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) = \underline{v_2}$, $T\left(\underline{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) = \underline{v_1}$
  This transforming mapping is called orthonormal transformation (can create also negative sign)

- For $\underline{X} = \underline{X'} + \eta_{\underline{X}} = \begin{bmatrix} {X'}_1 + \eta_{X_1} \\ {X'}_2 + \eta_{X_2} \end{bmatrix}$ we get constant shift $\begin{bmatrix} \eta_{X_1} \\ \eta_{X_2} \end{bmatrix}$.

Proves from the conclusion of the definition:

1. The immediate conclusion of the definition of Gaussian random variable if we choose unit vector in one of the coordinates, for example for $\underline{a} = \underline{e}_1 = \left(0, ..., 0, \underset{i-th\ place}{1}, 0, ..., 0\right)$, we get $Y = \underline{a} \cdot \underline{Xx}$, i.e. $X_i$ is a Gaussian random variable.

2. We check the demand of the definition for random vector $\underline{Y}$:

   $$Z = \langle \underline{a}, \underline{Y} \rangle = \underline{a}^T \cdot \underline{Y} = \underline{a}^T \cdot (\underline{\underline{A}} \cdot \underline{X} + \underline{b}) = \underline{a}^T \cdot \underline{\underline{A}} \cdot \underline{X} + \underline{a}^T \cdot \underline{b} \overset{45}{=} \underline{a}'^T \cdot \underline{X} + \underline{a}^T \cdot \underline{b}$$

   And we get that $\underline{Y}$ is Gaussian random variable (because $\underline{a}'^T \cdot \underline{X}$ is a Gaussian random variable because $\underline{X}$ Gaussian random vector and a linear combination of its elements is a a Gaussian random variable and adding the constant $\underline{a}^T \cdot \underline{b}$ doesn't matter.)

3. For $\underline{\omega} \in \mathbb{R}^N$ we will look at the random variable $Y = \underline{\omega} \cdot \underline{X}$. Because $\underline{X}$ is a Gaussian random vector, $Y$ is a Gaussian random variable.
   CF of a Gaussian random variable is given as:

   $$\Phi_Y(u) = \mathbb{E}[e^{juY}] = e^{j \cdot u \cdot \eta_Y - \frac{1}{2} \cdot u^2 \cdot \sigma_Y^2}$$

[47] $a'^T = a^T \cdot \underline{\underline{A}}$

From the connection between the CF of $\underline{X}$ and of $Y$ we deduce:

$$\Phi_{\underline{X}}(\underline{\omega}) = \mathbb{E}[e^{j\cdot\underline{\omega}^T\cdot\underline{X}}] = \mathbb{E}[e^{j\cdot Y}] = \Phi_Y(1) = e^{j\cdot\eta_Y - \frac{1}{2}\sigma_Y^2}$$

We know (from the transformation) that:

$$\eta_Y = \underline{\omega}^T \cdot \eta_{\underline{X}},\ \sigma_Y^2 = \underline{\omega}^T \cdot C_{\underline{XX}} \cdot \underline{\omega}$$

After replacing in the equation we get the needed expression

4. **First way:** We can prove by inverse transformation of the jCF $\Phi_{\underline{X}}(\underline{\omega})$
**Second way:** First step - we assume $\underline{X}$ Gaussian uncorrelated random vector (hence independent). we are talking about the trivial case:

$$\begin{aligned}
f_{\underline{X}}(\underline{x}) &= \prod_{i=1}^{N} f_{X_i}(x_i) = \prod_{i=1}^{N}\left(\frac{1}{\sqrt{2\pi\cdot\sigma_{X_i}}}e^{-\frac{1}{2}\cdot\frac{(x_i-\eta_{X_i})^2}{\sigma_{X_i}^2}}\right)\\
&= \prod_{i=1}^{N}\left(\frac{1}{\sqrt{2\pi\cdot\sigma_{X_i}}}e^{-\frac{1}{2}\cdot(x_i-\eta_{X_i})^T\cdot\frac{1}{\sigma_{X_i}}\cdot(x_i-\eta_{X_i})}\right)\\
&= \frac{1}{\sqrt{(2\pi)^N|C_{\underline{XX}}|}}\cdot exp\left(-\frac{1}{2}\cdot(\underline{x}-\eta_{\underline{X}})^T\cdot\begin{bmatrix}\frac{1}{\sigma_{X_1}^2} & 0 & \dots & 0\\ 0 & \ddots & & \vdots\\ \vdots & & \ddots & 0\\ 0 & \dots & 0 & \frac{1}{\sigma_{X_N}^2}\end{bmatrix}\cdot(\underline{x}-\eta_{\underline{X}})\right)\\
&= \frac{1}{\sqrt{(2\pi)^N|C_{\underline{XX}}|}}\cdot e^{-\frac{1}{2}(\underline{x}-\eta_{\underline{X}})^T\cdot(C_{\underline{\underline{XX}}})^{-1}\cdot(\underline{x}-\eta_{\underline{X}})}
\end{aligned}$$

Second step - We assume $\underline{X}$ general Gaussian random vector.
We would like to transform $\underline{X}$ in such a way, that it will become uncorrelated/independent, Namely we would like to white the Gaussian random vector $\underline{X} \sim N(\eta_{\underline{X}}, C_{\underline{XX}})$: $\underline{Y} = T(\underline{X}) \sim N(0, \underline{\underline{\Lambda}})$ (in fact, the demand that the mean will be zero is not necessary as we have seen in step 1). The transformation that matches general Gaussian random vector whitening is $\underline{Y} = \underline{\underline{P}}^T \cdot (\underline{X} - \eta_{\underline{X}})$. After the whitening process,$\underline{\underline{\Lambda}}$ is a diagonal matrix that contains the eigen values of $C_{\underline{XX}}$.
We know that the jPDF of $\underline{Y}$ is given as:

$$f_{\underline{Y}}(\underline{y}) = \frac{1}{\sqrt{(2\pi)^N|C_{\underline{XX}}|}}\cdot e^{-\frac{1}{2}(\underline{y}-\eta_{\underline{Y}})^T\cdot(C_{\underline{YY}})^{-1}\cdot(\underline{y}-\eta_{\underline{Y}})}$$

We examine the jPDF of $\underline{X}$ (using the connection between the jPDFs after transformation):

$$\begin{aligned}
f_{\underline{X}}(\underline{x}) &\overset{46}{=} \frac{1}{|\underline{\underline{P}}^T|} \cdot f_{\underline{Y}}\left(\underline{\underline{P}}^T \cdot (\underline{x} - \eta_{\underline{X}})\right) \\
&= \frac{1}{\sqrt{(2\pi)^N |C_{\underline{YY}}|}} \cdot exp\left(-\frac{1}{2}\left(\underline{\underline{P}}^T \cdot (\underline{x} - \eta_{\underline{X}})\right)^T \cdot (C_{\underline{YY}})^{-1} \cdot \left(\underline{\underline{P}}^T \cdot (\underline{x} - \eta_{\underline{X}})\right)\right) \\
&\overset{47}{=} \frac{1}{\sqrt{(2\pi)^N |C_{\underline{YY}}|}} \cdot exp\left(-\frac{1}{2}\left(\underline{x} - \eta_{\underline{X}}\right)^T \cdot \left(\underline{\underline{P}} \cdot (C_{\underline{YY}})^{-1} \cdot \underline{\underline{P}}^T\right)\left(\underline{x} - \eta_{\underline{X}}\right)\right) \\
&\overset{48}{=} \frac{1}{\sqrt{(2\pi)^N |C_{\underline{YY}}|}} \cdot exp\left(-\frac{1}{2}\left(\underline{x} - \eta_{\underline{X}}\right)^T \cdot (C_{\underline{YY}})^{-1}\left(\underline{x} - \eta_{\underline{X}}\right)\right)
\end{aligned}$$

5. We don’t prove it here,it follows from the properties of 2-nd order statistics. We will prove it in the estimation chapter.

## Estimation

Consider two random variables $X, Y$ (or a random vector $(X, Y)$) where we observe $Y$ but are interested in $X$. Thus, given $Y$, we would like to estimate, that is “guess”, the value of $X$, the desired variable.

**Example:** We transmit an information signal $X$ with a known distribution, for example, $X$ is a discrete signal:

$$X = \begin{cases} 1, \; w.p. \; 0.5 \\ -1, \; w.p. \; 0.5 \end{cases}$$

over a channel corrupted by additive Gaussian noise $Z \sim N(0,1)$. We receive the signal $Y = X + Z$ and we would like to estimate what signal $X$ was transmitted. We may estimate $X$ given $Y$ using a function, which we denote as $\hat{X} = g(y)$.

### Error Criteria

We would like to get $X = \hat{X}$, but we can’t always guaranty that. It is possible only if the connection between $X$ and $Y$ is one-to-one. We have to compromise and define estimation error.

[48] $\underline{a}'^T = \underline{a}^T \cdot \underline{\underline{A}}$

[49] using the connection $\underline{Y} = \underline{\underline{A}} \cdot \underline{X} + \underline{b} \underset{|\underline{\underline{A}}| \neq 0}{\Rightarrow} f_{\underline{Y}}(\underline{y}) = \frac{1}{|\underline{\underline{A}}|} \cdot f_{\underline{X}}\left(\underline{\underline{A}}^{-1} \cdot (\underline{y} - \underline{b})\right)$

[50] $|C_{\underline{YY}}| = |C_{\underline{XX}}|$, because $C_{\underline{YY}} = \underline{\underline{P}}^T \cdot C_{\underline{XX}} \cdot \underline{\underline{P}}$, and $\underline{\underline{P}}$ is diagonal unitary matrix, therefore its eigen values are 1, hence $|\underline{\underline{P}}| = 1$

**Estimation Error:** We define the estimation error to be the difference between the desired r.v. and its estimation:

$$e = X - \hat{X} = X - g(Y).$$

Note that as both $X$ and $\hat{X}$ are random variables, the estimation error $e$ is also a random variable.

**Distortion Measure:** We would like to design a system $g(\cdot)$ so as to attain a small estimation error. But as $e$ is a random variable, it is not clear how one should quantify its magnitude. We may do so by defining a distortion measure. A function $d(e)$ will be called a distortion measure if $\forall e,\ d(e) \geq 0$ and $d(0) = 0$. Notice that since $e$ is a random variable, $d(e)$ is also a random variable.

Of course, there is no “right” distortion measure that will be appropriate for all systems. Rather, the choice of distortion measure is dependent on the particular system being considered.

**Examples:**

1. 0/1 measure (for discrete random variable only), also named “probability of error measure”

$$d(e) = \begin{cases} 0, & e = 0 \\ 1, & e \neq 0 \end{cases}$$

2. square error measure

$$d(e) = e^2$$

3. absolute error measure

$$d(e) = |e|$$

**Average Distortion:** As said, also $d(e)$ is a random variable. We define the average distortion (or the expected distortion) as:

$$D = \mathbb{E}\left[d(e)\right] = \mathbb{E}\left[d(X - \hat{X})\right] = \mathbb{E}\left[d(X - g(Y))\right].$$

The latter is a deterministic quantity which assigns a single number as a measure of the ”goodness” of estimation.

Note: Usually, we design the estimator under the assumption that it will be applied to many realizations of the pair of random variables $(X, Y)$.

Hence, the law of large numbers will apply and the performance of the system (i.e., the empirical average distortion) will approach its expected value. This justifies the design of the estimator so as to minimize the expected distortion. We will discuss this further when the notion of ergodicity is introdued.

Goal: Design a system $g(\cdot)$ that will minimize $D$. Note that the optimal such system $g(\cdot)$ implicitly depends on the choice of distortion measure $d(e)$.

### Optimal Estimation

Goal: Given a distortion measure $d(e)$, we would like to find a system $g(\cdot)$ that will minimize $D$:

$$D = \mathbb{E}\left[d\left(X - g(Y)\right)\right] \to \min$$

Solution: We choose a system such that for each value $Y = y$ "guesses" $\alpha$ that minimizes $D$, i.e.,

$$g^{d}_{opt}(Y = y) = \arg\min_{\alpha} \left(\mathbb{E}\left[d(X - \alpha) \mid Y = y\right]\right).$$

Note: Notice that given $y$ $\mathbb{E}\left[d(X - \alpha) \mid Y = y\right]$ is a function of $\alpha$ only.
Proof:

$$D = \mathbb{E}\left[d(X - g(Y))\right] \overset{49}{=} E_Y\left[\mathbb{E}\left[d(X - g(Y)) \mid Y = y\right]\right] = E_Y\left[h(\alpha, Y)\right]$$

$$D = \int f_Y(y) \left(\int d(X - g(Y)) f_{X|Y}(x, y) dx\right) dy$$

We would like to minimize $D$ as a function of $\alpha$, so we pick:

$$g^{d(\alpha)}_{opt}(Y = y) = \alpha^* = \arg\min_{\alpha} \left(\mathbb{E}\left[d(X - \alpha) \mid Y = y\right]\right)$$

i.e. look through all the options and pick the best solution.

**Example:** For a discrete random variable $X$, we may choose the distortion measure to be the "probability of error measure":

$$d(e) = \begin{cases} 0, & e = 0 \\ 1, & e \neq 0 \end{cases}$$

[51] From the law of total expectation, and definition of $h(\alpha, Y) = \mathbb{E}\left[d(X - g(Y)) \mid Y = y\right]$

If we do so, we obtain:

$$\begin{aligned} D = \mathbb{E}[d(e)] &= \Pr(e=0)\cdot e + \Pr(e=1)\cdot e \\ &= \Pr(e=0)\cdot 0 + \Pr(e=1)\cdot 1 \\ &= \Pr(e=1) = \Pr\left(\hat{X} \neq X\right) \end{aligned}$$

From the latter, the measure takes its name, as $D$ is the probability of error (Pe).

### Optimal Estimator under the Probability of Error Measure

**Claim:** The optimal estimator under the 0/1 (probability of error) measure is given by:

$$X_{opt}^{Pe}(Y=y) = g_{opt}^{Pe}(Y=y) = \arg\max_{\alpha} \Pr(X=\alpha|Y=y).$$

This solution is called the MAP (maximum a posteriori probability) rule, because we take the estimated value to be the one with maximum probability after "seeing" $Y$.
<u>Proof</u>:

$$\begin{aligned} X_{opt}^{Pe}(Y=y) &= g_{opt}^{Pe}(Y=y) \\ &= \arg\min_{\alpha} \left(\mathbb{E}\left[d(X-\alpha) \mid Y=y\right]\right) \\ &\overset{50}{=} \arg\min_{\alpha} \left(\Pr\left(X \neq \alpha \mid Y=y\right)\right) \\ &= \arg\max_{\alpha} \left(\Pr\left(X = \alpha \mid Y=y\right)\right) \end{aligned}$$

**Example:** Consider the discrete (binary) signal $X$,

$$X = \begin{cases} 1, & w.p.\ p \\ -1, & w.p.\ 1-p \end{cases},$$

corrupted by additive Gaussian noise $N \sim N(0,\sigma^2)$, so that the received signal is $Y = X + N$,

[52] $\mathbb{E}\left[d(X-\alpha) \mid Y=y\right] = \Pr\left(d(X-\alpha)=0|Y=y\right)\cdot d(X-\alpha) + \Pr\left(d(X-\alpha)\neq 0|Y=y\right)\cdot d(X-\alpha) = \Pr\left(d(X-\alpha)=0|Y=y\right)\cdot 0 + \Pr\left(d(X-\alpha)\neq 0|Y=y\right)\cdot 1 = \Pr\left(d(X-\alpha)\neq 0|Y=y\right) = \Pr\left(X-\alpha \neq 0|Y=y\right) = \Pr\left(X\neq\alpha|Y=y\right)$

If we use the 0/1 (Pe) measure,

$$d(e) = \begin{cases} 0, & e = 0 \\ 1, & e \neq 0 \end{cases},$$

then intuitively:

- if $p = 0.5$ then we should decide using a threshold "0": if $Y \geq 0$ we guess $X = 1$, else we guess $X = -1$.
- if $p > 0.5$, then we should prioritize $X = 1$ and decide using a negative threshold.

More formally and explicitly, we decide optimally using the MAP rule:

$$\hat{X}^{Pe}_{opt}(Y = y) = g^{Pe}_{opt}(Y = y) = \arg\max_{\alpha} \ \Pr(X = \alpha | Y = y)$$

Namely, for each value $Y = y$ we calculate the conditional probability $\Pr(X = 1|Y = y), \Pr(X = -1|Y = y)$ and choose the more probable one. In order to carry out this calculation, we need to find the these two probabilities explicitly.

We identify that the random variable $Y|X = x$ is Gaussian because the noise $N$ is a Gaussian noise, and given $X$ we are just shifting it. Hence:

$$Y|X = x \sim N(x, \sigma^2) \Rightarrow \begin{cases} Y|X = 1 \sim N(1, \sigma^2) \\ Y|X = -1 \sim N(-1, \sigma^2) \end{cases}$$

In order to find the conditional distribution of $X$ given $Y$, we next use Bayes' theorem:

$$\begin{aligned} \Pr(X = 1|Y = y) &\overset{51}{=} \frac{f_{Y|X=1}(y|1) \cdot \Pr(x = 1)}{f_Y(y)} \\ &= \frac{f_N(y-1) \cdot p}{p \cdot f_N(y-1) + (1-p) \cdot f_N(y+1)} \\ &= \frac{p \cdot e^{-\frac{(y-1)^2}{2\sigma^2}}}{p \cdot e^{-\frac{(y-1)^2}{2\sigma^2}} + (1-p) \cdot e^{-\frac{(y+1)^2}{2\sigma^2}}} \end{aligned}$$

$$\begin{aligned} \Pr(X = -1|Y = y) &\overset{52}{=} \frac{f_{Y|X=-1}(y|-1) \cdot \Pr(x = -1)}{f_Y(y)} = \ldots \\ &= \frac{(1-p) \cdot e^{-\frac{(y+1)^2}{2\sigma^2}}}{p \cdot e^{-\frac{(y-1)^2}{2\sigma^2}} + (1-p) \cdot e^{-\frac{(y+1)^2}{2\sigma^2}}} \end{aligned}$$

We want to find the said threshold, i.e. check for each $y$ we get $\Pr(X = 1|Y = y) > \Pr(X = -1|Y = y)$ and decide $\hat{X}^{Pe}_{opt} = 1$, and for each $y$ we get $\Pr(X = 1|Y = y) < \Pr(X = -1|Y = y)$ and decide $\hat{X}^{Pe}_{opt} = -1$. In other words, we find the threshold by considering the following equivalent relations:

$$\begin{aligned}
\Pr(X = 1|Y = y) &\overset{1}{\underset{-1}{><}} \Pr(X = -1|Y = y) \\
p \cdot e^{-\frac{(y-1)^2}{2\sigma^2}} &\overset{1}{\underset{-1}{><}} (1-p) \cdot e^{-\frac{(y+1)^2}{2\sigma^2}} \\
e^{-\frac{(y-1)^2}{2\sigma^2}+\frac{(y+1)^2}{2\sigma^2}} &\overset{1}{\underset{-1}{><}} \frac{1-p}{p} \\
e^{\frac{4y}{2\sigma^2}} &\overset{1}{\underset{-1}{><}} \frac{1-p}{p}
\end{aligned}$$

Because the exponential function is a monotonically increasing function, we can apply $\ln(\cdot)$ to both sides of the equation to obtain:

$$\begin{aligned}
\frac{4y}{2\sigma^2} &\overset{1}{\underset{-1}{><}} \ln\left(\frac{1-p}{p}\right) \\
y &\overset{1}{\underset{-1}{><}} \frac{1}{2} \cdot \sigma^2 \cdot \ln(\frac{1-p}{p}) = T(p)
\end{aligned}$$

And in a graphical presentation:

### Minimum Mean Square Error (MMSE) Estimation

We have seen an optimal estimation using Pe, Namely an estimation the minimize the chances of error.
This estimation has some disadvantages:

1. The distortion measure is binary, hence it belongs only when we use discrete random variables.

2. There is no control over the size of the error when it exists, and we just minimize the probability of the error.

3. The system is operating using the MAP rule, i.e. an optimal value is calculated using the received value.

[53] Bayes' theorem

[54] See previous footnote.

If we use the distortion measure of the square error $d(e) = e^2$, then the mean of the distortion is defined as:

$$D = \mathbb{E}[d(e)] = \mathbb{E}\left[\left(\hat{X} - X\right)^2\right]$$

This is the mean square error (MSE) measure. We would like to find optimal MSE estimation, i.e. we would like to find a system $g(\cdot)$ that minimize the mean of the distortion (minimum MSE- MMSE):

$$D = \mathbb{E}[d(e)] = \mathbb{E}\left[\left(\hat{X} - X\right)^2\right] \rightarrow \ \min$$

We notice that the mean of distortion represents the power of the error.

**Optimal Estimator Under MSE Index:** Claim:
The optimal estimator under MSE measure is:

$$\hat{X}^{MSE}_{opt}(Y = y) = g^{MSE}_{opt}(Y = y) = \mathbb{E}\left[X \mid Y = y\right]$$

Namely the value of the optimal estimator of $X$ for a certain $Y$ is the conditional mean of $X$ given the same Y.
Proof:

$$\begin{aligned}\hat{X}^{MSE}_{opt}(Y = y) &= g^{MSE}_{opt}(Y = y)\\ &= \arg\min_{\alpha}\left(\mathbb{E}\left[d(X-\alpha) \mid Y = y\right]\right)\\ &= \arg\min_{\alpha}\left(\mathbb{E}\left[(X-\alpha)^2 \mid Y = y\right]\right)\end{aligned}$$

again, the argument $\arg\min_{\alpha}$:

$$h(\alpha, Y) = \mathbb{E}\left[d(X - g(Y)) \mid Y = y\right] = \mathbb{E}\left[(X-\alpha)^2 \mid Y = y\right] = \int_{-\infty}^{\infty}(x-\alpha)^2 \cdot f_{X|Y}(x|y)dx$$

We take the partial derivative by $\alpha$ and compare to zero to find the extremum point (because this is a quadratic equation of $X$ we can see that this is an upside parabola, hence it is a minimum point)

$$\frac{\partial}{\partial\alpha}h(\alpha, Y) = -2 \cdot \int_{-\infty}^{\infty}(x-\alpha) \cdot f_{X|Y}(x|y)dx \Rightarrow -2 \cdot \int_{-\infty}^{\infty}(x-\alpha^*) \cdot f_{X|Y}(x|y)dx = 0$$

$$\int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y)dx = \int_{-\infty}^{\infty} \alpha^* \cdot f_{X|Y}(x|y)dx$$

$$\int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y)dx \overset{53}{=} \alpha^* \cdot \int_{-\infty}^{\infty} f_{X|Y}(x|y)dx$$

$$\mathbb{E}\left[X \mid Y = y\right] \overset{54}{=} \alpha^* \cdot 1 = \alpha^*$$

**Example:** Discrete signal $X$ and noise $N$ received: $Y = X + N$ ($X, N$ are independent)

$$X = \begin{cases} 1, \ w.p. \ p \\ -1, \ w.p. \ 1-p \end{cases}, \ N \sim N(0, \sigma^2), \ d(e) = e^2, \ Y = X + N$$

If we use the square error measure (as usual for differential $e$):

$$\begin{aligned}
\hat{X}_{opt}^{MSE}(Y=y) &= g_{opt}^{MSE}(Y=y) = \mathbb{E}\left[X \mid Y=y\right] \\
&= 1 \cdot Pr\left([X=1 \mid Y=y]\right) + (-1) \cdot Pr\left([X=-1 \mid Y=y]\right) \\
&= Pr\left([X=1 \mid Y=y]\right) - Pr\left([X=-1 \mid Y=y]\right) \\
&= \frac{p \cdot e^{-\frac{(y-1)^2}{2\sigma^2}}}{p \cdot e^{-\frac{(y-1)^2}{2\sigma^2}} + (1-p) \cdot e^{-\frac{(y+1)^2}{2\sigma^2}}} - \frac{(1-p) \cdot e^{-\frac{(y+1)^2}{2\sigma^2}}}{p \cdot e^{-\frac{(y-1)^2}{2\sigma^2}} + (1-p) \cdot e^{-\frac{(y+1)^2}{2\sigma^2}}} \\
&= \frac{p \cdot e^{-\frac{(y-1)^2}{2\sigma^2}} - (1-p) \cdot e^{-\frac{(y+1)^2}{2\sigma^2}}}{p \cdot e^{-\frac{(y-1)^2}{2\sigma^2}} + (1-p) \cdot e^{-\frac{(y+1)^2}{2\sigma^2}}} = \cdots = \\
&= \frac{p \cdot e^{\frac{y}{\sigma^2}} - (1-p) \cdot e^{-\frac{y}{\sigma^2}}}{p \cdot e^{\frac{y}{\sigma^2}} + (1-p) \cdot e^{-\frac{y}{\sigma^2}}}
\end{aligned}$$

If $p = 0.5$ we get the system/ estimator:

$$\hat{X}_{opt}^{MSE}(Y=y) = g_{opt}^{MSE}(Y=y) \overset{55}{=} \frac{e^{\frac{y}{\sigma^2}} - e^{-\frac{y}{\sigma^2}}}{e^{\frac{y}{\sigma^2}} + e^{-\frac{y}{\sigma^2}}} = \tanh\left(\frac{y}{\sigma^2}\right)$$

We will show the decision rule graphically, when $P = 0.5$ for different values of $\sigma$:

We notice that the decision rule using the Pe measure was less smoother then the decision rule using the MSE measure.
For $\sigma \to 0$ (or even $\sigma \ll 1$) we converge to the decision rule of Pe. This makes sense, because if the noise has no variance it mean it converges around its mean, which is 0, and if there is no noise the decision is very easy around the zero. Another way to look at it is to take a very little $\sigma$, then the argument is big and we are “thrown into the edges of the graph”. For $\sigma \ll 1$ the graph is smoother because the noise is more scattered.

[55] We can take the constant $\alpha^*$ out of the integral on $x$.

[56] Integral on the domain $(-\infty, \infty)$ for every PDF function is 1.

[55] $p = 0.5$

**Interim Summary - Optimal Estimators using Pe and MSE Indexes:** We are looking only at differential error, i.e. $e = X - \hat{X} = X - g(Y)$. given a certain distortion measure $d(e)$, the optimal estimator:

$$g_{opt}^{d(e)}(Y = y) = \alpha^* = \underset{\alpha}{\arg\min}\ \left(\mathbb{E}\left[d(X - \alpha) \mid Y = y\right]\right)$$

For different distortion measures $d(e)$ we get different optimal systems $g_{opt}^{d(e)}(Y)$:

1. Pe measure and MAP-based system (the mean of the distortion describes the probability of the error)

$$d(e) = \begin{cases} 0, & e = 0 \\ 1, & e \neq 0 \end{cases} \Rightarrow \hat{X}_{opt}^{d(e)}(Y = y) = g_{opt}^{d(e)}(Y = y) = \underset{\alpha}{arg\max}\ \Pr(X = \alpha | Y = y),$$

$$D = \Pr(\hat{X} \neq X) = Pe$$

2. MSE and conditional mean-based systsem (the mean of the distortion describes the probability of the error)

$$d(e) = e^2 \Rightarrow \hat{X}_{opt}^{d(e)}(Y = y) = g_{opt}^{d(e)}(Y = y) = \mathbb{E}[X|Y = y],$$

$$D = \mathbb{E}\left[(\hat{X} - X)^2\right] = \mathbb{E}[e^2]$$

**Optimal Linear MSE Estimation (LMMSE):** We know how to plan an MSE optimal estimation system using:

$$\hat{X}_{opt}^{d(e)}(Y = y) = g_{opt}^{d(e)}(Y = y) = \mathbb{E}[X|Y = y]$$

Sometimes we want to plan an estimation system $\hat{X}(Y) = g(Y)$ that will be linear, namely $g(Y) = a \cdot Y + b$. Why would we want to compromise and limit the estimator to be linear if we can find the optimal estimator?

1. Simpler implementation (very easy to build a linear system)
2. The system depends (and affected by) statistics up to 2-nd order of the vector $(X, Y)$
3. A system that based only on lower order moments are more robust.

We notice from 2 that for a Gaussian random vector, the MSE optimal estimator is linear, because the only non-zero moments are up to 2-nd order. Therefore, the more the data is similar to Gaussian, the "loss" from linear estimation is smaller.

**Notations:** From now on, we will discuss only optimal or linear optimal MSE estimators. For convenience we define the next notations:

- For general MSE optimal estimator:

  $$\hat{X}_{opt}^{MSE}(Y) = \hat{X}_{opt}(Y) = \hat{X}_{MMSE}(Y)$$

  Where MMSE=Minimum Mean Squared Error.

- For linear MSE optimal estimator:

  $$\hat{X}_{opt}^{Linear\ MSE}(Y) = \hat{X}_{LMMSE}(Y) = \hat{X}_{BLE}(Y)$$

  Where BLE=Best linear estimation, and LMMSE=Linear MMSE.

**Linear MSE Optimal Estimator:** We would like a linear optimal estimating system, i.e. we want a linear function of $Y$ that will estimate $X$ optimally. We denote the linear optimal estimating system by:

$$\hat{X}_{LMMSE}(Y) = a_{LMMSE} \cdot Y + b_{LMMSE}$$

Where:

$$(a_{LMMSE}, b_{LMMSE}) = \underset{a,b}{arg\ min}\ \left(\mathbb{E}\left[(X - a \cdot Y - b) \mid Y = y\right]\right)$$

We get that:

$$(a_{LMMSE}, b_{LMMSE}) = \left(\frac{\sigma_{XY}}{\sigma_Y^2}, \eta_X - \frac{\sigma_{XY}}{\sigma_Y^2} \cdot \eta_Y\right)$$

From these parameters, a linear optimal estimating system is given as:

$$\hat{X}_{LMMSE}(Y) = \eta_X + \frac{\sigma_{XY}}{\sigma_Y^2}(Y - \eta_Y)$$

And the mean of mean square error (reminder: $\sigma_{XY} = \text{Cov}(X, Y)$):

$$\mathbb{E}[e_{LMMSE}^2] = \sigma_X^2(1 - \rho_{XY}^2) = \sigma_X^2 - \frac{\sigma_X^2 \cdot \sigma_{XY}^2}{\sigma_X^2 \cdot \sigma_Y^2} = \sigma_X^2 - \frac{\sigma_{XY}^2}{\sigma_Y^2}$$

We get that the mean of distortion is the power of the square error, that is the mean of the squared distortion:

$$D = \mathbb{E}\left[(X - \hat{X})^2\right] = \mathbb{E}[e^2] = MSE$$

Proof (based on properties of LMMSE estimators):
Thee proof to the estimator formula is in the proving of the orthogonality property of LMMSE estimator in the direction optimal → orthogonality.
We prove the expression for the mean of the squared error:

$$\begin{aligned}\mathbb{E}[e_{LMMSE}^2] &= \mathbb{E}\left[(X-\hat{X}_{LMMSE})^2\right]\\ &= \mathbb{E}[X^2] - 2\cdot\mathbb{E}[X\cdot\hat{X}_{LMMSE}] + \mathbb{E}[\hat{X}_{LMMSE}^2]\\ &= \mathbb{E}[X^2] - 2\cdot\mathbb{E}[(X-\hat{X}_{LMMSE}+\hat{X}_{LMMSE})\cdot\hat{X}_{LMMSE}] + \mathbb{E}[\hat{X}_{LMMSE}^2]\\ &= \mathbb{E}[X^2] - 2\cdot\mathbb{E}[(e_{LMMSE}+\hat{X}_{LMMSE})\cdot\hat{X}_{LMMSE}] + \mathbb{E}[\hat{X}_{LMMSE}^2]\\ &= \mathbb{E}[X^2] - 2\cdot\mathbb{E}[e_{LMMSE}\cdot\hat{X}_{LMMSE}] - 2\cdot\mathbb{E}[\hat{X}_{LMMSE}^2] + \mathbb{E}[\hat{X}_{LMMSE}^2]\\ &\overset{56}{=} \mathbb{E}[X^2] - 2\cdot 0 - \mathbb{E}[\hat{X}_{LMMSE}^2]\\ &= \mathbb{E}[X^2] - \mathbb{E}[\hat{X}_{LMMSE}^2]\end{aligned}$$

We notice that:

$$\begin{aligned}\mathrm{Var}(\hat{X}_{LMMSE}) &= \mathbb{E}\left[(\hat{X}_{LMMSE}-\eta_{LMMSE})^2\right]\\ &\overset{57}{=} \mathbb{E}\left[(\hat{X}_{LMMSE}-\eta_X)^2\right]\\ &= \mathbb{E}\left[\left((\eta_X+\frac{\sigma_{XY}}{\sigma_Y^2}(Y-\eta_Y))-\eta_X\right)^2\right]\\ &= \mathbb{E}\left[\left(\frac{\sigma_{XY}}{\sigma_Y^2}(Y-\eta_Y)\right)^2\right]\\ &= \left(\frac{\sigma_{XY}}{\sigma_Y^2}\right)^2\cdot\mathbb{E}[(Y-\eta_Y)^2]\\ &= \left(\frac{\sigma_{XY}}{\sigma_Y^2}\right)^2\cdot\sigma_Y^2 = \frac{\sigma_{XY}^2}{\sigma_Y^2}\end{aligned}$$

Hence, we get:

$$\mathbb{E}[e_{LMMSE}^2] = \sigma_X^2 - \frac{\sigma_{XY}^2}{\sigma_Y^2} = \sigma_X^2 - \frac{\sigma_X^2\cdot\sigma_{XY}^2}{\sigma_X^2\cdot\sigma_Y^2} = \sigma_X^2(1-\rho_{XY}^2)$$

**Conclusions from L/MMSE Estimators(Based on properties of L/MMSE estimators):**

[56] $e \perp \hat{X}(Y)$, will be proved later.

[57] $\mathbb{E}[X] = \mathbb{E}[\hat{X}_{LMMSE}]$

1. We know that $e = X - \hat{X}$, notice we can write $\hat{X} = X - e$.
   From the properties of L/MMSE estimators, we know that $e, X$ are orthogonal (because the error $e$ is orthogonal to every function of the received signal $Y$). In addition we know from the properties that $\mathbb{E}[e] = 0$
   .
   Therefore we can deduce that $e, X$ are uncorrelated (for L/MMSE estimations):
   $$\text{Cov}(e, \hat{X}) \triangleq \mathbb{E}\left[(e - \mathbb{E}[e]) \cdot (\hat{X} - \mathbb{E}[\hat{X}])\right] = \mathbb{E}[e \cdot \hat{X}] - \mathbb{E}[e] \cdot \mathbb{E}[\hat{X}] = 0 - 0 \cdot \mathbb{E}[\hat{X}] = 0$$

2. We have seen that for the mean of the squared error(for L/MMSE):
   $$\mathbb{E}[e^2_{LMMSE}] = \sigma^2_X(1 - \rho_{XY})$$
   Identify that for correlation coefficient $|\rho_{XY}| \to 1$, the mean of the squared error of the estimation will near zero. In addition, we know that the mean of the estimation error LMMSE is always zero. Hence we get that the error is 0:
   $$|\rho_{XY}| = 1 \Rightarrow \mathbb{E}[e^2_{LMMSE}] = 0, \ \mathbb{E}[e_{LMMSE}] = 0 \Rightarrow e \equiv 0$$
   This makes sense, because if $\rho_{XY} = \pm 1$ then there is a linear connection of stretching between $X$ and $Y$ (deterministic linear dependency). Because the estimator $\hat{X}$ is a linear function of Y then there is a linear connection of stretching between $X, \hat{X}$. In fact, because in this state $e \equiv 0$ then $\hat{X} \equiv X$.

