Probability and Random Processes

`eecs501 f21`, `reference and reading list`

> - Probability, Statistics, and Random Processes for Electrical Engineering, Third Edition, Alberto Leon-Garcia
> - Probability and Random Process for Electrical and Computer engineers, John A. Gubner
> - Some random variables
> - MA3K0 - High-Dimensional Probability, Stefan Adams
> - A reference for proof (includes the SLLN proof)
> - Chernoff bounds, and some applications

Set theory and introduction to probability

Sets. A set is a well-defined collection of objects.

- how set is defined? (A set is usually defined in one of the 3 following ways: (a) Statement: LetXbe the set of all natural numbers less than 5. (b) Roster:X={1,2,3,4} (c) Set-builder)
- Operations on sets
  - Universal set Ω ;
  - Complement
  - Union
  - Intersection
  - Infinite unions. $\bigcup_{n=1}^{\infty} A_n$
  - Infinite intersection. $\bigcap_{n=1}^{\infty} A_n$
  - Difference: $A - B = A \cap B^c$
  - Symmetric difference: $A \triangle B = (A - B) \cup (B - A)$.
  - **Power set**. $2^A = pow(A)$
- *Properties of unions and intersections*
  - Commutative: $A \cap B = B \cap A$. $A \cup B = B \cup A$.
  - Associative: $(A \cap B) \cap C = A \cap (B \cap C)$.
  - Distributive laws: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$, $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
  - De Morgan's laws: $(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$
- **Cardinality**: The cardinality of set A is the number of elementsin the set A and isdenoted as |A|.
- **Countable set**: a set A is said to be countable if it has finite number of elements or its magnitude is the same as that of natural numbers $\mathbb{N}$. (Exists one-to-one map from A to $\mathbb{N}$.)

Random Experiment, Outcome, Sample Space.

- Outcomes `vs.` Event(collection of outcomes)($E \subseteq \Omega, E \in 2^{\Omega}$) `vs.` $\mathcal{F}$-field(collection of events)
- **σ-Algebra**: a non-empty collection of events that is closed under **complementation** and **countable union**.
  1. Ω is in $\mathcal{F}$
  2. if A in $\mathcal{F}$, A^c in $\mathcal{F}$
  3. $E_n \in \mathcal{F}, \forall n \in \mathbb{N} \implies \bigcup_{n=1}^{\infty} E_n \in \mathcal{F}$
  - propositions
    1. $\varnothing, \Omega$ are a part of any $\mathcal{F}$.
    2. σ-feld is also closed under **countable intersection**. `proof`
  - examples. #toreview
- **Probability measure**.
  - Definition: A probability measure P on a σ-algebra F is a function $P : \mathcal{F} \to [0, 1]$ which satisfies the following **three axioms** of probability:
    1. Normalization: $P(\Omega) = 1$
    2. Non-negativity
    3. **Countable additivity**. #todo $E_1, E_2, \dots$ is a countable collection of events from $\mathcal{F}$ such that they are **pairwise-disjoint**, → then $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i) = \lim_{n\to\infty} \sum_{i=1}^{N} P(E_i)$
  - properties
    - Measure of empty set: $P(\varnothing) = 0$
    - Law of complements: $P(E^c) = 1 - P(E)$
    - Inclusion-Exclusion Principle: $\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$
    - Union Bound: $P(A \cup B) \le P(A) + P(B)$, $A \subseteq B \implies P(A) \le P(B)$.
    - Total Probability Theorem: $P(A) = \sum_{i=1}^{m} P(A \cap B_i)$ where B_i is an *even space*(forms a partition).
  - **Continuity** of probability measure #tolearn `proof` #tounderstand
    1. $A_1 \subseteq A_2 \subseteq \cdots A_n \subseteq \cdots$ non-decreasing sequence of events. $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{N\to\infty} P(A_N)$ `proof`$A = A_1 \cup (A_1 \setminus A_2) \cup (A_2 \setminus A_3) \cup \cdots$, then $P(A) = \dots$
    2. $A_1 \supseteq A_2 \supseteq \cdots A_n \supseteq \cdots$ non-increasing sequence of events. $P\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{N\to\infty} P(A_N)$

- **probability space**. $(\Omega, \mathcal{F}, P)$. #todo
  - Ω: sample space. P: probability measure defined on $\mathcal{F}$.

`notes` General principle of **inclusion-exclusion** for finite sets:

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{i=1}^{n} |A_i| - \sum_{1 \leqslant i<j \leqslant n} |A_i \cap A_j| + \sum_{1 \leqslant i<j<k \leqslant n} |A_i \cap A_j \cap A_k| - \cdots + (-1)^{n-1} |A_1 \cap \cdots \cap A_n|.$$

Conditional Probability and Independence

Conditional probability & independence

- Conditional probability
  1. **conditional probability**. Given a probability space $(\Omega, \mathcal{F}, P)$ and an event $B \in F$ $(P(B) \neq 0)$, we can define a new probability measure P_B on F as

$$\boxed{P_B(A) := P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \text{ if } P(B) \neq 0}$$

  2. This definition leads to a new probability space $(\Omega, \mathcal{F}, P_B)$. Also can thick of the changing of σ-algebra. The probability measure of each single-outcome event in B increases by factor $1/P(B)$.
  3. Same to probability, those probability properties holds. e.g. $P(A^c|B) = 1 - P(A|B)$. #toreview
  4. Conditional probability also satisfies 3 probability measure axioms. (normalization, non-negativity, countable additivity.)`proof` #toreview
- Independence
  1. **Independence** of events: A and B are independent if $P(A|B) = P(A)$ or $P(B) = 0$. $\rightarrow P(A \cap B) = P(A)P(B)$.
  2. The main point is the **probability measure of A remains**. (B does not provide **infomation** about A)
  - `comments` *Exclusivity and independence.*
- **Conditional Independence**: given (event) C if $P(C) > 0$ and $P(A \cap B|C) = P(A|C)P(B|C)$. $\longleftrightarrow P_C(A \cap B) = P_C(A)P_C(B)$
  - This **is independence under conditional measure**! #tounderstand
  - independence $\nRightarrow$ Conditional independence. Conditional independence $\nRightarrow$ independence. #toreview `examples`
  - Inpendence property: independence of A, B, A^c, B^c.
- Independence of multiple events
  - for events $\{A_i, i \in I\}$, $J \subseteq I$: $P(\bigcap_{i \in J} A_i) = \prod_{i \in J} P(A_i)$.

> **law of total probability**. if $B_1, B_2, \dots B_n, \dots$ is a partition, $P(B_i) > 0$, for any A : $\boxed{P(A) = \sum_j P(B_j)P(A|B_j)}$
>
> $$P(A) = P(A|B)P(B) + P(A|B^C)P(B^C)$$

> **Bayes' Theorem** (**Bayes' rule**)
>
> $$\boxed{P(B|A) = \frac{P(A|B)P(B) = P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^C)P(B^C)}}$$
>
> $$\boxed{P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_j P(A|B_j)}P(B_j)}$$
>
> strictly speaking, B_i is not required to be a partition on the sample space.
>
> 1. $$\boxed{P(A|BC) = \frac{P(C|AB)P(A|B)}{P(C|B)}}$$
> 2. Bayes' theorem can be used to calculate **posterior probabilities**. #tolearn

> - prior probability: $P(A)$
> - posterior probability: $P(A|X)$ "with observation/condition"

- Equally likely outcomes: how to calculate the probability of events? #toreview

Combintorics

$$\boxed{\binom{n}{k} := \frac{n!}{k!(n-k)!} = C_n^k}$$

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n$$

$$\binom{n}{k-1} + \binom{n}{k} = \binom{n+1}{k}$$

$$\sum_{k=m}^{n}\binom{k}{m}=\binom{n+1}{m+1},\quad \sum_{k=m}^{n+1}\binom{k}{m}=\binom{n+2}{m+1}$$

$$\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{N-k}=1$$

multinomial coefficient

$$k_0+k_1+\ldots+k_{m-1}=n,\quad \binom{n}{k_0,k1,\cdots,k_{m-1}}:=\frac{n!}{k_0!k_1!\cdots k_{m-1}!}$$

MAP versus ML(maximum likelihood)

- `Radar problem` #toreview what is the reliability of a system?

MAP Rules(maximum a posteriori probability).
It turns out that no decision rule can have a smaller **probability of error** than the maximum a posteriori probability (MAP) rule. Having observed $Y=j$, the MAP rule says to decide $X=1$ if

$$P(X=1|Y=j)\geq P(X=0|Y=j)$$

$\longrightarrow$ the posterior probability of $X=1$ given the observation $Y=j$ need to be greater than the posterior probability of $X=0$ given the observation $Y=j$.

$$\frac{P(Y=j|X=1)P(X=1)}{P(Y=j)}\geq\frac{P(Y=j|X=0)P(X=0)}{P(Y=j)}$$

Fact: MAP rule is optimal!