**Properties of L/MMSE Estimator**

*For convenience,when we want to introduce a property that apply to both MMSE and LMMSE estimators, the index will be named L/MMSE. It means that the property applies to both of them, but only to one of them at a time - i.e. you cant mix them in the property or the equation.*

1. Orthogonality:
   (a) MMSE estimator
   Estimator $\hat{X}(Y) = g(Y)$ is MMSE iff the error $e = X - g(Y)$ is $e \perp h(Y)$ for all functions $h(Y)$. The error is orthogonal to every function of the measurements, namely $\mathbb{E}[e \cdot h(Y)] = 0 \ \forall h : \mathbb{R} \to \mathbb{R}$
   (b) LMMSE estimator
   Linear estimator $\hat{X}(Y) = g(Y) = a \cdot Y + b$ is LMMSE iff its

error $e = X - g(Y) = X - (a \cdot X + b)$ is orthogonal to every linear function $h(Y)$: $e \perp h(Y)$ (In fact, we will demand that $\forall c, d \in \mathbb{R}: \ e \perp (c \cdot Y + d)$). The error is orthogonal to every linear function of the measurements, namely: $\forall c, d \in \mathbb{R}$, $\mathbb{E}[e \cdot (c \cdot Y + d)] = \mathbb{E}\left[(X - (a \cdot Y + b)) \cdot (c \cdot Y + d)\right] = 0$

2. Unbiasedness
   For estimator L/MMSE, the mean of the error goes to zero. In fact, the mean of the estimator equals the mean of the estimated variable $X$

   $$\mathbb{E}[e_{L/MMSE}] = 0 \Rightarrow \mathbb{E}[X - \hat{X}_{L/MMSE}] = 0 \Rightarrow \mathbb{E}[\hat{X}_{L/MMSE}] = \mathbb{E}[X] = \eta_X$$

3. Pythagoras
   The power of the random variable we want to estimate is equal to the power of the random variable plus the error (as said before, for MSE measure the error power is D: $D = MSE$)

   $$\mathbb{E}[X^2] = \mathbb{E}[\hat{X}^2_{L/MMSE}] + [e^2_{L/MMSE}]$$

   We notice that if we use linear estimator (limit our options), the power of the error will get bigger on the account of the power of the estimator.

4. For Gaussian random vector $\begin{bmatrix} X \\ Y \end{bmatrix}$: $\hat{X}_{MMSE} = \hat{X}_{LMMSE}$, i.e. LMMSE estimator is optimal.
   If $\begin{bmatrix} X \\ Y \end{bmatrix}$ is a Gaussian random vector then the MMSE optimal estimator is linear, and: $\hat{X}_{MMSE} = \mathbb{E}[X|Y = y] \overset{58}{=} \eta_X + \frac{\sigma_{XY}}{\sigma_Y^2}(y - \eta_Y) = \hat{X}_{LMMSE}$

Proof:

1. Orthogonality:
   First direction: orthogonality ⇒ MSE optimal

   (a) MMSE estimator
   Denote estimator $\hat{X}(Y) = g(Y)$, and $e \perp h(Y)$ for all functions $h(Y)$. Assume that $\hat{X}(Y) = g(Y)$ is not optimal, and there is a better estimator $\hat{\hat{X}} = \tilde{g}(Y)$. We want to show that the power of

[58] Because the vector is a Gaussian random vector.

error of the “better” estimator is not smaller than the orthogonal one:

$$\mathbb{E}[e^2] \leq \mathbb{E}[\tilde{e}^2] \iff \mathbb{E}\left[(X-\hat{X})^2\right] \leq \mathbb{E}\left[(X-\hat{\hat{X}})^2\right]$$

Calculation the power of the error of the “better” estimator:

$$\begin{aligned}
\mathbb{E}[\tilde{e}^2] &= \mathbb{E}\left[(X-\hat{\hat{X}})^2\right] \\
&= \mathbb{E}\left[\left[(X-\hat{X})+(\hat{X}-\hat{\hat{X}})\right]^2\right] \\
&= \mathbb{E}\left[(X-\hat{X})\right] + 2\cdot\mathbb{E}\left[(X-\hat{X})\cdot(\hat{X}-\hat{\hat{X}})\right] + \mathbb{E}\left[(\hat{X}-\hat{\hat{X}})^2\right] \\
&= \mathbb{E}[e^2] + 2\cdot\mathbb{E}\left[e\cdot(\hat{X}-\hat{\hat{X}})\right] + \mathbb{E}\left[(\hat{X}-\hat{\hat{X}})^2\right] \\
&\overset{59}{=} \mathbb{E}[e^2] + \mathbb{E}\left[(\hat{X}-\hat{\hat{X}})^2\right] \overset{60}{\geq} \mathbb{E}[e^2]
\end{aligned}$$

(b) LMMSE estimator The proof is the same because $(\hat{X}-\hat{\hat{X}})$ is a linear function (a subtract of linear functions) and $\mathbb{E}\left[e\cdot(\hat{X}-\hat{\hat{X}})\right] = 0$ because of the orthogonality of $e$.

Second direction: MSE optimal ⇒ orthogonality

(a) MMSE estimator
Denote MSE optimal estimator $\hat{X}(Y) = g(Y)$. We need to show that $\forall h(Y): e \perp h(Y)$. We need to find explicit expression of the estimator $\hat{X} = g(Y)$:

$$\forall h(Y): e \perp h(Y) \iff \mathbb{E}\left[e\cdot h(Y)\right] = 0 \iff \mathbb{E}\left[(X-g(Y))\cdot h(Y)\right] = 0$$

$$\begin{aligned}
E_Y\left[E_{X|Y}\left[(X-g(y))\cdot h(y) \mid Y=y\right]\right] &= 0 \quad \forall h(Y) \\
E_Y\left[h(Y)\cdot E_{X|Y}\left[(X-g(y)) \mid Y=y\right]\right] &= 0 \quad \forall h(Y) \\
E_Y\left[h(Y)\cdot (E_{X|Y}\left[X \mid Y=y\right]-g(y))\right] &= 0 \quad \forall h(Y)
\end{aligned}$$

We would like the equation to apply for every function $h(Y)$, hence we demand (we can also develop by the definition of the mean using integrals):

$$E_{X|Y}\left[X \mid Y=y\right] - g(Y) = 0 \Rightarrow g(Y=y) = E_{X|Y}\left[X \mid Y=y\right] = \mathbb{E}[X|Y]$$

[61] $e \perp h(y)$, and both estimators are functions of $y$

[60] $\mathbb{E}\left[(\hat{X}-\hat{\hat{X}})^2\right]$ is non-negative.

And indeed we get that $g(Y)$ us denoted by the conditional variance, i.e. the definition of the MMSE estimator.

(b) LMMSE estimator
We will do the same thing for an estimator that looks like $\hat{X}(Y) = g(Y) = a \cdot Y + b$:

$$\forall c, d \in \mathbb{R} : \quad e \perp (c \cdot Y + d) \iff \mathbb{E}\left[e \cdot (c \cdot Y + d)\right] = 0$$

We notice that it is enough to demand (because they are sufficient to span the function space of $Y$):

i. $e \perp 1$ (or every other constant in $\mathbb{R}$)
ii. $e \perp (Y - \eta_Y)$

$$\begin{aligned}\mathbb{E}\left[e \cdot (c \cdot Y + d)\right] &= \mathbb{E}\left[e \cdot (c \cdot Y + c \cdot \eta_Y - c \cdot \eta_Y + d)\right]\\ &= \mathbb{E}\left[e \cdot c \cdot (Y - \eta_Y) + e \cdot (c \cdot \eta_Y + d)\right]\\ &= c \cdot \mathbb{E}\left[e \cdot (Y - \eta_Y)\right] + (c \cdot \eta_Y + d) \cdot \mathbb{E}[e \cdot 1]\\ &= c \cdot 0 + (c \cdot \eta_Y + d) \cdot 0 = 0\end{aligned}$$

Now we will see what is the meaning of both these demands:

i.

$$\begin{aligned}e \perp 1 \iff & \mathbb{E}[e \cdot 1] = 0\\ & \rightarrow \mathbb{E}[X - a \cdot Y - b] = 0\\ & \rightarrow \eta_X - a \cdot \eta_Y - b = 0\\ & \rightarrow b(a) = \eta_X - a \cdot \eta_Y\end{aligned}$$

ii.

$$\begin{aligned}e \perp (Y - \eta_Y) \iff & \mathbb{E}[e \cdot (Y - \eta_Y)] = 0\\ & \rightarrow \mathbb{E}[(X - a \cdot Y - b) \cdot (Y - \eta_Y)] = 0\\ & \overset{61}{\rightarrow} \mathbb{E}[(X - a \cdot Y - \eta_X + a \cdot \eta_Y) \cdot (Y - \eta_Y)] = 0\\ & \rightarrow \mathbb{E}[((X - \eta_X) - a \cdot (Y - eta_Y)) \cdot ((Y - \eta_Y))]\sigma_{XY}\\ & \qquad - a \cdot \sigma_Y^2 = 0\\ & \Rightarrow a = \frac{\sigma_{XY}}{\sigma_Y^2}\end{aligned}$$

[61] $b(a) = \eta_X - a \cdot \eta_Y$

Now, we put it in the original equation (and get the same $X_{LMMSE}$ as we described before):

$$\begin{aligned}\hat{X}(Y) &= a \cdot Y + b \\ &= \begin{bmatrix} a = \frac{\sigma_{XY}}{\sigma_Y^2} \\ b(a) = \eta_X - a \cdot \eta_Y \end{bmatrix} \\ &= \frac{\sigma_{XY}}{\sigma_Y^2} Y + \eta_X - \eta_Y \frac{\sigma_{XY}}{\sigma_Y^2} \\ &= \eta_X + \frac{\sigma_{XY}}{\sigma_Y^2}(Y - \eta_Y)\end{aligned}$$

2. Unbiasedness

   (a) MMSE estimator
   We will show that: $\mathbb{E}[\hat{X}_{MMSE}] = \mathbb{E}[X] = \eta_X$:

   $$\mathbb{E}[\hat{X}_{MMSE}] = E_Y[\hat{X}_{MMSE}] = E_Y[E_{X|Y}[X|Y=y]] \overset{62}{=} \mathbb{E}[X]$$

   (b) LMMSE estimator
   In this case, the proof is trivial because we demanded that $e \perp 1$ hence $\mathbb{E}[E_{LMMSE}] = 0$

3. Pythagoras

   $$\begin{aligned}\mathbb{E}[X^2] &= \mathbb{E}\left[\left((X - \hat{X}_{L/MMSE}) + \hat{X}_{L/MMSE}\right)^2\right] \\ &= \mathbb{E}\left[(e + \hat{X}_{L/MMSE})^2\right] \\ &= \mathbb{E}[e^2] + 2 \cdot \mathbb{E}[e \cdot \hat{X}_{L/MMSE}] + \mathbb{E}[\hat{X}_{L/MMSE}] \\ &\overset{63}{=} \mathbb{E}[\hat{X}^2_{L/MMSE}] + e^2_{L/MMSE}\end{aligned}$$

   The element $\mathbb{E}[e \cdot \hat{X}_{L/MMSE}]$ is zero for general or linear optimal case (because of the orthogonality of the error).

4. For a Gaussian random vector $\begin{bmatrix} X \\ Y \end{bmatrix}$: $\hat{X}_{LMMSE} = \hat{X}_{MMSE}$, i.e. the LMMSE estimator is optimal. We have seen (conclusions from L/MMSE estimator) That the pair $e, \hat{X}$ is uncorrelated ($\text{Cov}(e, \hat{X}) = 0$), because

[62] Law of total expectation

[63] $e \perp \hat{X}_{L/MMSE}$

the vector is a Gaussian random vector, so also the vector $\begin{bmatrix} e_{LMMSE} \\ \hat{X}_{LMMSE} \end{bmatrix}$ is a Gaussian random vector, therefore the pair $e_{LMMSE}, \hat{X}_{LMMSE}$ is not only uncorrelated, but also independent.

In fact we notice that the vector $\begin{bmatrix} X \\ Y \\ \hat{X}_{LMMSE} \\ e_{LMMSE} \end{bmatrix}$ is a Gaussian random vector, so we conclude:

- The pair $e_{LMMSE}, \hat{X}_{LMMSE}$ is independent.
- The pair $e_{LMMSE}, Y$ is independent
- We can fully characterize the error: $e_{LLMSE} \sim N(0, \sigma_X^2 \cdot (1 - \rho_{XY}^2))$
- The conditional distribution is known $(X|Y = y)$:

$$(X|Y = y) = \hat{X}_{LMMSE}(y) + e_{LMMSE}(y)$$

$$\hat{X}_{LMMSE}(y) = \eta_X + \frac{\sigma_{XY}}{\sigma_Y^2}(y - \eta_Y) \sim N(\eta_X + \frac{\sigma_{XY}}{\sigma_Y^2}(y - \eta_Y), 0)$$

$$e_{LMMSE}(y) \sim N(0, \sigma_X^2 \cdot (1 - \rho_{XY}^2))$$

$$\Rightarrow (X|Y = y) \sim N((\eta_X + \frac{\sigma_{XY}}{\sigma_Y^2}(y - \eta_Y), \sigma_X^2 \cdot (1 - \rho_{XY}^2))$$

We also know how to calculate $\hat{X}_{MMSE}(Y)$, therefore (together with knowing the conditional distribution):

$$\hat{X}_{MMSE}(Y = y) = \mathbb{E}[X|Y = y] = \eta_X + \frac{\sigma_{XY}}{\sigma_Y^2}(Y - \eta_Y) = \hat{X}_{LMMSE}(Y = y)$$

### Geometric Image of L/MMSE estimators

**Inner Product Axioms:** Given a vector space, the definition of the inner product:

1. $\langle \underline{X}, \underline{Y} \rangle = \langle \underline{Y}, \underline{X} \rangle^*$
2. $\langle a \cdot \underline{X}, \underline{Y} \rangle = a \cdot \langle \underline{X}, \underline{Y} \rangle = \langle \underline{X}, a \cdot \underline{Y} \rangle$
3. $\langle \underline{X} + \underline{Y}, \underline{Z} \rangle = \langle \underline{X}, \underline{Z} \rangle + \langle \underline{Y}, \underline{Z} \rangle$
4. $\langle \underline{X}, \underline{X} \rangle \geq 0, \quad \langle \underline{X}, \underline{X} \rangle = 0 \iff \underline{X} = 0$

**Conclusions:**

- We notice that a collection of random variables (with finite variance, i.e. finite 2-nd order moment) is a vector space because it is always close under linear combination (every linear combination we do will create a random variable)
- For a random variable $X$, the collection of all the functions $g(X)$ is a vector sub-space of the vector space of random variables. In particular for the collection of all the linear functions of the random variable $X$

**Definition of Inner Product in Random Variables Space:** We define an inner product on the random vector space by: $\langle X.Y\rangle = \mathbb{E}[X \cdot Y]$. Examine the Axioms:

1. Trivial
2. Trivial
3. Trivial
4. Of course $\langle X, X\rangle \geq 0$ because $\langle X, X\rangle = \mathbb{E}[X \cdot X] = \mathbb{E}[X^2]$, and $X^2$ is non-negative.
   We can also see that $\langle 0, 0\rangle = \mathbb{E}[0] = 0$ (first direction) and if $\langle X, X\rangle = \mathbb{E}[X^2] = 0$ then necessarily $X = 0$ from the definition of the integral (second direction).

**Definition of The Norm of a Random Variable:** We define a norm of a random variable by the inner product of it with itself:

$$\|X\| = \langle X, X\rangle = \mathbb{E}[X^2]$$

Using the definition of norm we can define:

$$\cos(\angle(X,Y)) = \frac{\langle X, Y\rangle}{\|X\| \cdot \|Y\|} = \frac{\mathbb{E}[X \cdot Y]}{\sqrt{\mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]}} = r(X,Y)$$

$$\cos(\angle(X-\eta_X, Y-\eta_Y)) = \cdots = \frac{\mathbb{E}[(X-\eta_X) \cdot (Y-\eta_Y)]}{\sqrt{\mathbb{E}[(X-\eta_X)^2] \cdot \mathbb{E}[(Y-\eta_Y)^2]}} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \rho(X,Y)$$

As you recall, $|r(X,Y)| \leq 1$ (and the same goes for $\rho(X,Y)$). It makes sense, since $|cos(\cdot)| \leq 1$.

**Orthogonality of Random Variables:** We Say that two random variables $X, Y$ are orthogonal ($X \perp Y$) iff $\langle X, Y \rangle = 0$ (namely, $\mathbb{E}[X \cdot Y] = 0$)

**Geometric Image of L/MMSE Estimator:** The plane represents the collection (space) of all the possible functions of $Y$, namely the collection of all the functions h(Y).
Inside this space there is a sub-space of all the linear functions of $Y$.
The random variable $X$ is not a function of $Y$, and so is not on the plane. The estimator $\hat{X}_{MMSE}$ is a function of $Y$ hence it is on the plane. The estimator $\hat{X}_{LMMSE}$ is a linear function of $Y$ and so it is on the sub-space pf all the linear functions of $Y$. Because the estimators are optimal, the error must be as small as possible, therefore it is the altitude to the plane. The estimator is the projection of $X$ on the function space (general or linear, in accordance to the type of the estimator). We can also notice that because the errors $e_{L/MMSE}$ is denoted by definition as $e_{L/MMSE} = X - \hat{X}_{L/MMSE}$, and because we treat the random variables as a vector space, the errors we make are just a vector subtraction and indeed we are orthogonal to the respective function space. In the same way we can limit the estimator to be a polynomial function up to 2-nd order (for example), and then take the projection and the altitude to the sub-space.

### Estimation of a Random Vector Using a Random Vector

We have learned how to estimate a random variable using another random variable, and now we would like to generalize our discussion to estimating a random vector out of a random vector - i.e., we are given a random vector $\underline{Y}$ and we would like to estimate another random vector $\underline{X}$. Note that these vectors don't have to be with the same dimensions! (but $\underline{X}$ and $\hat{\underline{X}}$ do.)
The system $\underline{g}(\cdot)$ is a vector function the gets as an input an N-dimension vector. $\underline{g}(\cdot)$ is in fact built from $N$ different functions, each one for another elements in the vector $\underline{X}$, and every one of the functions use all the elements of vector $Y$. Hence we can write (and recognize the system is composed of N estimators):

$$\hat{\underline{X}}_{N\times 1}(\underline{Y}) = \underline{g}(\underline{Y}) = \begin{bmatrix} g1(\underline{Y}) \\ \vdots \\ g_N(\underline{Y}) \end{bmatrix} = \begin{bmatrix} g_1(Y_1, ...Y_M) \\ \vdots \\ g_N(Y_1, ...Y_M) \end{bmatrix} = \begin{bmatrix} X_1(Y_1, ...Y_M) \\ \vdots \\ X_N(Y_1, ...Y_M) \end{bmatrix} = \begin{bmatrix} X_1(\underline{Y}) \\ \vdots \\ X_N(\underline{Y}) \end{bmatrix}$$

**Estimation Error Vector** Now we define the estimation error vector in a similar manner to the one we defined in a single random variable (as usual, we work with differential error):

$$\underline{e} = \underline{X} - \underline{\hat{X}} = \underline{X} - \underline{g}(\underline{Y})$$

The estimation error vector $\underline{e}$ is a random vector, because both $\underline{X}, \underline{\hat{X}}$ are random vectors (and $e$ is a function of both of them).

**Vector Distortion Measure** For random variable estimation we defined a distortion measure function $d : \mathbb{R} \to \mathbb{R}$ and applied to it several demands. For a random vector estimation, the function will map each error vector $\underline{e}$ a single number, namely $d : \mathbb{R}^N \to \mathbb{R}$. Function $d(\underline{e})$ will be called distortion measure if $d(\underline{e}) \geq 0 \; \forall \underline{e}, \;\; d(\underline{0}) = 0$. Note that because $\underline{e}$ is a random variable, then $d(\underline{e})$ is a random variable.

**Decomposable Distortion Measure** Given a scalar distortion measure $d : \mathbb{R} \to \mathbb{R}$, it induces a vector distortion measure as follows:

$$d(\underline{e}) = \sum_i d(e_i)$$

**Examples:**

1. scalar distortion measure $d(e) = e^2$ (MSE measure) is a vector distortion measure:
$$d(\underline{e}) = \sum_i d(e_i) = \sum_i e_i^2 = \|\underline{e}\|^2$$
2. Scalar distortion measure “Probability of Error measure” (also called SER measure) gives us a vector distortion measure:
$$d(e) = \begin{cases} 0, & e = 0 \\ 1, & e \neq 0 \end{cases} \quad \Rightarrow \; d(\underline{e}) = \sum_i d(e_i) = \#(Errors)$$

Example of a distortion measure that is not decomposable:
We define a binary distortion measure that represents whether there is an error in the estimation:

$$d_{FER}(\underline{e}) = \begin{cases} 0, & \underline{e} = 0 \\ 1, & \underline{e} \neq 0 \end{cases}$$

This measure suits cases where even a small error in the information must be detected (like a damaged bit in a file).

This measure is usually named FER=Frame Error Rate. For each of the elements we can define separately the SER(Symbol Error Rate) measure:

$$d_{SER}(e) = \begin{cases} 0, & e = 0, \\ 1, & e \neq 0 \end{cases}$$

We want to represent $d_{FER}(\underline{e})$ using $d_{SER}(e)$ on each separate element:

$$d_{FER}(\underline{e}) = 1 - (1 - d_{SER}(e_1)) \cdot ... \cdot (1 - d_{SER}(e_N)) = 1 - \prod_{i=1}^{N}(1 - d_{SER}(e_i))$$

We notice it is not decomposable, but multiplicative.

**Average Distortion/ Mean of the Distortion** The average distortion (or the mean of the distortion) will be calculated the same:

$$D = \mathbb{E}[d(\underline{e})] = \mathbb{E}[d(\underline{X} - \underline{\hat{X}})] = \mathbb{E}[d(\underline{X} - \underline{g}(\underline{Y}))]$$

**Optimal Estimator** Claim:
We choose the system that for every value of $Y$ "guesses" $\underline{\alpha}$ that will minimize $D$:

$$g_{opt}^{d(\underline{e})}(\underline{Y} = \underline{y}) = \underline{\alpha}^* = \underset{\underline{\alpha} \in \mathbb{R}^N}{arg} \; min(\mathbb{E}[d(\underline{X} - \underline{\alpha})|\underline{Y} = \underline{y}])$$

Proof:
The same way as a single random variable estimating (using the law of total expectation)

**Optimal Estimation for Additive Distortion Indexes:** If we narrow ourselves to decomposable distortion measures (hence composite), we can represent the mean of the distortion by:

$$D = \mathbb{E}[d(\underline{e})] = \mathbb{E}\left[\sum_i d(e_i)\right] \overset{64}{=} \sum_i \mathbb{E}[d(e_i)] \overset{65}{=} \sum_i D_i$$

Note: We can do the same for multiplicative $d(\underline{e})$, but in order to separate them we will use log.
Conclusion:
In order to minimize $D$ we can minimize each $D_i$ separately (because in

[66] Because mean (and integral are linear for addition)

[65] $D_i = \mathbb{E}[d(e_i)]$

order to minimize a sum of elements, we need to minimize each element by itself). So we can write the estimation function of each coordinate in the next manner:

$$g_{opt}^{d(\underline{e})}(\underline{Y}=\underline{y})=\alpha^*=\underset{\alpha}{arg\ min}(\mathbb{E}[D(X_i-\alpha)|\underline{Y}=\underline{y}])$$

Hence the full estimator will look like:

$$\underline{X}_{opt}^{d(\underline{e})}(\underline{Y}=\underline{y})=\underline{g}_{opt}^{d(\underline{e})}(\underline{Y}=\underline{y})=\begin{bmatrix} g_{opt,1}^{d(\underline{e})}(\underline{Y}=\underline{y}) \\ \vdots \\ g_{opt,N}^{d(\underline{e})}(\underline{Y}=\underline{y}) \end{bmatrix}$$

Conclusion:
We can, from now on, use the process of estimating a random variable out of a random vector, because estimating a random vector out of a random vector is actually just $N$ estimation problems of estimating a random variable out of a random vector, when the distortion measures is composite.

**Estimating a Random Variable Out of a Random Vector** Now we want to estimate a random variable out of a random vector:
We discuss this case only for composite distortion measure as we have seen (as before, using differential estimate error):

$$D_i=\mathbb{E}\left[d(X_i-\hat{X}_i)\right]=\mathbb{E}\left[d(X_i-g_i(\underline{Y}))\right]$$

When we have seen that optimal estimation (again, for composite distortion measure) is being done by the system:

$$X_{opt,i}^{d(\underline{e})}(\underline{Y}=\underline{y})=g_{opt,i}^{d(\underline{e})}(\underline{Y}=\underline{y})=\alpha^*=\underset{\alpha}{arg\ min}\mathbb{E}\left(\mathbb{E}\left[d(X_i-\alpha)\mid\underline{Y}=\underline{y}\right]\right)$$

That is why $X_i$ depends solely on $X_i, \underline{Y}$ and not on the rest of the elements of $\underline{X}$

**Optimally Estimating a Random Variable out of a Random Vector Under a Pe Index** For a Pe measure (or SER), we would like to plan the system that will optimally estimate a random variable out of a random vector, i.e. will minimize the mean of distortion:

$$d_{Pe}(e)=\begin{cases}0, & e=0\\ 1, & e\neq 0\end{cases} \qquad D=\mathbb{E}[d_{Pe}(e)]\overset{66}{==}\Pr(\hat{X}\neq X)$$

Claim:
The optimal estimator of a random variable out of a random vector under Pe measure is:

$$X^{Pe}_{opt}(\underline{Y}=\underline{y}) = g^{Pe}_{opt}(\underline{Y}=\underline{y}) = \underset{\alpha}{arg\max}\Pr(X=\alpha|\underline{Y}=\underline{y})$$

This is a choice using the MAP (maximum a posteriori probability) rule because we decide the estimated value after "seeing" $\underline{Y}$
Proof:
The same as the case of estimating a random variable out of a random variable.

**Optimally Estimating a Random Variable Out of a Random Vector Under MSE Index** We have seen that the squared error distortion measure $d(e) = e^2$ (MSE measure) is giving a vector distortion measure: $d(\underline{e}) = \|\underline{e}\|^2$.
We would like to plan the system that will optimally estimate a random variable out of a random vector. i.e. minimize the distortion measure.
Optimal estimating a random variable out of a random vector under MMSE:
Claim:
The optimal estimator of a random variable out of a random vector under MSE measure is:

$$\hat{X}_{MMSE}(\underline{Y}=\underline{y}) = \hat{X}^{MSE}_{opt}(\underline{Y}=\underline{y}) = g^{MSE}_{opt}(\underline{Y}=\underline{y}) = \mathbb{E}[X|\underline{Y}=\underline{y}]$$

This is a choice using the MAP (maximum a posteriori probability) rule because we decide the estimated value after "seeing" $\underline{Y}$
Proof:
The same as the case of estimating a random variable out of a random variable. Optimal estimating a random variable out of a random vector under LMMSE:
Now we want to limit our estimator to be a linear function of $\underline{Y}$, i.e. the estimator will look like:

$$\hat{X}(\underline{Y}) = \underline{a}^T \cdot \underline{Y} + b = \sum_{i=1}^{M} a_i \cdot Y_i + b$$

The estimation error will be:

$$e = X - \hat{X}(\underline{Y}) = X - \underline{a}^T \cdot \underline{Y} - b$$

[66] As we have seen on optimal estimating a random variable out of a random variable under Pe measure

We would like to find ($\underline{a}$,b) that will minimize the mean of the distortion, namely $D = \mathbb{E}[d(e)] = \mathbb{E}[e^2] \to min.$
Claim:
The optimal linear estimator of a random variable out of a random vector under MSE measure is:

$$\hat{X}_{LMMSE}(\underline{Y}) = \hat{X}_{opt}^{LinearMSE}(\underline{Y}) = g_{opt}^{LinearMSE}(\underline{Y}) = \eta_X + C_{X\underline{Y}}C_{\underline{YY}}^{-1}(\underline{Y} - \eta_{\underline{Y}})$$

And for the error:

$$\mathbb{E}[e^2_{LMMSE}] = \sigma_X^2 - C_{X\underline{Y}} \cdot C_{\underline{YY}}^{-1} \cdot C_{\underline{Y}X}$$

Proof:
The proof of the estimator formula is in the proof of the orthogonality property for a LMMSE estimator of a random variable out of a random vector. The proof of the squared mean is in the same way as the proof for a LMMSE estimator of a random variable out of a random variable.
Properties of MMSE/LMMSE forestimator of a random variable out of a random vector:

1. Orthogonality
   (a) MMSE estimator:
   $\hat{X}(\underline{Y}) = g(\underline{Y})$ estimator is MMSE iff its error is orthogonal to every function $h(\underline{Y})$, namely: $e \perp h(\underline{Y})$. In fact, the error is orthogonal to every function of the measurements $\underline{Y}$, i.e. $\mathbb{E}[e \cdot h(\underline{Y})] = 0 \ \forall h : \mathbb{R}^M \to \mathbb{R}$
   (b) LMMSE estimator:
   Linear estimator $\hat{X}(\underline{Y}) = \underline{a}^T \cdot \underline{Y} + b$ is LMMSE iff its error $e = X - g(\underline{Y}) = X - (\underline{a}^T \cdot \underline{Y} + b)$ is orthogonal to every linear function $h(\underline{Y})$: $e \perp h(\underline{Y}) \Rightarrow e \perp (\underline{c}^T \cdot \underline{Y} + d),\ \forall \underline{c} \in \mathbb{R}^M, d \in \mathbb{R}$. In fact, the error is orthogonal to every linear function of the measurements: $\mathbb{E}[e \cdot (\underline{c}^T \cdot \underline{Y} + d)] = \mathbb{E}[X - (\underline{a}^T \cdot \underline{Y} + b) \cdot (\underline{c}^T \cdot \underline{Y} + d)] = 0,\ \forall \underline{c} \in \mathbb{R}^M, d \in \mathbb{R}$
2. Unbiasedness
   For L/MMSE estimator, the mean of the error is zeroing, or the mean of the estimator equal the mean of the estimated variable $X$:

$$\begin{aligned} \mathbb{E}[e_{L/MMSE}] &= 0 \\ \mathbb{E}[X - \hat{X}_{L/MMSE}] &= 0 \\ \mathbb{E}[\hat{X}_{L/MMSE}] &= \mathbb{E}[X] = \eta_X \end{aligned}$$

3. Pythagoras
   The power of the random variable we want to estimate is equal to the power of the estimator plus the power of the error (As said for MSE measure the error power is $D = MSE$)

   $$\mathbb{E}[X^2] = \mathbb{E}[\hat{X}^2_{L/MMSE}] + \mathbb{E}[e^2_{L/MMSE}]$$

   We notice that if we limit ourselves to a linear estimator, the power of the error will grow on account of the power of the estimator.

4. For a Gaussian random vector $\begin{bmatrix} X \\ \underline{Y} \end{bmatrix}$: $\hat{X}_{MMSE} = \hat{X}_{LMMSE}$, i.e. the LMMSE estimator is the optimal estimator.
   IF $\begin{bmatrix} X \\ \underline{Y} \end{bmatrix}$ us a Gaussian random vector,then the optimal MMSE estimator is linear, and:

   $$\hat{X}_{MMSE} = \mathbb{E}[X|\underline{Y} = \underline{y}] \overset{67}{=} \eta_{\underline{X}} + C_{\underline{XY}} \cdot C^{-1}_{\underline{Y}\underline{y}} \cdot (\underline{Y} - \eta_{\underline{Y}}) = \hat{X}_{LMMSE}$$

Proof:
We only prove the existence of a linear optimal estimator that its error orthogonal to every linear function of the measurements, because the rest of the proofs are similar to the proofs of L/MMSE optimal estimators of random variables out of random variables.
We would like $(\underline{a}, b)$ so that:$(X - (\underline{a}^T \cdot \underline{Y} + b)) \perp (\underline{c}^T \cdot \underline{Y} + d)$, $\forall \underline{c} \in \mathbb{R}^M, d \in \mathbb{R}$
Claim:
Orthogonality is equivalent to the next demands:

1. $e \perp 1$ (or every constant in $\mathbb{R}$).

2. $$\mathbb{E}[e \cdot (\underline{Y} - \eta_{\underline{Y}})^T] = 0 \iff e \perp (Y_i - \eta_{Y_i}),\ \forall i = 1, ..., M \iff \begin{cases} e \perp (Y_1 - \eta_{Y_1}) \\ \vdots \\ e \perp (Y_M - \eta_{Y_M}) \end{cases}$$

We now show the meaning of these demands:

[67] Because the vector is a Gaussian random vector

1.

$$\begin{aligned} e \perp 1 \iff & \mathbb{E}[e \cdot 1] = 0 \\ & \mathbb{E}[X - \underline{a}^T \cdot \underline{Y} - b] = 0 \\ & \eta_X - \underline{a}^T \cdot \eta_{\underline{Y}} - b = 0 \\ & b(\underline{a}) = \eta_X - \underline{a}^T \cdot \eta_{\underline{Y}} \end{aligned}$$

2.

$$\begin{aligned} & \mathbb{E}[e \cdot (\underline{Y} - \eta_{\underline{Y}})^T] = 0 \\ & \mathbb{E}[(X - \underline{a}^T \cdot \underline{Y - b}) \cdot (\underline{Y} - \eta_{\underline{Y}})^T] = 0 \end{aligned}$$

We use $b(\underline{a})$ that we got earlier:

$$\begin{aligned} \mathbb{E}\left[\left(X - \underline{a}^T \cdot \underline{Y} - \eta_X + \underline{a}^T \eta_{\underline{Y}}\right) \cdot (\underline{Y} - \eta_{\underline{Y}})^T\right] = 0 \\ \mathbb{E}\left[\left((X - \eta_X) - \underline{a}^T \cdot (\underline{Y} - \eta_{\underline{Y}})\right) \cdot (\underline{Y} - \eta_{\underline{Y}})\right] = 0 \\ C_{X\underline{Y}} - \underline{a}^T \cdot C_{\underline{YY}} = 0 \Rightarrow \underline{a}^T = C_{X\underline{Y}} \cdot C_{\underline{YY}}^{-1} \end{aligned}$$

And now we replace the estimator we found (and find out this is $X_{LMMSE}$ as we defined):

$$\begin{aligned} \hat{X}(\underline{Y}) &= \underline{a}^T \cdot \underline{Y} + b \\ &= \begin{bmatrix} \underline{a}^T = C_{X\underline{Y}} \cdot C_{\underline{YY}}^{-1} \\ b(\underline{a}) = \eta_X - \underline{a}^T \cdot \eta_{\underline{Y}} \end{bmatrix} \\ &= C_{X\underline{Y}} \cdot C_{\underline{YY}}^{-1} \cdot \underline{Y} + \eta_X - C_{X\underline{Y}} \cdot C_{\underline{YY}}^{-1} \cdot \eta_{\underline{Y}} \\ &= \eta_{\underline{X}} + C_{\underline{XY}} \cdot C_{\underline{YY}}^{-1} \cdot (\underline{Y} - \eta_{\underline{Y}}) \end{aligned}$$

**MSE Optimal Estimation of a Random Vector Out of a Random Vector** We return to discuss estimation of a random vector out of a random vector, when we reduce our problem to using differential error using MSE distortion measure
MMSE estimator:

$$\hat{X}(\underline{Y} = \underline{y}) = \mathbb{E}[\underline{X} | \underline{Y} = \underline{y}] = \begin{bmatrix} \mathbb{E}[X_1 | \underline{Y} = \underline{y}] \\ \vdots \\ \mathbb{E}[X_N | \underline{Y} = \underline{y}] \end{bmatrix}$$

LMMSE estimator:

$$\hat{X}_{LMMSE}(\underline{Y}) = \hat{X}_{opt}^{LinearMSE}(\underline{Y}) = g_{opt}^{LinearMSE}(\underline{Y}) = \eta_{\underline{X}} + C_{\underline{XY}} \cdot C_{\underline{YY}}^{-1} \cdot (\underline{Y} - \eta_{\underline{Y}})$$

$$C_{\underline{ee}} = C_{\underline{XX}} - C_{\underline{XY}} \cdot C_{\underline{YY}}^{-1} C_{\underline{YX}}$$

Properties of Covariance matrix of the error $C_{\underline{ee}}$:

1. Let there be another estimator $\underline{\hat{\tilde{X}}} = \underline{h}(\underline{Y})$ with error $\underline{\tilde{E}} = \underline{X} - \underline{\hat{\tilde{X}}}$, then: $C_{\underline{\tilde{E}}\underline{\tilde{E}}} - C_{\underline{e}_{LMMSE}\underline{e}_{LMMSE}} \geq 0$
2. Let there be a second estimator $\underline{\hat{\tilde{X}}} = \underline{c}^T \cdot \underline{Y} + \underline{d}$ with error $\underline{\tilde{E}} = \underline{X} - \underline{\hat{\tilde{X}}}$ then the matrix $C_{\underline{\tilde{E}}\underline{\tilde{E}}} - C_{\underline{e}_{LMMSE}\underline{e}_{LMMSE}}$ is PSD (defined, non-negative matrix)

Note:
The Covariance matrix of the error vector is defined as:

$$C_{\underline{e}_{LMMSE}\underline{e}_{LMMSE}} = C_{\underline{X}\underline{X}} - C_{\underline{\hat{X}}_{L/MMSE}\underline{\hat{X}}_{L/MMSE}}$$

# Part B: Random Processes and Operations

## Introduction, Definitions and Properties

Random processes are a generalization of random vectors, processes occur on time/space hence the random variable index gets the appropriate meaning. In addition, the order of the random variables is meaningful (the indexes represent the development of the process).
We will deal with random processes that develop in time. Random process (RP) can be in continuous or discrete time, and get continuous or discrete values.
If we think of RP as an infinite random vector, then:

- RP in continuous time will have infinite “size” (cardinality) that can be well-ordered ($\aleph$)
- RP in discrete time with countable infinite cardinality ($\aleph_0$)

Because RP are infinite, sometimes instead of $\{X_n\}_{n=1}^{\infty}$, we write $X_n$ (even though we **don’t** mean only single specific $n$) for discrete time, and for continuous time, instead of $\{X(t)\},\ t \in [0, \infty)$, we write $X(t)$ (even though we **don’t** mean only single specific $t$).

**Examples:**

1. The series of bytes we get when downloading a file - $B[n]$
   This is a discrete RP in discrete time - it can be only discrete values (1 or 0) for every time difference.

2. Time of arrival of the n-th bus to the station - $T[n]$
   This is a continuous RP in discrete “time” - the time of arrival is continuous, but the number representing the bus is discrete.

3. Blood pressure measurement as a function of time - $P(t)$
   This is a continuous RP in a continuous time - blood pressure can have any value (in a specific range) and is measured all the time.

4. Number of people in station as a function of time - $N(t)$
   This is a discrete RP in continuous time - the number of people in the station is discrete, but measured continuously.

**RP in Continuous Time** RP in continuous time $X(\omega, t)$ is a function that maps every continuous time $t$ and experiment outcome $\omega \in \Omega$ a real number:

$$X : \Omega \times \mathbb{R} \to \mathbb{R}$$

Note:
A different representation is $X : \Omega \to \mathbb{R}^{\mathbb{R}}$, i.e. for every experiment outcome $\omega_0 \in \Omega$ we get infinite number of pairs $(t, x)$ such that $t, x \in \mathbb{R}$. For every value of $t_0$ there is a pair with a value $x_0$ that matches to $X(\omega_0, t_0)$.
As said, we use the notation for process as $X(t)$ for simplicity.

**RP in Discrete Time** RP in discrete time $X[\omega, n]$ is a function that maps every discrete time $n$ and experiment outcome $\omega \in \Omega$ a real number:

$$X : \Omega \times \mathbb{Z} \to \mathbb{R} \leftrightarrow X : \Omega \to \mathbb{R}^{\mathbb{Z}}$$

As said, we we use the notation for process as $X_n$ for simplicity.

**Realization of a RP** Realization of a RP is the outcomes of the sampling of the RP for $\omega = \omega_0$ an outcome of an experiment. Of course, for this outcome of experiment, the PR gets a value for every point in time. Namely, for $\omega = \omega_0$ (i.e. circling the $\omega$ axis) we get function $X(\omega_0, t)$ which is a mapping of the experiment outcome in accordance with time.
If in addition we pick $t = t_0$, we get a random variable $X(\omega_0, t_0)$ (it is a random variable because $X$ is a RP and not deterministic.)

### Motivation for RP

In order to understand the creation of RP, we will see some simple examples. In the future we will build non-trivial RP and we will have to identify them and simplify them.

**i.i.d RP** Discrete RP (only for demonstration, can ve continuous too) in time $X_n$ will be an i.i.d. RP if:

1. All the samples of the process have the same distribution, namely: $\Pr(X_i = x) = \Pr(X_j = x)$, $\forall i, j \in \mathbb{Z}$, $x \in \mathbb{R}$ (or with PDF for continuous time)
2. The samples of the process $X_1, X_2, ...$ are mutually independent (not only pairwise)
   statistical independence of pair of processes: $X \perp\!\!\!\perp Y$
   Hence we demand: $X_1, X_2, ..., X_{n-1} \perp\!\!\!\perp X_n$
   Note: pairwise independence $\nRightarrow$ mutual independence.

**Simple Constructive Examples of RP:**

1. i.i.d RP:
   i.i.d. RP is a process that all of its samples are i.i.d. For demonstration we will be more specific and show i.i.d RP with Bernoulli samples with parameter $p$:

$$W_n = \begin{cases} 1, \ p \\ 0, \ q = 1 - p \end{cases} \sim Ber(p), \quad \{W_n\} \ i.i.d$$

   We want to test the statistics of the process for $k$ points in time: $i_1, ..., i_k$ The samples are i.i.d, and so independent, so we can know the full statistics of the RP for every such set of times by:

$$\Pr(W_{i_1} = \omega_1, ..., W_{i_k} = \omega_k) \overset{68}{=} \prod_{j=1}^{k} \Pr(W_{i_j} = \omega_j)$$

2. Counting RP:
   We take the earlier i.i.d Bernoulli RP $\{W_n\}$ and insert it into a sum:

$$X_n = \sum_{i=1}^{n} W_i$$

   Notice we can represent the process as a recursive process:

$$X_n = X_{n-1} + W_n, \ X_0 = 0, \ W_n = \begin{cases} 1, \ p \\ 0, \ q = 1 - p \end{cases}, \ \{W_n\} \ i.i.d$$

   This is the definition of Auto Regressive Process.
   RP $\{X_n\}$ is not i.i.d even though it was built using an i.i.d RP. In addition, you can notice that even without using the connection between samples, its samples don’t distribute in the same way. There is a connection between samples - a pair of near samples have a difference of maximum 1, and the series is increasing.
   This is in fact a counting process of Bernoulli samples, hence this is a Binomial random variable $X_n \sim Bin(n, p)$. We will see that:

$$\mathbb{E}[X_n] = \mathbb{E}\left[\sum_{i=1}^{n} W_i\right] = \sum_{i=1}^{n} p = n \cdot p$$

[68] Independence

$$\text{Var}(X_n) = \mathbb{E}[X_n^2] - \mathbb{E}^2[X_n] = ... = n \cdot p \cdot (1-p)$$

We notice this is a RP with a lot of memory - each sample distributes differently, and there is dependence between samples.
Notice that id $p = 0, 1$ the Variance is 0 - because then the process is deterministic.

3. XOR RP: On the same i.i.d Bernoulli RP $\{W_n\}$, we build the next RP:

$$X'_n = mod_2\left(\sum_{i=1}^{n} W_i\right)$$

Notice we can represent the RP as a recursive RP:

$$X'_n = X_{n-1} \oplus W_n,\ X'_0 = 0,\ W_n = \begin{cases} 1, & p \\ 0, & q = 1-p \end{cases},\ \{W_n\}\ i.i.d,\ A \oplus B = \begin{cases} 1, & A \neq B \\ 0, & A = B \end{cases}$$

Namely, if $W_n = 1$ then the value of $X'_n$ will be the opposite of the previous sample, else it will be identical ($X'_n \in \{0, 1\}$).
For $p = 0.5$ we get that all the samples of the RP are i.i.d (meaning $X'_n$ will get 0 or 1 in equal probability, independently of its other samples).

4. Random walk RP:
We would like to create a RP that represent a movement to one of each two possible directions in any point in time. This is different from Bernoulli distribution, because in Bernoulli you either made the step ($W_n = 1$) ot didn’t ($W_n = 0$).
For that, we use the i.i.d Bernoulli RP $\{W_n\}$ to build the next RP:

$$W''_n = 2 \cdot W_n - 1 = \begin{cases} 1, & p \\ -1, & q = 1-p \end{cases},\ W_n \sim Ber(p),\ \{W_n\}\ i.i.d$$

Now we sum the RP:

$$X''_n = \sum_{i=1}^{n} W''_i$$

Notice we can represent the process as a recursive RP:

$$X''_n = X''_{n-1} + W''_n,\ X''_0 = 0,\ W''_n = \begin{cases} 1, & p \\ -1, & 1-p \end{cases},\ \{W''_n\}\ i.i.d$$

Analyze the statistics of the RP $\{X''_n\}$:

$$\mathbb{E}[X''_n] = \mathbb{E}\left[\sum_{i=1}^{n} W''_i\right] = \sum_{i=1}^{n} \mathbb{E}[W''_i] = \sum_{i=1}^{n} \mathbb{E}[2 \cdot W_i - 1] = \sum_{i=1}^{n} (2 \cdot p - 1) = n \cdot (2 \cdot p - 1)$$

$$\mathrm{Var}(X''_n) \overset{69}{=} 4 \cdot \mathrm{Var}(x_n) = ... = 4 \cdot n \cdot p \cdot (1-p)$$

For $P = 0.5$ we get $\mathbb{E}[X''_n] = 0$, this makes sense because the random walk can take each direction with the same probability.
Note: This RP is similar to Brownian Motion, which we will discuss later.

## Full Statistical information About RP

For a random vector $\underline{X}$, knowing the full statistics of the vector was equal to knowing its jCDF (or jPDF):

$$F_{\underline{X}}(\underline{x}) = \Pr(X_1 \leq x_1, ..., X_n \leq x_n) = \Pr(\underline{X} \leq \underline{x})$$

For a RP, the vector is of infinite size so we can't write or calculate the function in the same way.
Hence we will represent RP statistically using its jCDF(or jPDF) for a collection of points in time. Namely, we will choose a finite collection of the infinite time axis, narrow our problem to random vectors, and we will want to calculate the jCDF.
For a continuous time RP we would like to know its jCDF for each set of time $\{t_1, ...t_N\}$:

$$\begin{aligned} F(t_1, ..., t_N, x_1, ..., x_n) &= F_{X(t_1),...,X(t_N)}(x_1, ..., x_N) \\ &= F_{t_1,...,t_N}(\underline{x}) \\ &= \Pr(X(t_1) \leq x_1, ..., X(t_N) \leq x_N) \end{aligned}$$

For a discrete time RP we would like to know its jCDF for each set of time $\{k_1, ...k_N\}$:

$$\begin{aligned} F(k_1, ..., k_N, x_1, ..., x_n) &= F_{X(k_1),...,X(k_N)}(x_1, ..., x_N) \\ &= F_{k_1,...,k_N}(\underline{x}) \\ &= \Pr(X(k_1) \leq x_1, ..., X(k_N) \leq x_N) \end{aligned}$$

Usually we use the next notations: $X(k_1) = X_{k_1},\ X(t_1) = X_{t_1}$
**Important**: knowing the full statistical information about a RP meaning knowing this function for every set of times.
Notes:

[69] Properties of Variance: $\mathrm{Var}(a \cdot X + b) = a^2 \cdot \mathrm{Var}(X)$

- Because we want to know the jCDF (or jPDF) for every set of times, it will still be difficult to write the expression to the function, unless we know other things about the process (Gaussian, i.i.d, etc.)
- The jCDF for different sets of times will have to be consistent. Notice that if we take the two sets $\{t_1, t_2\}$, $\{t_2, t_3\}$ then we get the next jCDF (for simplicity,$X_{t_i} = X_i$):

  $$F_{X_1,X_2}(x_1, x_2),\ F_{X_2,X_3}(x_2, x_3)$$

  Using the jCDF properties:

  $$F_{X_2}(x) = F_{X_1,X_2}(\infty, x),\ F_{X_2}(x) = F_{X_2,X_3}(x, \infty)$$

  Meaning there is a connection between the jCDF of random vector that created using overlapped times. So for every two sets of overlapping time, there is a demand that need to be noted, which is hard to define

## Gaussian RP in Continuous Time

RP $X(t)$ will be named a Gaussian RP (or GRP) in continuous time if for every set of its samples (for every time set to be chosen), the group os jointly Gaussian (i.e. the samples can create a Gaussian random vector).
Namely, for every time set $\{t_1, ..., t_N\}$ the sample of the process $\{X(t_1), ..., X(t_2)\}$ can create a Gaussian random variable using their sum: $Y = \sum_{i=1}^{N} \alpha_i \cdot X(t_i)$ is a Gaussian random variable for every $\{\alpha_i\}_{i=1}^{N}$, $\alpha_i \in \mathbb{R}$.
As said, in order to fully describe a RP statistically we have to know some other information about it.
For GRP $X(t)$ we can write the jCDF for every time set $\{t_1, ..., t_N\}$ by:

$$F_{X(t_1),...,X(t_N)}(x_1, ..., x_N) = F_{t_1,...,t_N}(\underline{x}) = \frac{1}{\sqrt{(2\pi)^N \cdot C_{\underline{tt}}}} \cdot e^{-\frac{1}{2}\cdot(\underline{x}-\eta_{\underline{t}})^T \cdot C_{\underline{tt}}^{-1} \cdot((\underline{x}-\eta_{\underline{t}})}$$

When by using the notation $(\eta_{\underline{t}}, C_{\underline{tt}})$ we mean $(\eta_{\underline{X}_t}, C_{\underline{X}_t\underline{X}_t})$.
The matrix $C_{\underline{tt}}$ has to be consist. This matrix is just the covariance matrix of the samples of the process, and is calculated by:

$$C_{\underline{tt}} = \begin{bmatrix} \sigma^2_{X(t_1)} & \dots & \mathrm{Cov}(X(t_1), X(t_N)) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X(t_N), X(t_1)) & \cdots & \sigma^2_{X(t_N)} \end{bmatrix} = \begin{bmatrix} \sigma^2_{X_{t_1}} & \dots & \mathrm{Cov}(X_{t_1}, X_{t_N}) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_{t_N}, X_{t_1}) & \cdots & \sigma^2_{X_{t_N}} \end{bmatrix}$$

Note: Reminder- if we want to say a RP is a GRP, it is not enough that the jPDF is Gaussian.

For a random variable, Gaussian distribution is defined by the mean and the variance. For a random vector, Gaussian distribution is defined by the mean vector and the covariance matrix. In both cases, this is a 2-nd order statistics.
GRP is defined by a 2-nd order statistics of a RP, i.e. mean function and autocorrelation function or auto-covariance function.

## 2-nd Order Statistics of RP

2-nd order statistics of a RP $X(t)$ or $X_N$ is the collection of all the 2-nd order moments of the process:
*Continuous:* $\forall t, t_1, t_2 \in \mathbb{R}$

$$\begin{gathered}\eta_X(t) = \mathbb{E}[X_t] = \mathbb{E}[X(t)] \\ R_{XX}(t_1,t_2) = \mathbb{E}[X(t_1)\cdot X(t_2)] \\ C_{XX}(t_1,t_2) = \mathbb{E}[(X(t_1)-\eta_X(t_1))\cdot(X(t_2)-\eta_X(t_2))] = \mathrm{Cov}(X(t_1),X(t_2)) \\ C_{XX}(t_1,t_2) = R_{XX}(t_1,t_2) - \eta_X(t_1)\cdot\eta_X(t_2)\end{gathered}$$

*Discrete:* $\forall n, m \in \mathbb{Z}$

$$\begin{gathered}\eta_X(n) = \mathbb{E}[X_n] = \mathbb{E}[X(n)] \\ R_{XX}(n,m) = \mathbb{E}[X_n\cdot X_m] \\ C_{XX}(n,m) = \mathbb{E}[(X_n-\eta_X(n))\cdot(X_m-\eta_X(m))] = \mathrm{Cov}(X_n,X_m) \\ C_{XX}(n,m) = R_{XX}(n,m) - \eta_X(n)\cdot\eta_X(m)\end{gathered}$$

Hence knowing the 2-nd order statistics means knowing all of the above (as we said, knowing the regular or centralized moments is equal).
Notes:

- For RP, the correlation and covariance matrices are infinite. The notation above mean addressing a certain cell. Of course, the demand of knowing all the statistics is equal to knowing every cell.

$$C_{\underline{X}_t\underline{X}_t} = \begin{bmatrix} C_{\underline{X}_t\underline{X}_t}(t_1,t_1) & \dots & C_{\underline{X}_t\underline{X}_t}(t_1,t_N) \\ \vdots & \ddots & \vdots \\ C_{\underline{X}_t\underline{X}_t}(t_N,t_1) & \dots & C_{\underline{X}_t\underline{X}_t}(t_N,t_N)\end{bmatrix}$$

$$R_{\underline{X}_t\underline{X}_t} = \begin{bmatrix} R_{\underline{X}_t\underline{X}_t}(t_1,t_1) & \dots & R_{\underline{X}_t\underline{X}_t}(t_1,t_N) \\ \vdots & \ddots & \vdots \\ R_{\underline{X}_t\underline{X}_t}(t_N,t_1) & \dots & R_{\underline{X}_t\underline{X}_t}(t_N,t_N)\end{bmatrix}$$

- Matrices $C_{\underline{X}_t\underline{X}_t},\ R_{\underline{X}_t\underline{X}_t}$ are symmetric, meaning $A(i,j)=A(j,i)$
- Notice the diagonal of the 2-nd order moments matrices:

$$R_{XX}(t,t)=\mathbb{E}[X(t)\cdot X(t)]=\mathbb{E}[X^2(t)]$$

$$C_{XX}=\mathbb{E}[(X(t)-\eta_X(t))\cdot(X(t)-\eta_X(t))]=\mathrm{Var}(X(t))$$

- Matrices $C_{\underline{X}_t\underline{X}_t},\ R_{\underline{X}_t\underline{X}_t}$ are PSD matrices, hence:

$$\forall \underline{a},\underline{t}\in\mathbb{R}^N,\ \underline{a}^T\cdot R_{\underline{XX}}\cdot\underline{a}\geq 0\Rightarrow\sum_{i=1}^{N}\sum_{j=1}^{N}a_i\cdot a_j\cdot R_{\underline{XX}}(t_i,t_j)\geq 0$$

Or in the integral limit:

$$\int_{i=-\infty}^{\infty}\int_{j=-\infty}^{\infty}a(i)\cdot a(j)\cdot R_{\underline{XX}}(i,j)\cdot dj\cdot di\geq 0$$

**Linear Correlation Coefficients Between Samples of RP**

We define the correlation coefficients between a pair of samples of the RP $X(t)$ as:

$$r(t_1,t_2)=\frac{R_{XX}(t_1,t_2)}{\sqrt{R_{XX}(t_1,t_1)\cdot R_{XX}(t_2,t_2)}}=\frac{\mathbb{E}[X(t_1)\cdot X(t_2)]}{\sqrt{\mathbb{E}[X^2(t_1)]\cdot\mathbb{E}[X^2(t_2)]}}$$

$$\rho(t_1,t_2)=\frac{C_{XX}(t_1,t_2)}{\sqrt{C_{XX}(t_1,t_1)\cdot C_{XX}(t_2,t_2)}}=\frac{\mathrm{Cov}(X(t_1),X(t_2))}{\sqrt{\mathrm{Var}(X^2(t_1))\cdot\mathrm{Var}(X^2(t_2))}}$$

And as before: $-1\leq r(t_1,t_2),\rho(t_1,t_2)\leq 1$
We can get the next inequality:

$$C^2_{XX}(t_1,t_2)\leq C_{XX}(t_1,t_1)\cdot C_{XX}(t_2,t_2)\iff Cov^2(X(t_1),X(t_2))\leq\mathrm{Var}(X(t_1))\cdot\mathrm{Var}(X(t_2))$$

Meaning the squared variance of two samples of a RP is not bigger then the multiplication of the variance of each sample.

**Strict Sense Stationary Process (SSS) and Wide Sense Stationary Process (WSS)**

RP have stochastic properties (stochastic processes), namely the development of the process depends on random elements (non-deterministic). Meaning from the initial state of the system there are several different states the system can get (some might have bigger probability).
There is a sub set of the set of RP that have the stationary property.

**Strict Sense Stationary (SSS)** RP will be strict sense stationary (SSS) if its marginal distribution (for every set of samples) is independent in time. Namely, RP $X(t)$ will be SSS if for every set of times $\{t_1, ..., t_N\}$ and every shift in time $\Delta$:

$$f_{X(t_1),...,X(t_N)}(x_1, ..., x_N) = f_{X(t_1+\Delta),...,X(t_N+\Delta)}(x_1, ..., X_N)$$

Especially if the marginal distribution is independent of time, i.e. $f - X(t)(x) = f_{X(t+\Delta)}(x)$ for every $t, \Delta$

**Wide Sense Stationary (WSS)** RP $X(t)$ is wide sense stationary(WSS) if:

1. The mean of the process is independent of time
2. Correlation function between each pair of samples is independent in shift in time (only depends on the difference of times)

Namely, RP $X(t)$ is WSS if:

1. $\forall t \in \mathbb{R},\ \eta_X(t) = \eta_X$
2. $\forall t_1, t_2, \Delta \in \mathbb{R},\ R_{XX}(t_1, t_2) = R_{XX}(t_1 + \Delta, t_2 + \Delta)$

Note:
If we present condition 2 with $\Delta = -t_2$, we get: $R_{XX}(t_1, t_2) = R_{XX}(t_1 - t_2, t_2 - t_2) = R_{XX}(t_1 - t_2, 0)$, meaning the correlation function depends only on the time difference $t_1 - t_2$. We denote $\tau = t_1 - t_2$ and write:

$$R_{XX}(t_1, t_2) = R_{XX}(t_1 - t_2, 0) \overset{70}{=} R_{XX}(\tau, 0) \overset{71}{=} R_{XX}(\tau)$$

**Properties of WSS Correlation Function** If $X(t)$ is a WSS RP, then:

1. $R_{XX}(t_1, t_2) = R_{XX}(t_1 - t_2, 0) \Rightarrow \mathbb{E}[X(t_1) \cdot X(t-2)] = \mathbb{E}[X(t_1 - t_2) \cdot X(0)]$
2. $R_{XX}(0)$ is the power of the RP
   Te time difference is zeroed for a pair of identical samples, meaning both times were identical. We get:

$$R_{XX}(t, t) = \mathbb{E}[X(t) \cdot X(t)] = \mathbb{E}[X^2(t)] = Power(X(t))$$

[72] $\tau = t_1 - t_2$

[71] We presents the function $R_{XX}(\cdot, \cdot)$ as a function of one variable, even though it is a function of two variables.

We can show on other way:

$$R_{XX}(0) = R_{XX}(0,0) \overset{72}{=} R_{XX}(t,t) = \mathbb{E}[X(t)\cdot X(t)] = \mathbb{E}[x^2(t)] = Power(X(t))$$

3. Correlation function $R_{XX}(\tau)$ is a symmetric function.
From the definition of the mean we can deduce that the correlation function of WSS RP is a symmetric function:

$$\begin{aligned} R_{XX}(\tau) &= R_{XX}(\tau,0) \overset{73}{=} R_{XX}(0,-\tau) \\ &= \mathbb{E}[X(0)\cdot X(-\tau)] = \mathbb{E}[X(-tau)\cdot X(0)] \\ &= R_{XX}(\tau,0) = R_{XX}(-\tau) \end{aligned}$$

4. $\forall\tau \in \mathbb{R} \quad |R_{XX}(\tau)| \leq R_{XX}(0)$ Namely, the function gets a maximal value (size) for time difference of zero.
As remembered, the definition of Pearson's correlation coefficient states that:

$$\begin{aligned} r(t_1,t_2) &= \frac{R_{XX}(t_1,t_2)}{\sqrt{R_{XX}(t_1,t_2)\cdot R_{XX}(t_2,t_2)}} \\ &\overset{74}{=} \frac{R_{XX}\tau,0)}{\sqrt{R_{XX}(0,0)\cdot R_{XX}(0,0)}} \\ &= \frac{R_{XX}(\tau)}{R^2_{XX}(\tau)} = \frac{R_{XX}(\tau)}{|R_{XX}(0)} \end{aligned}$$

Hence, we get:

$$R_{XX}(\tau)r(t_1,t_2)\cdot|R_{XX}(0) \Rightarrow |R_{XX}(\tau)| = |r(t_1,t_2)|\cdot|R_{XX}(0)|$$

We know that $|r(t_1,t_2)| \leq 1$ and $|R_{XX}(0)|$ is non-negative (property 2 - power), so:

$$|R_{XX}(\tau)| \leq R_{XX}(0)$$

5. $$\int_{i=-\infty}^{\infty}\int_{j=-\infty}^{\infty} a(i)\cdot a(j)\cdot R_{\underline{XX}}(i-j)\cdot dj\cdot di \geq 0$$

[72] From property 2 of WSS RP: Correlation function is independent in time.

[73] See precious foot note.

[74] $\tau = t_1 - t_2$

We have seen that because the correlation matrix $R_{\underline{XX}}$ is a PSD matrix:

$$\int_{i=-\infty}^{\infty}\int_{j=-\infty}^{\infty} a(i)\cdot a(j)\cdot R_{\underline{XX}}(i,j)\cdot dj\cdot di \geq 0$$

And because $X(t)$ is a WSS RP: $R_{\underline{XX}}(i,j) = R_{\underline{XX}}(i-j,0) = R_{\underline{XX}}(i-j)$, so:

$$\int_{i=-\infty}^{\infty}\int_{j=-\infty}^{\infty} a(i)\cdot a(j)\cdot R_{\underline{XX}}(i-j)\cdot dj\cdot di \geq 0$$

**Asymptotic Stationarity** Sometime, a RP will not meet the terms for being WSS, but will when going asymptotically (in the limit)

**Asymptotic SSS:** RP $X(t)$ will be asymptotic SSS if its marginal distribution (for every set of samples $\{t_1, ..., t_N\}$) the limit exists:

$$\lim_{\Delta\to\infty}(f_{X(t_1+\Delta),...,X(t_N+\Delta)}(x_1,...,x_N))$$

We denote the limit as the stationary distribution: $f^{Stationary}_{t_1,...,t_N}(x_1,...,x_n)$

**Asymptotic WSS:** RP $X(t)$ will be asymptotic WSS if its marginal distribution (for every set of samples $\{t_1, ..., t_N\}$) the limit exists:

$$\lim_{\Delta\to\infty}(f_{X(t_1+\Delta),...,X(t_N+\Delta)}(x_1,...,x_N))$$

We denote the limit as the stationary distribution: $f^{Stationary}_{t_1,...,t_N}(x_1,...,x_n)$

### Jointly SSS (jSSS) and Jointly WSS (jWSS)

We explained that RP have the stochastic property. For a pair ofo RP, each one of the is stochastic and we would like to know if we can say that together they have a stationarity property.

**jSSS** A pair of RP $X(t), Y(t)$ will be jSSS (jointly strict sense stationary) if their jPDF (for every set of samples) is independent in time.
Meaning, EP $X(t), Y(t)$ will be jSSS if for every set of times $\{t_1, ..., t_N, t_{N+1}, ..., t_{N+M}\}$ and every shift in time $\Delta$:

$$f_{X(t_1),...,X(t_N),Y(t_{N+1}),...,Y(t_{N+M})}(x_1,...,x_N,y_1,...y_M) =$$
$$f_{X(t_1+\Delta),...,X(t_N+\Delta),Y(t_{N+}+\Delta),...,Y(t_{N+M}+\Delta)}(x_1,...,x_N,y_1,...y_M)$$

And the same goes for the jCDF.
If a pair of RP $X(t), Y(t)$ is jSSS, then each of them is SSS separately (Notice that the condition is also for the jCDF and we can make a marginal jCDF by putting $t_1 = \infty$ for every time of the process to nullify it).

**jWSS** A pair of RP $X(t), Y(t)$ will be jWSS (jointly wide sense stationary) if:

1. Each of the RP is WSS, so:
   (a) $\forall t \in \mathbb{R}: \ \eta_X(t) = \eta_X, \eta_Y(t) = \eta_Y,$
   (b) $\forall t_1, t_2, \Delta \in \mathbb{R}: \ R_{XX}(t_1, t_2) = R_{XX}(t_1 + \Delta, t_2 + \Delta)$, $R_{YY}(t_1, t_2) = R_{YY}(t_1 + \Delta, t_2 + \Delta)$
2. Cross-correlation function between each pair of samples is independent in shifting in time:
   $\forall t_1, t_2, \Delta \in \mathbb{R}: \ R_{XY}(t_1, t_2) = R_{XY}(t_1 + \Delta, t_2 + \Delta)$

Note:
For the cross-correlation function of two jWSS RP, we can define:

$$R_{XY}(t_1, t_2) = R_{XY}(t_1 - t_2, 0) \overset{75}{=} R_{XY}(\tau, 0) = R_{XY}(\tau)$$

## Auto Regressive Random Process

Denote $\{W_n\}$ i.i.d RP with a known PDF $f_W(w)$ and initial condition a random variable $X_0$ with known PDF $f_{X_0}(x)$ (especially if $X_0$ is deterministic), where the initial condition $X_0$ and $\{W_n\}$ are independent. ( i.e. for all $\{W_1, W_2, ...\}$ together and not only each separately).
Deterministic function $g(\bullet, \bullet)$ will define $\{X_n\}$ A.R. RP by:

$$X_n = g(X_{n-1}, W - n)$$

The function can be any deterministic function, for example:

$$g_1(u, v) = u + v, \ g_2(u, v) = u \oplus v$$

### Markovianity of A.R RP

We will discuss Markovianity later, but know we will go through it briefly. A system is defined as a Markovian system if the sum of the information in the present is enough to know the future, i.e. there is no need to remember the past in order to be "smarter", because the current sample provides the same information.

[75] $\tau = t_1 - t_2$

Claim: A.R. RP is Markovian. Explanation: In every iteration, $X_n$ depends on $X_{n-1}, W_n$. We know that $\{W_n\}$ is an i.i.d RP, therefore its distribution is independent in past values $(W_1, ...W_{n-1})$ and also independent in past $X_n$ values $(X_1, ...X_{n-1})$. Hence $X_n$ is based only on $X_{n-1}$, meaning $X_{n-1}$ contains all the relevant past data.

### Linear A.R. RP

A.R. RP $\{X_n\}$ is a linear A.R. RP if $g(\bullet, \bullet)$ is a linear deterministic function like: $g(u, v) = \alpha \cdot u + v$.
Therefore, The defenition of a linear A.R. RP $\{X_n\}$ will look like:

$$X_n = \alpha \cdot X_{n-1} + W_n$$

In our course, we will work only with linear A.R. RP.
Note:
We notice from the constructive examples that all the processes are A.R. processes, but not all of them are linear:

1. $\{W_n\}$ i.i.d, $X_0 = 0$ $X_n = 0 \cdot X_{n-1} + W_n$
2. $\{W_n\}$ i.i.d, $X_0 = 0$ $X_n = 1 \cdot X_{n-1} + W_n$
3. $\{W_n\}$ i.i.d, $X'_0 = 0$ $X'_n = X'_{n-1} \oplus W_n$
4. $\{W''_n\}$ i.i.d, $X''_0 = 0$ $X''_n = 1 \cdot X''_{n-1} + W''_n$

**Stationarity of Linear A.R. RP** Claim:
For a linear A.R. RP $\{X_n\}$ that defined by:

$$X_n = \alpha \cdot X_{n-1} + W_n,\ X_0 \sim f_{X_0}(x),\ W_i \sim f_W(w)\ \forall i,\ \{W_n\}\ i.i.d.$$

The next apply:

- if $|\alpha| \geq 1$ then the RP is not stationary.
  Notice that for $|\alpha| \geq 1$: $\text{Var}(X_n) \underset{n\to\infty}{\rightarrow} \infty$ which is not true for WSS/SSS RP.
- if $|\alpha| < 1$ then the RP is asymptotic SSS (hence also asymptotic WSS) We can show that the limit exists because in time "infinity" $\alpha^n \to 0$
- if $|\alpha| < 1$ **and** the initial condition is matched, then the RP is stationary according to the initial condition.

- Matched initial condition for SSS is $f_{X_0}(x) = f_X^{Stat}(x)$ (meaning the PDF of the initial moment will be the same as the stationary distribution)
- Matched RP for WSS is $\eta_{X_0} = \eta_{Stat}$, $\sigma^2_{X_0} = \sigma^2_{Stat}$, $(\eta_{Stat}, \sigma^2_{Stat})$ out of the 2-nd order statistics of the process (We demand that the variance and the mean of the initial moment will be identical to stationary).