ML rule(maximum-likelihood) #tolearn
Sometimes we do not know the prior probabilities $P(X=i)$.

$$P(Y=j|X=1)\geq P(Y=j|X=0)$$

In this context, $P(Y=j|X=i)$ is called the **likelihood** of $Y=j$. The maximum-likelihood rule decides $X=i$ if i maximizes the likelihood of the observation $Y=j$.

Relating the MAP rule and ML rule. From MAP rule's equation, we can get a form:

$$likelihood\ ratio=\frac{P(Y=j|X=1)}{P(Y=j|X=0)}\geq\frac{P(X=0)}{P(X=1)}$$

topic

> **the Monty Hall Problem**

Discrete random variables

Random variable X : a function that assigns a real number to each outcome.
Range: All possible values.
A random variable is called **discrete** if its range is a **countable** set.

discrete random variables

1. Probability Mass Function (PMF): $P_X(x)=P(X=x)$
   - Theorem of PMF. For a discrete r.v. X with PMF and range S
     1. non-negativity
     2. $\sum P_X(x)=1$
     3. $\forall B\subseteq S, P(B)=\sum_{X\in B}P_X(x)$
   - `some R.V examples`: *Bernoulli r.v., Binomial, Geometric.*
2. Cumulative Distribution Function (CDF): $F_X(x)=P(X\leq x)$
   - CDF properties
     1. $F_X(-\infty)=0, F_X(\infty)=1.$
     2. for all $x'\geq x$, $F_X(x')\geq F_X(x)$.
     3. $F_X(b)-F_X(a)=P(a<X\leq b)$.
3. Expectation (or mean) $\boxed{\mathbb{E}[X]=\sum_x xP_X(x)}$
   - `example`: *answer games strategy.*
   - > Let X be a non-negative integer valued random variable.
     >
     > $$\boxed{E[X]=\sum_{n=0}^{\infty}P(X>n),\qquad \frac{1}{2}(E(X^2)-E(X))=\sum_{n=0}^{\infty}nP(X>n)}$$
     >
     > @JY ♥

4. Function of random variables (derived random variable). $Y = g(X)$.
   - **LOTUS**(Law of the unconscious statistician). $E[Y] = E[g(X)] = \sum_i g(x_i)P_X(x_i)$. `proof`
5. Linearity of expectation. $E[aX+b] = aE[X]+b$, $E[aX+bY] = aE[X]+bE[Y]$.

Two(multiple) random variables

1. Joint PMF. $P_{XY}(x,y) = P(\{X=x\}\cap\{Y=y\})$.
   - Marginalization. $P_X(x) = \sum_y P_{XY}(x,y)$, $P_Y(y) = \sum_x P_{XY}(x,y)$.
2. Functions of two random variables. (**LOTUS**) $E[Z] = E[g(X,Y)] = \sum_x \sum_y g(x,y)P(X=x,Y=y)$.
3. Linearity of expectation. $E[aX+bY+c] = aE[X]+bE[Y]+c$.
4. Conditional PMF. (event as condition; r.v. as condition;)

$$P_{X|Y}(x|y) = P(\{X=x\}|Y=y) = \frac{P(\{X=x\}\cap\{Y=y\})}{P(\{Y=y\})} = \frac{P_{XY}(x,y)}{P_Y(y)}.$$

$$P_{XY}(xy) = P_Y(y)P_{X|Y}(x|y).$$

5. Conditional Expectation. (conditioned on event:)$E[X|A] = \sum_x xP_{X|A}(x)$. (R.V. as condition:)$E[X|Y=y] = \sum_x xP_{X|Y}(x|y)$. And $E[X|Y] = E[X|Y](y)$.
6. Independence
   - independence of a random variable from a event. $P_{X|A} = P_X$ #toreview
   - independence of two R.V.s. $P_{XY}(x,y) = P_X(x)P_Y(y),\ \forall x,y \implies P_{X|Y}(x|y) = P_X(x)$.
   - If X and Y are independent, $\rightarrow$ `proof` #todo
     1. $E[XY] = E[X]E[Y]$.
     2. $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$.
     3. $E[X+Y] = E[X]+E[Y]$ (holds with/without independence).
   - Independence of several R.V.s. $P_{XYZ}(x,y,z) = P_X(x)P_Y(y)P_Z(z),\ \forall x,y,z$. (different from definition of indep. of events!) #tounderstand `proof` `reason`
7. **LOTE**(Law of total expectation). let $(\Omega,\mathcal{F},P)$ be a probability space and $B_1,B_2,\ldots,B_n$ be a partition of the Ω. `proof` #todo

$$\boxed{\mathbb{E}[X] = \sum_{i=1}^{n} P(B_i)\mathbb{E}[X|B_i].}$$

8. **Gem** `proof` #todo

> 1. Smoothing. $\boxed{E[E[Y|X]] = E[Y]}$. **the law of iterated expectation**.
> 2. $\mathbb{E}[h(X)|X] = h(X)$.
> 3. Substitution. $E[g(X,Y)|X=x] = E[g(x,Y)|X=x]$.
> 4. $E[g(X)Y|X] = g(X)E[Y|X]$.
> 5. Towering. $E[\ E[X|Y,Z]|Z\ ] = E[X|Z]$.

9. **Variance**. $Var(X) = E[\ (X-E[X])^2\ ] =\ E[X^2]-E^2[X]$.
   - $Var(X) \geq 0$
   - $Var(aX+b) = .. = a^2Var(X)$.
10. **conditional variance**. #toreview
    - $Var(X|A) = E[(X-E[X|A])^2|A] = \sum_x (x-E[X|A])^2 P_{X|A}(x) = E[X^2|A]-E^2[X|A]$ (respect to event)
    - $Var(X|Y=y) = Var(X|\{Y=y\}) = E[(X-E[X|Y=y])^2|Y=y]$
    - $Var(X|Y)(y) = Var(X|Y=y)$
    - $Var(X|Y) = E[X^2|Y]-E^2[X|Y]$
11. **LOTV**(Law of total variance). `proof` #toreview $\boxed{Var(X) = \mathbb{E}[Var(X|Y)]+Var(\mathbb{E}[X|Y])}$

> `proof` #todo

`note`
some important R.V.

Continuous Random Variable

continuous r.v.

💡Assign probability to an interval !! (and a legitimate probability law) `example` continuous probability models.

1. continuous probability space. $(\Omega,\mathcal{F},P)$ where Ω is **uncountable**.
2. **Borel σ-algebra** $\mathcal{B}$. #tolearn #toreview
3. Measure theory. #tolearn
   - **measure**.
   - measurable space.
   - measure space.
   2. measurable random variable

4. (CDF)Cumulative density function. $F_X(x) = P_X([-\infty, x]) = P(X \leq x)$
    - properties: (1) non-decreasing (2) right-continuous $\lim_{x \to x_o^+} F_X(x) = F_X(x_0)$. (3) $F_X(-\infty) = 0$. (4) $F_X(\infty) = 1$. (▶ if F has those properties, -> F can be a CDF.)
    - *CDF for three types of R.V. (discrete; continuous; mixed;)*
5. *continuous and mixed random variables.*
6. (PDF)Probability density function. $f_X(x) := \frac{\partial F_X(x)}{\partial x}$
    - $f_X(x)$ to be a valid pdf:
        1. $f_X(x) \geq 0, \forall x$
        2. $\int_{-\infty}^{\infty} f_X(x) = 1$
7. Expectation & variance. $E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$
    1. **LOTUS**. $E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$
    2. Conditional expectation. $E[X|A] = \int_{-\infty}^{\infty} x f_{X|A}(x) dx$
    3. **LOTE**. $E[X] = \int_{-\infty}^{\infty} E[X|Y = y] f_Y(y) dy$
    4. Variance. $Var(X) = E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x) dx$
    5. Linearity. $Y = aX + b \implies E[Y] = aE[X] + b, Var(Y) = a^2 Var(X)$
8. **transformations(functions) of continuous r.variables**. #todo
    - **Find PDF**. Let X be a random variable and $Y = g(X)$. Given the PDF of X, find the PDF of Y.
        1. special case(a): g is a strictly **increasing** function of X. $\rightarrow F_Y(y) = F_X(h(y))$. $\rightarrow f_Y(y) = \ldots = f_X(h(y)) \frac{dh(y)}{dy}, h = g^{-1}$.
        2. special case(b): g is a strictly **decreasing** function of X. $\rightarrow F_Y(y) = 1 - F_X(h(y))$. $\rightarrow f_Y(y) = \ldots = -f_X(h(y)) \frac{dh(y)}{dy}, h = g^{-1}$.
        3. Special case: g is a strictly **monotonic** function of X. $\rightarrow \boxed{f_Y(y) = f_X(h(y)) \left| \frac{dh(y)}{dy} \right|}, h = g^{-1}$.
        - `examples` ▶ *Y=aX+b* #toreview the result is ... #todo ▶ $Y = X^2, X \sim U[-1, 1]$
    - **find transformation to match the PDF**. #toreview
        - Uniqueness can be guaranteed only if we assume g to be monotone non-decreasing or monotone non-increasing.
        - `example` ▶ $X \sim U[0, 1], Y \sim Exp(\lambda), Y = g(X) = ?$ #toreview