**Example:** We examine the next linear A.R. RP:

$$X_n = \frac{1}{2} \cdot X_{n-1} + W_n,\ X_0 = 0,\ W_n = \begin{cases} 1, & w.p.\ \frac{1}{2} \\ 0, & w.p.\ \frac{1}{2} \end{cases},\ \{W_n\}\ i.i.d.$$

We see that $\alpha = \frac{1}{2}$, hence in worse case scenario the process is asymptotically SSS.
The initial condition is $X_0 = 0$, meaning that it is known deterministically (its value is 0 with probability 1), so $f_{X_0}(x) = \delta(x)$
We calculate for the next samples:

$$X_1 = \frac{1}{2} \cdot X_0 + W_1 = \frac{1}{2} \cdot 0 + W = W = \begin{cases} 1, & w.p.\ \frac{1}{2} \\ 0, & w.p.\ \frac{1}{2} \end{cases}$$
$$\Rightarrow f_{X_1}(x) = \frac{1}{2} \cdot \delta(x) + \frac{1}{2} \cdot \delta(x-1)$$

$$X_2 = \frac{1}{2} \cdot X_1 + W_2 = \frac{1}{2} \cdot X_1 + W = \begin{cases} 1.5, & w.p.\ \frac{1}{4} \\ 1, & w.p.\ \frac{1}{4} \\ 0.5, & w.p.\ \frac{1}{4} \\ 0, & w.p.\ \frac{1}{4} \end{cases}$$
$$\Rightarrow f_{X_1}(x) = \frac{1}{4} \cdot [\delta(x) + \delta(x-0.5) + \delta(x-1) + \delta(x-1.5)]$$

We identify that in the limit $n \to \infty$ the distribution of the sample $X_n$ will go at the limit to the uniform distribution $U(0, 2)$
We can verify this by running the iteration $(n+1)$ for $n \to$:

$$X_{n+1} = \frac{1}{2} \cdot X_n + W_{n+1}$$

$$\begin{aligned}
f_{X_{n+1}}(x) &= \left(f_{\frac{1}{2}\cdot X_n}(x)\right) * (f_W(x)) = (2 \cdot f_{X_n}(2x)) * (f_W(x)) \\
&= \left(2 \cdot \begin{cases} 0.5, & 0 \le 2x \le 2 \\ 0, & else \end{cases}\right) * \left(\frac{1}{2} \cdot \delta(x) + \frac{1}{2} \cdot \delta(x-1)\right) \\
&= \frac{1}{2} \cdot \left(\begin{cases} 1, & 0 \le x \le 1 \\ 0, & else \end{cases}\right) * (\delta(x) + \delta(x-1)) \\
&= \frac{1}{2}\left[\left(\begin{cases} 1, & 0 \le x \le 1 \\ 0, & else \end{cases}\right) * \delta(x) + \left(\begin{cases} 1, & 0 \le x \le 1 \\ 0, & else \end{cases}\right) * \delta(x-1)\right] \\
&= \frac{1}{2}\left[\left(\begin{cases} 1, & 0 \le x \le 1 \\ 0, & else \end{cases}\right) + \left(\begin{cases} 1, & 0 \le x \le 1 \\ 0, & else \end{cases}\right)\right] \\
&\sim \frac{1}{2} \cdot [U(0,1) + U(1,2)] = U(0,2)
\end{aligned}$$

And indeed we got that the distribution of two consecutive samples at infinity does not change, hence this is a stationary distribution.
We can also find a recursive equation for the process:

$$\begin{aligned}
X_n &= \frac{1}{2} \cdot X_{n-1} + W_n \\
&= \frac{1}{2}\left(\frac{1}{2} \cdot X_{n-2} + W_{n-1}\right) + W_n \\
&= \left(\frac{1}{2}\right)^2 \cdot X_{n-2} + \frac{1}{2} \cdot W_{n-1} + W_n = \ldots = \\
&= \left(\frac{1}{2}\right)^n \cdot X_0 + \sum_{i=1}^{n} \left(\frac{1}{2}\right)^{n-i} \cdot W_i
\end{aligned}$$

We identify that the head is exponentially decaying in $n$, so only the sum is relevant. The RP will be S.S.S. if $X_0$ is matched to $X_n, n \to \infty$, i.e. a distribution $U(0,2)$ is matched. In our case, it is not.
Note: In the calculation of the stationary distribution we did (from thetransforms of a random variable):

$$X_n = \frac{1}{2} \cdot X_{n-1} + W_n,\ f_X(x) = f_{\frac{1}{2}X} * f_W(x) = (2 \cdot f_X(2x)) * f_W(x)$$

## Markov Chains

Markov chains are a toll to describe discrete time Markovian processes. We will define a process Markovian, but will use only discrete time.(usually we will deal with homogenus,discrete time, finite number of events Markovian processes).
Markov chains present a development of processes as a series of event that can be used to calculate the probability of a certain set of events, divide the system to events type and more.

### Markovian Process

Markovian RP are a simple and intuitive subset of the random (stochastic) processes.
RP $X(t)$ will be a Markovian process if the distribution of the future states depends solely on the present and is independent in past states.
In fact, RP $X(t)$ will be a Markovian process if the conditional distribution of the future, given the present and the past is equal to the conditional distribution of the future given only the present.(since it is enough to condition the probability in the present value of the process, we call this value the “state” of the process.)
For a continuous time RP (for every set of times $t_1 < t_2 < ... < t_N \in \mathbb{R}$):

$$Pr\left(X(t_N) \in S \mid X(t_{N-1}) = x_{N-1}, ..., X(t_1) = x_1\right) = Pr\left(X(t_N) \in S \mid X(t_{N-1}) = x_{N-1}\right)$$

Or in another way:

$$\begin{aligned} Pr\left(X(t_N) \in S \mid \{X(t), t < t_N\}\right) &= Pr\left(X(t_N) \in S \mid \{X(t), t < t_{N-1}\}\right) \\ &= Pr\left(X(t_N) \in S \mid X(t_{N-1}) = x_{N-1}\right) \end{aligned}$$

And for discrete time RP, $\forall k_1 < k_2 < ... < k_N \in \mathbb{Z}$:

$$Pr\left(X_{k_N} \in S \mid X_{k_{N-1}} = x_{N-1}, ..., X_{k_1} = x_1\right) = Pr\left(X_{k_N} \in S \mid X_{k_{N-1}} = x_{N-1}\right)$$

**Chain Rule (for Markovian Chains)** For discrete time Markovian RP (for an increasing set of times $k_1 < k_2 < ... < k_N \in \mathbb{Z}$):

$$\begin{aligned}
\Pr(X_{k_1} = x_1, X_{k_2} = x_2, ..., X_{k_N} = x_N)& \\
&\overset{76}{=} \Pr(X_{k_1} = x_1) \cdot \Pr(X_{k_2} = x_2, ..., X_{k_N} = x_N | X_{k_1} = x_1) \\
&\overset{77}{=} \Pr(X_{k_1} = x_1) \cdot \Pr(X_{k_2} = x_2 | X_{k_1} = x_1) \\
&\qquad \cdot \Pr(X_{k_3} = x_3, ..., X_{k_N} = x_N | X_{k_2} = x_2, X_{k_1} = x_1) \\
&\overset{78}{=} \Pr(X_{k_1} = x_1) \cdot \Pr(X_{k_2} = x_2 | X_{k_1} = x_1) \\
&\qquad \cdot \Pr(X_{k_3} = x_3, ..., X_{k_N} = x_N | X_{k_2} = x_2) \\
&= ... = \Pr(X_{k_1} = x_1) \cdot \Pr(X_{k_2} = x_2 | X_{k_1} = x_1) \\
&\qquad \cdot \Pr(X_{k_3} = x_3 | X_{k_3} = x_3) \cdot \Pr(X_{k_4} = x_4, ..., X_{k_N} = x_N | X_{k_3} = x_3) \\
&= ... = \Pr(X_{k_1} = x_1) \cdot \prod_{i=2}^{N} \Pr(X_{k_i} | X_{k_{i-1}} = x_{i-1})
\end{aligned}$$

**Chapman Kolmogorov Equation** For a discrete time Markovian RP (for $k < m < n \in \mathbb{Z}$):

$$\Pr(X_n = x_n | X_k = x_k) = \sum_{m} \left[ Pr\left(X_n = x_n \mid X_m = x_m\right) \cdot Pr\left(X_m = x_m \mid X_k = x_k\right) \right]$$

Meaning the probability to get to a specific state $x_n$ at moment $n$ if it is known that in moment $k$ we were at a state $x_k$ is the sum of all the probabilities to get from state $x_k$ to state $x_n$ through a mid-state $x_m$.

We can represent the equation also by summing the possible mid-states, instead of possible mid-time points:

$$\Pr(X_n = x_n | X_k = x_k) = \sum_{X_m} \left[ \Pr(X_n = x_n | X_m = x_m) \cdot \Pr(X_m = x_m | X_k = x_k) \right]$$

[78] By conditional probability

[79] See previous footnote.

[78] Markovian process

Proof:
We can conclude that from the law of total probability:

$$\begin{aligned}\Pr(X_n = x_n | X_k = x_k) &\overset{79}{=} \sum_{X_m} \Pr(X_n = x_n, X_m = x_m | X_k = x_k) \\ &\overset{80}{=} \sum_{X_m} \Pr(X_n = x_n | X_m = x_m, X_k = x_k) \cdot \Pr(X_m = x_m | X_k = x_k) \\ &\overset{81}{=} \sum_{X_m} \Pr(X_n = x_n | X_m = x_m) \cdot \Pr(X_m = x_m | X_k = x_k)\end{aligned}$$

### Finite Markov Chain

Markov chain (discrete time Markovian RP) will be finite if it has a finite (or countable) number of states, i.e. a state can get a value from a finite (or countable) set.

$$X_n \in \{x_1, x_2, ..., x - j\} \overset{82}{=} \{1, 2, ..., J\},\ \forall n \in \mathbb{Z}$$

*Note: From now we will discuss only finite Markov chains.*

### Marginal Distribution of a Markov Chain

For a Markov chain, we would like to know how the state of the process distributes at a certain moment $n$. If the chain is finite, we can present the marginal distribution of $X_n$ by:

$$\underline{\pi}(n) = (\pi_1(n), ..., \pi_J(n)) = (\Pr(X_n = 1), ..., \Pr(X_n = J))$$

### Transformation Matrix of a Markov Chain

For a Markov chain, we would like to know at every point in time the transform probability between states. If the chain is finite, we can present these probabilities usinf a matrix:

$$\underline{\underline{P}}(n) = \begin{bmatrix} P_{11}(n) & \dots & P_{1J}(n) \\ \vdots & \ddots & \vdots \\ P_{J1}(n) & \dots & P_{JJ}(n) \end{bmatrix},\ P_{ij}(n) = \Pr(X_n = j | X_{n-1} = i)$$

[81] Law of total probability.

[82] Bayes’ theorem.

[81] Markovian RP.

[82] For convenience only, we nark the states using numbers.

The matrix is composed of cells $P_{ij}(n)$. The value $P_{ij}(n)$ presents the probability to jet in time $n$ to state $j$ if it is known that in the prior time unit, $n-1$, we were at state $i$.
Identify that row $i$ represents the probability to get from state $i$ in time $n-1$ to every other cells by the location of the column at time $n$.
Identify that column $j$ represents the probability to get from state $j$ at time $n$ to every other states by the location of the row at time $n-1$.
All the elements in the matrix are non-negative(they are probabilities) and the sum of each row is 1 (the sum of the probabilities to transform from a certain state) hence matrix $\underline{\underline{P}}(n)$ is a stochastic matrix.
Conclusion:
A full statistical description of a finite Markov chain up to time $N$ is denoted by the marginal distribution of the initial state and the transformation matrices for every time $n = 1, ..., N$, meaning by:

1. $\underline{\pi}(0)$
2. $\underline{\underline{P}}(n), \quad \forall n = 1, ..., N$

or generally by $\underline{\pi}(0),\ \underline{\underline{P}}(n),\ \ \forall n \in \mathbb{Z}$
Explanation:
Using the law of total probability, we get the marginal distribution of each state using the law of total probability:

$$\underline{\pi}(n) = \underline{\pi}(n-1) \cdot \underline{\underline{P}}(n)$$

And in particular, $\underline{\pi}(1) = \underline{\pi}(0) \cdot \underline{\underline{P}}(1)$ (intuitive about the distribution of the state after the first transformation)
We will show that:

$$\underline{\pi}(n) = \underline{\pi}(n-1)\cdot\underline{\underline{P}}(n) = \underline{\pi}(n-2)\cdot\underline{\underline{P}}(n-1)\cdot\underline{\underline{P}}(n) = ... = \underline{\pi}(0)\prod_{i=0}^{n-1}\underline{\underline{P}}(n-i) = \underline{\pi}(0)\prod_{i=1}^{n}\underline{\underline{P}}(i)$$

### Trellis Diagram

Trellis diagram is a graphical tool for displaying states of Markov chains in time.
In each time column we will dedicate a point for each possible state of the chain. we connect two states $i \rightarrow j$ in consecutive times $n-1 \rightarrow n$ if the probability to transform is positive ($P_{ij}(n) > 0$) and write on connection the transformation probability.

**Example:** We present Trellis diagram fpr a denoted Markov chain by:

$$X_n = mod_4(X_{n-1} + W_n),\ X_0 = 0,\ \ W_n = \begin{cases} 0 & n\ is\ odd \\ \begin{cases} 1 & w.p.\ \frac{1}{2} \\ 0 & w.p.\ \frac{1}{2} \end{cases} & n\ is\ even \end{cases}$$

We can see that $X_n \in \{0, 1, 2, 3\}$ so the chain is finite.
For odd times there is no transformation at all, because $W_n = 0$ for each option, and from the recurrence relation and the finite chain; $X_n = X_{n-1}$.
For even times, there is probability of half to get $W_n = 0$ (and so like before there is no change) and a probability of half to get $W_n = 1$ (and then there is $+1, mod_4$, to get to the same finite group).
it is denoted that the initial state of the chain is $X_0 = 0$, hence the chain looks like:

We notice the chain “knows” the states as it progresses. Trellis diagram for an advanced time (for even times) in thee chain:

We identify that the chain is periodic.
Note:
The probability values aren’t written on the connections (all the transformation to odd times arer 1 - deterministic, and all the transformations to even times are probability half).

### Homogeneous Markov Chains and States Diagram

Markov chain is homogeneous (in time) if:

$$\Pr(X_n = x_D | X_{n-1} = x_S) = \Pr(X_m = x_D | X_{m-1} = x_s),\ \forall m, n \in \mathbb{Z}$$

Meaning, the conditional distribution of the transformation in one unit of time is constant in time. The meaning is that transformation from an arbitrary initial state to another state is independent in time.
specifically, we will show that the probability of the transformation is equal to the probability of the first transformation of the system:

$$\Pr(X_n = x_D | X_{n-1} = x_S) = \Pr(X_1 = x_D | X_0 = x_S)\ \forall n \in \mathbb{Z}$$

Notice that if a Markov chain is homogeneous, then its transformation matrix is independent in time $\underline{\underline{P}}(n) = \underline{\underline{P}}$.
Also: $\underline{\pi}(n) = \underline{\pi}(0) \cdot (\underline{\underline{P}})^n$

*Note: From now, we will discuss only finite homogeneous Markov chains.*

Notice that for a finite homogeneous Markov chain, we don’t have to draw on Trellis diagram more than a single trail (from the final state for an advanced time) because there is no change in the probability of the transformations in time. Hence finite homogeneous Markov chains are represented/denoted by state diagrams and not Trellis diagrams. In fact, we can transform a single trail of Trellis diagram to state diagram by “folding” it.

**Exmaple:** We continue with this example, and treat is as a case of a “hoping beetle”.
When the beetle is in position 1, it has the same probability to stay and to switch states, the beetle is not comfortable in state 2 hence it will go back to state 1.
Assume we know that in time 0 we positioned the beetle in state 1, meaning $\underline{\pi}(0) = (1, 0)$.
From the state diagram we build the transformation matrix and use the transformation marginal distribution rule:

$$\underline{\underline{P}}(n) = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{bmatrix},\ \underline{\pi}(n) = \underline{\pi}(n-1)\cdot \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{bmatrix} = \underline{\pi}(0)\cdot \left( \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{bmatrix} \right)^n = (1,0)\cdot \left( \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{bmatrix} \right)^n$$

We show the marginal distributions of the states of the processes in time:

$$\underline{\pi}(1) = (\frac{1}{2}, \frac{1}{2}),\ \underline{\pi}(2) = (\frac{3}{4}, \frac{1}{4}),\ \underline{\pi}(3) = (\frac{5}{8}, \frac{3}{8}),\ \underline{\pi}(4) = (\frac{11}{16}, \frac{5}{16})$$

We would like to know the marginal distribution of the beetle state for $n \to \infty$ (if it exist)
We can calculate ot by taking $\underline{\underline{P}}(n)$ in power $n$, where $n \to \infty$ and hope for divergence. We get:

$$\left(\underline{\underline{P}}(n)\right)^n = \left( \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{bmatrix} \right)^n \underset{n\to\infty}{\to} \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix}$$

Notice that:

$$\lim_{n\to\infty} \underline{\pi}(n) = \lim_{n\to\infty} \left( \underline{\pi}(0) \cdot \left( \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{bmatrix} \right)^n \right) = (1,0) \cdot \begin{bmatrix} \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} \end{bmatrix} = \left( \frac{2}{3}, \frac{1}{3} \right)$$

We notice that in fact the initial state did not matter. If all the elements of each column in the matrix $(\underline{\underline{P(n)}})^n$ are identical the sum of all the initial vector will be 1 (even if the initial stat has a non-deterministic distribution)

then the distribution of the final state will be a row vector from the matrix $(\underline{\underline{P(n)}})^n$ - doesn't depend on the marginal distribution $\underline{\pi}(0)$.
In another way with a little intuition, we can identify the direction of convergence and guess that the probability in "infinity" to be at state 1 is $\frac{2}{3}$, meaning $\lim\limits_{n\to\infty} \pi_1(n) = \frac{2}{3}$ From the transformation rule:

$$\underline{\pi}(n) = \underline{\pi}(n-1) \cdot \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{bmatrix} \Rightarrow \pi_1(n) = \frac{1}{2} \cdot \pi_1(n-1) + 1 \cdot \pi_2(n-1)$$

If the assumption holds, then for a very large $n$:

$$\pi_1(n) = \pi_1(n-1) = \frac{2}{3},\ \pi_2(n-1) = 1 - \pi_1(n-1) = \frac{1}{3}$$

We insert it back to the equation to check ourselves:

$$\frac{2}{3} \overset{?}{=} \frac{1}{2} \cdot \frac{2}{3} + 1 \cdot \frac{1}{3} \Rightarrow TRUE$$

### Stationary Distribution

As we have seen in the beetle case, there are chains where the marginal distribution of the chain converge (meaning for every constant $n \geq k$ starting a certain $k$)
We say that a finite homogeneous Markov chain has stationary distribution is its distribution is stationary, i.e.:

$$\exists k : \underline{\pi}(n) = \underline{\pi}(k),\ \forall n \geq k$$

**Stationary Vector** Vector $\underline{x}$ is a stationary vector for matrix $\underline{\underline{A}}$ if $\underline{x} = \underline{x} \cdot \underline{\underline{A}}$
Note: Notice that matrix $\underline{\underline{A}}$ nust be a square matrix of the same dimension of $\underline{x}$.

**Vector of Stationary Distribution** We denote the vector of stationary distribution (we call it the stationary distribution) of a Markov chain by $\underline{\pi}'$ or $\underline{\pi}^\infty$. In a finite, homogeneous Markov chain, we can say that for the stationary distribution:$\underline{\pi}' = \underline{\pi}' \cdot \underline{\underline{P}}$

**Stationarity in Markov Chains** As you recall, for SSS the demand is that the marginal distribution will be independent in time. In Markov chains, if the state has a marginal distribution identical to the stationary distribution, then all the states that will follow will have the same marginal distribution.
For finite,homogeneous Markov chains we define:

1. if $\underline{\pi}(k) = \underline{\pi}'$ then for all $n \geq k$: $\underline{\pi}(n) = \underline{\pi}'$, and we call this chain stationary asymptotic.
   The meaning is that if in certain time $k$ the marginal distribution of the state $X_k$ is identical to the stationary distribution, then the chain made it to the stationary state, and the marginal distribution of all the next states $X_n$, $\forall n \geq k$ will not change.

2. If $\underline{\pi}(0) = \underline{\pi}'$ , then $\forall n \in \mathbb{Z}:$ $\underline{\pi}(n) = \underline{\pi}'$ and we call the chain stationary. This is SSS, because we get that the marginal distribution is independent in time because in every time point the marginal distribution of the state is constant - it starts from the stationary distribution.
   This demand is equivalent to the demand that the initial distribution $\underline{\pi}(0)$ will be a left eigenvector of the matrix $\underline{\underline{P}}$ with eigenvalue $\lambda = 1$

Claim:
If for a Markov chain with stationary distribution $\underline{\pi}'$ exists, the limit $\lim\limits_{n\to\infty} \underline{\pi}(n) = \underline{\pi}''$, then $\underline{\pi}''$ will converge to the stationary vector of the chain, i.e.:$\underline{\pi}' = \underline{\pi}''$.
Proof:
From the existence of the limit $\lim\limits_{n\to\infty} \underline{\pi}(n) = \underline{\pi}''$, we can say that $\lim\limits_{n\to\infty} \underline{\pi}(n+1) = \underline{\pi}''$, hence:

$$\lim_{n\to\infty} \underline{\pi}(n) = \lim_{n\to\infty} \underline{\pi}(n+1)$$

$$\lim_{n\to\infty} \underline{\pi}(n) = \lim_{n\to\infty} \left[\underline{\pi}(n) \cdot \underline{\underline{P}}\right]$$

$$\lim_{n\to\infty} \underline{\pi}(n) = \lim_{n\to\infty} \underline{\pi}(n) \cdot \underline{\underline{P}}$$

$$\underline{\pi}'' = \underline{\pi}'' \cdot \underline{\underline{P}}$$

We get that $\underline{\pi}''$ is a stationary vector of matrix $\underline{\underline{P}}$ and from the uniqueness of the limit $\underline{\pi}' = \underline{\pi}''$

### Characterization of Markov Chains

We would like to answer the natural question: when will a Markov chain behave in a way that there is no dependence between the initial state (like the beetle), namely we would like to know when the chain will converge into a stationary marginal distribution. Of course we will discuss only finite homogeneous Markov chains. For the characterization of the chain we will classify the states of the system by states diagram (or transformation matrix)
Definitions of classification:

1. Accessibility: We say that state $j$ is accessible from state $i$, $(i \rightarrow j)$ if there is a way in the state diagram to move, in a finite number of steps, from state $i$ to state $j$ (i.e., if there is a directed path from $i$ to $j$).

2. Communicating states: We say that state $i$ and state $j$ communicate $(i \leftrightarrow j)$ if $i \rightarrow j$ and $j \rightarrow i$
   Note - communication is a transitive, meaning if $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$

3. Class: A class is a maximal set of connected states.
   In a class, all the states are connected to each other and there is no state that connected to a state that is not in the class.

4. Recurrent state: A state that is accessible with every state that can be accessible with it (a state that is accessible to all all the states it can communicate with), meaning a state we can go back to if we left it.

5. Transient state: A state that is not recurrent. Meaning a state that we can go from it to another state, but can't go back from it.
   A transient state will exist only a finite amount of times.

6. Recurrent class: A class that all of its elements are recurrent. In order to check if a class is recurrent, it is enough to check if a single state in the class is recurrent.

7. Transient class: A class that is not recurrent. A class that all of its states are transient. In order to check if a class is transient, it is enough to check if a single state in the class is transient.

8. Return times: Return times of a state $i$ are the collection of all the lengths of the returning paths that go out of state $i$ and back to it.
   Return times (if exist) are an infinite series and marked by $n_1^i, n_2^i, ...$

9. Period: The period of state $i$ is the greatest common divider of the return times of the state, i.e.: $d(i) = GCD\left(\{n_k^i\}_{k=1}^{\infty}\right)$

Note: we can classify classes (recurrent/transient) according to the transform matrix in the next ways:

1. Identify recurrent classes by identifying square sub-matrices so that in the same rows there are only zeroes, except the elements of the sum-matrix (from the states of rows of sub-matrices you can't go to states outside the sub-matrix).

2. Identify transient classes by identifying square sub-matrices so that in the same columns there are only zeros other than the elements of the sub-matrix(can't get to any row state of the sub-matrix from the outside)

**Periodic Classes** A class will be a periodic class if for state $i$ in the class, $d(i) > 1$.
Claim:
All the states of the same class have the same period.
Proof:
We take two states $i, j$ from the same class. We denote $d(i), d(j)$.
We will show that the period $d(i)$ is a divider of $d(j)$, and the opposite, meaning $d(j) = d(i)$
We say there is a path from state $i$ to state $j$ with length $s$, and that there is also an opposite pate with length $t$.
From the definition of a state, we know that $d(i)$ is a divider of all the return times of the state $i$, hence:

$$d(i)|s+t,\ d(i)|s+t+n_k^f\ \forall k \Rightarrow d(i)|n_k^f\ \forall k \Rightarrow d(i)|d(j)$$

If we switch roles between $i, j$, we get $d(j)|d(i)$.
Because we get $d(i)|d(j)$ and $d(j)|d(i)$ (meaning they divide each other), then $d(i) = d(j)$.

**Example 1:** Denote a state diagram (or transformation matrix)

$$\begin{bmatrix} 0 & \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

We will classify by classes (recall that in transient class all the states transient states and in recurrent class all the states are recurrent):
Classes (Identified by square sub-matrices with zero row/column elements):

- $T_1 = \{1\}$ - transient class
- $T_2 = \{2, 3\}$ - recurrent class
- $T_3 = \{4, 5\}$ - recurrent class

Both the sub-matrices of the recurrent classes are identical. In addition to these sub-matrices we have a stationary distribution (not for the whole chain):

$$\lim_{n\to\infty}\left(\left(\begin{bmatrix}0 & 1\\ \frac{1}{2} & \frac{1}{2}\end{bmatrix}\right)^n\right)=\begin{bmatrix}\frac{1}{3} & \frac{2}{3}\\ \frac{1}{3} & \frac{2}{3}\end{bmatrix}$$

And for the whole matrix:

$$\lim_{n\to\infty}\left(\left(\underline{\underline{P}}\right)^n\right)=\begin{bmatrix}0 & \frac{1}{9} & \frac{2}{9} & \frac{2}{9} & \frac{4}{9}\\ 0 & \frac{1}{3} & \frac{2}{3} & 0 & 0\\ 0 & \frac{1}{3} & \frac{2}{3} & 0 & 0\\ 0 & 0 & 0 & \frac{1}{3} & \frac{2}{3}\\ 0 & 0 & 0 & \frac{1}{3} & \frac{2}{3}\end{bmatrix}$$

The value of the first row vector is equal to (in accordance to the probability of moving to each class and their stationary distribution):

$$\frac{1}{3}\cdot\underline{\pi}^{\infty}(T_2)+\frac{2}{3}\cdot\underline{\pi}^{\infty}(T_3)$$

Both recurrent classes are not periodic, since there is a path of length one (from state 3 to itself and so for state 5)

**Example 2 (calculation of periods)** Denote a state diagram:
We identify that all the states are communicating and are all one recurrent class. We calculate path for one step, say for state 1:

$$n_1^1=4,\ n_2^1=8,\ n_3^1=10,\ n_4^1=12,\ n_5^1=14,\ ...$$

We identify that the return times will continue to be even, so:

$$d(1)=GCD(4,8,10,12,14)=2$$

Meaning the period of the state is 2, hence the period of the whole class is 2, this class is all the states in chain, so this is also the period of the chain.

### "Forgetting the Past"

We have seen in the beetle example that there is no importance to the initial state. Meaning there is no dependence in the initial condition.
We define this type of behavior as "forgetting the past" (no memory of the past states)
We say a Markovian chain forgets the past (Ergodicity) if:

$$\lim_{n\to\infty}\Pr(X_n=j|X_0=i)=\pi_j^{\infty}$$

This condition has to apply for every possible state $j \in \{1, ..., J\}$ in the chain and initial state $i$.
Hence, if the chain is Ergodic, or:

$$\lim_{m\to\infty}(\underline{\pi}(n)) = \lim_{n\to\infty}(\underline{\pi}(0)\cdot(\underline{\underline{P}})^n) = \underline{\pi}^\infty = (\underline{\pi}_1^\infty, ..., \underline{\pi}_J^\infty)$$

Where $\underline{\pi}(0)$ is the vector of initial conditions. Most of the times we will know the initial state of the system deterministically, and then $\underline{\pi}(0) = \underline{e}_i$, meaning the marginal distribution of the initial state is a unit vector of dimension $J$ (zeros vector with 1 in the $i$-th place, that represents the probability 1 to be in the $i$-th place at time 0). It can also be non-deterministic distribution.
If the said limit exists, then necessarily the property of the stationary vector of the cain apply, meaning: $\underline{\pi}^\infty = \underline{\pi}^\infty \cdot \underline{\underline{P}}$ If the chain is ergodic, then the limit apply for every initial distribution $\underline{\pi}(0)$, especially for the initial distribution that represents deterministic knowledge of the initial state - meaning the unit vector from the standard base $\underline{e}_i$ , so:

$$\lim_{n\to\infty}(\underline{e}_i\cdot(\underline{\underline{P}})^n) = \underline{\pi}^\infty$$

The meaning of the multiplication of the unit vector $\underline{e}_i$ with the matrix is getting the $i$-th row of the matrix. From this we can deduce that all the rows of $(\underline{\underline{P}})^n$ have to be equal to the stationary vector $\underline{\pi}^\infty$.
We say that a finite homogeneous Markov chain is ergodic (forgets the past) iff we cab say about its transform matrix $\underline{\underline{P}}$:

$$\lim_{n\to\infty}((\underline{\underline{P}})^n) = \begin{bmatrix} \leftarrow & \underline{\pi}^\infty & \rightarrow \\ & \vdots & \\ \leftarrow & \underline{\pi}^\infty & \rightarrow \end{bmatrix}$$

Notice it doesn't prevent from the marginal distribution of the initial state to be a different vector from the unit vector, because the sum of it elements is one and the rows of the matrix $(\underline{\underline{P}})^n$ are all identical if the chain is ergodic. We can also see it applies to the beetle example.

## Ergodicity

A RP will be ergodic if for a long enough sample, all time statistics function will converge into the statistics of the process. In other words, RP is ergodic if we can learn its full distribution (i.e. all its statistical parameters) from looking at the (long enough) single sample function.
In this way, for an ergodic RP, we can deduce statistical properties by the

mean of the sample and with no initial statistical knowledge on the process. We stress that for an ergodic RP we get convergence from a single sample. In general, processing of RP (or system planning) is based usually on knowing their statics.
For a RP with no known (or insufficiently known) statistics, if it is ergodic we can "measure" its statistics using a long enough sample.

### Ergodicity in Markov Chains

We have seen that there are some Markov chain that are ergodic, and some that are not. We discuss the topics of ergodicity on (finite, homogeneous) Markov chains.

**Perron-Frobenius Theorem** For a finite, homogeneous Markov chain with transform matrix $\underline{\underline{P}}$:

1. There is always (at least one) statistical solution to the equation $\underline{\pi} = \underline{\pi} \cdot \underline{\underline{P}}$. Meaning there will always be for the transform matrix $\underline{\underline{P}}$ a non-negative left eigenvector with an eigenvalue $\lambda = 1$ (from the properties of a stochastic matrix)

2. If the chain has only one recurrent class, then the solution to the equation $\underline{\pi} = \underline{\pi} \cdot \underline{\underline{P}}$ is unique.

3. If the chain has $r$ recurrent classes, then the equation $\underline{\pi} = \underline{\pi} \cdot \underline{\underline{P}}$ has $r$ independent linear solutions.
   Every statistical combination (linear combination with non-negative coefficient with the sum of 1) of these solutions (the stationary vectors) is a eigenvector of the matrix, the meaning is that these stationary vectors span all the eigenvectors of the matrix $\underline{\underline{P}}$.
   As said, these stationary vectors are linearly independent, hence will look like (we arranged the matrix $\underline{\underline{P}}$ for convenience):

$$\underline{\underline{P}} = \begin{bmatrix} * & * & 0 & 0 & 0 \\ * & * & 0 & 0 & 0 \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{bmatrix}, \ \underline{\pi}_1^{Stat} = (*, *, 0, 0, 0), \ \underline{\pi}_2^{Stat} = (0, 0, *, *, *)$$

4. The chain is ergodic if it has only one class, and it is recurrent and not periodic ($d = 1$).
   In this case, the chain is asymptotically stationary, and converges to the distribution $\underline{\pi}^{\infty}$ for every $\underline{\pi}(0)$

5. The chain is ergodic and with transient phenomenon if it has one recurrent class (which is not periodic) and in addition the chain has transient states (or transient classes).
   In this state it is guaranteed that if we start with a recurrent class we stay in it and be reduced to case 4, and if we start with a transient state then in some point in time we will eventually get to recurrent class. In this case we can learn from a single sample function only on the recurrent class (and about some of the transient states but not clearly enough for statistics).

   - For an ergodic chain (transient or not):

   $$\lim_{n\to\infty}\left((\underline{\underline{P}})^n\right) = \begin{bmatrix} \leftarrow & \underline{\pi}^{\infty} & \rightarrow \\ & \vdots & \\ \leftarrow & \underline{\pi}^{\infty} & \rightarrow \end{bmatrix}$$

   We go into details of the theroem for identification of non-ergodic chains:

6. If a chain has two recurrent classes or more, then it is not ergodic.
   In this case, If we start at a recurrent state, then we stay in it. If we start at a transient state, then we get to one of the recurrent classes that communicate with it. Hence we can't deduce the statistics of all the states from one sample.
   If In this case, all the recurrent classes are not periodic, then $(\underline{\underline{P}})^n$ converges, but its rows are not identical.

7. If a chain has one periodic recurrent class $d$, then it is not ergodic. To this case we can also add transient effect.
   Because the recurrent class is periodic then $(\underline{\underline{P}})^n$ does not converges (the limit $\lim n \to \infty \left((\underline{\underline{P}})^n\right)$ does not exists). The phenomenon of periodically prevents stationarity, but we get that $\forall i = 0, 1, \ldots, d-1$, $(\underline{\underline{P}})^{d\cdot n+i}$ does converge.

**Examples of Non-Ergodic Markov Chains:**

1. Denote the next periodic chain:
   The chain has to states with certain transitions. It is clear that it does not converge, because it "jumps" from state to state, and there is importance to the parity of time. If $\underline{\pi}(0) = (1, 0)$, then on odd times we will be in state 2 and in even times in state 1:

   $$\underline{\underline{P}} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

The chain has a single recurrent state with period of 2, hence by PF theorem the chain is not ergodic.
The transformation matrix shows us that the limit $\lim\limits_{n\to\infty}\left(\left(\underline{\underline{P}}\right)^{n}\right)$, but the the limit exists for $(\underline{\underline{P}})^{2\cdot n+1}, (\underline{\underline{P}})^{2\cdot n}$

$$\lim_{n\to\infty}\left((\underline{\underline{P}})^{2\cdot n}\right) = \lim_{n\to\infty}\left(\left(\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\right)^{2\cdot n}\right) = \lim_{n\to\infty}\left(\left(\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\right)^{n}\right) = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

$$\lim_{n\to\infty}\left((\underline{\underline{P}})^{2\cdot n+1}\right) = \lim_{n\to\infty}\left(\left(\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\right)^{2\cdot n+1}\right) = \lim_{n\to\infty}\left(\left(\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\right)^{2\cdot n}\right)\cdot\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} = \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}$$

And the limits are different. In any way, the solution to the equation $\underline{\pi} = \underline{\pi}\cdot\underline{\underline{P}}$ exists, and $\underline{\pi}^{Stat} = \left(\frac{1}{2},\frac{1}{2}\right)$

2. Denote the next chain:

$$\underline{\underline{P}} = \begin{bmatrix}0 & 1/5 & 0 & 1/5 & 0 & 3/5 & 0\\ 0 & 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0 & 0 & 1 & 0\end{bmatrix}$$

We identify classes using the transfer matrix (recurrent classes,transient class):

$$T_1 = \{1\},\ T_2 = \{2,3\},\ T_3 = \{4,5\},\ T_4 = \{6,7\}$$

We calculate periods,and get: $d(T_2) = 1,\ d(T_3) = 1,\ d(T_4) = 2$
This is a chain with transient phenomenon and a number of recurrent classes. In addition, one of the recurrent classes have a period.
Because of the period, the limit lim $\lim\limits_{n\to\infty}\left((\underline{\underline{P}})^{n}\right)$ doesn’t exist for the matrix $\underline{\underline{P}}$.
We have three recurrent classes, hence the equation $\underline{\pi} = \underline{\pi}\cdot\underline{\underline{P}}$ has three solutions, one for each class, and we call them by the name of the classes $\underline{\pi}_{T_2}^{Stat}, \underline{\pi}_{T_3}^{Stat}, \underline{\pi}_{T_4}^{Stat}$ (but the value of $\underline{\pi}_{T_4}^{Stat}$ is dependent on time, so its value is different) and the vector of the stationary distribution is a combination of them, meaning every vector like:

$$\underline{\pi}^{Stat} = p_1\cdot\underline{\pi}_{T_2}^{Stat} + p_2\cdot\underline{\pi}_{T_3}^{Stat} + p_4\cdot\underline{\pi}_{T_4}^{Stat},\ p_1,p_2,p_3 \geq 0,\ p_1+p_2+p_3 = 1$$

We will not want this stationary distribution for every set $(p_1, p_2, p_3)$, it depends of the coefficients and the initail conditions (the meaning is

that every set like that will make a vector that solves equation $\underline{\pi} = \underline{\pi} \cdot \underline{\underline{P}}$)
We show the limits for $(\underline{\underline{P}})^n$

$$\lim_{n\to\infty}\left((\underline{\underline{P}})^{2\cdot n}\right) = \left[\begin{array}{ccccccc} & & 1/5\cdot\underline{\pi}_{T_2}^{Stat} + & 1/5\cdot\underline{\pi}_{T_3}^{Stat} + & 3/5\cdot\underline{\pi}_{T_4}^{Stat,Even} & & \\ 0 & 1/3 & 2/3 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 2/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 2/3 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 2/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right],$$

$$\lim_{n\to\infty}\left((\underline{\underline{P}})^{2\cdot n+1}\right) = \left[\begin{array}{ccccccc} & & 1/5\cdot\underline{\pi}_{T_2}^{Stat} + & 1/5\cdot\underline{\pi}_{T_3}^{Stat} + & 3/5\cdot\underline{\pi}_{T_4}^{Stat,Even} & & \\ 0 & 1/3 & 2/3 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 2/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 2/3 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 2/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right]$$

Notice the movement of class $T_4$, hence also in its stationary distribution.

**Formal Definition for Ergodicity** We have seen that RP is ergodic if it joint distribution(all of its statistical parameters) can be fully learned from looking at a long enough single sample function.
In fact, a RP is ergodic iff the time average of any function applied to the process converges to the expected value of the function, meaning iff the time average is equal to the statistical mean for any function $g : \mathbb{R}^M \to \mathbb{R}$ and every set of times $k_1, ..., k_M \in \mathbb{Z}$, meaning:

$$\lim_{N\to\infty}\left(\frac{1}{N}\cdot\sum_{i=0}^{N-1} g(X_{k_1+i}, ..., X_{k_M+i})\right) \overset{83}{=} \lim_{N\to\infty}\left(\mathbb{E}\left[g(X_{k_1+N}, ..., X_{k_M+N})\right]\right) \overset{84}{=} \mathbb{E}\left[g(X_{k_1}, ..., X_{k_M})\right]$$

And for a continuous time RP it will apply for every time set $t_1, ..., t_M \in \mathbb{R}$:

$$\lim_{N\to\infty}\left(\frac{1}{T}\cdot\int_{\tau=0}^{T} g(X(t_1+\tau), ..., X(t_M+\tau))d\tau\right) \overset{85}{=} \lim_{N\to\infty}\left(\mathbb{E}\left[g(X(t_1+\tau), ..., X(t_M+\tau))\right]\right)$$
$$\overset{86}{=} \mathbb{E}\left[g(X(t_1), ..., X(t_M))\right]$$

[85] Asymptotic stationary
[84] Stationary

In fact, Ergodicity is the Law of Large Numbers (LLN) because the mean of time converges into the statistical mean. For example, for a system with probability of error of 0.1 per bit, where the series of bits is ergodic, we can guarantee 10% total error in time (for a long enough period of time).

**Law of Large Numbers (LLN)** For a sequence $\{X_n\}$ RP i.i.d.:

$$\lim_{n\to\infty}\left(\frac{1}{N}\cdot\sum_{i=0}^{N-1}X_n\right)=\mathbb{E}[X_n]$$

Notice that LLN is a private case of Ergodicity where $g(x)=x$ (and then the content of the limit is a time average which approaches the mean).
In order to check if the sequence $\{X_n\}$ must be i.i.d. for LLN, we look at the function $g(x,y)=x\cdot y$. For a $\{X_n\}$ RP ergodic and stationary, we choose set of times $k_1=1,k_2=3$, and get:

$$\begin{aligned}\lim_{N\to\infty}\left(\frac{1}{N}\cdot\sum_{i=0}^{N-1}(X_{1+n}\cdot X_{3+n})\right)&=\lim_{N\to\infty}\left(\frac{1}{N}\cdot\sum_{i=0}^{N-1}(X_{n+1}\cdot X_{n+3})\right)\\&\overset{87}{=}\mathbb{E}\left[(X_1\cdot X_3)\right]=R_{XX}(n+1,n+3)\\&\overset{88}{=}R_{XX}(-2)\overset{89}{=}R_{XX}(2)\end{aligned}$$

In general, for a set of times $k_1,k_2$:

$$\lim_{N\to\infty}\left(\frac{1}{N}\cdot\sum_{i=0}^{N-1}(X_{k_1+n}\cdot X_{k_2+n})\right)=R_{XX}(|k_2-k_1|)$$

We get the property of memory, hence we can lower our demand and the existence of Ergodicity will alow us convergence of the time average to the the statistical mean even for non-i.i.d. RP.

**Finding Statistics of an Ergodic Markov Chain** From the definition of Ergodicity for an ergodic RP we can find all of its statistical parameters from knowing a single sample function, long as we want. We would like to

[87] We have assume here that the process is asymptotically stationary w.r.t. the function $g$.

[86] Assuming that the process is stationary w.r.t. the function $g$.

[89] Ergodicity

[90] Stationary

[91] Properties of correlation function - even function.

learn a certain order distribution for an ergodic Markov chain from looking at a sample function.
We will see how to estimate the stationary distribution and the transformation matrix given a sample function of an ergodic Markov chain.
We will do so using the beetle chain (which we know is ergodic and stationary), reminder:

$$\underline{\underline{P}}(n) = \begin{bmatrix} 1/2 & 1/2 \\ 1 & 0 \end{bmatrix}, \ \underline{\pi}^{stat} = \left(\frac{2}{3}, \frac{1}{3}\right)$$

And assume an arbitrary sample function (the elements are lower case letters because this is a realization of the process): $\{X_n\}_{n=0}^{\infty} = 1, 1, 2, 1, 2, 1, 2, 1, 1, 2, ...$

**Estimating the Stationary Distribution of an Ergodic Markov Chain** We would like to estimate the stationary distribution $\underline{\pi}^{Stat}$, we will calculate the elements of the vector cell-by-cell. the probability of each cell will be the frequency of the state in the sample function (we use the indicator function $I_a(x)$):

$$\pi_a^{stat}(N) = \frac{1}{N}\sum_{i=0}^{N-1} I_a(x_i), \quad I_a(x) = \begin{cases} 1 & x = a \\ 0 & else \end{cases}$$

In this way, for an estimation $\pi_a^{stat}(N)$ we count the number of times state $a$ apeered in $N$ samples, and divide by $N$, i.e. averaging over the amount it appears.

**Estimating Transformation Matrix of (Homogeneous) Ergodic Markov Chain** We would like to estimate the transformation matrix $\underline{\underline{P}}$, so we calculate the elements of the matrix cell-by-cell.
We estimate the probability of transition from state $a$ to state $b$, $\underline{\underline{P}}(a, b)$ in the next way:

1. Estimate the joint distribution $\Pr(X_n = a, X_{n+1} = b)$ by summing the appearances of states $b$ after state $a$, notice that for a sequence with size $N$ there are $N - 1$ passes (we use the double indicator function $I_{ab}(x)$)

$$p(a, b)(N) = \frac{1}{N-1}\sum_{i=0}^{N-2} I_{ab}(x_i, x_{i+1}), \quad I_{ab}(x, y) = \begin{cases} 1 & x = a, y = b \\ 0 & else \end{cases}$$

2. Calculate from the above estimate probability the conditional distribution $\Pr(X_{n+1} = b|X_n = a)$ (that based on the estimation) using the

equation for conditional distribution (and the probability estimation for state $a$):

$$\Pr(X_{n+1}=b|X_n=a)=\frac{\Pr(X_{n+1}=b\cap X_n=a)}{\Pr(X_n=a)}=\frac{\Pr(X_n=a,X_{n+1}=b)}{\Pr(X_n=a)}$$

$$\Rightarrow P_{ab}(N)=\underline{\underline{P}}(a,b)(N)=\frac{p(a,b)(N)}{\pi_a^{stat}(N)}$$

**Example:** For the beetle example with the (small- $N=10$) sample function above we estimate the stationary distribution:

$$\underline{\pi}^{stat}(10)=\left(\pi_1^{stat}(10),\pi_2^{stat}(10)\right)=\left(\frac{6}{10},\frac{4}{10}\right)=\left(\frac{3}{5},\frac{2}{5}\right)=(0.6,0.4)$$

Calculate the probability to pass for a single cell (that represents the probability of passing from state 2 to itself):

$$p_{2,2}(10)=\frac{p(2,2)(10)}{\pi_2^{stat}(10)}=\frac{0/9}{0.4}=0$$

For the stationary distribution indeed $(0.6,0.4)\approx(\frac{2}{3},\frac{1}{3})$, and there is certainly probability 0 to pass.
Notes:

- If there is only one recurrent class, then the joint distribution of any order can be deduced (at least asymptotically) from a single sample function, and:

  $$\lim_{N\to\infty}\left(\pi_a^{stat}(N)\right)=\pi_a^{stat},\ \forall a\in\{1,...,J\}\Rightarrow\lim_{N\to\infty}\left(\underline{\pi}^{stat}(N)\right)=\underline{\pi}^{stat}$$

  $$\lim_{N\to\infty}\left(P_{ab}(N)\right)=\underline{\underline{P}}(a,b)\ \ \forall a,b\in\{1,...,J\}$$

  If there is no (at least) asymptotic stationary, then we won’t get such a convergence. Assume for a recurrent class with two states with a junp with period of 2 (from state one to state two, and from state two to state one) there is no convergence to a general $N$ but a distinction between odd and even.

- If the chain has more than one recurrent class we can’t learn from it by looking at a single sample function because we will visit, at most, one recurrent class so we will remain oblivious of all others.[90]

- In non-ergodic chain we can never learn the statistics by looking at a single realization.

[90] “At most”, since we may never enter a recurrent class if there are transient states.

**The Connection Between The Intuitive and the Formal Definition of Ergodicity** We have seen that Ergodicity, intuitively, is the ability to learn the full statistics of a RP by using a single sample function. After that, we have also seen a formal definition where Ergodicity was the time average of a certain time dependent function on a RP to the mean of the same function and its samples. We will see that the definitions are equal.
First direction:
Denote that the time average converges into the mean, we would like to show that we can learn the statistics of the process. Demonstration for a 2-nd order marginal distribution(two point in time of the process, we can do the same for every order similarly), by using the double indicator function:

$$\frac{1}{N}\sum_{i=0}^{N-1} I_{ab}(X_{k_1+i}, X_{k_2+i}) \to \mathbb{E}\left[I_{ab}(X_{k_1}, X_{k_2})\right] = \Pr(X_{k_1} = a, X_{k_2} = b)$$

We have seen that using the fact the time average converges to the mean we can use the Ergodicity properties on the function $g(x,y)I_{ab}(x,y)$ and get the marginal distribution $\Pr(x = a, y = b)$
Second direction:
Denote that we can learn the statistics of the process, we want to prove that the time average converges to the mean.
We assemble a certain function (again we demonstrate for a function of two samples of the RP) using the indicator function (we demonstrate for a finite RP for convenience, the same holds for integrals instead of sums) by:

$$g(x,y) = \sum_{a=1}^{J}\sum_{b=1}^{J} g(a,b)I_{ab}(x,y)$$

This will suite any function because we sum all the possible value times the indication only for relevant values, meaning only for $(x,y) = (a,b)$, then the sum of elements will not be zero, but $g(a,b)\cdot 1 = g(x,y)$
By knowing that the statistics converges, we would like to see that averaging in time converges to the mean of the function:

$$\begin{aligned}\frac{1}{N}\sum_{i=0}^{N-1} g(X_{k_1+i}, X_{k_2+i}) &= \frac{1}{N}\sum_{i=0}^{N-1}\left[\sum_{a=1}^{J}\sum_{b=1}^{J} g(a,b) I_{ab}(X_{k_1+i}, X_{k_2+i})\right] \\ &= \sum_{i=0}^{N-1}\left[\sum_{a=1}^{J}\sum_{b=1}^{J} g(a,b)\frac{I_{ab}(X_{k_1+i}, X_{k_2+i})}{N}\right] \\ &\to \sum_{i=0}^{N-1}\left[\sum_{a=1}^{J}\sum_{b=1}^{J} g(a,b)\Pr(X_{k_1}=a, X_{k_2}=b)\right] \\ &= \mathbb{E}\left[g(X_{k_1}, X_{k_2})\right]\end{aligned}$$

And we get the wanted convergence.

**Weak Ergodicity** As remembered, the formal definition of Ergodicity is convergence of the time average of every function to the its mean. There are processes that don’t have this trait for every function, but only for some of them (weak Ergodicity).
We say that a RP $X(t) or X_n$ is Ergodic for a specific function $g : \mathbb{R}^M \to \mathbb{R}$ if the Ergodicity property apply for this function, meaning:

$$\lim_{T\to\infty}\left(\frac{1}{T}\int_{\tau=0}^{T} g(X(t_1+\tau), ..., X(t_M+\tau))d\tau\right) = \mathbb{E}\left[g(X(t_1), ..., X(t_M))\right] \ \ \forall t_1, t_M \in \mathbb{R}$$

For this convergence we will demand that process to be at least asymptotic stationary.
Because not every process is fully Ergodic, we can estimate certain sized in an inaccurate way.

**Estimating Parameters of a RP** Assume we try to estimate a specific parameter of a RP ($\alpha$ coefficient of a linear A.R. RP, mean of RP, phase, etc.) we denote the estimated parameter as $\Theta$ and its estimator as $\Theta(T)$. We identify that $\Theta$ is a deterministic value (we just don’t know its value) while $\Theta(T)$ is a RP (based on the samples of the RP in time $[0, T]$).

**Existence of the MSE limit For a Random Sequence** We would like to understand better the existence of the limit for a random sequence and understand the meaning of a random sequence that converges into a constant.
In fact, we try to understand the meaning of:

$$\lim_{T\to\infty}\left(\Theta(T)\right) = \Theta$$

We say a RP $X(t), X[n]$ approach to a constant $\alpha$ as $N \to \infty$ if:

$$\lim_{N\to\infty} \mathbb{E}\left[(X[N]-\alpha)^2\right] = 0$$

and we mark: $X[n] \overset{MSE}{\to} \alpha$
Note:
MSE approaching → statistically approaching:
Assume that $X[n] \overset{MSE}{\to} \alpha$, so:

$$\Pr(|X[N]-\alpha| \geq \epsilon) = \Pr((X[N]-\alpha)^2 \geq \epsilon^2) \overset{Markov'sInequality}{\leq} \frac{E\left[(X[N]-\alpha)^2\right]}{\epsilon^2} \underset{N\to\infty}{\to} 0$$

We get approach to 0 for every $\epsilon$, so this is in fact statistically approach.

**Asymptotically Unbiased Estimator** We shall say that the estimator $\Theta(T)$ of the parameter $\Theta$ is asymptotically unbiased if $\mathbb{E}[\Theta(T)] \underset{T\to\infty}{\to} \Theta$.

**Consistent Estimator** We shall say that the estimator $\Theta(T)$ of the parameter $\Theta$ is consistent if:

1. $\Theta(T)$ is an asymptotic unbiased estimator
2. $\mathrm{Var}(\Theta(T)) \underset{T\to\infty}{\to} 0$

The meaning of the definition is that $\lim_{T\to\infty}\left(f_{\Theta(T)}(x)\right) = \delta(x-\Theta)$, meaning that as $T$ gets bigger, the distribution is getting closer to a delta around the deterministic value of the parameter.

**Ergodic RP in Parameter** We say that a RP is Ergodic in parameter $\Theta$ if exists an estimator $\Theta(T)$ that is a consistent estimator of $\Theta$.

**Ergodicity of Mean and Correlation** We expand the discussion of weak Ergodicity for two specific functions. We demand that the RP will be stationary at least asymptotically for the existence of the limit (marginal distribution at infinity independent in time).

**Ergodicity in the Mean** RP $X(t)$ asymptotically stationary, will be ergodic in mean if it is ergodic for a specific function $g(x) = x$,meaning:

$$\lim_{T\to\infty}\left(\eta_X(t_1)(T)\right) = \lim_{T\to\infty}\left(\frac{1}{T}\int_{\tau=0}^{T} X(t_1+\tau)d\tau\right) = \mathbb{E}[X(t_1)] \underset{T\to\infty}{=} \eta_X^{stat}$$

Because the process is stationary, we can write:

$$\lim_{T\to\infty}\left(\eta_X(T)\right) = \lim_{T\to\infty}\left(\frac{1}{T}\int_0^T X(t)dt\right) = \mathbb{E}[X(t)] \underset{T\to\infty}{=} \eta_X^{stat}$$

**Auto-correlation Ergodicity** RP $X(t)$ asymptotically stationary, will be auto-correlation ergodic if it is ergodic for a specific function $g(x,y) = x \cdot y$,meaning:

$$\begin{aligned}\lim_{T\to\infty}\left(R_{XX}(t_1,t_2)(T)\right) &= \lim_{T\to\infty}\left(\frac{1}{T}\int_{\tau=0}^T X(t_1+\tau)X(t_2+\tau)d\tau\right)\\ &= \mathbb{E}[X(t_1)X(t_2)] = R_{XX}^{stat}(t_1,t_2)\\ &= R_{XX}^{stat}(|t_2-t_1|)\end{aligned}$$

Because the process is stationary, we can write:

$$\lim_{T\to\infty}\left(R_{XX}(T)\right) = \lim_{T\to\infty}\left(\frac{1}{T}\int_0^T X(t)\cdot X(t+\tau)dt\right) = R_{XX}^{stat}(\tau)$$

**Conditions to Ergodicity in Mean** We introduce sufficient and necessary conditions for ergodicity in mean of a RP:
*Sufficient condition:*
*A sufficient condition is when the the if A occurs, then B must occur too. The notation is $A \Rightarrow B$, meaning it is enough to know that $x \in A$ to determine that $x \in B$. Equivalent claim for $A \Rightarrow B$ is $\neg B \Rightarrow \neg A$*
*Necessary condition:*
*A necessary condition is when in order to determine whether claim A occurs, B must occur. For example for $x \in A$ must also be that $x \in B$. We notice that a necessary condition is not sufficient - for example, if $x \in B$ doesn't mean $x \in A$. If claim B is a necessary claim to claim A, then the sufficient claim can be deduced $A \Rightarrow B$*
*Sufficient and Necessary Condition*
*The existence of both conditions (A is a sufficient and necessary condition for B) is in fact a logical equivalence between claim A and claim B. It happens because from A we can deduce B (sufficient) and from B we can deduce A (necessary), meaning $A \iff B$*
We mark the RP as $X(t)$ or $X[n]$. Assume that the process is stationary.
We have seen that in order to demand convergence, we need to demand that the estimator will be consistent (The mean will converge to the deterministic estimated value and the variance will converge to zero) we discuss the

parameter of the mean.
The mean estimator is built on the function $g(x) = x$, hence it is:

$$\eta_X(T) = \frac{1}{T}\int_0^T X(t)dt$$

We see that the unbiasedness is guaranteed for every T (with no approaching to infinity $T \to \infty$) if the process is stationary (mean independent in time):

$$\begin{aligned}
\mathbb{E}[\eta_X(T)] &= \mathbb{E}[\frac{1}{T}\int_0^T X(t)dt] = \frac{1}{T}\cdot\mathbb{E}[\int_0^T X(t)dt] \\
&= \frac{1}{T}\int_0^T \mathbb{E}[X(t)]dt = \frac{1}{T}\int_0^T \eta_X(t)dt \\
&\overset{91}{=} \frac{1}{T}\int_0^T \eta_X dt = \frac{1}{T}\cdot\eta_X\cdot\int_0^T 1\cdot dt \\
&= \eta_X\cdot\frac{1}{T}\cdot T = \eta_X
\end{aligned}$$

Now we need to check the condition of convergence of the variance to zero. We would want to know when $\lim_{T\to\infty} \mathrm{Var}\left(\eta_X(T)\right) = 0$:

$$\mathrm{Var}\left(\eta_X(T)\right) = \mathrm{Var}\left(\frac{1}{T}\int_0^T X(t)dt\right) = ... = \frac{1}{T}\int_{-T}^T C_{XX}(t)\cdot\left(1-\frac{|t|}{T}\right)dt$$

[91] Stationarity

proof (for discrete time RP for convenience):

$$\eta_X(T) = \frac{1}{N}\sum_{n=0}^{N-1} X[n]$$

$$\begin{aligned}
\mathrm{Var}\left(\eta_X(T)\right) &= \mathbb{E}\left[\left(\eta_X(T) - \mathbb{E}[\eta_X(T)]\right)^2\right] \overset{92}{=} \mathbb{E}\left[\left(\eta_X(T) - \eta_X\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{1}{N}\sum_{n=0}^{N-1} X[n] - \eta_X\right)^2\right] \overset{93}{=} \mathbb{E}\left[\left(\frac{1}{N}\sum_{n=0}^{N-1}(X[n] - \eta_X)\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{N^2}\left(\sum_{n=0}^{N-1}(X[n] - \eta_X)\right)^2\right] \\
&= \frac{1}{N^2}\mathbb{E}\left[\left(\sum_{n_1=0}^{N-1}(X[n_1] - \eta_X)\right)\left(\sum_{n_2=0}^{N-1}(X[n_2] - \eta_X)\right)\right] \\
&= \frac{1}{N^2}\sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1}\mathbb{E}\left[(X[n_1] - \eta_X)(X[n_2] - \eta_X)\right] \\
&= \frac{1}{N^2}\sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1} C_{XX}(n_1, n_2) \\
&\overset{94}{=} \frac{1}{N^2}\sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1} C_{XX}(n_1 - n_2)
\end{aligned}$$

Hence we need to sum $C_{XX}(n_1 - n_2)$ for every possible combination of $(n_1, n_2),\ n_1, n_2 \in \mathbb{Z}$.
We chose to separately sum all the combinations with the same difference. These point can be seen graphically as diagonals (movement from the main diagonal reduces the number of elements in the diagonal by 1). In the main diagonal there are $N$ elements.

$$\begin{aligned}
\frac{1}{N^2}\sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1} C_{XX}(n_1 - n_2) &\overset{n=n_1-n_2}{=} \frac{1}{N^2}\sum_{n=-(N-1)}^{N-1} C_{XX}(n)\cdot(N - |n|) \\
&= \frac{1}{N}\sum_{n=-(N-1)}^{N-1} C_{XX}(n)\cdot\left(1 - \frac{|n|}{N}\right)
\end{aligned}$$

[94] Unbiasedness of the mean of a stationary RP

[95] We insert the constant of the mean to the sum, hence it will be divided in the length of the sum, so we can get the fraction out as a common divider.

[96] Stationarity

Claim A - a set of conditions:

1. Ergodicity of the mean iff (necessary and sufficient):

   $$\frac{1}{N}\sum_{n=-(N-1)}^{N-1} C_{XX}(n)\cdot\left(1-\frac{|n|}{N}\right)\underset{N\to\infty}{\to}0$$

   Or

   $$\frac{1}{T}\int_{-T}^{T} C_{XX}(t)\cdot\left(1-\frac{|t|}{T}\right)dt\underset{T\to\infty}{\to}0$$

   Because the covariance function is an even function, an equivalent condition is:

   $$\frac{2}{N}\sum_{n=0}^{N-1} C_{XX}(n)\cdot\left(1-\frac{|n|}{N}\right)\underset{N\to\infty}{\to}0$$

   Or

   $$\frac{2}{T}\int_{0}^{T} C_{XX}(t)\cdot\left(1-\frac{|t|}{T}\right)dt\underset{T\to\infty}{\to}0$$

2. Ergodicity of the mean if (sufficient condition):

   $$\frac{1}{N}\sum_{n=-(N-1)}^{N-1} |C_{XX}(n)|\underset{N\to\infty}{\to}0$$

   Or

   $$\frac{1}{T}\int_{-T}^{T} |C_{XX}(t)|dt\underset{T\to\infty}{\to}0$$

   Because the correlation coefficient $\rho_{XX}(\tau)$ is similar to the correlation function up to a factor of normalization in the power of the signal, we can write the sufficient condition in the next matter:

   $$\rho_{XX}(\tau)=\frac{C_{XX}(\tau)}{C_{XX}(0)}\Rightarrow\frac{1}{N}\sum_{n=-(N-1)}^{N-1}|\rho_{XX}(n)|\underset{N\to\infty}{\to}0,\ \frac{1}{T}\sum_{-T}^{T}|\rho_{XX}(\tau)d\tau|\underset{T\to\infty}{\to}0$$

3. Ergodicity of the mean apply if (sufficient condition) the correlation coefficient is:

   (a) $\rho_{XX}(\tau)=\frac{C_{XX}(\tau)}{C_{XX}(0)}\underset{\tau\to\infty}{\to}0$

(b) Non-divergence of the integral:

$$\int_0^\infty |\rho_{XX}(\tau)| d\tau < \infty$$

Notes:
Condition 1 is a demand that comes from the definition of the consistency of of the estimator (from the expression of the variance of the estimator)
Condition 2 is a weaker version of condition 1:

$$\frac{1}{T}\int_{-T}^{T} C_{XX}(t)\cdot\left(1-\frac{|t|}{T}\right)dt \le \frac{1}{T}\int_{-T}^{T} C_{XX}(t)\cdot 1\cdot dt \le \frac{1}{T}\int_{-T}^{T} |C_{XX}(t)|dt$$

Condition 3a comes from the definition of the correlation coefficient and from the definition of finite power of signal. Condition 3b. Comes from the non-divergence of the integral (we can deduce from it condition 2, that we explained):

$$\exists A : \int_0^\infty |\rho_{XX}(\tau)| d\tau = A \Rightarrow \frac{1}{N}\int_0^\infty |\rho_{XX}(\tau)| d\tau = \frac{A}{N} \underset{N\to\infty}{\to} 0$$

Claim B - Slutsky's Theorem
For a stationary RP, it will be Ergodic of the mean by Slutsky's theorem iff(necessary and sufficient):

$$\frac{1}{N}\sum_{n=-(N-1)}^{N-1} C_{XX}(n) \underset{N\to\infty}{\to} 0, \quad \frac{1}{T}\int_{-T}^{T} C_{XX}(t)dt \underset{T\to\infty}{\to} 0$$

In claim A condition 1, we notice that we take the covariance function, and multiply it with a triangular window with width of $2T$ and maximum height 1, i.e. the signal is being reduced, hence if the expression converges to zero without reduction it is a sufficient condition.

**Discussion in Ergodicity of Mean of a Markov Chain** We want to show that for a stationary Markov chain, forgetting the past is equal to the convergence of the time average to the mean. So we want to show that if:

$$\lim_{n\to\infty}\left((\underline{\underline{P}})^n\right) = \begin{bmatrix} \leftarrow & \underline{\pi}^\infty & \rightarrow \\ & \vdots & \\ \leftarrow & \underline{\pi}^\infty & \rightarrow \end{bmatrix}$$

Then:

$$\lim_{N\to\infty}\left(\underline{\pi}^{stat}(N)\right)=\underline{\pi}^{stat},\quad \lim_{N\to\infty}\left(P_{ab}(N)\right)=\underline{\underline{P}}(a,b),\ \forall a,b\in\{1,...,J\}$$

It is enough to show the convergence of the value of the transition probability(because its formula includes the marginal distribution of the initial state - comes from conditional distribution and opens the topic of finding the statistics of an Ergodic Markov chain)
From the expression for the transition probability estimator, we see that we have to show that the estimated marginal distribution converges to the statistical marginal distribution:

$$p(a,b)(N)=\frac{1}{N-1}\sum_{i=0}^{N-2}I_{ab}(x_i,x_{i+1})\underset{N\to\infty}{\to}\Pr(x_i=a,x_{i+1}=b)$$

Hence we need to show that the indicator function $I_{ab}(x,y)$ is Ergodic of mean. We will show that by using condition 3A, meaning showing the correlation coefficient of two samples approaches zero:
We discuss Markov chains, so for us the indicator function always being used for two consecutive times, meaning $I_{ab}(k)=I_{ab}(X_k,X_{k+1})$:

$$\mathbb{E}[I_{ab}(X_i,X_{i+1})]=\Pr(X_i=a,X_{i+1}=b)=\underline{\pi}_a^{stat}\cdot\underline{\underline{P}}(a,b)$$

$$R_{I_{ab}I_{ab}}(k)=R_{I_{ab}I_{ab}}(i+k,i)=\mathbb{E}[I_{ab}(X_{i+k},X_{i+k+1})\cdot I_{ab}(X_i,X_{i+1})]$$

we know that $I_{ab}(X_{i+k},X_{i+k+1})\cdot I_{ab}(X_i,X_{i+1})$ iff $X_i=X_{i+k}=a,X_{i+1}=X_{i+k+1}=b$:

$$\begin{aligned}R_{I_{ab}I_{ab}}(k)&=1\cdot\pi_a^{stat}\cdot\underline{\underline{P}}(a,b)\cdot\Pr(X_{i+1}=b,X_{i+k}=a)\cdot\underline{\underline{P}}(a,b)\\&=\pi_a^{stat}\cdot\left(\underline{\underline{P}}(a,b)\right)^2\cdot\Pr(X_{i+1}=b,X_{i+k}=a)\cdot\end{aligned}$$

Therefore, the Covariance function is:

$$\begin{aligned}C_{I_{ab}I_{ab}}&=R_{I_{ab}I_{ab}}(k)-\mathbb{E}[I_{ab}(X_{i+k},X_{i+k+1})]\cdot\mathbb{E}[I_{ab}(X_{i+k},X_{i+k+1})]\\&=\pi_a^{stat}\cdot\left(\underline{\underline{P}}(a,b)\right)^2\cdot\Pr(X_{i+1}=b,X_{i+k}=a)-\pi_a^{stat}\cdot\underline{\underline{P}}(a,b)\cdot\pi_a^{stat}\cdot\underline{\underline{P}}(a,b)\\&=\pi_a^{stat}\cdot\left(\underline{\underline{P}}(a,b)\right)^2\cdot\Pr(X_{i+1}=b,X_{i+k}=a)-(\pi_a^{stat})^2\cdot\left(\underline{\underline{P}}(a,b)\right)^2\\&=\pi_a^{stat}\cdot\left(\underline{\underline{P}}(a,b)\right)^2\cdot\left[\Pr(X_{i+1}=b,X_{i+k}=a)-\pi_a^{stat}\right]\end{aligned}$$

And from forgetting the past we get that $\lim_{k\to\infty}\left(\Pr(X_{i+1}=b,X_{i+k}=a)\right)=\pi_a^{stat}$, therefore $C_{I_{ab}I_{ab}}\underset{k\to\infty}{\to}0$ meaning Ergodicity.

## Power Spectral Density (PSD)

The Power Spectral Density (PSD) is a non-negative function that represent the power for a frequency unit for signals in time (deterministic of stochastic). We calculate PSD for WSS RP only.

### Calculation of PSD

The PSD is a Fourier transform of the auto correlation of the process.
As known, the auto correlation of a RP is a deterministic function, therefore the PSD is also deterministic.

**PSD of WSS RP in Continuous Time** The PSD of a WSS RP in continuous time $X(t)$ with auto correlation function $R_{XX}(\tau)$ is:

$$S_{XX}(\omega) = \mathcal{F}\{R_{XX}(\tau)\} = \int_{-\infty}^{\infty} R_{XX}(\tau) \cdot e^{-j\omega\tau} d\tau$$

**PSD of WSS RP in Discrete Time** The PSD of a WSS RP in discrete time $X[n]$ with auto correlation function $R_{XX}(k)$ is:

$$S_{XX}(e^{j\omega}) = \mathcal{F}\{R_{XX}(k)\} = \sum_{k=-\infty}^{\infty} R_{XX}(k) \cdot e^{-j\omega k}$$

**Cross-Spectrum PSD of a Pair of jWSS RP** The PSD of a pair of jWSS RP in continuous time $X(t), Y(t)$ with joint correlation function $R_{XY}(\tau)$ is:

$$S_{XY}(\omega) = \mathcal{F}\{R_{XY}(\tau)\} = \int_{-\infty}^{\infty} R_{XY}(\tau) \cdot e^{-j\omega\tau} d\tau$$

We can define the same equation for discrete time RP.

**Properties of PSD for a Real RP** In this course we only deal with real processes (no complex values), so the change to the spectral dimension is simple.

**Properties of the Auto Correlation - Reminder**

- $R_{XX}(\tau)$ is an even function, meaning $R_{XX}(\tau) = R_{XX}(-\tau) \ \ \forall\tau$
- For a WSS process: $R_{XX}(0) \geq |R_{XX}(\tau)| \ \ \forall\tau$

**Properties of the PSD**

1. $S_{XX}(\omega)$ is a real function. Proof:

$$\begin{aligned} S_{XX}(\omega) &= \int_{-\infty}^{\infty} R_{XX}(\tau) \cdot e^{-j\omega\tau} d\tau \\ &= \int_{-\infty}^{\infty} R_{XX}(\tau) \cdot (\cos\omega t + j \cdot \sin\omega\tau) d\tau \\ &= \int_{-\infty}^{\infty} R_{XX}(\tau) \cdot \cos\omega\tau \ d\tau + j \cdot \int_{-\infty}^{\infty} R_{XX}(\tau) \cdot \sin\omega\tau \ d\tau \\ &\overset{95}{=} \int_{-\infty}^{\infty} R_{XX}(\tau) \cdot \cos\omega\tau \ d\tau \end{aligned}$$

2. $S_{XX}(\omega)$ is an even function ($R_{XX}$, cos are even) in a symmetric interval.

3. $S_{XX}(\omega)$ is a non-negative function, meaning $S_{XX}(\omega)$ $\forall\omega$ because the Fourier transform of a defined non-negative function is a non-negative function.

### White Noise in Continuous Time

White noise is a random process with a constant PSD in every frequency. Of course, a white noise with a infinite bandwidth is only theoretical, else we would get a RP with an infinite energy.
In white noise, every pair of different samples are independent, even though the white noise is a continuous time RP,and it doesn’t matter how close were the samples (identical to i.i.d.). It means that a white noise RP is non-continuous in any point.
White noise is a model of a thermal noise, therefore it is very useful.In reality, the samples of a thermal noise do depend on each other (it is a continuous RP) but in such a fast rate, our system don’t sample dependent samples. It doesn’t matter how fast we sample a thermal noise, it will be slow enough for the samples to be i.i.d., so can be modeled as a white noise.
We say that a continuous random process $X(t)$ is a white noise (in continuous time) if it is WSS, and its auto-correlation function is shaped like $R_{XX}(\tau) = A \cdot \delta(\tau)$ (meaning a delta with height $A$ in $\tau = 0$). It means that each pair of samples at different time are uncorrelated. Its PSD is a constant:

$$S_{XX}(\omega) = \mathcal{F}\{A \cdot \delta(\tau)\} = A \cdot \mathcal{F}\{\delta(\tau)\} = A$$

We usually denote $A = \frac{N_0}{2}$ (because every frequency has a “negative” frequency).

[95] Odd integral over a symmetric interval (sin is odd × $R_{XX}$ is even = odd function)

## WSS RP as Input of Linear Systems

We have learned how to analyze the case of a random vector that passes through a linear system. We also learned to calculate the statistics of the output and the statistics of the cross correlation elements given the statistics of the input and the characterization of the system (up to 2-nd order because this is a linear system)
We would like to expand this discussion for a passing of a RP through a linear system.