`notes`

> **some important continuous R.V.** and their features. `hided`
> (Expectation, variance, CDF, moment generating function, charateristic function,)
> ▶ *Uniform r.v.; Exponential(**memoryless property** #toreview); Gaussian;*

> **Realting exponential r.v. and geometric r.v.**. $k = \lceil X \rceil$. #toreview

Multiple random variables and their relationships

1. Joint CDF and PDF.

$$F_{XY}(x, y) = P((X, Y) \in [-\infty, x] \times [-\infty, y])$$
$$f_{XY}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x, y)$$

> *`example Question`: #todo #toreview Given CDFs of X and Y, find joint CDF of $U = max(X, Y)$ and $V = min(X, Y)$?*
> ▶ $\{U \leq u\} = \{X \leq u, Y \leq u\}$, $\{V \leq v\} = \{X \leq v, Y \leq v\} \cup \{X \leq v, Y > v\} \cup \{X > v, Y \leq v\} = \{X \leq v\} \cup \{Y \leq v\}$ based on these and further use the set operation to simplify!
>
> $$\begin{aligned} F_{UV} &= P(U \leq u, V \leq v) \\ &= P(\{\} \cap \{\}) \end{aligned}$$

2. Marginal CDF. $F_X(x) = P(X \leq x) = P(X \leq x, -\infty \leq Y \leq \infty) = F_{XY}(x, \infty) := \lim_{y \to \infty} F_{XY}(x, y)$.
3. **Conditional CDF** and **conditional PDF**.

$$F_{Y|X}(y|x) = \lim_{\delta x \to 0} P(Y \leq y | x < X \leq x + \delta x)$$
$$f_{Y|X}(y|x) = \frac{d}{dy} F_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

4. A **total probability theorem**. $A_1, \ldots, A_n$ form a partition of the sample space

$$\boxed{f_X(x) = \sum_{i=1}^{n} P(A_i) f_{X|A_i}(x)}$$

$$\boxed{F_X(x) = \int_{-\infty}^{x} f_X(t) dt = P(X \leq x) = \sum_{i=1}^{n} P(A_i) \int_{-\infty}^{x} f_{X|A_i}(t) dt}$$

5. **Independence**
    - two RV independent: if any one of the following **equivalent** statement holds `proof`

1. $P((X,Y)\in A\times B)=P(X\in A)P(X\in B),\forall A,B$
2. $\iff f_{XY}(x,y)=f_X(x)f_Y(y),\forall x,y$
3. $\iff F_{XY}(x,y)=F_X(x)F_Y(y),\forall x,y$

- If X and Y are independent R.V. $\Longrightarrow$
  1. $E[XY]=E[X]E[Y]$
  2. $U=g(X)$ and $V=h(Y)$ are independent.
  3. $Var(X+Y)=Var(X)+Var(Y)$.

6. **Sum** of two random variables. $Z=X+Y$ `derive: Leibniz rule`

$$F_Z(z)=P(X+Y\le z)=\int_{-\infty}^{\infty}\int_{-\infty}^{z-x}f_{XY}(x,y)dydx$$
$$f_Z(z)=\frac{d}{dz}F_Z(z)=\int_{-\infty}^{\infty}f_{XY}(x,z-x)dx$$

If X and Y independent: $f_Z(z)=\int_{-\infty}^{\infty}f_X(x)f_Y(z-x)dx=(f_X*f_Y)(z)$.

7. **Covariance and correlation**. #toreview
   - Covariance: $\boxed{Cov(X,Y)=E[(X-EX)(Y-EY)]=E[XY]-E[X]E[Y]}$
   - Correlation: #todo
   - $independent\Rightarrow uncorrelated.$
   - $independent\not\Leftarrow uncorrelated.$
8. **Moment generating function (MGF)**.$M(t):\mathbb{R}\to[0,\infty)$: If $M(t)\le\infty$ on some open interval containing the origin,

$\boxed{M(t)=E[e^{tX}]=\int e^{tx}f_X(x)dx=\int e^{tx}dF_X(x)}$

   - ▶ $M'(0)=E[X]$. ▶ $M^{(k)}(0)=E[X^k]$.
   - if $X_1,\dots,X_n$ are independent, $W=X_1+\dots+X_n\implies M_W(t)=\prod_i M_{X_i}(t)$
9. **Characteristic function**. $\phi:\mathbb{R}\to\mathbb{C}$: $\boxed{\phi(t)=E[e^{itX}]=\int e^{itx}f_X(x)dx=E[\cos tX]+iE[\sin tX]}$
   - properties.
     1. $\phi(0)=E[1]=1$.
     2. $|\phi(t)|\le\int|e^{itX}|f_X(x)dx=1.\to$ So $\phi(t)$ exists while $M(t)$ may not.
     3. if X and Y are independent, $\longrightarrow\ \phi_{X+Y}(t)=\phi_X(t)\phi_Y(t)$
     4. $aX+b\longrightarrow e^{itb}\phi_X(at)$
   - `characteristic func. R.V. Examples:` #toreview `Bernoulli, exponential, Gaussian`
10. **joint characteristic function** of X and Y. $\phi_{X,Y}:\mathbb{R}^2\to\mathbb{C}$: $\boxed{\phi_{X,Y}(s,t)=E[e^{isX}e^{itY}]=E[e^{i(sX+tY)}]=\phi_{sX+tY}(1)}$
    - X and Y are independent, **iff** $\phi_{XY}(s,t)=\phi_X(s)\phi_Y(t)$

- `notes` `examples` *MGF and chara. func. for: Bernoulli; Exponential; Gaussian;* #toreview

Gaussian R.V.

Joint Gaussian random variables: the combination $\boxed{\sum_{k\in K}a_kX_k}$ is still Gaussian. $X=\begin{pmatrix}X_1\\ \vdots\\ X_n\end{pmatrix}$ is Gaussian r.vector if $X_1,\dots,X_n$ are jointly Gaussian.

- `comments`:
  1. $X_1,\dots,X_n$ independent Gaussian R.variables $\implies$ jointly Gaussian. `proof(using chara.func.)`
  2. If jointly Gaussian and C_X diagonal($Cov(X_i,X_j)=0,i\ne j$) (uncorrelated) $\implies$ independent. `proof` $\phi_{X_1,X_2,..}(u_1,u_2,..)=\phi_{u_1X_1+u_2X_2+..}(1)$ #tounderstand
- **PDF of jointly Gaussian**: for C **nonsingular**,

$$\boxed{f_x(\mathbf{x})=\frac{1}{(\sqrt{2\pi})^n\sqrt{det(C)}}exp\left[-\frac{1}{2}(x-\mu)'C^{-1}(x-\mu)\right]},\mu=\begin{pmatrix}\mu_1\\ \vdots\\ \mu_n\end{pmatrix}$$

> `proof` $f_x(x)$ is a valid pdf! #todo
> $\Sigma(or\ C)$ is $z^T\Sigma z>0,\forall z\ne 0$. There exists $\Sigma=U\Lambda U^T,U^TU=I.\ det(\Lambda)=det(\Sigma)$.
>
> $$define\ \boxed{Y=U^T(\mathbf{x}-\mu)},d\mathbf{y}=det(U^T)d\mathbf{x}$$
>
> Then, $Cov(Y)=U^TCov(X)U=\Lambda$, Y is uncorelated and Y is independent, $\implies f_Y(y)=\prod_{i=1}^n f_{Y_i}(y_i)=\frac{1}{(\sqrt{2\pi})^n\sqrt{det(\Lambda)}}e^{-\frac{1}{2}y'\Lambda^{-1}y}$
>
> then, $\implies\frac{dF_X}{dx}=\frac{dF_Y(U^T(\mathbf{x}-\mu))}{d\mathbf{x}}=f_Y(U^T(\mathbf{x}-\mu))\frac{dy}{dx}=f_Y(U^T(\mathbf{x}-\mu))=\text{general case.}$
> (Y is just a linear function of X.) #toreview #tounderstand

- **Linear transformation** of jointly Gaussian r.v.: $X\longrightarrow Y=AX$ is still Gaussian R.V. #toreview
- Transferring X into independent Gaussian R.variable: $X\to Y=U^T(X-\mu)$

MMSE & LMMSE & MMAE

Minimum Mean Square Error Estimation (MMSE)

Use Y to estimate X, by minimize $\boxed{E[(X-g(Y))^T(X-g(Y))]} \implies \boxed{g(Y)=E[X|Y]}$

`simple case proof` Assume both X and Y are scalars, then

$$\begin{aligned} E[(X-g(Y))^2] &= E[E[(X-g(Y))^2|Y]] \\ &= \int E[(X-g(Y))^2|Y=y]f_Y(y)dy \\ &= \int\int (x-g(y))^2 f_{X|Y}(x|y)dx f_Y(y)dy \\ &= \int\left(\int x^2 f_{X|Y}(x|y)dx - 2g(y)\int x f_{X|Y}(x|y)dx + g(y)^2\right) f_Y(y)dy \\ &= E[X^2|Y=y] - 2g(y)E[X|Y=y] + g(y)^2 \\ &= (g(y)-E[X|Y=y])^2 + E[X^2|Y=y] - (E[X|Y=y])^2 \\ &\geq E[X^2|Y=y] - (E[X|Y=y])^2 \end{aligned}$$

And with equality by choosing $g(y)=E[X|Y=y] \implies g(Y)=E[X|Y]$.