### Passing of a RP Through a Linear System

A system for processes cant be defined using matrices or being of a finite dimensions, as the input is a RP - an infinite sequence. In reality, a sequence will enter the system and a sequence will get out. In fact, this system is defined using a transfer function it is usually denoted as a filter (even though it doesn't actually don't actually filter)
We will discuss only WSS processes that go through linear systems that don't change in time (Linear Time Invariant-LTI), meaning the effect of time on the system is only through the effect of time on the input.
A general structure of such system will look like this (with $X[n], Y[n], h[n]$ for discrete time processes):
For LTI systems, the output is denoted by convolution of the input and the filter function of the system:

$$Y(t) = X(t) * h(t)$$

We didn't add constants (neither for the input nor for the outputs) for simplification of the discussion. Even if we add them, constants can only change the mean because they are necessarily independent in time, as we talk about LTI systems.
We would like to know the 2-nd order statistics of the output given the 2-nd order statistics of the input. In order to do so, we calculate the statistics for a general model of passing a RP in a LTI system, and then we can deduce to our "simple" system.

**Help Model for Analysis of the Connection Between Input and Output in LTI System** Denote two systems $h_1(t), h_2(t)$, which get as an input two RP $W(t), Z(t)$ respectively. These processes are jWSS. The 2-nd order statistics of them ($R_{WW}(\tau), R_{WZ}(\tau), R_{ZZ}(\tau)$) is known. We would like to know the joint 2-nd order moment of the output, meaning $R_{W'Z'}(\tau)$ (if they are jWSS at all)

$$\begin{aligned}
R_{W'Z'}(t+\tau,t) &= \mathbb{E}[W'(t+\tau)\cdot Z'(t)] \\
&= \mathbb{E}[(W(t+\tau)*h_1(t))\cdot(Z(t)*h_2(t))] \\
&= \mathbb{E}\left[\left(\int_{-\infty}^{\infty} h_1(\alpha)\cdot W(t+\tau-\alpha)\cdot d\alpha\right)\cdot\left(\int_{-\infty}^{\infty} h_2(\beta)\cdot Z(t-\beta)\cdot d\beta\right)\right] \\
&\overset{96}{=} \int_{\beta=-\infty}^{\infty}\int_{\alpha=-\infty}^{\infty} h_1(\alpha)\cdot h_2(\beta)\cdot \mathbb{E}[W(t+\tau-\alpha)\cdot Z(t-\beta)]\ d\alpha\ d\beta \\
&= \int_{\beta=-\infty}^{\infty}\int_{\alpha=-\infty}^{\infty} h_1(\alpha)\cdot h_2(\beta)\cdot R_{WZ}(t+\tau-\alpha,t-\beta)\ d\alpha\ d\beta \\
&\overset{97}{=} \int_{\beta=-\infty}^{\infty}\int_{\alpha=-\infty}^{\infty} h_1(\alpha)\cdot h_2(\beta)\cdot R_{WZ}(\tau-\alpha+\beta)\ d\alpha\ d\beta \\
&= \int_{\beta=-\infty}^{\infty} h_2(\beta)\cdot\left(\int_{\beta=-\infty}^{\infty} h_1(\alpha)R_{WZ}(\tau+\beta-\alpha)\ d\alpha\right)\ d\beta \\
&= \int_{\beta=-\infty}^{\infty} h_2(\beta)\cdot[h_1(\tau+\beta)*R_{WZ}(\tau+\beta)]\ d\beta \\
&\overset{98}{=} \int_{\gamma=-\infty}^{\infty} h_2(-\gamma)\cdot[h_1(\tau-\gamma)*R_{WZ}(\tau-\gamma)]\ d\gamma \\
&= \int_{\gamma=-\infty}^{\infty} h_2(-\gamma)\cdot[h_1*R_{WZ}](\tau-\gamma)\ d\gamma \\
&= h_1(\tau)*R_{WZ}(\tau)*h_2(-\tau) = R_{W'Z'}(\tau)
\end{aligned}$$

We get that the outputs $W'(t), Z'(t)$ are jWSS, meaning that for every jWSS processes that go through a LTI system, the outputs will be jWSS.

**Statistics of the Output of a LTI System with jWSS Input** For $X(t)$, a WSS RP with a known 2-nd order statistics $R_{XX}(\tau)$ and a LTI system with a known transfer function $h(t)$ we denote the output of the system as $Y(t)$, and we know that $Y(t) = X(t)*h(t)$.
We want to know that 2-nd order statistics of the output and the joint statistics, i.e. $R_{YY}(\tau), R_{XY}(\tau)$ respectively.
For that we use the help model we developed: $R_{W'Z'}(\tau) = h_1(\tau)*R_{WZ}(\tau)*h_2(-\tau)$.
We identify that we can use the said model in two different configurations that will give us the wanted statistics.

[98] Linearity of the mean and integral

[99] $W(t), Z(t)$ are jWSS

[100] $\gamma = -\beta, d\gamma = -d\beta$

We show both developments for the pair of configurations:
(Left):

$$\begin{gathered}W(t) = X(t), h_1(t) = h(t) \Rightarrow W'(t) = W(t) * h_1(t) = X(t) * h(t) = Y(t)\\ Z(t) = X(t), h_2(t) = h(t) \Rightarrow Z'(t) = Z(t) * h_2(t) = X(t) * h(t) = Y(t)\\ R_{W'Z'}(\tau) = h_1(\tau) * R_{WZ} * h_2(-\tau) \Rightarrow R_{YY}(\tau) = h(\tau) * R_{XX}(\tau) * h(-\tau)\end{gathered}$$

(Right):

$$\begin{gathered}W(t) = X(t), h_1(t) = \delta(t) \Rightarrow W'(t) = W(t) * h_1(t) = X(t) * \delta(t) = X(t)\\ Z(t) = X(t), h_2(t) = h(t) \Rightarrow Z'(t) = Z(t) * h_2(t) = X(t) * h(t) = Y(t)\\ R_{W'Z'}(\tau) = h_1(\tau) * R_{WZ} * h_2(-\tau) \Rightarrow R_{YY}(\tau) = \delta(\tau) * R_{XX}(\tau) * h(-\tau)\end{gathered}$$

Overall we get:

$$\begin{gathered}R_{XY}(\tau) = R_{XX}(\tau) * h(-\tau)\\ R_{YX}(\tau) = R_{XX}(\tau) * h(\tau)\\ R_{YY} = h(\tau) * R_{XX}(\tau) * h(-\tau)\end{gathered}$$

We calculate the PSD of these signals (remember that convolution in time $\leftrightarrow$ multiplication in frequency), by the definition of Fourier transform of the correlation function we get:

$$\begin{gathered}S_{XY}(\omega) = S_{XX}(\omega) \cdot H^*(\omega)\\ S_{YX}(\omega) = S_{XX}(\omega) \cdot H(\omega)\\ S_{YY}(\omega) = H(\omega) \cdot S_{XX} \cdot H^*(\omega) = S_{XX}(\omega) \cdot |H(\omega)|^2\end{gathered}$$

In the same manner, for WSS discrete time RP we get:

$$\begin{gathered}R_{XY}[n] = R_{XX}[n] * h[-n] \Leftrightarrow S_{XY}(e^{j\omega}) = S_{XX}(e^{j\omega}) \cdot H^*(e^{j\omega})\\ R_{YX}[n] = R_{XX}[n] * h[n] \Leftrightarrow S_{YX}(e^{j\omega}) = S_{XX}(e^{j\omega}) \cdot H(e^{j\omega})\\ R_{YY}[n] = h[n] * R_{XX}[n] * h[-n] \Leftrightarrow S_{YY}(e^{j\omega}) \cdot |H(e^{j\omega})|^2\end{gathered}$$

We would like to deduce from these expressions to the covariance, Meaning show that the connection above also work for $C_{XX}(\tau), C_{XY}(\tau), C_{YY}(\tau)$.
For the processes $X(t), Y(t)$ above (the input and output of the system $h(t)$), we define the next two processes:

$$X'(t) = X(t) - \eta_X, \ Y'(t) = Y(t) - \eta_Y$$

We input the process $X'(t)$ into the system $h(t)$ and see what we get as an output:

$$X'(t) * h(t) = (X(t) - \eta_X) * h(t) = X(t) * h(t) - \eta_X * h(t) = Y(t) - \eta_Y = Y'(t)$$

We get $Y'(t)$ as an output. We can deduce that the above formulas for the auto correlation function are valid for the processes $X'(t), Y'(t)$.
Because of the way we defined $X'(t), Y'(t)$:

$$R_{X'Y'}(\tau) = C_{XY}(\tau),\ R_{X'X'}(\tau) = C_{XX}(\tau),\ R_{Y'Y'}(\tau) = C_{YY}(\tau),$$

We add notation tot the transform of the covariance so we can distinguish:

$$\begin{gathered} C_{XY}(\tau) = C_{XX}(\tau) * h(-\tau) \Leftrightarrow S^C_{XY}(\omega) = S^C_{XX}(\omega) \cdot H^*(\omega) \\ C_{YX}(\tau) = C_{XX}(\tau) * h(\tau) \Leftrightarrow S^C_{YX}(\omega) = S^C_{XX}(\omega) \cdot H(\omega) \\ C_{YY}(\tau) = h(\tau) * C_{XX}(\tau) * h(-\tau) \Leftrightarrow S^C_{YY}(\omega) \cdot |H(\omega)|^2 \end{gathered}$$

For WSS, discrete time RP:

$$\begin{gathered} C_{XY}[n] = C_{XX}[n] * h[-n] \Leftrightarrow S^C_{XY}(e^{j\omega}) = S^C_{XX}(e^{j\omega}) \cdot H^*(e^{j\omega}) \\ C_{YX}[n] = C_{XX}[n] * h[n] \Leftrightarrow S^C_{YX}(e^{j\omega}) = S^C_{XX}(e^{j\omega}) \cdot H(e^{j\omega}) \\ C_{YY}[n] = h[n] * C_{XX}[n] * h[-n] \Leftrightarrow S^C_{YY}(e^{j\omega}) \cdot |H(e^{j\omega})|^2 \end{gathered}$$

**Creating a Stationary Gaussian Random Process** We have learned that we can create a Gaussian random vector as we wish by passing an i.i.d. Gaussian random vector in a linear system. We call this action "coloring" a Gaussian random vector (and we also learned about "whitening" a Gaussian random vector). We have started with a Gaussian random vector with i.i.d elements $\underline{Z} \sim N(\underline{0}, \underline{\underline{I}})$ and we got in the output a Gaussian random vector of the same dimensions with distribution: $\underline{X} \sim N(\eta_{\underline{X}}, C_{\underline{XX}})$
The system (where $\underline{\underline{P}}$ is the diagonalizing matrix of $C_{\underline{XX}}$)
For RP we can create a stationary Gaussian random process (GRP). We notice from the help model that the auto-correlation of the output of a LTI system with a filter $h(t)$ and WSS input $X(t)$ is:

$$R_{YY}(\tau) = h(\tau) * R_{XX}(\tau) * h(-\tau)$$

And we can deduce that the output is stationary. We want to control and "shape" the output. If we take a process source with a time-independent correlation function (white noise), we get:

$$R_{YY}(\tau) = h(\tau) * \left(\frac{N_0}{2} \cdot \delta(\tau)\right) * h(-\tau) = \frac{N_0}{2} \cdot h(\tau) * h(-\tau)$$

Hence, if we correctly design our filter, we can get almost any correlation function.
If we want the output to be a GRP, then the input will have to be a GRP

(because it passes through a LTI system). hence if the input will be a white noise with normal distribution and mean of zero (a white Gaussian noise) then we will get a stationary GRP in the output.
The same goes for a discrete time stationary GRP: $R_{YY}[n] = \frac{N_0}{2} \cdot h[n] * h[-n]$

### Narrow Band Pass Filter (NBPF)

We discuss the case of a WSS real RP that passes through a narrow band pass filter (NBPF)
We pass a continuous time WSS real RP $X(t)$ through a BPF with a filter function in frequency $H_{\omega_{ij},\Delta}(\omega)$.
Notice that we mark the output as $X_{\omega_{ij},\Delta}(t)$, because it is an intersection of $X(t)$ in the frequency domain with the filter, i.e. around $\omega_0$. It is important to note that the center frequency is marked using the angular frequency $\omega$, while the width of the filter in $f$, hence (using $\omega = 2\pi f$):

$$H_{\omega_{ij},\Delta} = \begin{cases} 1, & f \in (f_0 - \frac{\Delta}{2}, f_0 = \frac{\Delta}{2}) \cup (-f_0 - \frac{\Delta}{2}, -f_0 + \frac{\Delta}{2}) \\ 0, & else \end{cases}, \quad f_0 = \frac{\omega_0}{2\pi}$$

Assume that we know the PSD of the process $X(t)$, so we can calculate the PSD of the output by:

$$S_{X_{\omega_{ij},\Delta},X_{\omega_{ij},\Delta}}(\omega) = S_{XX}(\omega) \cdot |H_{\omega_{ij},\Delta}(\omega)|^2$$

Of course, if the filter is not with amplitude of 1, then we would get in the output different heights of the PSD, but with the same behavior as if the filter was constant(we can use another filter that “fixes” the problems). We would like to calculate the power of the output signal:

$$\begin{aligned} POWER(X_{\omega_0,\Delta}(t)) &= \mathbb{E}\left[(X_{\omega_0,\Delta}(t)^2)\right] = R_{X_{\omega_0,\Delta},X_{\omega_0,\Delta}}(0) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{X_{\omega_0,\Delta},X_{\omega_0,\Delta}}(\omega) e^{j\omega \cdot 0} d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{X_{\omega_0,\Delta},X_{\omega_0,\Delta}}(\omega) d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{XX}(\omega)|H_{\omega_0,\Delta}|^2 d\omega \\ &= 2 \cdot \frac{1}{2\pi}\int_{2\pi(f_0-\frac{\Delta}{2})}^{2\pi(f_0+\frac{\Delta}{2})} S_{XX}(\omega) d\omega \end{aligned}$$

Meaning we sum the PSD for the frequencies where the filter is not zero.
We can now use Riemann integral (the width of the integral times the value

of the integrand in the middle):

$$POWER(X_{\omega_0,\Delta}(t)) = \frac{1}{\pi}\int_{2\pi(f_0-\frac{\Delta}{2})}^{2\pi(f_0+\frac{\Delta}{2})} S_{XX}(\omega)d\omega$$
$$\xrightarrow[\Delta<<1]{Riemann} \frac{1}{\pi}\cdot 2\pi\Delta\cdot S_{XX}(2\pi f_0) = 2\Delta\cdot S_{XX}(\omega_0)$$

Notice that if we pass a continuous time white noise with half PSD we get in the output a signal with power of:

$$N(t) = \frac{N_0}{2}\delta(t) \Rightarrow POWER(X_{\omega_0,\Delta}(t)) = 2\Delta\cdot S_{NN}(\omega_0) = 2\Delta\cdot\frac{N_0}{2} = \Delta\cdot N_0$$

Note:
For auto-correlation function, its PSD will always be non-negative.It works backward too - For every even non-negative function in the frequency domain, we can define a correlation function by using the inverse Fourier transform.

**Uncorrelatedness Between Frequency Bands**

We discuss continuous time RP, but the development is identical for discrete time RP.

**Uncorrelatedness Between Exclusive Frequency Bands of a WSS RP** We would like to show that for a WSS RP in continuous time $X(t)$, two signals that are created from its intersection with BPF with disjoint(exclusive) frequency bands are always uncorrelated.
Using the next system:

Claim:for intersection in disjoint frequency bands, meaning for $\Delta < |f_1 - f_2|$, the output signals $X_{\omega_1,\Delta}(t), X_{\omega_2,\Delta}(t)$ are jWSS (as we have proved) and uncorrelated. i.e. for intersection of disjoint frequency bands:

$$R_{X_{\omega_1,\Delta},X_{\omega_2,\Delta}}(\tau) = \mathbb{E}[X_{\omega_1,\Delta}(t+\tau)\cdot X_{\omega_2,\Delta}(t)] = 0,\ \forall\tau$$

Proof: WE identify we can use the help model to build a configuration for this system:

$$W(t) = X(t), h_1(t) = h_{\omega_1,\Delta}(t) \Rightarrow W'(t) = W(t) * h_1(t) = X(t) * h_{\omega_1,\Delta}(t) = X_{\omega_1,\Delta}(t)$$
$$Z(t) = X(t), h_2(t) = h_{\omega_2,\Delta}(t) \Rightarrow Z'(t) = Z(t) * h_2(t) = X(t) * h_{\omega_2,\Delta}(t) = X_{\omega_2,\Delta}(t)$$
$$R_{W'Z'}(\tau) = h_1(\tau) * R_{WZ}(\tau) * h_2(-\tau) \Rightarrow R_{X_{\omega_1,\Delta},X_{\omega_2,\Delta}}(\tau) = h_{\omega_1,\Delta}(\tau) * R_{XX}(\tau) * h_{\omega_2,\Delta}(-\tau)$$

99

$$\Rightarrow S_{X_{\omega_1,\Delta},X_{\omega_2,\Delta}}(\omega) = S_{XX}(\omega)\cdot H_{\omega_1,\Delta}(\omega)\cdot H^*_{\omega_2,\Delta}(\omega) = 0$$

Hence the PSD is 0 for every frequency, hence the auto-correlation is also constant zero:

$$R_{X_{\omega_1,\Delta},X_{\omega_2,\Delta}}(\tau) = \mathcal{F}^{-1}\left\{S_{X_{\omega_1,\Delta},X_{\omega_2,\Delta}}(\omega)\right\} = \mathcal{F}^{-1}\{0\} = 0$$

Notes:

- If the output is Gaussian, then uncorrelated $\Rightarrow$ independent.
  Hence in fact each frequency by itself can be modulated as a sin times a Gaussian random variable.
- If the output is a general signal, then we can cut out (using LPF, BPF, HPF)some irrelevant parts of the frequency domain. That way we can prevent interference with other signals or plan ahead and use the frequencies for multiple needs.

**Uncorrelatedness Between Exclusive Frequency Bands of jWSS RP**
We would like to show that for two continuous time RP $X(t), Y(t)$, which are jWSS,the two signals that are created from cutting each of them using an exclusive frequency band are always uncorrelated.
Using the next system:
Claim:
For cutting using exclusive frequency bands, i.e. for $\Delta = |f_1 - f_2|$, The output signal $X_{\omega_1,\Delta}(t), Y_{\omega_2,\Delta}(t)$ (which are jWSS, as we have seen) are uncorrelated.
Meaning for different frequency bands cutting:

$$R_{X_{\omega_1,\Delta},Y_{\omega_2,\Delta}}(\tau) = \mathbb{E}[X_{\omega_1,\Delta}(t+\tau) \cdot Y_{\omega_2,\Delta}(t)] = 0$$

Proof:
We identify we can use the help model here (because the inputs are jWSS), so:

$$W(t) = X(t), h_1(t) = h_{\omega_1,\Delta}(t) \Rightarrow W'(t) = W(t) * h_1(t) = X(t) * h_{\omega_1,\Delta}(t) = X_{\omega_1,\Delta}(t)$$

$$Z(t) = Y(t), h_2(t) = h_{\omega_2,\Delta}(t) \Rightarrow Z'(t) = Z(t) * h_2(t) = Y(t) * h_{\omega_2,\Delta}(t) = Y_{\omega_2,\Delta}(t)$$

$$R_{W'Z'}(\tau) = h_1(\tau) * R_{WZ}(\tau) * h_2(-\tau) \Rightarrow R_{X_{\omega_1,\Delta},X_{\omega_2,\Delta}}(\tau) = h_{\omega_1,\Delta}(\tau) * R_{XX}(\tau) * h_{\omega_2,\Delta}(-\tau)$$

$$\Rightarrow S_{X_{\omega_1,\Delta},Y_{\omega_2,\Delta}}(\omega) = S_{XY}(\omega) \cdot H_{\omega_1,\Delta}(\omega) \cdot H^*_{\omega_2,\Delta}(\omega) \overset{100}{=} 0$$

So if the PSD is zero for every frequency, then:

$$R_{X_{\omega_1,\Delta},Y_{\omega_2,\Delta}}(\tau) = \mathcal{F}^{-1}\left\{S_{X_{\omega_1,\Delta},Y_{\omega_2,\Delta}}(\omega)\right\} = \mathcal{F}^{-1}\{0\} = 0$$

[99] Because in the frequency domain, the filters don’t overlap.

[100] See previous footnote.

Note:
In order to have correlation between the signals, there must be an overlap in the frequency domain between the frequency bands (a necessary but not a sufficient condition because there might be a signal that its properties zeros the correlation regardless of the overlap, but in our discussion we will ignore this case). In other words, if we use two overlapped frequency bands, the processes will be correlated, which allows us to use linear estimation (because we have 2-nd order statistics)
This is not true in the time domain, because each time sample has some correlation to another time sample. Conclusion:
Linear estimation between two RP is valid if we sample one of the signals in a single frequency, and then estimate in the same frequency. Meaning gather information of a RP $X(t)$ in frequency $\omega_0$ and from it deduce information on the random process $Y(t)$ in the same frequency $\omega_0$

**Parallel Processing of Frequency Bands** We get that there is no correlation between two exclusive frequency bands. Notice it happens for every pair of exclusive frequency bands. For example, if we take a signal and pass it through $M$ filters with no frequency bands overlapping, then every pair of the M signals will be uncorrelated.

### MSE Optimal Linear Estimation of a RP (Wiener Filter)

Given a pair of RP $X[n], Y[n]$, we would like to estimate te RP $X[n]$ using the samples of $Y[n]$ (MSE)
When we estimated a random vector $\underline{X}$ from random vector $\underline{Y}$ we estimated the elements of $\underline{X}$ one by one, where every element $X_i = \underline{X}(i)$ was estimated using the whole vector $\underline{Y}$.
Now we would like to estimate a RP out of a RP. Also now we estimate every element $X_i$ from the sequence $\{X_n\}$ from the information we have of the sequence $\{Y_n\}$.
For the estimation of each element, which we mark as $X_i$, we look at some samples from a certain time window:

$$\{Y_n\}_{n=i-k}^{i+k} = \{Y_{i-k}, Y_{i-k+1}, ..., Y_{i-1}, Y_i, Y_{i+1}, ..., Y_{i+k-1}, Y_{i+k}\}$$

Given these samples we can optimally MSE estimate it, as we estimated a random variable out of a random vector (because the number of elements is finite) using the conditional mean estimator.
The MSE optimal estimator of $X_i$ given a time window is:

$$X_i^{opt,MSE} = X_i^{MMSE} = \mathbb{E}[X_i | Y_{i-k}, ..., Y_{i+k}]$$

Using that way, we can use any time window we want (not necessarily symmetric around $n = i$) but as the window gets bigger, then matrices also get bigger, the calculation will take longer but the estimation will be more precise.
If we know that the RP $X[n], Y[n]$ are jWSS, then we can use linear estimation (this estimation will not necessarily be MMSE, we will learn only LMMSE estimators).

**MMSE Linear Estimation of a RP From a jWSS RP** For a pair of jWSS RP $X[n], Y[n]$, LMMSE Estimator is optimal and denoted by the next LTI filter (Wiener Filter):

$$X[n] = X^{LMMSE}[n] = Y[n] * h_{LMMSE}[n] + b_{LMMSE}$$

$$H_{LMMSE}(e^{jw\omega}) = \frac{S^C_{XY}(e^{j\omega})}{S^C_{YY}(e^{j\omega})},\ S^C_{YY}(e^{j\omega}) \neq 0$$

$$b_{LMMSE} = \eta_X - \eta_Y \cdot \sum_{n=-\infty}^{\infty} h[n]$$

From the definition of Fourier transform, we can identify that the sum is actually $\mathcal{F}\{h[n]\}|_{\omega=0}$ hence:

$$b_{LMMSE} = \eta_X - \eta_Y \cdot H_{LMMSE}(\omega = 0)$$

So overall:

$$X^{LMMSE}[n] = Y[n] * h_{LMMSE}[n] + b_{LMMSE}$$

$$H_{LMMSE}(e^{j\omega}) = \frac{S^C_{XY}(e^{j\omega})}{S^C_{YY}(e^{j\omega})}$$

$$b_{LMMSE} = \eta_X - \eta_Y \cdot H_{LMMSE}(\omega = 0)$$

Or:

$$X^{LMMSE}(t) = Y(t) * h_{LMMSE}(t) + b_{LMMSE}$$

$$H_{LMMSE}(\omega) = \frac{S^C_{XY}(\omega)}{S^C_{YY}(\omega)}$$

$$b_{LMMSE} = \eta_X - \eta_Y \cdot H_{LMMSE}(\omega = 0)$$

Proof:
We prove using a claim, that says that the linear estimator is optimal iff the error is perpendicular to every linear function of the samples.

Meaning iff $e_n \perp f^{Linear}(\{Y_n\},1)$ (the proof is the same to random vectors or random variables), where:

$$f^{Linear}(\{Y_n\},1) = \sum_i a_i \cdot Y_i + c,\ \forall n, a_i, c$$

This perpendicularity is equal to the next demands:

1. $e_n \perp 1$ (or every other constant in $\mathbb{R}$)
2. $e_n \perp (Y_m - \eta_Y)\ \forall n, m$

Hence we would like to choose the linear filter with parameters $h[n], b$ that will answer these demands.
Now we will see what is the meaning of this pair of demands (as usual we use the differential error):

1. $$\begin{aligned} e_n \perp 1 \iff & \mathbb{E}[e_n \cdot 1] = 0 \\ & \mathbb{E}[X[n] - Y[n] * h[n] - b] = 0 \\ & \eta_X - \eta_Y \cdot \sum_{n=-\infty}^{\infty} h[n] - b = 0 \\ & b(h) = \eta_X - \eta_Y \cdot \sum_{n=-\infty}^{\infty} h[n] \end{aligned}$$

2. $$\mathbb{E}[e_n \cdot (Y_m = \eta_Y)] = 0$$
$$\mathbb{E}[(X[n] - Y[n] * h[n] - b) \cdot (Y_m - \eta_Y)] = 0$$

   Insert $b(h)$:

$$\mathbb{E}\left[\left(X[n] - Y[n] * h[n] - \eta_X + \eta_Y \cdot \sum_{n=-\infty}^{\infty} h[n]\right) \cdot (Y_m - \eta_Y)\right] = 0$$
$$\mathbb{E}\left[\left(X[n] - \sum_{k=-\infty}^{\infty} h[k] \cdot Y[n-k] - \eta_X + \eta_Y \cdot \sum_{k=-\infty}^{\infty} h[k]\right) \cdot (Y_m - \eta_Y)\right] = 0$$
$$\mathbb{E}\left[\left((X[n] - \eta_X) - \sum_{k=-\infty}^{\infty} (h[k] \cdot (Y[n-k] - \eta_Y))\right) \cdot (Y_m - \eta_Y)\right] = 0$$
$$C_{XY}(n-m) - \sum_{k=-\infty}^{\infty} h[k] \cdot C_{YY}(n-k-m) = 0$$

But remember it has to apply for $\forall n, m$:

$$\begin{aligned}
&C_{XY}[n] - \sum_{k=-\infty}^{\infty} h[k] \cdot C_{YY}[n-k] = 0 \\
&C_{XY}[n] - h[n] * C_{YY}[n] = 0 \\
&C_{XY} = h[n] * C_{YY}[n] \\
&S_{XY}^{C}(e^{j\omega}) = H(e^{j\omega}) \cdot S_{YY}^{C}(e^{j\omega}) \Rightarrow H(e^{j\omega}) = \frac{S_{XY}^{C}(e^{j\omega})}{S_{YY}^{C}(e^{j\omega})}
\end{aligned}$$

And we get the same coefficient we described before.

**A System for LMMSE Estimation of Two jWSS RP** We have seen that:

$$X^{LMMSE}[n] = Y[n] * h_{LMMSE}[n] + \eta_X - \eta_Y \cdot \sum_{n=-\infty}^{\infty} h[n]$$

$$H_{LMMSE}(e^{j\omega}) = \frac{S_{XY}^{C}(e^{j\omega})}{S_{YY}^{C}(e^{j\omega})},\ b_{LMMSE} = \eta_X - \eta_Y \cdot H_{LMMSE}(\omega = 0) = \eta_X - \eta_Y \cdot \sum_{n=-\infty}^{\infty} h[n]$$

Insert $b_{LMMSE}$ and:

$$X^{LMMSE}[n] = Y[n] * h_{LMMSE}[n] + \eta_X - \eta_Y \cdot \sum_{n=-\infty}^{\infty} h[n]$$

$$\begin{aligned}
Y[n] * h_{LMMSE}[n] + \eta_X - \eta_Y \cdot \sum_{n=-\infty}^{\infty} h[n] &= X^{LMMSE}[n] \\
Y[n] * h_{LMMSE}[n] + \eta_X - \eta_Y * h_{LMMSE}[n] &= X^{LMMSE}[n] \\
(Y[n] - \eta_Y) * h_{LMMSE}[n] &= X^{LMMSE}[n] - \eta_X
\end{aligned}$$

Meaning, if we input $Y[n] - \eta_Y$ into the system $h_{LMMSE}[n]$, we get as output of the LMMSE estimation $X[n] - \eta_X$. Hence the LMMSE estimation system of the jWSS RP looks like:
Of course we can also use the “basic” system:

**Example:** Denote a LTI system $g$, that looks like:

$$Y[n] = X[n] * g[n] + Z[n]$$

This equation describe a system with additive noise $Z[n]$. We would like to recover (estimate) the process $X[n]$.
It is given that the WSS RP $X[n]$ with $\eta_X = 0$ and PSD $S_{XX}(e^{j\omega})$. It is given as well that the RP $Z[n]$ is WSS with $\eta_Z = 0$ and PSD $S_{ZZ}(e^{j\omega})$. We also know that $X[n], Z[n]$ are independent.
Because the mean of both RP is 0, then also the mean of the output process $Y[n]$ will be 0 because we don't add any non-zero constant or any RP with non-zero mean.
Hence for all the processes, the auto-correlation and cross-correlation, and the auto-covariance and the cross-covariance are the same (it would happen even if only one of $\eta_X, \eta_Z$ was equal to zero, but then we can't state that $\eta_Y = 0$). We show that in general for two RP with zero mean:

$$C_{AB}(t_1,t_2) \triangleq \mathbb{E}\left[(A(t_1) - \eta_A) \cdot (B(t_2) - \eta_B)\right] = \mathbb{E}\left[A(t_1) \cdot B(t_2)\right] \triangleq R_{AB}(t_1,t_2)$$

We also know that both processes $X[n], Z[n]$ are WSS each, hence also jWSS. Calculate their cross-correlation:

$$R_{XZ}(t_1,t_2) = \mathbb{E}\left[X(t_1) \cdot Z(t_2)\right] \overset{101}{=} \mathbb{E}[X(t_1)] \cdot \mathbb{E}[Z(t_2)] = \eta_X \cdot \eta_Z = 0 \cdot 0 = 0$$

We get that it is equal to zero, hence we can say it is independent in time. Because both RP are WSS separately and the cross-correlation is independent in time, they are jWSS.
Because for these processes $C(k) = R(k)$, hence $S(e^{j\omega}) = S^C(e^{j\omega})$.
We would like to show that the processes $X[n], Y[n]$ are jWSS so we could use Weiner filter (it is LMMSE filter). The process $Y[n]$ is WSS because it is WSS process that pass through a LTI system with adding of another WSS process.
Both processes $X[n], Y[n]$ are jWSS as we have seen in the help model.
Hence we can use Weiner filter (MSE linear optimal).

$$X^{LMMSE}[n] = Y[n] * h_{LMMSE}[n] + b_{LMMSE}$$

$$H_{LMMSE}(e^{j\omega}) = \frac{S^C_{XY}(e^{j\omega})}{S^C_{YY}(e^{j\omega})}$$

$$b_{LMMSE} = \eta_X - \eta_Y \cdot H_{LMMSE}(\omega = 0)$$

$$b_{LMMSE} = \eta_X - \eta_Y \cdot \sum_{n=-infty}^{\infty} h[n]$$

We can say the transmission function of the filter is:

$$H_{LMMSE}(e^{j\omega}) = \frac{S^C_{XY}(e^{j\omega})}{S^C_{YY}(e^{j\omega})} \overset{102}{=} \frac{S_{XY}(e^{j\omega})}{S_{YY}(e^{j\omega})}$$

[101] Independence

Find the auto-correlation in order to find the PSD for the filter, using the help model. we mark $Y'[n] = X[n] * g[n]$ and get:

$$R_{XY}(k) = R_{XY'+XZ}(k) = R_{XY'}(k) + R_{XZ}(k) = R_{XY'}(k) = \delta(k) * R_{XX}(k) * g(-k) = R_{XX}(k) * g(-k)$$

$$R_{YY}(k) = R_{(Y'+Z)(Y'+Z)}(k) = R_{YY'}(k) + R_{ZY'}(k) + R_{ZZ}(k)$$

But $Z[n], Y'[n]$ are independent because $X[n], Z[n]$ are independent and we get $Y[n]$ from a RP that passes through an LTI system. In addition the mean of $Z[n]$ (and of $Y'[n]$) is 0, hence $R_{YZ}(k) = 0$. We can also show:

$$R_{YZ}(k) \triangleq \mathbb{E}[(X[n+k] * g[n+k]) \cdot Z[n]] \overset{103}{=} \mathbb{E}[X[n+k] * g[n+k]] \cdot \mathbb{E}[Z[n]] = 0$$

Hence the auto-correlation function of $Y[n]$ will be:

$$R_{YY}(k) = R_{Y'Y'}(k) + R_{ZZ}(k) = g(k) * R_{XX}(k) * g(-k) + R_{ZZ}(k)$$

So the PSD will be:

$$S_{XY}(e^{j\omega}) = \mathcal{F}\{R_{XY}(k)\} = S_{XX}(e^{j\omega}) \cdot G^*(e^{j\omega})$$

$$S_{YY}(e^{j\omega}) = \mathcal{F}\{R_{YY}(k)\} = S_{XX}(e^{j\omega}) \cdot |G(e^{j\omega})|^2 + S_{ZZ}(e^{j\omega})$$

and the filter will be:

$$H_{LMMSE}(e^{j\omega}) = \frac{S_{XX}(e^{j\omega}) \cdot G^*(e^{j\omega})}{S_{XX}(e^{j\omega}) \cdot |G(e^{j\omega})|^2 + S_{ZZ}(e^{j\omega})}$$

We test the system using the Signal to Noise Ratio (SNR):

- If $SNR \gg 1$, then the noise power ($S_{ZZ}(e^{j\omega})$) can be neglected compared to the signal, and:

  $$H_{LMMSE}(e^{j\omega}) = \frac{S_{XX}(e^{j\omega}) \cdot G^*(e^{j\omega})}{S_{XX}(e^{j\omega}) \cdot |G(e^{j\omega})|^2 + S_{ZZ}(e^{j\omega})} \approx \frac{1}{G(e^{j\omega})}$$

  It makes sense, since the noise can be neglected , then the estimation system only needs to “undo” what the system $g(\cdot)$

- If $SNR \ll 1$, then :

  $$H_{LMMSE}(e^{j\omega}) = \frac{S_{XX}(e^{j\omega}) \cdot G^*(e^{j\omega})}{S_{XX}(e^{j\omega}) \cdot |G(e^{j\omega})|^2 + S_{ZZ}(e^{j\omega})} \approx$$

  This happens because if the noise is big compared to the signal, then the process $Y[n]$ that we use to estimate $X[n]$ is too “contaminated” and therefore needed to be “deleted”. In this case, we estimate only using the additive constant that determines the mean.

[102] $C_{XY} = R_{XY}, C_{YY} = R_{YY}$

[103] Independence

**Wiener Filter Error** In random variables estimation, we have seen that the error is a random variable itself. The same happens for random vectors , and for RP.
The estimation error (process) is given by:

$$e[n] = X[n] - \hat{X}[n]$$

We would like to test the error process.
Claim:
The error process $e[n]$ is WSS and:

1. $\mathbb{E}\left[e[n]\right],\ \forall n$, i.e. the error mean for every sample is zero.
2. $R_{ee}(k) = R_{XX}(k) - R_{\hat{X}\hat{X}}(k) = C_{XX}(k) - C_{\hat{X}\hat{X}}(k)$

Proof:

1. it happens because $e_n \perp 1$
2. We show that it is WSS and express the auto-correlation function of the process.
   To do that we start with the auto-correlation function of the source process $X[n]$:

$$\begin{aligned}
R_{XX}(k) &\triangleq \mathbb{E}[X(n_k)\cdot X(n)] = \mathbb{E}\left[(X[n+k]+e[n+k])\cdot(X[n]+e[n])\right]\\
&= \mathbb{E}\left[X[n+k]\cdot X[n]\right] + \mathbb{E}\left[X[n+k]\cdot e[n]\right] + \mathbb{E}\left[e[n+k]\cdot X[n]\right] + \mathbb{E}\left[e[n+k]\cdot e[n]\right]\\
&\overset{104}{=} \mathbb{E}\left[X[n+k]\cdot X[n]\right] + \mathbb{E}\left[e[n+k]\cdot e[n]\right]\\
&\overset{105}{=} R_{XX}(k) + R_{ee}(n+k,n)
\end{aligned}$$

From this equation we can get the auto-correlation of the error:

$$R_{ee}(n+k,n) = R_{XX}(k) - R_{\hat{X}\hat{X}}(k)$$

We get that the auto-correlation of the error depends only on the time difference, so WSS.
From claim 1, $\mathbb{E}[e[n]] = 0$, hence:

$$\mathbb{E}[e[n]] = 0 \Rightarrow \mathbb{E}[X[n] - \hat{X}[n]] = 0 \Rightarrow \mathbb{E}[X[n]] - \mathbb{E}[\hat{X}[n]] \Rightarrow \eta_X = \eta_{\hat{X}}$$

[106] $X = f^{Linear}(\{Y_n\}, 1)$

[105] Estimator $X$ is an output of a LTI system, which its input is $Y[n]$ - a WSS RP, hence $X[n]$ is also WSS.

We use the connection between the correlation to the covariance and get:

$$C_{AB} = R_{AB}(k) + \eta_A \cdot \eta_B$$

So we can deduce:

$$\begin{aligned} R_{ee}(k) &= R_{ee}(n+k,n) = R_{XX}(k) - R_{\hat{X}\hat{X}}(k) \\ &\overset{106}{=} R_{XX}(k) - R_{\hat{X}\hat{X}}(k) + \eta_X^2 - \eta_{\hat{X}}^2 \\ &= R_{XX}(k) + \eta_X^2 - \left(R_{\hat{X}\hat{X}}(k) + \eta_{\hat{X}}^2\right) \\ &= C_{XX}(k) - C_{\hat{X}\hat{X}}(k) \end{aligned}$$

Conclusion:
The PSD of the error process is given by (Fourier transform of the auto-correlation function of a WSS process):

$$S_{ee}(e^{j\omega}) = \mathcal{F}\{R_{ee}(k)\} = S_{XX}(e^{j\omega}) - S_{\hat{X}\hat{X}}(e^{j\omega}) = S^C_{XX}(e^{j\omega}) - S^C_{\hat{X}\hat{X}}(e^{j\omega})$$

But the estimator is given as a passing of a process $Y[n]$ in the filter $h_{LMMSE}[n]$ (LTI system, hence convolution in time domain is multiplication in frequency domain), so:

$$S_{ee}(e^{j\omega}) = S^C_{XX}(e^{j\omega}) - S^C_{\hat{X}\hat{X}}(e^{j\omega}) = S^C_{XX}(e^{j\omega}) - |H_{LMMSE}(e^{j\omega})|^2 \cdot S^C_{YY}(e^{j\omega})$$

By inserting the expression of the filter we get:

$$S_{ee}(e^{j\omega}) = S^C_{XX}(e^{j\omega}) - \left|\frac{S^C_{XY}(e^{j\omega})}{S^C_{YY}(e^{j\omega})}\right|^2 \cdot S^C_{YY}(e^{j\omega}) = S^C_{XX}(e^{j\omega}) - \frac{|S^C_{XY}(e^{j\omega})|^2}{S^C_{YY}(e^{j\omega})}$$

Hence the mean square error(MSE) is:

$$\begin{aligned} MSE \triangleq \mathbb{E}\left[e^2[n]\right] &= R_{ee}(0) = \mathcal{F}^{-1}\{S_{ee}(e^{j\omega})\}|_{n=0} \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{ee}(e^{j\omega}) \cdot e^{j\omega 0} d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{ee}(e^{j\omega}) d\omega \end{aligned}$$

Note:
Notice that this result reminds us of the estimator of a random variable from random variable (variance ↔ spectrum). The meaning of the error PSD is that every frequency is evaluated separately. In practice, for every frequency we use a scalar estimator.

[106] $\eta_X = \eta_{\hat{X}}$

## Advanced Random Processes

In this part we introduce two familiar, simple RP - Wiener process and Poisson process. Both are processes with independent increments.

**RP With Independent Increments** A RP will be with independent increments if its increments are independent:
We mark an increment of the process $X[n]$ from time $k_1$ to time $k_2$ by:

$$X(k_1, k_2) = X[k_2] - X[k_1]$$

For a RP with independent increments, the increments $X(k_1, k_2), x(k_3, k_4)$ are independent if the intervals $(k_1, k_2), (k_3, k_4)$ are disjoint intervals (if the intervals are not disjoint, we can say anything about the increments).

### Wiener Process/ Brownian Motion

Wiener process is a mathematical model for describing Brownian motion that we met when we wanted to see the motivation for RP. Brownian motion is a spatial motion which in every moment (continuous time process) can be a movement to each direction in each axis. In order to develop the model we start with random walk in one dimension in discrete time.

**One Dimensional Random Walk in Discrete Time** One dimensional random walk in discrete time is given by:

$$X[n] = X[n-1] + d \cdot W[n], \quad X[0] = 0$$

$$W[n] = W_n = \begin{cases} 1, \ w.p.0.5 \\ -1, \ w.p.0.5 \end{cases}, \quad \{W_n\} \ i.i.d.$$

In each moment a “decision” is made which determines where to go. The steps are of constant size $d$. In fact:

$$X[n] = d \cdot \sum_{k=1}^{n} W_k$$

We can identify using the regression equation that this is a A.R. linear process with coefficient $\alpha = 1$ therefore it is not stationary.
Notice that random walk process is an independent increments process (introduce a pair of increments $X(i, j), X(l, m)$ when we assume $l < m, i < j$):

$$X(l, m) = X[l] - X[m] = d \sum_{k=1}^{l} W_k - d \sum_{k=1}^{m} W_k = d \sum_{k=l+1}^{m} W_k$$

$$X(i,j) = X[i] - X[j] = d\sum_{k=1}^{i} W_k - d\sum_{k=1}^{j} W_k = d\sum_{k=i+1}^{j} W_k$$

Because $\{W_n\}$ is an i.i.d. sequence, then its elements are independent. if the intervals $(i,j),(l,m)$ are disjoint, then in the respect increments there are no elements from the sequence $\{W_n\}$.

The increments $X(l,m)$ is a function (sum) of elements $W_{l+1},...,W_m$, which are mutually independent. The increments $X(i,j)$ is a function (sum) of elements $W_{i+1},...,W_j$, which are mutually independent. Because if variables ae independent, then their functions are independent (and because the intervals are disjoint, there is no common element $W_k$) then we get that $X(l,m) \perp\!\!\!\perp X(i,j)$.

**Up to 2-nd Order Statistics of a One-Dimensional Random Walk in Discrete Time** The process is given by:

$$X[n] = d\sum_{k=1}^{n} W_k$$

The mean of the process is 0 for every $n$:

$$\mathbb{E}\left[X[n]\right] = \mathbb{E}\left[d\sum_{k=1}^{n} W_k\right] = d\cdot\mathbb{E}\left[\sum_{k=1}^{n} W_k\right] = d\cdot\sum_{k=1}^{n}\mathbb{E}[W_k] = d\sum_{k=1}^{n}\left(\frac{1}{2}\cdot 1 + \frac{1}{2}\cdot(-1)\right) = 0$$

The correlation function is given by (assume $n > m$):

$$\begin{aligned}
R_{XX}(n,m) &\triangleq \mathbb{E}\left[X[n]\cdot X[m]\right] \\
&\overset{107}{=} \mathbb{E}\left[(X[m] + X(m,n))\cdot X[m]\right] \\
&= \mathbb{E}\left[(X[m])^2\right] + \mathbb{E}\left[X(m,n)\cdot X[m]\right] \\
&\overset{108}{=} \mathbb{E}\left[(X[m])^2\right] + \mathbb{E}\left[X(m,n)\cdot X(0,m)\right] \\
&\overset{109}{=} \mathbb{E}\left[(X[m])^2\right] + \mathbb{E}\left[X(m,n)\right]\cdot\mathbb{E}\left[X(0,m)\right] \\
&= \mathbb{E}\left[(X[m])^2\right] + 0\cdot 0 \\
&= \mathbb{E}\left[(X[m])^2\right] \\
&= \mathbb{E}\left[\left(d\sum_{k=1}^{m} W_k\right)\right] \\
&= d^2\mathbb{E}\left[\left(\sum_{k=1}^{m} W_k\right)^2\right]
\end{aligned}$$

For $\mathbb{E}\left[\left(\sum_{k=1}^{m} W_k\right)^2\right]$, we can see that the squared elements of the sequence are also i.i.d. Almost all the multiplication will be zeroes because $\mathbb{E}\left[W_i \cdot W_j\right] \overset{Independence}{=} \mathbb{E}[W_i] \cdot \mathbb{E}[W_j = 0]$, hence we get:

$$R_{XX}(n,m) \triangleq d^2 \mathbb{E}\left[\left(\sum_{k=1}^{m} W_k\right)^2\right] \overset{110}{=} d^2 \mathbb{E}\left[\sum_{k=1}^{m} \left(W_k\right)^2\right] = d^2 \cdot m$$

The correlation function is not dependent only on time difference, therefore the process is not stationary.
In general, for general $m, n$ we get:

$$R_{XX}(n,m) \triangleq d^2 \cdot \min(n,m)$$

**One-Dimensional Random Walk in Continuous Time** We would like to make the discrete time random walk RP “become continuous”, so that the step will be made (with probability of half) in times that are an integer multiplication of $T$, meaning in times $nT$ (and won’t move for other times). For this process, the realization will look like:

We mark the process $X_{T,d}(t)$. We identify that up to time $t$, there are $\lfloor \frac{t}{T} \rfloor$ steps. We mark this quantity as $N(t) = \lfloor \frac{t}{T} \rfloor$.
A mathematical expression for the random walk is:

$$X_{T,d}(t) = X[N(t)] = X\left[\left\lfloor \frac{t}{T} \right\rfloor\right] = d\sum_{k=1}^{N(t)} W_k = d\sum_{k=1}^{\lfloor \frac{t}{T} \rfloor} W_k = d\sum_{k=1}^{\infty} W_k \cdot U(t-kT)$$

Where $U(t)$ is Heaviside step function.
The step function “ignores” decision that occur after time $t$ (considers only the first $N(t)$ first decisions), because it is zero for decisions $k$ that are bigger than decision number $N(t)$.

**Statistics Up To 2-nd Order of One-Dimensional Random Walk in Continuous Time** The process is given as $X_{T,d}(t) = d\sum_{k=1}^{\infty} W_k \cdot U(t-kT)$.
The mean of the process is zero for each $n$ from the same reasons: $\mathbb{E}[X_{T,d}(t)] =$

[109] Because $X[n] = X[n] + X[m] - X[m] = X[m] + X(m,n)$
[110] Because $X[m] = X[m] + X[0] - X[0] = X[0] + X(0,m) = X(0,m)$
[111] Independence
[110] $\{W_n\}$ *i.i.d.*

0.
The correlation function is given as (for general $n, m$):

$$R_{XX}(t_1, t_2) = d^2 \cdot \min(N(t_1), N(t_2)) = d^2 \cdot \min\left(\left\lfloor \frac{t_1}{T} \right\rfloor, \left\lfloor \frac{t_2}{T} \right\rfloor\right)$$

The process is not stationary (as the process in discrete time), if $t_1, t_2 \gg T$,then (because $\left\lfloor \frac{x}{A} \right\rfloor \underset{x \gg A}{\approx} \frac{x}{A}$):

$$R_{XX}(t_1, t_2) \underset{t_1, t_2 \gg T}{\approx} d^2 \cdot \min(\frac{t_1}{T}, \frac{t_2}{T})$$

**Wiener Process - Development** We would like to show the random walk process as a continuous time RP. We associate the random walk to a particle that "chooses" infinitesimal movements every time period that goes to zero. We actually want the limit of the process where $T \to 0, d \to 0$ (which will get us a continuous RP) in such a way that the 2-nd order statistics will converge to a finite limit.
The condition for that is the limit existence:

$$\lim_{T \to 0, d \to 0} \left(d^2 \cdot \frac{1}{T}\right)$$

It happens if the expression in the limit is a finite positive constant, meaning $\frac{d^2}{T} = \alpha,\ \alpha > 0$.
Hence we can say that $d^2 = \alpha \cdot T$, so if one of the parameters goes to zero, so is the other one.
Now we can define Wiener process: Wiener process, denoted as $X_a(t)$, is:

$$X_a(t) = \lim_{T \to 0, d^2 = \alpha \cdot T} (X_{T,d}(t))$$

**Up to 2-nd Order Statistics of Wiener Process** The process is given as $X_a(t) = \lim\limits_{T \to 0, d^2 = \alpha \cdot T} (X_{T,d}(t))$
Statistics up to 2-nd order is denoted by:

$$\mathbb{E}[X_a(t)] = 0 \ \forall t$$
$$R_{XX}(t_1, t_2) = \alpha \cdot \min(t_1, t_2)$$

**Wiener Process Distribution** We would like to find the distribution of Wiener process. To do that, we look at the process before the limit,

meaning $X_{T,d}(t)$
By using the equation $d^2 = \alpha \cdot T$, We can just use the notation $X_T(t)$:

$$\begin{aligned}
X_T(t) = X_{T,d}(t)|_{d^2=\alpha \cdot T} &= \sqrt{\alpha \cdot T} \cdot \sum_{k=1}^{N(t)} W_k \\
&= \sqrt{\alpha \cdot T} \cdot \sqrt{\frac{t}{t}} \cdot \sum_{k+1}^{N(t)} W_k \\
&= \sqrt{\alpha \cdot t} \cdot \sqrt{\frac{T}{t}} \cdot \sum_{k=1}^{N(t) W_k} \\
&= \sqrt{\alpha \cdot t} \cdot \frac{1}{\sqrt{t/T}} \cdot \frac{\sqrt{N(t)}}{\sqrt{N(t)}} \cdot \sum_{k=1}^{N(t)} W_k \\
&= \sqrt{\alpha \cdot t} \cdot \frac{\sqrt{N(t)}}{\sqrt{t/T}} \cdot \frac{1}{\sqrt{N(t)}} \cdot \sum_{k=1}^{N(t)} W_k \\
&= \sqrt{\alpha \cdot t} \cdot \frac{\sqrt{\lfloor t/T \rfloor}}{\sqrt{t/T}} \cdot \frac{1}{\sqrt{N(t)}} \cdot \sum_{k=1}^{N(t)} W_k
\end{aligned}$$

Now we use the limit $T \to 0$ because we want to discuss Wiener process.
If the limit exists, we can calculate it in part using the multiplication:

1. $\frac{\sqrt{\lfloor t/T \rfloor}}{\sqrt{t/T}} \underset{T \to 0}{\to} 1$ (because $\left\lfloor \frac{x}{A} \right\rfloor \underset{x \gg A}{\approx} \frac{x}{A}$)
2. $\lim\limits_{T \to 0} \left( \frac{1}{\sqrt{N(t)}} \cdot \sum_{k=1}^{N(t)} W_k \right) \sim N(0,1)$ (using the central limit theorem)

Hence we get that:

$$X_a(t) = \lim_{T \to 0, d^2 = \alpha \cdot T} (X_{T,d}(t)) \sim \sqrt{\alpha \cdot t} \cdot N(0,1) = N(0, \alpha \cdot t)$$

Overall, we get that Wiener process has a normal distribution with mean $eta_X = 0$ and variance of $\sigma_X^2 = \alpha \cdot t$.
Just like the random walk, Wiener process is an independent increments RP, so:

$$X_a(t_1, t_2) = X_a(t_2) - X_a(t_1) \overset{111}{\sim} X_a(t_2 - t_1) \sim N(0, \alpha \cdot (t_2 - t_1))$$

From this we can deduce that the increments also have Gaussian distribution.
Now we will see that Wiener process is a GRP (reminder: GRP is a RP which

[111] The samples difference $X_a(t_2) - X_a(t_1)$ distributes like the samples of the process $X_a(t_2 - t_1)$ because this is an independent increments process, hence the difference represents a similar behavior to that of the process from time $t_1$ to time $t_2$.

every set of its samples are jointly Gaussian - every linear combination of it is a Gaussian random variable).
We look at a pair of samples of the process in times $t_1, t_2$ (assuming $t_1 < t_2$) and notice that:

$$X_a(t_1) = X_a(0, t_1),\ X_a(t_2) = X_a(t_1) + X_a(t_1, t_2) = X_a(0, t_1) + X_a(t_1, t_2)$$

$$\Rightarrow \begin{bmatrix} X_a(t_1) \\ X_a(t_2) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} X_a(0, t_1) \\ x_a(t_1, t_2) \end{bmatrix}$$

We know that the pair of increments $X_a(0, t_1), X_a(t_1, t_2)$ are independent because they are on disjoint time intervals, and in addition we know that each of them distributes Gaussian as we have seen.
Because the vector $\begin{bmatrix} X_a(t_1) & X_a(t_2) \end{bmatrix}^T$ is created from a linear transformation of a Gaussian random vector, it is also a Gaussian random vector, therefore all of its elements are Gaussian. We can do so for every set of time , hence all of the samples of the process can create a Gaussian random vector - meaning the process is a GRP.

**Wiener Proces - Formal Definition** Wiener process $X(t)$ with coefficient $\alpha$ is a RP with Gaussian independent increments that:

$$X(t_1, t_2) \sim N(0, \alpha \cdot (t_2 - t_1))$$

With the initial condition $X(0) = 0$.
The statistics of the process (this is a GRP, so 2-nd order statistics defines the process):

$$\mathbb{E}[X_a(t)] = 0,\ \forall t$$

$$R_{XX}(t_1, t_2) = \alpha \cdot \min(t_1, t_2)$$

Notice the process is not stationary (its correlation function depends on time itself). We can explain intuitively the 2-nd order statistics (assume $t_1 < t_2$):

$$\begin{aligned} R_{XX}(t_1, t_2) &\triangleq \mathbb{E}[X(t_1 \cdot X(t_2))] = \mathbb{E}[X(t_1) \cdot (X(t_1) + X(t_1, t_2))] \\ &= \mathbb{E}[X^2(t_1)] + \mathbb{E}[X(t_1 \cdot X(t_1, t_2))] \\ &= \mathbb{E}[X^2(t_1)] + \mathbb{E}[X(0, t_1) \cdot X(t_1, t_2)] \\ &\overset{112}{=} \mathbb{E}[X^2(t_1)] + \mathbb{E}[X(0, t_1)] \cdot \mathbb{E}[X(t_1, t_2)] \\ &= \mathbb{E}[X^2(t_1)] + 0 \cdot 0 = \mathbb{E}[X^2(t_1)] \\ &\overset{113}{=} \mathrm{Var}(X(t_1)) \overset{114}{=} \sigma_X^2 = \alpha \cdot t_1 \end{aligned}$$

This is intuitive because for each pair of samples of the process, there is correlation only in the overlap, hence this is actually a auto-correlation in this interval, the overlap is an increment hence it has a Gaussian distribution. auto-correlation of a Gaussian random variable is the variance, so we get $R_{XX}(t_1, t_2) = \alpha \cdot \min(t_1, t_2)$

**Derivative Process of Wiener Process** We would like to look at the derivative process of the Wiener process. To do that, we will first look at the slope process of the Wiener process, and define:

$$X_\epsilon(t) = \frac{X(t+\epsilon) - X(t)}{\epsilon} = \frac{X(t, t+\epsilon)}{\epsilon}$$

The slope process is a RP (its is a function of a RP), and it is Gaussian( linear combination of number of samples of a GRP - all the samples of a GRP are jointly Gaussian).
So the derivative process is:

$$X'(t) = \frac{dX(t)}{dt} = \lim_{\epsilon \to 0} X_\epsilon(t)$$

The derivative process is also a GRP because it is a limit of a GRP (this is not a formal proof, but the claim is true).
Because the derivative process $X'(t)$ is a GRP, we need to find the statistics up to 2-nd order in order to characterize it.
Firsts we find the statistics for the slope process, and then use the limit.
The mean of the process:

$$\mathbb{E}[X_\epsilon(t)] = \mathbb{E}\left[\frac{X(t+\epsilon) - X(t)}{\epsilon}\right] = \frac{1}{\epsilon}\left(\mathbb{E}[X(t+\epsilon)] - \mathbb{E}[X(t)]\right) \overset{115}{=} \frac{1}{\epsilon}(0{-}0) = 0$$

The correlation function of the process (for simplicity, assume $\tau > 0$):

$$\begin{aligned} R_{X_\epsilon X_\epsilon}(t+\tau, t) &\triangleq \mathbb{E}[X_\epsilon(t+\tau) \cdot X_\epsilon(t)] \\ &= \mathbb{E}\left[\frac{X(t+\tau+\epsilon) - X(t+\tau)}{\epsilon} \cdot \frac{X(t+\epsilon) - X(\epsilon)}{\epsilon}\right] \\ &= \frac{1}{\epsilon^2}\mathbb{E}[X(t+\tau, t+\tau+\epsilon) \cdot X(t, t+\epsilon)] \end{aligned}$$

[114] This is an independent increments RP, and the intervals $(0, t_1), (t_1, t_2)$ are disjoint.

[115] Because $\eta_X = 0$: $\sigma_X^2 = \mathrm{Var}(X(t_1)) = \mathbb{E}[X^2(t_1)] - \eta_X^2 = \mathbb{E}[X^2(t_1)]$

[116] $X(t_1) = X(0, t_1)$

[115] For Wiener process sample: $\eta_X = 0$

Notice we represented the correlation function by a pair of increments of the Wiener process (which is a process of independent increments).
Divide the development to 2 cases:

1. $0 < \epsilon < \tau$ (there is no overlap in the times of two different derivative processes):
   In this case the intervals are disjoint. Hence the increments are independent, each with mean of zero.

2. $0 < \tau < \epsilon$ (there is an overlap between the times of a pair of derivative processes):
   In this case, the intervals won't be disjointed and they overlap in the interval $(t + \tau, t + \epsilon)$.
   Divide the increments to separate parts:

$$X(t, t + \epsilon) = X(t, t + \tau) + X(t + \tau, t + \epsilon)$$

$$X(t + \tau, t + \tau + \epsilon) = X(t + \tau, t + \epsilon) + X(t + \epsilon, t + \tau + \epsilon)$$

   After the multiplication, every increment of disjoint intervals are becoming zero, because they are independent increment with mean of 0, and we get:

$$\begin{aligned} R_{X_e X_e}(t + \tau, t) &= \frac{1}{\epsilon^2} \cdot \mathbb{E}[X(t + \tau, t + \tau + \epsilon) \cdot X(t, t + \epsilon)] \\ &= \frac{1}{\epsilon^2} \cdot \mathbb{E}[X^2(t + \tau, t + \epsilon)] \\ &\overset{116}{=} \frac{1}{\epsilon^2} \cdot \mathbb{E}[X^2((t + \epsilon) - (t + \tau))] \\ &= \frac{1}{\epsilon^2} \cdot \mathbb{E}[X^2(\epsilon - \tau) \\ &\overset{117}{=} \frac{1}{\epsilon^2} \cdot \alpha \cdot (\epsilon - \tau) \end{aligned}$$

We developed for $\tau > 0$, if we look at cases where $\tau < 0$ whe intervals will look like:

Also here, in the first case, there is no overlap - so the correlation function will be zero. In the second case, there is an overlap in the interval $(t, t + \tau + \epsilon)$ - the length of this interval is $\epsilon + \tau$, while for $\tau > 0$ we got an interval of length $\epsilon - \tau$. In our case $\tau$ us negative, so in both cases we can

[118] Wiener process properties - independent increments and their distribution

[117] See previous footnote

equivalently write the overlap interval length as $\epsilon - |\tau|$.
In both cases we get that the correlation function depends only on the time difference, hence the process is WSS.
We can summarize and write:

$$R_{X_eX_e}(t+\tau,t) = R_{X_eX_e}(\tau) = \begin{cases} 0, & |\tau| > \epsilon \\ \frac{1}{\epsilon^2} \cdot \alpha \cdot (\epsilon - |\tau|), & |\tau| < \epsilon \end{cases}$$

We can also show the correlation function graphically:

Identify that the area of the triangle is $\alpha$ and doesn't depend on $\epsilon$ but in the ratio of the parameters $T, d$.
In the derivative limit, where $\epsilon \to 0$, we get a delta of height $\alpha$, i.e.: $R_{X'X'}(t+\tau,t) = \alpha \cdot \delta(\tau)$.
It is important to note that this is not the usual delta function, as the area is finite.
Then we get:

$$R_{X'X'}(t+\tau,t) = R_{X'X'}(\tau) = \begin{cases} 0, & |\tau| > \epsilon \\ \alpha \cdot \delta(\tau), & |\tau| < \epsilon \end{cases}$$

As you recall, such a correlation function fits a white noise (there is no correlation between a pair of samples, doesn't matter how close they are) with PSD $\alpha$. we usually denote $\alpha = \frac{N_0}{2}$ in order to represent PSD of the positive frequencies.
White noise has a constant power (the amplitude of *delta*) for all frequencies. this big power is coming from the original random walk, a process that contains a lot of energy.

### Poisson Random Process

Poisson RP is the simplest counting process, the process counts the number of events in time for events that happen randomly and independently. This is a continuous time random process with discrete values.

**Examples of natural Poisson RP:**

- Number of rain drops that fall into a cup (or any defined circular cross section)
- Radioactive decays

- Lines (in first order approximation)

In Poisson RP the time distribution between two consecutive events is exponential distribution and the distribution of two events that happens in a time interval is Poisson distribution.

**Exponential Distribution** The exponential distribution is a continuous distribution over the non-negative numbers. this distribution has a key property of being memorylessness, hence it describes random event that the probability of them to happen is constant in time. This distribution has one parameter that represents the decay rate.

$$F_X(x) = \begin{cases} 1 - e^{-\lambda \cdot x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

$$f_X(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

**Memorylessness Property of an Exponential RP** As said, exponential random variable is memoryless, i.e. for $X \sim exp(\lambda)$:

$$\Pr(X > x_1 + x_2 | X > x_1) = \Pr(X > x_2)$$

Proof:

$$\begin{aligned} \Pr(X > x_1 + x_2 | X > x_1) &= \frac{\Pr(X > x_1 + x_2 \cap X > x_1)}{\Pr(X > x_1)} \\ &= \frac{\Pr(X > x_1 + x_2)}{\Pr(X > x_1)} \\ &\overset{118}{=} \frac{1 - \left[1 - e^{-(x_1 + x_2) \cdot \lambda}\right]}{1 - \left[1 - e^{-x_1 \cdot \lambda}\right]} \\ &= \frac{e^{-(x_1 + x_2) \cdot \lambda}}{e^{-x_1 \cdot \lambda}} \\ &= e^{-x_1 \cdot \lambda} = \Pr(X > x_2) \end{aligned}$$

[118] $X \sim exp(\lambda)$

**Poisson Distribution** Poisson distribution is a distribution of a discrete RP. the distribution describes the probability of number of events to occur in a certain time interval, when the events happen in a defined rate and independently. The distribution is defined using a single parameter that represents the rate of events.
Poisson random variable with parameter $\lambda$ is denoted by $X \sim Poisson(\lambda)$
Probability function of Poisson random variable (the probability to get $k$ events):

$$\Pr(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

**Building Poisson RP** We build Poisson RP using thw limit of a discrete RP. We start with an i.i.d. Bernoulli RP.

**”Balloon Life” RP in Discrete Time** Balloon life happens until the event of popping the balloon (the popping is the event we are interested in - for example, a drop that falls into a cup).
We define a RP $W_n$ of an “annoying kid” that in every time interval tries to pop the balloon. Every attempt has a probability of success $p$ and the event will be denoted as $W_n = 1$. If the boy don’t succeed in popping it at a certain attempt, it doesn’t affect the balloon, hence the attempts sequence is i.i.d.

$$W_n = \begin{cases} 1, & w.p.\ p \\ 0, & w.p.\ 1-p \end{cases}, \{W_n\}\ i.i.d,\ W_n \sim Ber(p)$$

In order to define the experiment, we assu,e the kid continues to try and pop the balloon even after it already popped.
We denote the lifetime of the balloon by $S^D$, where lifetime of the balloon is the number of attempts until the first success. Because all the attempts have the same distribution $Ber(p)$ and i.i.d., then the time of the first “pop” (first success in terms of Bernoulli distribution) distributes geometrically with parameter $p$. So:

$$S^D = \min_n (W_n = 1) \sim Geo(p)$$

$$\Rightarrow \mathbb{E}[S^D] = \frac{1}{p},\ \text{Var}(S^D) = \frac{1-p}{p^2}$$

If we want a first success at time $n$, it means there have to be $n-1$ unsuccessful attempts and then a success:

$$\Pr(S^D = n) = (1-p)^{n-1} \cdot p$$

We define the process "Balloon life" in discrete time by:

$$X^D[n] = U(n - S^D)$$

Note:
Notice that like a Wiener RP, we created the process from a Bernoulli random variable, but in here we took the minimal time between changes, while in Wiener RP we summed the changes (in addition, in the development of Wiener RP "failure" had an equal probability and a value of $-1$).

**"Balloon Life" RP in Continuous Time** We would like to turn the "balloon life" RP from discrete time to continuous time. Assume the attempts to pop the balloon happen every time interval $T$.
Hence the time of pop will be $S^C = T \cdot S^D$. Now the "pop success density" is the success probability $p$ for time $T$, so the pop time distributes $S^C \sim Geo(\frac{p}{T})$. We define the RP "balloon life" in continuous time:

$$X^C_{p,T}(t) = X^C(t) = U(n - T \cdot S^D) = U(n - S^C)$$

**"Balloon Life" RP in Limit** We would like to minimize the time intervals between popping attempts, so that the boy will try to pop the balloon "all the time", but in a way that the success rate for a time unit will stay constant. To do so, we have to minimize both $T$ and $p$ in the same manner (if we minimize only one, the balloon will pop immediately or will not pop for ever).
In every time unit there are $\frac{1}{T}$ popping attempts with probability $p$, hence the mean of success for a time unit is $\frac{p}{T}$ (sum of $\{W_n\}$ i.i.d.). So we denote the mean of number of successful pops in time unit (density) by $\lambda = \frac{p}{T}$.
In order to follow the demand, we maintain the relation $p = \lambda \cdot T$. In this way we can make one of the parameters $p, T$ go to zero and the other one will go to zero at the same time, thus maintaining the relation (from this we can set the time interval $T$ if $p$ is known and the wanted $\lambda$ coefficient).

$$X(t) = \lim_{T \to 0, p = \lambda \cdot T} (X^C_{p,T}(t))$$

We know that the discrete geometrical distribution goes to continuous exponential distribution in this limit. For that, we calculate the probability that the first pop will not occur until a certain moment, meaning will happen only after that moment:

$$\begin{aligned}
\Pr(S^C > t) &= \Pr(T\cdot S^D > t) = \Pr\left(S^D > \frac{t}{T}\right)\\
&= \Pr(S^D > \left\lfloor \frac{t}{T} \right\rfloor) \overset{119}{=} (1-p)^{\left\lfloor \frac{t}{T} \right\rfloor}\\
&\overset{120}{=} (1-\lambda\cdot T)^{\left\lfloor \frac{t}{T} \right\rfloor} = (1-\lambda\cdot T)^{\left\lfloor \frac{t}{T} \right\rfloor\cdot\frac{t/T}{t/T}}\\
&= (1-\lambda\cdot T)^{\frac{\lfloor t/T \rfloor}{t/T}\cdot(t/T)} \underset{T\to 0}{\rightarrow} (1-\lambda\cdot T)^{1\cdot(t/T)}\\
&= \left(\left[(1-\lambda\cdot T)^{1/T}\right]\Big|_{T\to 0}\right)^{t}\\
&\overset{121}{=} (e^{-\lambda})^{t} = e^{-\lambda\cdot t}
\end{aligned}$$

Hence the process:

$$X(t) = \lim_{T\to 0, p=\lambda\cdot T} (X^C_{p,t}(t)) = U(t-S)$$

$$S \sim exp(\lambda),\ \lambda = \frac{p}{T}$$

**”N Balloons Life” RP** This time we take a collection of $N$ independent balloons (the times of pop are independent) and denote the popping time of each balloon:

$$S_1, ..., S_N \sim exp(\lambda),\ \{S_i\}_{i=1}^{N} i.i.d.$$

Now there are also $N$ “annoying kids”. If a certain lids pops his/hers balloon, they don’t get a new one, thus effectively removed from the “game”. The rest of the kids continue to “play” independently until all of them pop their balloons.
Hence the process will look like (like a staircase):

$$X(t) = \sum_{i=1}^{N} U(t-S_i)$$

We would like to know the odds of the time of the first balloon pop, the second and so on. In general we want to know the odds of the time of the

[121] $S^D \sim Geo(p)$, and we are looking for the opposite of CDF, so: $1 - F_{S^D}\left(\left\lfloor \frac{t}{T} \right\rfloor\right) = 1 - [1-(1-p)^{\left\lfloor \frac{t}{T} \right\rfloor}] = (1-p)^{\left\lfloor \frac{t}{T} \right\rfloor}$

[122] $p = \lambda \cdot T$

[123] For $n = \frac{1}{T},\ x = -\lambda$, the limit is: $\lim\limits_{n\to\infty}\left[\left(1+\frac{x}{n}\right)^{n}\right] = e^{x}$

i-th pop (we don't want to know when the i-th balloon popped, because this is just exponential distribution since they are i.i.d., but to know when was the i-th pop out of $N$ balloons). We want to know these time because we are not interested in when a certain deop will fall to the cup, but when the next drop will fall. We denote the sequence of the popping times of all the $N$ balloons by (this is an example of Order Statistics):

$$S^1, ..., S^N$$

Notice that:

$$S^1 = \min(S_1, ..., S_N),\ S^2 = \min(S_1, ..., S_N \backslash S^1), ...$$

We find the distribution of the time up to the first pop, meaning $S^1$:

$$\begin{aligned} \Pr(S^1 < t) &\overset{122}{=} \Pr(S_1 > t \cap S_2 > t\cap, ..., \cap S_N > t) \\ &\overset{123}{=} \prod_{i=1}^{N} \Pr(S_i > t) \overset{124}{=} [\Pr(S_i > t)]^N \\ &\overset{125}{=} e^{\lambda \cdot N} \end{aligned}$$

We define an effective coefficient $\lambda = \lambda_{eff} = \lambda \cdot N$, notice that as there are more "kids" the effective coefficient gets bigger, this makes sense since if there are more kids, the first pop will happen faster.
Overall we got that the distribution of the first pop is:

$$S^1 \sim exp(\lambda),\ \lambda = \lambda_{eff} = \lambda \cdot N,\ \lambda = \frac{p}{T}$$

The meaning of memorylessness of an exponential random variable in our experiment is that the probability the pop didn't happen till time $t_1 + t_2$ given that it didn't happen till time $t_1$ is equal to the probability it won't pop until time $t_2$. Meaning it has to survive $t_2$ more time, and the fact that it already "survived" time $t_1$ doesn't matter. Meaning:

$$\Pr(S^1 > t_1 + t_2 | S^1 > t_1) = \Pr(S^1 > t_2)$$

A first balloon popped. we don't know which balloon, but we know the pop happened in time $S^1$. We would like to look at the state of balloon $S_i$.