LMMSE(Linear estimator): $g(Y)=AY+b$.

$$\boxed{\begin{aligned} A &= C_{XY}C_Y^{-1} \\ b &= E[X]-AE[Y] \end{aligned}}$$

- **Scalar case**: $g(Y)$ is LMMSE, **iff** $\boxed{\begin{aligned} &E[g(Y)]=E[X], \\ &Cov(X-g(Y),Y)=0. \end{aligned}}$

$$\boxed{Z=aY+b=E[X]+\frac{Cov(X,Y)}{Var(Y)}(Y-E[Y]).}$$

$$E[|X-\hat{X}|^2]=E[X^2]-E[\hat{X}^2]$$

`proof` Let $Z=aY+b$, $V=cY+d$, then

$$\begin{aligned} E[(X-V)^2] &= E[(X-Z+Z-V)^2] \\ &= E[(X-Z)^2+(Z-V)^2+2(X-Z)(Z-V)] \\ &= E[(X-Z)^2]+E[(Z-V)^2]+2E[(X-Z)(Z-V)] \\ &= E[(X-Z)^2]+E[(Z-V)^2]+2(E[(X-Z)(aY+b-cY-d)]) \\ &= E[(X-Z)^2]+E[(Z-V)^2]+2((b-d)E[(X-Z)]+(a-c)E[(X-Z)Y]) \\ Cov(X-Z) &= E[(X-Z)Y]-E[X-Z]E[Y]=E[(X-Z)Y] \end{aligned}$$

According the condition on Z, $E[(X-Z)(Z-V)]=(a-c)Cov(X-Z,Y)=0$,

$$\implies E[(X-V)^2]=E[(X-Z)^2]+E[(Z-V)^2] \geq E[(X-Z)^2]$$

From the conditions, we have $E[Z]=aE[Y]+b=E[X]$, $Cov(X-Z,Y)=Cov(X-aY-b,Y)=Cov(X,Y)-aVar(Y)=0$

$$\implies Z=aY+b=E[X]+\frac{Cov(X,Y)}{Var(Y)}(Y-E[Y]).$$

- **LMMSE** for **Gaussian R.Variables**. If (X,Y) are **jointly Gaussian**, then

$$\underbrace{E[X|Y]}_{MMSE}=\underbrace{E[X]+\frac{Cov(X,Y)}{Var(Y)}(Y-E[Y])}_{LMMSE}$$

`proof` #todo #toreview Define LMMSE as $Z=E[X]+\frac{Cov(X,Y)}{Var(Y)}(Y-E[Y])$, then we have $E[X-Z]=0$, $E[(X-Z)Y]=..=0$. Then $Cov(X-Z,Y)=E[(X-Z)Y]-E[X-Z]E[Y]=0$, so $(X-Z)$ and Y are uncorrelated, hence independent. $E[X|Y]=E[X-Z+Z|Y]=E[X-Z|Y]+E[Z|Y]=E[Z|Y]=Z$.

  - ▶ When X and Y are vectors and $Var(Y)$ is invertible,

$$E[X|Y]=E[X]+Cov(X,Y)Var(Y)^{-1}(Y-E[Y])$$

- **Linear innovations sequences**.
  - Assume all random variables have **finite 2nd moments**, **zero mean** $E[Y_i]=0$, and $E[Y_iY_j]=0, i\neq j$ (**orthogonal**), then

$$\boxed{\hat{E}[X|Y]=\hat{E}[X|Y_1,\ldots,Y_n]=E[X]+\sum_{i=1}^{n}\hat{E}[X-E[X]|Y_i]}, \{Y_i\}\textbf{linear innovations seq.}$$

  $Z=\sum_i a_iY_i+b$ is **LMMSE, if and only if** $E[Z]=E[X]$ and $E[(X-Z)Y_i]=0, \forall i$.

`proof` #todo #tounderstand define $Z=E[X]+\sum_{i=1}^{n}\hat{E}[X-E[X]|Y_i]$.

$$E[(X-Z)Y_i] = E\left[\left(X - E[X] - \sum_{j=1}^{n} \hat{E}[X-E[X]|Y_j]\right)Y_i\right]$$
$$= \underbrace{E[(X-E[X]-\hat{E}[X-E[X]|Y_i])Y_i]}_{=0 \text{ by property of LMMSE}} - \sum_{j\neq i} E[\hat{E}[X-E[X]|Y_j]Y_i]$$

Note that $\hat{E}[X-E[X]|Y_i] = B_j Y_j$ for some B_j , and $E[B_j Y_j Y_i] = B_j E[Y_j]E[Y_i] = 0$.

- *what if $Y_1, \ldots, Y_n$ are not orthogonal?* $\to$ **Orthogonalizing** to its **linear innovations sequence** $\tilde{Y}_1, \ldots, \tilde{Y}_n$.

$$\boxed{\begin{aligned}&\tilde{Y}_1 = Y_1 - E[Y_1]\\ &\tilde{Y}_i = Yi - E[Y_i] - \sum_{k=1}^{i-1} Cov(Y_i, \tilde{Y}_k)Var^{-1}(\tilde{Y}_k)\tilde{Y}_k, i \geq 2\end{aligned}}$$

(*view covariance as inner product*) #tounderstand

example Consider zero-mean random variables Y1, Y2 and X with correlation matrix $\begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.25 \\ 0 & 0.25 & 1 \end{pmatrix} \to$

$\tilde{Y}_1 = Y_1 - E[Y_1] = Y_1$
$\tilde{Y}_2 = Y2 - E[Y_2] - Cov(Y_2, \tilde{Y}_1)Var^{-1}(\tilde{Y}_1)\tilde{Y}_1 = Y_2 - \frac{1}{2}Y_1$

Transformation of multiple random variables proof $(x, y) = (g(u, v), h(u, v))$ then, with **Jacobian** matrix #toreview review about the integral with parameters!!

$$\iint_R f(x,y)dxdy = \iint_S f(g(u,v), h(u,v))|\det J(u,v)|dudv, \quad J(u,v) = \begin{bmatrix} \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \\ \frac{\partial h}{\partial u} & \frac{\partial h}{\partial v} \end{bmatrix} = \begin{bmatrix} g \\ h \end{bmatrix}[\frac{\partial}{\partial u}\ \frac{\partial}{\partial v}]$$
$$\implies f_{UV}(u,v) = f_{XY}(g(u,v), h(u,v))|detJ(u,v)|.$$

example $Y = AX + b$, A invertible, X has joint density f_X. Find f_Y. #toreview #tounderstand
$X = h(Y) = A^{-1}(Y-b), J(Y) = A^{-1}, \quad f_Y(Y) = X(A^{-1}(Y-b))|det\ J|$

Regression.

Linear regression. Using LMMSE: $E[X|Y] = E[X] + Cov(X,Y)Var(Y)^{-1}(Y - E[Y]) = \beta_0 + \sum_{i=1}^{\#features} \beta_i Y_i$

example #todo *Question: how to calculate the data's variance and covariance?*

Minimum Mean Absolute Error Estimation (MMAE)

$$\min_{\alpha} E[|X-\alpha|] \longrightarrow F_X(\alpha^*) = \frac{1}{2}$$
$$\min_{g(.)} E[|X-g(Y)|] \longrightarrow F_{X|Y}(g^*(Y)|Y) = \frac{1}{2}$$

proof #todo

markup

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} a & -b \\ -c & a \end{bmatrix}$$

Inequalities, Large numbers, & Bounds

- The **Markov Inequality**. If X is a **nonnegative** random variable, and for any ∀a > 0, $\boxed{P(X \geq a) \leq \frac{E[X]}{a}}$
  - proof constructing r.v. $Y = \begin{cases} 0 & if\ X < a \\ a & if\ X \geq a \end{cases}$, then $Y_a \leq X$, $E[X] \geq E[Y_a] = aP(Y_a = a) = aP(X \geq a)$. #tounderstand
- **Chebyshev's inequality**. X is a random variable with mean μ and variance σ^2, $\forall a > 0$, $\boxed{P(|X-\mu| \geq a) \leq \frac{\sigma^2}{a^2}}$
  - proof Define $Y = (X-\mu)^2$. Using the Markov inequality, $P(Y \geq a^2) \leq ..$
- A more general case. $(X > 0)$: $\boxed{P(X \geq a) \leq \frac{E[X^r]}{a^r}, r > 0}$
- **Weak Law of Large numbers (WLLN)**. $X_1, \ldots, X_n$ are **i.i.d.** mean μ, variance σ^2.

$$\text{For any} \epsilon > 0, \quad \boxed{\lim_{n\to\infty} P\left(\left|\frac{X_1 + .. + X_n}{n} - \mu\right| > \epsilon\right) = 0.} \quad \overline{X}_n \xrightarrow{p.} \mu$$

proof #toreview Define $\hat{\mu}_n = \frac{X1+\cdots+Xn}{n}$.

$$var(\hat{\mu}_n) = E[(\hat{\mu}_n - \mu)^2] = E\left[\left(\frac{X_1 + .. + X_n - n\mu}{n}\right)^2\right]$$
$$= \frac{1}{n^2}E[(X_1 + \cdots + X_n - n\mu)^2] = \frac{1}{n^2}Var(X_1 + .. + X_n) = \frac{\sigma^2}{n}$$

According to Chebyshev's inequality, $P\left(|\frac{X1 + \cdots + Xn}{n} - \mu| > \epsilon\right) \leq \frac{var(\hat{\mu}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$ as $n \to \infty$.