[124] Because the first pop will happen after time $t$, all the balloons must pop after time $t$.

[125] i.i.d., so independent.

[126] i.i.d.

[127] $S \sim exp(\lambda)$

Identify that the claim $S_i \neq S^1$ is identical to that $S_i > S^1$ because if the time it was popped is not the time of the first pop, then it didn't pop until that time.
Given this state we would like to know what the probability this balloon will “survive” $t$ more time, from memorylessness wee get that the answer is counting the time from the last time we know about - meaning the probability it will “survive” a period of time $t$:

$$\Pr(S_i > S^1 + t | S_i \neq S^1) = \Pr(S_i > S^1 + t | S_i > S^1) \overset{126}{=} \Pr(S_i > t) = e^{-\lambda \cdot t}$$

So the second pop time distribution, i.e. $S^2$, is:

$$\begin{aligned}
\Pr(S^2 > S^1 + t) &\overset{127}{=} \Pr((S_1 > S^1 + t | S_1 > S^1) \cap ... \cap (S_{N-1} > S^1 + t | S_{N-1} > S^1)) \\
&\overset{128}{=} \prod_{i=1}^{N-1} \Pr(S_i > S^1 + t | S_i > S^1) \\
&\overset{129}{=} \left[\Pr(S_i > S^1 + t | S_i > S^1)\right]^{N-1} \\
&\overset{130}{=} \left[\Pr(s_i > t)\right]^{N-1} \\
&\overset{131}{=} \left(e^{\lambda \cdot t}\right)^{N-1} = e^{-\lambda \cdot (N-1) \cdot t}
\end{aligned}$$

We get that the probability that the second pop will happen after more than time $t$ after the first pop (i.e. at least $S^1 + t$ time) is like the survival” of the first pop for time $t$ in a $N - 1$ balloons system.
Conclusion:

$$\begin{gathered}
S^1 \sim exp(\lambda \cdot N) \\
S^2 - S^1 \sim exp(\lambda \cdot (N-1)) \\
\vdots \\
S^N - S^{N-1} \sim exp(\lambda) \\
\Rightarrow \begin{cases} S^k - S^{k-1} \sim exp(\lambda \cdot (N+1-k)) \ \forall 1 \leq k \leq N \\ S^0 = 0 \end{cases}
\end{gathered}$$

[126] $S_i \sim exp(\lambda)$

[129] In order for the second pop to be after time $S^1 + t$, we demand that all the balloons left will survive the remaining time.

[130] i.i.d., so independent.

[131] i.i.d.

[132] $S \sim exp(\lambda)$

[133] $S_i \sim exp(\lambda)$

We see that for later pops there is a smaller delay constant than the earlier pops, for all the differences between pops growing bigger. This makes sense, since there are less kids playing hence smaller probability to pop (they are independent). This also works mathematically, since the mean of $X \sim exp(\lambda)$ is $\mathbb{E}[X] = \frac{1}{\lambda}$, meaning the smaller the decay constant, the longer it takes in average.
Now that we are given the pop time distribution we can use them to write the process (the equality holds since the sets $\{S^i\}_{i=1}^N, \{S_i\}_{i=1}^N$ have the same elements but in different order):

$$X(t) = \sum_{i=1}^{N} U(t - S_i) = \sum_{i=1}^{N} U(t - S^i)$$

Because we know the distribution, we can show the process graphically

**”N Balloon Life” RP With Returns** Now we assume that if a kid pops its balloon, it isn’t removed from the game but gets a new one and continues to play. Now the popping time wil distribute evenly. This game has no end and the process will continue forever (an eqaul example is a diode - we have constant voltages, and the events are electro “jumps”). We get:

$$S^k - S^{k-1} \sim exp(\lambda \cdot N)$$

It is equal to the natural case where $N \gg 1$ and then we get the $N-1 \approx N$, so the decay constant of the two first popping times are equal $\lambda \cdot n \approx \lambda \cdot (N-1)$, and in fact we can say that $N - k \approx N$ up to a certain $k$. This means that all the first $k$ pops have the same distribution:

$$S^k - S^{k-1} \underset{approx.}{\sim} exp(\lambda \cdot N),\ \frac{N-k}{N} \approx 1$$

**Poisson RP** We take $N \to \infty, \lambda \to 0$ so that their multiplication is a constant finite decays constant $\lambda = \lambda_{eff} = \lambda \cdot N$. We get that all the stairs gaps will distribute the same $exp(\lambda)$. This process will be denoted as $X(t) \sim Poissong(\lambda \cdot t)$ and :

$$X(t) = \sum_{i=1}^{\infty} U(t - S_i) = \sum_{i=1}^{\infty} U(t - S^i),\ S^i - S^{i-1} \sim exp(\lambda),\ S^0 = 0$$

Note:
We can treat the process as a process that is being built from the increments $S^i - S^{i-1}$:

$$S^i = S^{i-1} + Z_i,\ S^0 = 0,\ Z_i \sim exp(\lambda),\ \{Z_i\}_{i=1}^{\infty}\ i.i.d.$$

Where $\{Z_i\}$ are the pops gaps. this is an A.R. Linear RP with coefficient $\alpha = 1$ and so not stationary.

**Increments Process of a Poisson RP** For a Poisson RP with parameter $\lambda$, $X(t) \sim Poisson(\lambda \cdot t)$ We define an increments process $X(t_1, t_2) = X(t_2) - X(t_1)$. The increment $X(t_1, t_2)$ represents he number of pops in the time interval $[t_1, t_2]$ . This increments process distributes the same for the whole process (because of memorylessness and limit, in each point in time there is the same amount of kids with the same amount of balloons and the past successes/failures don't apply) and in addition it is an independent increments RP, so:

$$X(t_1, t_2) \sim Poisson(\lambda(t_2 - t_1)),\ X(t_1, t_1 + t) \sim Poisson(\lambda t)$$

**Distribution of an Independent Increments RP of a Poisson RP** We develop the distribution of the increment process of the Poisson RP using the "N balloons life" with returns where $N \to \infty$

$$\Pr(X(t_1, t_1 + t) = k) \overset{132}{=} \Pr(X(t) = k)$$

Meaning, the probability that in time interval of length $t$ $k$ balloons popped is independent in the "location" of time, i.e. there is no importance to whether we look in time interval $(0, t]$ or time interval $(t_1, t_1 + t]$.
In order to get exactly $k$ pops, we demand that $(N-k)$ balloons will "survive" (it doesn't matter which)

$$\begin{aligned}\Pr(X(t_1, t_1 + t) = k) &= \Pr(X(t) = k)\\ &= \binom{N}{k} \cdot (1 - e^{-\lambda \cdot t})^k \cdot (e^{-\lambda \cdot t})^{N-k}\\ &\overset{133}{=} \binom{N}{k} \cdot (1 - e^{\frac{\lambda}{N} t})^k \cdot (e^{-\frac{\lambda}{N} t})^{N-k}\end{aligned}$$

[132] Memorylessness

We would like to arrange the expression a little before $N \to \infty$, we get together expressions that will go to 1:

$$\begin{aligned}
&\binom{N}{k} \cdot (1 - e^{\frac{\lambda}{N}t})^k \cdot (e^{-\frac{\lambda}{N}t})^{N-k} \\
&\qquad = \frac{N!}{k! \cdot (N-k)!} \cdot (1 - e^{-\frac{\lambda t}{N}})^k \cdot \frac{(\lambda t/N)^k}{(\lambda t/N)^k} \cdot (e^{-\frac{\lambda t}{N}})^N \cdot (e^{-\frac{\lambda t}{N}})^{-k} \\
&\qquad = \frac{1}{k!} \cdot \frac{N \cdot ... \cdot (N-k+1)}{N^k} \cdot \frac{(N-k)!}{(N-k)!} \cdot \frac{(1 - e^{-\frac{\lambda t}{N}})^k}{(\lambda t/N)^k} \cdot (\lambda t)^k \cdot e^{-\lambda t} \cdot (e^{-\frac{\lambda t}{N}})^{-k}
\end{aligned}$$

Remember that $e^x = 1 + x + ...$, so $1 - e^x \approx -x$, so we get:

$$\begin{aligned}
\Pr(X(t_1, t_1 + t) = k) &= \frac{1}{k!} \cdot \frac{N \cdot ... \cdot (N-k+1)}{N^k} \cdot \frac{(\lambda t/N)^k}{(\lambda t/N)^k} \cdot (\lambda t)^k \cdot e^{-\lambda t} \cdot (e^{-\frac{\lambda t}{N}})^{-k} \\
&= \frac{1}{k!} \cdot \frac{N \cdot ... \cdot (N-k+1)}{N^k} \cdot (\lambda t)^k \cdot e^{-\lambda t} \cdot (e^{-\frac{\lambda t}{N}})^{-k} \\
&\underset{N\to\infty}{\rightarrow} \frac{1}{k!} \cdot 1 \cdot (\lambda t)^k \cdot e^{-\lambda t} \cdot 1 = \frac{1}{k!} \cdot (\lambda t)^k \cdot e^{-\lambda t}
\end{aligned}$$

And we get that the increment has the same distribution as expected.

**Possion RP - Formal Definition** We can “forget” the building process we did and “get” equivalent definitions to Poisson RP.
Poisson RP $X(t) \sim Poisson(\lambda t)$ (defined for every $t \geq 0$) is:

1. Built from an A.R. Process of the popping times (linear A.R. with coefficient $\alpha = 1$ so not stationary):

   $$S^i = S^{i-1} + Z_i,\ S^0 = 0,\ Z_i \sim exp(\lambda)\ \forall i,\ \{Z_i\}_{i=1}^{\infty}\ i.i.d.$$

   So the popping times build the process itself by:

   $$X(t) = \sum_{n=1}^{\infty} U(t - S^n),\ X(0) = 0$$

   Note:
   From the popping time we identify that

   $$S^n = \sum_{i=1}^{n} Z_i$$

[133] $\lambda = \lambda \cdot N$

The sequence is i.i.d., so the popping time is a sum of $n$ i.i.d. random variables, so:

$$f_{S^n} = f_{Z_1} * ... * f_{Z_n}$$

This is a sum of $n$ i.i.d. exponential random variables, this sum distributes with Erlang distribution:

$$f_{S^n} = \frac{\lambda^n \cdot t^{n-1} \cdot e^{-\lambda \cdot t}}{(n-1)!}$$

We can identify that for $n = 1$, Erlang distribution collapses into exponential distribution.

2. A process that satisfies:

   (a) Initial condition $X(0) = 0$

   (b) It is a non-negative integer-valued monotonically non-decreasing step-function

   (c) Has independent increments that are Poisson distributed: $X(t_1, t_2) \sim Poisson(\lambda \cdot (t_2 - t_1))$

   Note: the increment $X(t_1, t_2) = X(t_2) - X(t_1)$ represents the number of pops in a time interval $[t_1, t_2]$, the increments process distributes Poisson (at every point in time there is the same number of kids and the same number of balloons)

### Properties (without proof)

- $\mathbb{E}[X(t)] = \lambda \cdot t$, the mean is linear in time. We would expect the steps function to align with the line $y(t) = \lambda \cdot t$. This is a monotonic, non-decreasing process so the mean must be increasing, in addition the process acts the same for every time so we expect a line (that start at $(0, 0)$ like the initial value of the process).
- $\text{Var}(X(t)) = \lambda t$ the variance is equal to the mean. In fact, $\sqrt{\lambda t}$ is the "reasonable" dimension to get away form the mean line.
  Notice the process is not stationary (the derivative process is - as we will see), because the variance depends on time.
- The process is additive - will show later.

**Iterative Building of Poisson RP From Bernoulli Sum** We identify we can build Poisson RP by summing i.i.d. Bernoulli random variables, so we will be making decision really fast, with very low probabilities the size of a successful decision will be one. For the building we will show that in Poisson RP can happen only one event in an infinitesimal unit of time.
We take a RP $X(t) \sim Poisson(\lambda \cdot t)$ and analyze the number of pops in a short time. Remember:

$$\Pr(X(t) = k) = \frac{(\lambda t)^k \cdot e^{-\lambda t}}{k!}$$

We look for the probabilities to zero/one/more pops as $t = \Delta \ll 1$ and develop, where $O(\Delta^2)$ means negligible of $\Delta$ and represents the rest of the series (includes all the elements with $\Delta^2$ or more)

$$\begin{aligned}
\Pr(X(\Delta) = 0) &= \frac{(\lambda\Delta)^0 \cdot e^{-\lambda\Delta}}{0!} \overset{134}{=} 1 - \frac{\lambda\Delta}{1} + O(\Delta^2) = 1 - \lambda\Delta + O(\Delta^2) \\
Pr(X(\Delta) = 1) &= \frac{(\lambda\Delta)^1 \cdot e^{-\lambda\Delta}}{1!} = \lambda\Delta \cdot e^{-\lambda\Delta} = \lambda\Delta \cdot (1 - \lambda\Delta + O(\Delta^2)) = \lambda\Delta + O(\Delta^2) \\
\Pr(X(\Delta) > 1) &= 1 - (1 - \lambda\Delta + O(\Delta^2)) - (\lambda\Delta + O(\Delta^2)) = O(\Delta^2)
\end{aligned}$$

We get that in finite time, there can only be one event.
Notice that for time $t = \Delta \ll 1$:

$$\Pr(X(\Delta) = 0) \approx 1 - \lambda\Delta, \ \Pr(X(\Delta) = 1) \approx \lambda\Delta$$

This is Bernoulli distribution with success rate $p = \lambda\Delta$.
Hence we can build the process by:

$$X^T(t) = \sum_{n=1}^{\lfloor t/T \rfloor} B_n = \sum_{n=1}^{\infty} B_n \cdot U(t - nT), \ B_n \sim Ber(\lambda T), \ \{B_n\}_{n=1}^{\infty} \ i.i.d.$$

We sample the process every time $T$ (equivalent to that every time interval $T$ we decide if a balloon popped with a probability of $p = \lambda\Delta$)
We take the limit $T \to 0$ while maintaining the ratio $p = \lambda T$(so $p \to 0$ also)
We show that the distribution of the process $\lim\limits_{T \to 0, p = \lambda T}(X^T(t))$ goes to the distribution of Poisson RP:

$$\Pr(X^T(t) = k) = \Pr\left(\sum_{n=1}^{\lfloor t/T \rfloor} B_n = k\right) \overset{135}{=} \binom{\lfloor t/T \rfloor}{k} \cdot (1 - \lambda T)^{\lfloor t/T \rfloor - k} \cdot (\lambda T)^k$$

[134] Taylor series

We rearrange the terms, so we can take many part to one:

$$\begin{aligned}
\Pr(X^T(t) = k) = {} & \frac{\lfloor t/T \rfloor \cdot (\lfloor t/T \rfloor - 1) \cdots (\lfloor t/T \rfloor - k + 1)}{k!} \\
& \cdot (1 - \lambda T)^{\lfloor t/T \rfloor} \cdot \frac{1}{(1 - \lambda T)^k} \cdot (\lambda T)^k \\
= {} & \frac{1}{k!} \cdot (\lfloor t/T \rfloor)^k \cdot \frac{\lfloor t/T \rfloor \cdot (\lfloor t/T \rfloor - 1) \cdots (\lfloor t/T \rfloor - k + 1)}{(\lfloor t/T \rfloor)^k} \\
& \cdot \left(1 - \frac{\lambda}{1/T}\right)^{\lfloor t/T \rfloor} \cdot \left(\frac{\lambda T}{1 - \lambda T}\right)^k \\
\underset{T \to 0}{\longrightarrow} {} & \frac{1}{k!} \cdot \left(\frac{t}{T}\right)^k \cdot 1 \cdot e^{-\lambda t} \cdot (\lambda T)^k = \frac{1}{k!} \cdot e^{-\lambda t} \cdot (\lambda t)^k
\end{aligned}$$

Meaning we get: $X(t) \sim Poisson(\lambda t)$, $X(t) = \lim\limits_{T \to 0, p = \lambda T}$

Note:

Building the process as a sum of Bernoulli process reminds the building of Wiener process. Notice that in the development of the Wiener process with Bernoulli sequence, in every decision there was a step to some direction, while in the development of Poisson process with Bernoulli sequence, not every decision made a movement, the step is of a constant size and eaul to one but with low probability. The sums are very different.

**Simulation Options of a Poisson RP** Because Poisson process is a natural process we would like to simulate it, and for that we will need to randomly choose samples of the process. There are a few ways to simulate a Poisson process:

1. We take a series of i.i.d random variables $\{Z_n\}, Z_i \sim exp(\lambda)$, and use them to build $\{S^i\}$ using the recurrence relation.
   pros: by definition. cons: creating such $Z_n$ is not a simple task.

2. We take a serues of Bernoulli random variables $B_n$ using the alternative build and sum them $X(t) = \sum_n B_n$.
   pros: simple. cons:that is only an approximation.

3. To simulate a Poisson RP with parameter $\lambda$ with time $T_{tot}$ we take a Poisson random variable with distribution $Poisson(\lambda \cdot T_{tot})$ and denote

[135] $Bin(\lfloor t/T \rfloor, \lambda T)$

what we get as $N(T_{tot})$.
If $N(T_{tot}) = 0$, we finished.
Else, take $n = N(T_{tot})$ i.i.d. random variables with distribution $Uniform(0, T_{tot}]$, mark them as $u_1, ..., u_n$, sort them by increasing size $o_1, ..., o_n$ ($o_1 < ... < o_n$). The sequence $\{o_i\}_{i=1}^{N(T_{tot})}$ is a series of Poisson evets in the interval $(0, T_{tot}]$.
pros: precise and requires only one "though" randomness, cons: requires sorting.
Proof:

- $N(T_{tot}) = 0$, there are no event, the trivial event.
- $N(T_{tot}) = 1$, there is a single event, we will show that it distributes uniformly in the interval $(0, T_{tot}]$:

$$\begin{aligned}
\Pr(o_1 \le s | n = 1) &= \frac{\Pr(o_1 \le s \cap n = 1)}{\Pr(n = 1)} \\
&\overset{136}{=} \frac{\Pr(X(s) = 1 \cap X(T_{tot}) - X(s) = 0)}{\Pr(X(T_{tot}) = 1)} \\
&\overset{137}{=} \frac{\Pr(X(s) = 1) \cdot \Pr(X(T_{tot}) - X(s) = 0)}{\Pr(X(T_{tot}) = 1)} \\
&= \frac{e^{-\lambda s}(\lambda s)^1}{1!} \cdot \frac{e^{\lambda(T_{tot}-s)}(\lambda(T_{tot} - s))^0}{0!} \cdot \frac{1!}{e^{-\lambda T_{tot}}(\lambda T_{tot})^1} \\
&= \frac{s}{T_{tot}} = CDF(Unif(o, T_{tot}])(s)
\end{aligned}$$

- $N(T_{tot}) \ge 2$, there are at least two event. we will show that their location distributes $OrderedUniform(0, T_{tot}]$:

$$\Pr(t_1 = o_1, ..., t_n = o_n | N(T_{tot}) = n) = \frac{\Pr(t_1 = o_1, ..., t_n = o_n \cap N(T_{tot}) = n)}{\Pr(N(T_{tot}) = n)} = \mathcal{Y}$$

We replace the event on the sequence $t_1, ..., t_n$ to the ordered event on the difference sequence: $x_1 = o_1 - 0, x_2 = o_2 - o_1, ..., x_n = o_n - o_{n-1}$.
We are looking for an intersection with the event of having exactly

[138] The meaning is there was one event in the interval $(0, s)$ and zero events in the interval $(s, T_{tot})$

[137] In the numerator there is a union of disjoint interval that are independent, thus can be separated

$n$ events up to time $T_{tot}$, this means the “next” event occur only after time $T_{tot}$. Hence:

$$\begin{aligned}
\mathcal{Y} &= \frac{\Pr(x_1 = o_1 - 0, ..., x_n = o_n - o_{n-1}, x_{n+1} > T_{tot} - o_n)}{\Pr(N(T_{tot}) = n)} \\
&= \frac{\Pr(x_1 = o_1) \cdot ... \cdot \Pr(x_n = o_n - o_{n-1}) \cdot \Pr(x_{n+1} > T_{tot} - o_n)}{\Pr(N(T_{tot}) = n)} \\
&\overset{138}{=} \lambda e^{\lambda \cdot o_1} \cdot ... \cdot \lambda e^{\lambda \cdot (o_n - o_{n-1})} \cdot [1 - F_{exp(\lambda)}(T_{tot} - o_n)] \cdot \frac{n!}{e^{\lambda \cdot T_{tot}} \cdot (\lambda T_{tot})^n} \\
&= \lambda^n e^{-\lambda o_n} \cdot [1 - (1 - e^{-\lambda(T_{tot} - o_n)})] \cdot \frac{n!}{e^{\lambda \cdot T_{tot}} \cdot (\lambda T_{tot})^n} = \frac{n!}{T_{tot}^n} \\
&\overset{139}{=} OrderedUniform(0, T_{tot}]
\end{aligned}$$

**2-nd Order Statistics of Poisson RP** We show the statistics of the Poisson process with parameter $\lambda$ ($X(t) \sim Poisson(\lambda \cdot t)$) up to 2-nd order. As we already seen, the mean and variance:

$$\mathbb{E}[X(t)] = \lambda t, \ \mathrm{Var}(X(t)) = \lambda t$$

$$\Rightarrow \mathbb{E}[X^2(t)] = \mathrm{Var}(X(t)) + \mathbb{E}^2[X(t)] = \lambda t \cdot (1 + \lambda t)$$

We can develop the correlation function (assuming $t_1 \leq t_2$):

$$\begin{aligned}
R_{XX}(t_1, t_2) &\triangleq \mathbb{E}[X(t_1) \cdot X(t_2)] \\
&= \mathbb{E}[X(0, t_1) \cdot (X(0, t_1) + X(t_1, t_2))] \\
&= \mathbb{E}[X^2(0, t_1) + X(0, t_1) \cdot X(t_1, t_2)] \\
&= \mathbb{E}[X^2(0, t_1)] + \mathbb{E}[X(0, t_1) \cdot X(t_1, t_2)] \\
&= \mathbb{E}[X^2(0, t_1)] + \mathbb{E}[X(0, t_1)] \cdot \mathbb{E}[X(t_1, t_2)] \\
&= \lambda t_1 \cdot (1 + \lambda t_1) + \lambda t_1 \cdot \lambda (t_2 - t_1) \\
&= \lambda t_1 + \lambda^2 t_1^2 + \lambda^2 \cdot (t_1 t_2 - t_1^2) \\
&= \lambda t_1 + \lambda^2 t_1^2 + \lambda^2 \cdot t_1 t_2 - \lambda^2 t_1^2 \\
&= \lambda t_1 + \lambda^2 \cdot t_1 t_2
\end{aligned}$$

[140] $x_i \sim exp(\lambda)$

[139] Because we know that $\{u_i\}$ *i.i.d.*, $u_i \sim Unif(0, T_{tot}]$, so: $f_u(x) = \begin{cases} 1/T_{tot}, & x \in (0, T_{tot}] \\ 0, & else \end{cases}$, so for the PDF: $f_{u_1,...,u_n}(x_1, ..., x_n) \overset{i.i.d.}{=} f_{u_1}(x_1) \cdot ... \cdot f_{u_n}(x_n) = \prod_{i=1}^n f_u(x) = \left(\frac{1}{t_{tot}}\right)^n = \frac{1}{T_{tot}^n}$. But there are $n!$ option to choose before order, so: $f_{u_1,...,u_n}^{Ordered}(x_1, ..., x_n) = \frac{n!}{T_{tot}^n}$.

Develop the covariance function (assuming $t_1 \leq t_2$):

$$C_{XX}(t_1,t_2) = R_{XX}(t_1,t_2) - \mathbb{E}[X(t_1)] \cdot \mathbb{E}[X(t_2)] = \lambda t_1 + \lambda^2 t_1 t_2 - \lambda t_1 \cdot \lambda t_2 = \lambda t_1$$

And in general:

$$R_{XX}(t_1,t_2) = \lambda \min(t_1,t_2) + \lambda^2 t_1 t_2, \quad C_{XX}(t_1,t_2) = \lambda \min(t_1,t_2)$$

We identify that we got that both Poisson process and Wiener process have the same covariance (in Wiener process the mean in any point in time is equal to zero, so the correlation function and covariance of the process are identical) but they are totally different processes - Wiener process always moves but stays around zero, while Poisson process moves occasionally (at events times) but always increasing.
This means we can't deduce a lot about the processes from their 2-nd order statistics (or single moment or finite group) only.

**Merging and Division of Poisson Process** We remember the alternative build of Poisson process using Bernoulli sequence:

$$X^T(t) = \sum_{n=1}^{\lfloor \frac{t}{T} \rfloor} B_n = \sum_{n=1}^{\infty} B_N \cdot U(t-nT), B_n \sim Ber(\lambda T), \ \{B_n\} \ i.i.d.$$

$$X(t) = \lim_{T \to 0, p=\lambda T} (X^T(t)), \ X(t) \sim Poisson(\lambda t)$$

**Merging Poisson Processes** Look at the summing (merging) of two Poisson processes with different $\lambda$ constants:
We will show that the output is also a Poisson process with constant $\lambda_1 + \lambda_2$:

$$\begin{aligned} Y^T(t) &= X_1^T(t) + X_2^T(t) \\ &= \sum_{n=1}^{\infty} B_n^1 \cdot U(t-nT) + \sum_{n=1}^{\infty} B_n^2 \cdot U(t-nT) \\ &= \sum_{n=1}^{\infty} (B_n^1 + B_n^2) \cdot U(t-nT) \\ &= \sum_{n=1}^{\infty} B_n \cdot U(t-nT) \end{aligned}$$

$$B_n = B_n^1 + B_n^2 = \begin{cases} 0, & (1-\lambda_1 T) \cdot (1-\lambda_2 T) \\ 1, & \lambda_1 T \cdot (1-\lambda_2 T) + \lambda_2 T \cdot (1-\lambda_1 T) \\ 2, & \lambda_1 T \cdot \lambda_2 T \end{cases}$$

When the time step $T$ is very small $T \ll 1$, Then we can do a first order approximation:

$$\begin{aligned} B_n &= B_n^1 + B_n^2 \\ &= \begin{cases} 0, & (1-\lambda_1 T)\cdot(1-\lambda_2 T) \\ 1, & \lambda_1 T\cdot(1-\lambda_2 T) + \lambda_2 T\cdot(1-\lambda_1 T) \\ 2, & \lambda_1 T\cdot\lambda_2 T \end{cases} \\ &= \begin{cases} 0, & 1-(\lambda_1+\lambda_2)T + O(T^2) \\ 1, & (\lambda_1+\lambda_2)T + O(T^2) \\ 2, & 0 + O(T^2) \end{cases} \\ &\approx \begin{cases} 0, & 1-(\lambda_1+\lambda_2)T \\ 1, & (\lambda_1+\lambda_2)T \end{cases} \end{aligned}$$

And we get the coefficient $B_n$ is in fact a Bernoulli random variable with success rate $p = (\lambda_1+\lambda_2)T$ .

**Division of Poisson Process** Look at the next system:

The system gets a Poisson process with constant $\lambda$ as an input, and decide at which of the two output it will get out using a decider. We define a random variable for making decisions:

$$I_n = \begin{cases} 1, & p \\ 0, & 1-p \end{cases}$$

So we can show the output in the next way (only one of them gets the input with the respective probability):

$$X_1^T(t) = \sum_{n=1}^{\infty} I_n \cdot B_n \cdot U(t-nT),\ X_2^T(t) = \sum_{n=1}^{\infty} I_n \cdot B_n \cdot U(t-nT)$$

And so we can identify:

$$B_n^1 = I_n \cdot B_n = \begin{cases} 1, & p\cdot\lambda T \\ 0, & (1-p)\cdot\lambda T \end{cases},\ B_n^2 = I_n \cdot B_n = \begin{cases} 1, & (1-p)\cdot\lambda T \\ 0, & p\cdot\lambda T \end{cases}$$

And indeed we get that they are both Bernoulli random variables, so:

$$X_1(t) \sim Poisson(p\lambda t),\ X_2(t) \sim Poisson((1-p)\lambda t)$$

**Derivative Process of Poisson RP** As we did for Wiener process, we start by defining the slope process:

$$X_\epsilon(t) = \frac{X(t+\epsilon) - X(t)}{\epsilon} = \frac{X(t, t+\epsilon)}{\epsilon}$$

So the derivative process is $X'(t) = \lim\limits_{\epsilon \to 0} X_\epsilon(t)$
We would like to look at the 2-nd order statistics of the derivative process. Poisson process has the same covariance function as Wiener process (if ww replace $\alpha$ with $\lambda$), so also for Poisson process:

$$C_{X'X'}(\tau) = \lambda\delta(\tau)$$

We would also like to develop the correlation function of the derivative of Poisson process (and for that we have to find the mean of the derivative process).
First we show that the mean of the slope process is independent in $\epsilon$ so it is equal to the mean of the derivative process:

$$\mathbb{E}[X_\epsilon(t)] = \mathbb{E}[\frac{X(t, t+\epsilon)}{\epsilon}] = \frac{1}{\epsilon} \cdot \mathbb{E}[X(t, t+\epsilon)] = \frac{1}{\epsilon} \cdot \lambda(t+\epsilon-t) = \lambda \Rightarrow \mathbb{E}[X'(t)] = \lambda$$

So:

$$R_{X'X'}(\tau) = R_{X'X'}(\tau, 0) = C_{X'X'}(\tau) + \mathbb{E}[X'(t)] \cdot \mathbb{E}[X'(0)] = \lambda \cdot \delta(\tau) + \lambda^2$$

Notice that:

Meaning the derivative process of Poisson process have a flat PSD over all frequencies, meaning the derivative process is a white noise process.
This white noise is really different than the Gaussian white noise, because it doesn't exist all the time, but only on times of Poisson event (unlike Gaussian white noise that always exists). This is also not a "pure" white noise, as its mean is different than zero.
This white noise fits cases of a single particle, like a laser beam of width of a single photon, or a flow of an electron in a cross section with area of the same size scale of the electron itself.

**Poisson Process as a Renewal Process** There is a class of processes named Renewal Processes. Poisson RP is in this class and it is a special case of this class.
We looked at Poisson RP as an infinite assembly of balloons with pops (events) in parallel.

We would like to look at a counting RP but in turns, meaning there is only one balloon, and when it pops it is being replaced with another one.
For the discussion we discuss the process of replacing a light bulb (when a light bulb stops working, it is replaced with a new one). Of course the life of the bulb has a distribution, say $f_Z(z)$. So we can define the process by:

$$\{Z_i\}\ i.i.d.,\ S^i = S^{i-1} + Z_i,\ X(t) = \sum_{i=1}^{\infty} U(t - S^i)$$

Where the process $X(t)$ will represent in every point in time the number of bulbs that have been replaced up to that point.
Renewal processes lead us to the Queueing theory (for example, a single vender that sales to each person in a queue, and the time of each person with the vender has a certain distribution).
The character of the process depends on the sequence $\{Z_i\}$, and in general $f_Z(z)$ does not necessarily distributes exponentially. In a case where $Z_i \sim exp(\lambda)$, then we get a Poisson process $X(t) \sim Poisson(\lambda t)$. In this case we get that the process will have independent increments (the exponential distribution has the property of memorylessness, and this is not necessarily true for any other distribution. In addition we proved that for an exponential distribution of the increment the linear A.R. equation apply, but it doesn’t mean that if the linear A.R. equation apply then it will necessarily distributes exponentially).

**The Bulb Paradox** The bulb paradox shows an interesting issue about the mean of an exponential random variable.
We use the renewal process of replacing a light bulb in sequence. Assume the life of each light bulb distributes exponentially with $\frac{1}{\lambda} = 1\ month$. Hence from the moment it was plugged, its life expectancy (life mean) is one month. WE go inside a room and point at a light bulb that works. We want to know what is its life expectancy (from the moment it was replaced).
We get that the life expectancy (mean) of this bulb is 2 *months* Explanations:

- The exponential distribution is memoryless, so is the process. SO any time backward and forward distributes $exp(\lambda)$. From this we know we have a mean of one month backward or forward, hence it has a life mean of two months.
- This is a Poisson process. We know that events points distribute evenly on the axis (as we have seen in simulation option #3). So when we

get into the room, it is equal to a random point on the axis and it is more "probable" that we fall on a longer interval then on a shorter one (because they have a higher conditional probability because their length is longer):

- We show a RP that shows the life length of each bulb (in every moment, $t$ will show the life time of the current light bulb):

  This process is a squares process. Assume the process is long enough and contains a lot of light bulbs.
  We come at a random moment $T$. We want to know the average life of the light bulbs up to this time:

  $$\bar{Y}(T) = \frac{1}{T} \cdot \int_0^T T(t) \cdot dt$$

  This average is bounded ffrom both sides with the average at the beginning of the square and with the average at the end of the square:

  $$\frac{1}{T} \sum_{i=1}^{X(T)} Z_i^2 \le \bar{Y}(T) \le \frac{1}{T+1} \sum_{i=1}^{X(T)+1} Z_i^2 \underset{T+1>T}{\le} \frac{1}{T} \sum_{i=1}^{X(T)+1} Z_i^2$$

  $$\frac{1}{T} \cdot \frac{X(T)}{X(T)} \sum_{i=1}^{X(T)} Z_i^2 \le \bar{Y}(T) \le \frac{1}{T} \cdot \frac{X(T)}{X(T)} \sum_{i=1}^{X(T)+1} Z_i^2$$

  Look at a part of the expression:

  $$\frac{1}{T} \cdot \frac{X(T)}{X(T)} \sum_{i=1}^{X(T)} Z_i^2 = \left( \frac{1}{X(T)} \cdot \sum_{i=1}^{X(T)} Z_i^2 \right) \cdot \left( \frac{X(T)}{T} \right) \overset{LLN}{=} \mathbb{E}[Z_i^2] \cdot \left( \frac{X(T)}{T} \right) = \mathbb{E}[Z_i^2] \cdot \frac{1}{\mathbb{E}Z_i}$$

  Up to this point this was a general development for any renewal process, we insert $Z_i \sim exp(\lambda)$ and get:

  $$\frac{1}{T} \cdot \frac{X(T)}{X(T)} \sum_{i=1}^{X(T)} Z_i^2 = \mathbb{E}[Z_i^2] \cdot \frac{1}{\mathbb{E}Z_i} = \frac{1/\lambda^2 + 1/\lambda^2}{1/\lambda} = \frac{2}{\lambda} = 2 \cdot \mathbb{E}[Z_i]$$

  We proved that the mean is doubled.