- WLLN means **most of the sample paths** have the empirical mean **close to** the actual mean.
- WLLN does not $\hat{\mu}_n \to \mu$ mean every sample path, which requires SLLN.

- **Strong Law of Large Numbers (SLLN)**. #todo `the assumptions!`

$$\boxed{P\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1.} \quad \overline{X}_n \xrightarrow{a.s.} \mu$$

`proof` Assume: $E[X_i^4] < \infty$(**with loss of generality**), $\mu = 0$. #tounderstand Suppose $\lim_{n\to\infty} \hat{\mu}_n \neq 0$, for some sample path $w : X_1(w), X_2(w), ..$, then $\exists \epsilon > 0, for|\hat{\mu}_n(w)| > \epsilon$, for infinitely many n.

$$\text{define } A_n = \{w : |\hat{\mu}_n(w)| > \epsilon\}$$
$$P(A_n) = P(|\hat{\mu}_n(w)| > \epsilon) = P(|\hat{\mu}_n(w)|^4 > \epsilon^4) \leq \frac{E[(\hat{\mu}_n)^4]}{\epsilon^4}$$

Based on our assumptions, we further have

$$\begin{aligned}
&E[(X_1 + .. + X_n)^4] \\
&= nE[X_i^4] + \sum E[X_i^3 X_j] + \sum E[X_i^2 X_j X_k] + \sum E[X_i^2 X_j^2] + \sum E[X_i X_j X_k X_l] \\
&= nE[X_i^4] + 3n(n-1)\sigma^4 \\
&\leq cn^2 \text{ (for some constant c independent of n.)}
\end{aligned}$$

then

$$P(|\hat{\mu}_n(w)| > \epsilon) \leq \frac{c}{\epsilon^4 n^2}$$
$$\sum_n P(|\hat{\mu}_n(w)| > \epsilon) = (\leq) \sum_n \frac{c}{\epsilon^4 n^2} < \infty$$
$$\to P(|\hat{\mu}_n| > \epsilon \text{ infinitely often}) = 0. \forall \epsilon. \implies P(\lim_{n\to\infty} \hat{\mu}_n \neq 0) = 0$$
$$\to \text{therefore, SLLN holds}$$

The proof assumed that $E(X_i^4)$ and $E(X_i^2)$ are finite, it can be shown that the strong law of large numbers holds only under the assumption $E[|X_i|] < \infty$. Of course we are still taking X_i to be independent with common distribution. #tounderstand

- **Borel-Cantelli Lemma**. Let $A_1, A_2, \ldots$ be a sequence of events. Suppose$\sum_{i=1}^{\infty} P(A_i) < \infty$, then $\mathbb{P}(\underbrace{\{w : w \text{ in infinitely many } A_i\}}_{E}) = 0$

  `proof for above`

$$\to w \in E \implies w \in \bigcup_{j=i}^{\infty} A_j, \forall j, \text{ (}w\text{ has to appear in the union)}$$
$$\text{therefore, } P(E) \leq P(\bigcup_{j=i}^{\infty} A_j) \leq \sum_{j=i}^{\infty} P(A_j) \xrightarrow{as\ i\to\infty} 0$$
$$\implies P(E) = 0$$

- **Generalized WLLN** #tounderstand $Z_n = \frac{1}{n^\alpha}\sum_{i=1}^{n} X_i$. $(\alpha \geq 0)$. $\to E[Z_n] = \frac{1}{n^\alpha}n\mu = \frac{\mu}{n^{\alpha-1}}, Var(Z_n) = \frac{1}{n^{2\alpha}}(n\sigma^2) = \frac{\sigma^2}{n^{2\alpha-1}}$.

  Let $\boxed{\alpha > 1/2} \implies 2\alpha - 1 > 0, Var(Z_n) \xrightarrow{n\to\infty} 0$.

  Let $\epsilon > 0$, by Chebyshev's inequality:

$$P(|Z_n - E[Z_n]| > \epsilon) \leq \frac{Var(Z_n)}{\epsilon^2}$$
$$\Rightarrow \lim_{n\to\infty} P(|Z_n - E[Z_n]| > \epsilon) = 0 \quad \forall \epsilon > 0$$

  Think about $\alpha = 1/2$ ... #todo

- **Central limit theorem** (**CLT**). Collection $X_1, X_2, \ldots, X_n$. **i.i.d.** mean μ, variance σ^2.

$$\boxed{Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i - \mu}{\sigma} = \frac{(\sum_{i=1}^{n} X_i) - n\mu}{\sqrt{n\sigma^2}} = \frac{S - E[S]}{\sqrt{Var(S)}}}$$
$$CDF: \lim_{n\to\infty} F_n(z) = G(z) \to \text{CDF of standard Gaussian}$$

- `proof` #todo

> $$\begin{aligned} MGF: M_n(s) = E[e^{sZ_n}] &= E[e^{\frac{s}{\sqrt{n}}\sum_{i=1}^n X_i}] \\ &= E[\prod_{i=1}^n e^{\frac{s}{\sqrt{n}}X_i}] \\ &= \prod_{i=1}^n E[e^{\frac{s}{\sqrt{n}}X_i}] \ _{(independence)} \\ &= \left[M_X(\frac{s}{\sqrt{n}})\right]^n \end{aligned}$$
>
> $M_X(0) = 1$, $M_X'(0) = E[X] = 0$, and $M_X''(0) = E[X^2] = Var(X) = 1$ (zero mean).
>
> $$\begin{aligned} \lim_{n\to\infty} \log M_n(s) &= \lim_{n\to\infty} n \log M_X(\frac{s}{\sqrt{n}}) = \cdots \\ &= \cdots \\ &= \frac{s^2}{2} \implies \lim_{n\to\infty} M_n(s) = e^{s^2/2} \end{aligned}$$

- **Chernoff Bound**. Collection $X_1, X_2, \ldots, X_n$ i.i.d. mean μ. For any $\mathbf{x} > \mu$

$$\boxed{P(\sum_{i=1}^N X_i \geq Nx) \leq e^{-N\sup_{\theta>0}(\theta x - \Lambda(\theta))}}. \quad \Lambda(\theta) = \log E[e^{\theta X_i}] = \ln E[e^{\theta X_i}]$$

  - If $X_1, X_2, \ldots, X_n$ independent **Bernoulli** random variables: $P(\overline{X} \geq (1+\delta)\mu) \leq e^{-\delta^2\mu/3}$. $P(\overline{X} \leq (1-\delta)\mu) \leq e^{-\delta^2\mu/2}$.
  - `proof`

$$\begin{aligned} P(\sum_{i=1}^N X_i \geq Nx) &=^{(\theta>0)} P(\theta\sum_{i=1}^N X_i \geq \theta Nx) = P(e^{\theta\sum_{i=1}^N X_i} \geq e^{\theta Nx}) \ _{\Rightarrow\text{(Markov inequality)}} \\ &\leq \frac{E[e^{\theta\sum_{i=1}^N X_i}]}{e^{\theta Nx}} = \frac{(E[e^{\theta X_i}])^N}{e^{\theta Nx}} = \frac{e^{N\ln(E[e^{\theta X_i}])}}{e^{\theta Nx}} = e^{N\Lambda(\theta) - \theta Nx} = e^{-N(\theta x - \Lambda(\theta))} \end{aligned}$$

`Example for estimation using different laws and bounds.`

Convergence of random variables

> $X_1, \ldots, X_n, \ldots$ be a sequnce of **i.i.d.** $E[X_i] = \mu$, law of large number (LLN): $\frac{X_1 + \ldots + X_n}{n} \xrightarrow{as\ n\to\infty} \mu$.
> ▶ Why is SLLN stronger than WLLN? → to understand different notions of convergence.
> ▶ R.V. $X \geq Y$: e.g. $X_A \geq X_B$ almost surely or with probability 1 if $P(X_A > X_B) = 1$

- *Convergence of real numbers*. $\lim_{n\to\infty} x_n = x$ means that $\forall\epsilon > 0, \exists N_\epsilon$ such that $|x - x_n| \leq \epsilon, \forall n \geq N_\epsilon$.
- **Almost Sure Convergence**. $X_n \to X\ a.s.\ or\ X_n \xrightarrow{a.s.} X$ or with probability 1 $(w.p.1)$

$$\text{if } P(\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)) = 1. \quad \text{Given } \omega, \{X_n(\omega)\} \text{ are real numbers.}$$

- **Mean-Square Convergence**. $X_n \to X\ m.s.\ or\ X_n \xrightarrow{m.s.} X$

$$\text{if } \forall n, E[X_n^2] < \infty, \quad \text{and } \lim_{n\to\infty} E[(X_n - X)^2] = 0.$$

  - `note` so X has finite variance if $X_n \xrightarrow{m.s} X$.
- **Convergence in Probability**. $X_n \to X\ p.\ or\ X_n \xrightarrow{p.} X$

$$\text{if } \lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0, \forall\epsilon > 0$$

- **Convergence in Distribution**. $X_n \to X\ d.\ or\ X_n \xrightarrow{d.} X$

$$\text{if } \lim_{n\to\infty} F_{X_n}(x) = F_X(x), \quad \forall x,\ F_X(x) \text{ is continuous at } x.$$

  - `Example for uncontinuous point` #toreview $F_{X_n}(x) = \begin{cases} 0 \text{ if } x < \frac{1}{n} \\ 1 \text{ if } x \geq \frac{1}{n} \end{cases}$ #tounderstand
- Different Notions of Convergence. $\begin{array}{l} \text{m.s.} \searrow \\ \not\Downarrow\not\Uparrow \quad \text{in probability} \rightarrow \text{in distribution} \\ a.s. \nearrow \end{array}$
- If $X_n \to X$ in any **one** sense, $\implies$ then if it converges in any **other** sense, it must converge to the **same limit**. ("limit is unique")
  - *Example*. #todo
- Suppose X_n is **Gaussian** random variable for each n and $X_n \to X$ in any of the four sense $(a.s., m.s., d., p.)$, then X is a Gaussian random variable.

▶ `Example for illustration!!!` #toreview #todo Relationship between different type of convergence!

> `example` $W_0, W_1, \ldots$ are i.i.d. $\sim \mathcal{N}(0,1)$, $X_n = 0.9X_{n-1} + W_n, \quad n \geq 0$, $X_n \to ?$

$$
\begin{aligned}
&\quad Fact: P(W_n \geq 2) = P(W_n \leq -2) \geq 0.02 \\
&\rightarrow if\ P(|X_n - X|) \geq \epsilon) \rightarrow 0: \\
&\quad P(|X_n - X| \geq \epsilon) + P(|X_{n-1} - X| \geq \epsilon) \\
&\geq P(\ |X_n - X| \geq \epsilon\ \cup\ |X_{n-1} - X| \geq \epsilon\ ) \quad (union\ bound) e.g.\ \epsilon = 1 \\
&\geq P(|X_n - X_{n-1}| \geq 2) \quad (as\ a\ subset) \\
&= P(|0.1X_{n-1} - W_n| \geq 2) \\
&\geq P(X_n - 1 \geq 0 \cap W_n \leq -2) + P(X_{n-1} < 0 \cap W_n \geq 2) \\
&= P(X_n - 1 \geq 0) P(W_n \leq -2) + P(X_{n-1} < 0) P(W_n \geq 2) \\
&\geq 0.02 (P(X_{n-1} \geq 0) + P(X_{n-1} < 0)) \\
&= 0.02 \nrightarrow 0
\end{aligned}
$$

`notes`
The **Skorohod representation**.

Random Process

Intro random process

Random Process: **infinite**(countable/uncountable) collection of random variables.

- types
  1. Discrete-time random process:
  2. Continuous-time random process:
- **Sample path**: Let $\{X_t\}_{t \in I}$ be a random process. For each $\omega \in \Omega$, we get a sequence of a real numbers (discrete-time) $\{X_t(\omega)\}_{t \in I}$ which is called as a realization, a sample path or a sample function of the random process.
- `examples:` #toreview #todo
  1. Discrete-time: (Discrete-valued) Bernoulli(p) random process
  2. Discrete-time: (Continuous-valued) Amplifier
  3. Continuous-time: (Continuous-valued) Random phase-shifting
  4. Continuous-time: (Discrete-valued) Counting process

Markov chains

- **Discrete-Time Markov Chain** (**DTMC**):

$$P(X_k = i_k | X_{k-1} = i_{k-1}, X_{k-2} = i_{k-2}, \dots) = P(X_k = i_k | X_{k-1} = i_{k-1}). \quad i_j \in S$$

  State space: Let $\{X_k\}$ be a discrete-time random process that takes on values in a countable set S called the state space.
- **Time-Homogeneous** Markov chains (MC): if $P(X_k = j | X_{k-1} = i)$ does not depend on k.
  - matrix P with $P_{ij} = P(X_k = j | X_{k-1} = i)$ is called the transition probability matrix.
  - `e.g.` $P = \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix}$
- The probability of a sample path

$$P(X_0 = i_0, X_1 = i_1, \cdots, X_n = i_n) = P(X_0 = i_0) P_{i_0 i_1} P_{i_1 i_2} \cdots P_{i_{n-1} i_n}$$

- stationary distribution $\longrightarrow$ row vector π: $\pi = \pi P$
  - `thinking`: 1. if exist? 2. if unique? 3. limiting behavior? or convergence. `e.g.` $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
- `concepts`
  - Reachable: if exists finite time T, state j reachable from i, $P(X_T = j | X_0 = i) > 0$.
  - Irreducible: a Markov chain is irreducible if j is reachable from i, $\forall i, j$
  - Period: state i is said to have a period k if the MC returns to state i in T steps only if T is a multiple of k. #toreview
  - Aperiodic: a Markov chain is aperiodic if all states have period 1.
- ▶ Theorem. A **finite-state, irreducible** MC has a **unique** stationary distribution π such that $\pi P = \pi$.
  - `think` Does the distribution $p(k)$ converge to π as $k \to \infty$? e.g. $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$
- ▶ Lemma. Every state in an irreducible Markov chain has the same period. Thus, in an irreducible Markov chain, if one state is aperiodic, the Markov chain is aperiodic. #tounderstand
- ▶ Theorem. A **finite-state, aperiodic, irreducible** MC has a unique stationary distribution π such that $\pi P = \pi$. **Furthermore**, $\lim_{k \to \infty} p(k) = \pi$
- If the state space is **infinite**, the existence of a stationary distribution is **not guaranteed**, even if the Markov chain is irreducible.

> `example` $X_k = \begin{cases} X_k - 1 & Pr = 1/3 \\ X_k & Pr = 1/3 \\ X_k + 1 & Pr = 1/3 \end{cases}$ (irreducible and aperiodic)
>
> $$\pi_k = \frac{1}{3}\pi_{k-1} + \frac{1}{3}\pi_k + \frac{1}{3}\pi_{k+1} \quad \forall k$$
> $$\implies 2\pi_k = \pi_{k-1} + \pi_{k+1}, \quad \sum_{k=-\infty}^{\infty} \pi_k = 1, \quad \pi_k \geq 0, \quad \forall k$$
>
> $\longrightarrow$ but we **cannot** find a distribution that satifies above set of equations. Because #todo

$$\begin{aligned}\pi_{k+1} &= 2\pi_k - \pi_{k-1} \\ &\to \pi_k = (k-1)(\pi_1 - \pi_0) + \pi_1, \quad \forall k \geq 2 \\ \text{think if } \pi_1 &= \pi_0 > 0, \ldots \\ \text{if } \pi_1 &> \pi_0, \\ \text{if } \pi_1 &< \pi_0, \\ \text{if } \pi_1 &= \pi_0 = 0,\end{aligned}$$

A little thought shows that the last statement is also true for $k < 0$. Thus, a stationary distribution cannot exist. $\longrightarrow$ the need for more conditions beyond irreducibility to ensure the existence of stationary distributions in **countable-state-space** Markov chains.

- Recurrent or Transient
  1. **recurrence time** T_i of state i, $T_i = \min\{n \geq 1 : X_n = i \text{ } given \text{ } X_0 = i\}$ ("to leave first then go back")
  2. state i is **recurrent** if $P(T_i < \infty) = 1$ ("can return within finite time"). Otherwise, **transient** .
  3. **mean recurrence time** of state i: $M_i = E[T_i]$.
  4. positive recurrent state i: if $M_i < \infty$
  5. positive recurrent Markov chain: if all states are positive recurrent.
  - Suppose $\{X_k\}$ is **irreducible** and that **one** of its states is **positive recurrent**, **then all** of its states are positive recurrent. (The same statement holds if we replace positive recurrent by null recurrent or transient.) #tounderstand
  - If state i of a Markov chain is **aperiodic**, then $\lim_{k\to\infty} p_i(k) = 1/M_i$. (**This is true whether or not** $M_i < \infty$, and even for transient states by defining $M_i = \infty$ when state i is transient.)
- Uniqueness and Convergence
  - Theorem. Consider a **time-homogeneous** Markov chain which is **irreducible** and **aperiodic**. Then, the following results hold. #toreview
    1. if MC is **positive recurrent**, here exists a **unique** π such that $\pi = \pi P$ and $\lim_{k\to\infty} p(k) = \pi$. Further, $\pi_i = 1/M_i$. `"convergence"`
    2. if exists positive vector π that $\pi = \pi P$ and $\sum \pi_i = 1$, it must be the stationary distribution and $\lim_{k\to\infty} p(k) = \pi$. (from above, also means MC is positive recurrent) `"uniqueness"`
    3. if exists positive vector π that $\pi = \pi P$ and $\sum \pi_i = \infty$, then a stationary distribution does not exist, and $\lim_{k\to\infty} p_i(k) = 0$ for all i.

`example` **A simple model of a wireless link**. #toreview For a channel, number of packets served in time slot k is i.i.d. $s(k)$ (Bernoulli, mean μ); at beginning of time slot k, num of packets arrives $a(k)$ (Bernoulli, mean λ); assume $a(k)$ and $s(k)$ indepedent. let $q(k)$ be the number of packets waiting in the queue at the beginning of time slot k.

$$q(k) \to \text{Markov chain:} \quad q(k+1) = (q(k) + a(k) - s(k))^+.$$

`the graph` #todo

$$\begin{aligned}P_{ii} &= \lambda\mu + (1-\lambda)(1-\mu) \\ P_{i,i+1} &= \lambda(1-\mu)\end{aligned}$$

$$\begin{aligned}\pi_i &= \pi_{i-1}P_{i-1,i} + \pi_i P_{ii} + \pi_{i+1}P_{i+1,i} \quad i > 0, \\ \pi_0 &= \pi_0 P_{00} + \pi_1 P_{10}\end{aligned}$$

$$\to \boxed{\pi_i P_{i,i+1} = \pi_{i+1}P_{i+1,i} \quad \forall i} \text{ solves above equations, because: } P_{ii} + P_{i,i-1} + P_{i,i+1} = 1$$

$$\begin{aligned}&\to \pi_{i+1} = \frac{(1-\mu)\lambda}{(1-\lambda)\mu}\pi_i \to \pi_i = \left(\frac{(1-\mu)\lambda}{(1-\lambda)\mu}\right)^i \pi_0 \\ &also: \sum_{i\geq 0} \pi_i = 1 \to \pi_0 \sum_{i=0}^{\infty} \left(\frac{(1-\mu)\lambda}{(1-\lambda)\mu}\right)^i = 1\end{aligned}$$

if assume $\lambda < \mu$, $\to \frac{(1-\mu)\lambda}{(1-\lambda)\mu} < 1$, $\implies \pi_0 = 1 - \frac{(1-\mu)\lambda}{(1-\lambda)\mu} = 1 - \rho$ (ρ: the **workload**) #tounderstand `notice`: $\pi_i = \rho^i(1-\rho)$, $E[q(\infty)] = \frac{\rho}{1-\rho}$.

`example` **PageRank**. Markov chain perspective $\to$ stationary distribution

Random Walks and Gambler's Ruin

$$X_n = X_0 + W_1 + \cdots + W_n, \quad W_i = \begin{cases} 1 & Pr = p \\ -1 & Pr = 1-p \end{cases} (i.i.d)$$

Gambler's ruin problem. start with $X_0 = k$, the random process terminates when $X_n = 0$(ruined) or $X_n = b$(successful). Define S_b to be the event that the gambler is successful without being ruined first, then $P(S-b) = ?$ `graph` #todo

$$\begin{aligned}&\text{define: } s_k = P(S_b | X_0 = k). \\ &s_k = p s_{k+1} + (1-p)s_{k-1}, \quad s_0 = 0, s_b = 1\end{aligned}$$

case 1: $p = 1/2 \implies s_k = k/b$

case 2: $p \neq 1/2 \implies s_k = \dfrac{1 - (\frac{1-p}{p})^k}{1 - (\frac{1-p}{p})^b}$ #todo

and for $p > 1/2$: $\lim_{b\to\infty} s_k = 1 - \left(\dfrac{1-p}{p}\right)^k$ probability of ruin decreases geometrically with initial wealth k.

Kelly's Formula.
e.g. Bet a fixed fraction α. $\to$ what's the best fraction?

$$P(Z_n = 1+\alpha) = 0.6, P(Z_n = 1-\alpha) = 0.4$$
$$W_T = W_0 \prod_{n=1}^{T?} Z_n \longrightarrow \log W_T = \log W_0 + \sum \log Z_n$$
$$\text{LLN: } \frac{\log W_T}{T} \to 0.6\log(1+\alpha) + 0.4\log(1-\alpha) \quad (a.s.)$$
$$\text{then: } \max_{\alpha} 0.6\log(1+\alpha) + 0.4\log(1-\alpha) \to \alpha = 0.2$$

When betting x dollars, the gambler wins with probability p and gets Ax dollars and loses with probability $1-p$ and gets 0 dollars.

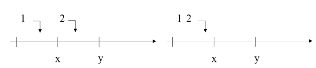
$$\to \max_{\alpha} p\log(1-\alpha+A\alpha) + (1-p)\log(1-\alpha)$$
$$\alpha = \frac{p(A-1)-(1-p)}{A-1} \quad \textbf{Kelly's formula}$$
$$Fraction = \frac{Edge}{Odd}$$

Edge: the fraction of money you win on average when betting a unit amount of money. **Odd**: when you win, the profit you make.

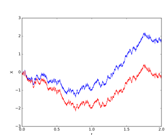
Poisson Process & indep. increment process

- Poisson process is a special type of counting process.
  - A **counting process** $\{N_t\}t \geq 0$ can be expressed in terms of arrival (or occurence) times Y_k . Y_k is the time of the kth arrival. $N_t = \sum_{k=1}^{\infty} \mathbb{I}\{Y_k \leq t\}$
- **Poisson process**: for a counting process $\{N_t\}_{t\geq 0}$ with following conditions:
  1. $N_0 = 0$ with probability 1.
  2. Independent increments. Events in disjoint intervals are independent.
  3. Time homogeneity + Poisson. Number of arrivals ($N(s) - N(t)$) in between $[t, s)$ is Poisson random variable with parameter $\lambda(s-t)$. (λ as the intensity of the process)
  - Poisson random variable: $P(N=k) = \frac{\lambda^k e^{-\lambda}}{k!}$. $E[N] = \lambda$.

- **Theorem**: **Interarrival times** of Poisson process are exponential random variables.
  Let T_i be the time between the ith arrival and the (i−1)th arrival. Then $\{T_i\}_{i\in N}$ are i.i.d. exponential(λ).

> `proof` #toreview For simplicity, consider T_1, T_2. Define $A_1 = T_1$, $A_2 = T_1 + T_2$.
>
> $$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} T_1 \\ T_2 \end{bmatrix}$$
>
> 
>
>
> $$\begin{aligned} F_{A_1A_2}(x,y) &= P(A_1 \leq x, A_2 \leq y) \\ &= P(N_x = 1, \underbrace{N_y - N_x \geq 1}_{\text{at least 1}}) + P(N_x \geq 2) \\ &= P(N_x = 1)P(1 - P(N_y - N_x = 0)) + P(N_x \geq 2) \\ &= e^{-\lambda x}\lambda x(1 - e^{-\lambda(y-x)}) + P(N_x \geq 2) \\ f_{A_1A_2}(x,y) &= \frac{\partial^2}{\partial x \partial y} F_{A_1A_2}(x,y) = \lambda e^{-\lambda x}\lambda e^{-\lambda(y-x)} \\ f_{T_1T_2}(t,s) &= f_{A_1A_2}(x,y)|\det(J)| = f_{A_1A_2}(x,y)|\det(\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix})| = \lambda e^{-\lambda t}\lambda e^{-\lambda s} \end{aligned}$$

- **Independent increment process**.
  - A random process $\{X_t\}$ is called an independent increment process if $X_{t_2} - X_{t_1}, X_{t_3} - X_{t_2}, \cdots, X_{t_n} - X_{t_{n-1}}$ are independent when $t_1 < t_2 < \cdots < t_{n-1} < t_n$.
- **Brownian Motion**. (a "**Gaussian process**")



  - W_t is a Brownian motion if :
    1. $W_0 = 0$ with probability 1.
    2. $W_{t_2} - W_{t_1}$ is a Gaussian random variable with mean $\mu(t_2 - t_1)$ and variance $\sigma^2(t_2 - t_1)$. (with zero-mean. → "**standard**")
    3. Independent increments.
    4. The sample paths are **continuous** with probability one.
  - `relating` CLT.? #tounderstand
  - for Brownian motion W_t with $\mu_t = 0$,

$$\begin{aligned} (\text{for } t > s):\ R_W(t,s) &= E[W_tW_s] = E[(W_t - W_s + W_s)W_s] \\ &= E[(W_t - W_s)W_s + W_s^2] \\ &= E[(W_t - W_s)]E[W_s] + E[W_s^2] \\ &= \sigma^2 s \\ \implies R_W(t,s) &= \sigma^2 \min(s,t) \end{aligned}$$

  - Brownian motion is **not stationary**.

More random process concepts

- Given a random process X_t:

- **mean function**. $\mu_t = E[X_t]$.
- **autocorrelation function**. $R_x(t_1, t_2) = E[X_{t_1} X_{t_2}]$.
- **autocovariance function**. $C_x(t_1, t_2) = R_x(t_1, t_2) - \mu_{t_1}\mu_{t_2}$.
- ▹ in general, **mean** and **autocorrelation** functions are **not sufficient** to define a random process. But they are **sufficient to** describe a **Gaussian process**.

- **Stationary**:
  - X is stationary process if $(X_{t_1}, \dots, X_{t_n})$ has the same joint distribution as $(X_{s+t_1}, \dots, X_{s+t_n})$, $\forall s$.
- **wide-sense stationary** (**WSS**):

$$\text{if} \quad \boxed{\mu_X(t) = \mu_X.} \quad \boxed{R_X(s+\tau, s) = R_X(\tau, 0).} \quad \longrightarrow WSS$$

  - if a process is WSS, we have $R_X(\tau) = E[X_\tau X_0]$
- note ($WSS \not\Rightarrow stationary, stationary \Rightarrow WSS$) For **Gaussian processes**: WSS $\implies$ stationary.

- example $X_t = A\cos(kt + \Theta)$, A, Θ are independent random variables such that $P(A > 0) = 1$, $E[A^2] < \infty$.
  Assume Θ is chosen uniformly from $[0, 2\pi]$, is X_t WSS? is X_t stationary?

$$\begin{aligned} \cos(kt+\Theta) &= \cos(kt)\cos(\Theta) - \sin(kt)\sin(\Theta). \\ \mu_{X_t} &= E[A](E[\cos(\Theta)]\cos(kt) - sin(kt)E[\sin(\Theta)]) = 0. \end{aligned}$$

$$\begin{aligned} R_X(s, s+t) &= E[A^2]E[\cos(ks+\Theta)cos(ks+kt+\Theta)] \\ &= E[A^2](\cos(kt) + E[\cos(k(2s+t)+2\Theta)]) \\ &= E[A^2]\cos(kt) \implies WSS \end{aligned}$$

  is stationary? joint distribution of $(X_t : t \in R)$ and joint distribution of $(X_{t+s}, t \in R)$:

$$\begin{aligned} X_{t+s} &= A\cos(k(t+s)+\Theta) = A\cos(kt+ks+\Theta) \\ &= A\cos(kt+\tilde{\Theta}) \qquad \tilde{\Theta} = (ks+\Theta) \bmod 2\pi \end{aligned}$$

  $\tilde{\Theta}$ also uniform over $[0, 2\pi]$, $\rightarrow$ same joint distribution, $\rightarrow$ stationary.

- example $X_t = A\cos(kt + \Theta)$, A, Θ are independent random variables such that $P(A > 0) = 1$, $E[A^2] < \infty$. #tounderstand
  Assume Θ takes $0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$ with equal probability, is X_t WSS? is X_t stationary?

$$\mu_{X_t} = .. = 0$$

$$P(X_0 = 0) = P(\Theta = \frac{\pi}{2} \text{or}\, \Theta = \frac{3\pi}{2}) = \frac{1}{2}$$

  Note that if kt is not an integer multiple of $\frac{\pi}{2}$, then $kt + \Theta$ cannot be an integer multiple of $\frac{\pi}{2}$. Therefore, #todo

$$P(X_t = 0) = 0 \implies \text{not stationary}$$

- properties of correlation function of a WSS process. proof #todo #toreview
  - $R_X(\tau)$ is symmetric
  - $R_X(\tau)$ is positive semidefinite. i.e. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a(t)R_X(t-s)a(s)dtds \geq 0$. $\sum_{m=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} a[m]R_X[m-n]a[n] \geq 0$. for all function a. #tounderstand
  - $R_X(\tau)$ is bounded: $|R_X(\tau)| \leq R_X(0)$.

  proof $R_X(\tau) = E[X_{t+\tau}X_t] = E[X_t X_{t+\tau}] = R_X(-\tau)$

$$\begin{aligned} \text{define } Y &= \int_{-\infty}^{\infty} a(t)X(t)dt \\ 0 \leq E[Y^2] &= E[\int_{-\infty}^{\infty} a(t)X(t)dt \int_{-\infty}^{\infty} a(s)X(s)ds] \\ &= E\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a(t)X(t)X(s)a(s)dtds\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a(t)E[X(t)X(s)]a(s)dtds \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a(t)R_X(t,s)a(s)dtds \end{aligned}$$

$$|R_X(\tau)| = |E[X_\tau X_0]| \overset{\text{Cauchy-Schwarz}}{\leq} \sqrt{E[X_\tau^2]E[X_0^2]} = \sqrt{R_X(0)R_X(0)} = R_X(0)$$

- **Mean Ergodicity**
  - X_t is WSS and $\mu_X = E[X_t]$, then X_t is mean ergodic if **in an appropriate sense**:

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t X_t dt = \mu_X \quad \text{or} \quad \lim_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K} X_k = \mu_X$$

  - example #toreview
    - $\{X_k\}$ i.i.d. with $E[X_k] = \mu, Var(X_k) < C$. $\rightarrow$ by SLLN is mean ergodic in a.s. sense.
    - $X_1 \sim U[0,1]$. $X_k = X_1 \text{ for } k > 1$. X_t is WSS, but not mean ergodic.
  - **Sufficient** conditions for mean ergodicity in the **m.s.** sense.
    X_t is WSS and X_t is mean ergodic in the m.s. sense if one of the following conditions holds:

$$\begin{array}{ll} 1 & \int_0^\infty |C_X(\tau)|d\tau < \infty \\ 2 & \int_0^\infty R_X(\tau)|d\tau < \infty \\ 3 & \lim_{\tau\to\infty} R_X(\tau) = 0 \\ 4 & \lim_{\tau\to\infty} C_X(\tau) = 0 \\ & (\text{2 or 3 imply } \mu_X = 0) \end{array}$$

> `proof` #tounderstand To prove mean ergodicity in the m.s. sense, we need $\lim_{T\to\infty} E[(\frac{1}{T}\int_0^T X_t dt - \mu_X)^2] = 0$.
>
> $$\begin{aligned} & E\left[(\frac{1}{T}\int_0^T X_t dt - \mu_X)^2\right] = E\left[\frac{1}{T^2}(\int_0^T (X_t-\mu_X)dt)^2\right] \\ =& \frac{1}{T^2}E\left[(\int_0^T (X_t-\mu_X)dt)(\int_0^T (X_s-\mu_X)ds)\right] = \frac{1}{T^2}E\left[\int_0^T\int_0^T (X_t-\mu_X)(X_s-\mu_X)dtds\right] \\ =& \frac{1}{T^2}\int_0^T\int_0^T C_X(t,s)dtds = \frac{1}{T^2}\int_0^T\int_0^T C_X(t-s)dtds\ (WSS) \\ =& \frac{1}{T^2}\int_{s=0}^T\int_{\tau=-s}^{T-s} C_X(\tau)d\tau ds \quad \leftarrow (\tau = t-s) \quad (\#\text{tounderstand}) \\ =& \frac{1}{T^2}\int_0^T\int_{s=0}^{T-\tau} C_X(\tau)dsd\tau + \frac{1}{T^2}\int_{\tau=-T}^0\int_{s=-\tau}^T C_X(\tau)dsd\tau \quad \leftarrow (\text{2 same part}) \\ =& \frac{2}{T^2}\int_0^T (T-\tau)C_X(\tau)d\tau \quad \xrightarrow{T\to\infty} 0 \text{ implies mean erdocicity in m.s. sense} \end{aligned}$$
>
> condition (1): $\frac{2}{T^2}\int_0^T \underbrace{(T-\tau)}_{\le T}C_X(\tau)d\tau \le |\frac{2}{T}\int_0^T C_X(\tau)d\tau| \le \frac{2}{T}\int_0^T |C_X(\tau)|d\tau$
>
> condition (4): For any $\epsilon > 0$, there exists T_ϵ that $|C_X(\tau)| \le \epsilon$ for $\tau > T_\epsilon$.
> $\lim_{T\to\infty}\frac{1}{T}\int_0^T(1-\frac{\tau}{T})C_X(\tau)d\tau \le \lim_{T\to\infty}\frac{1}{T}\int_0^{T_\epsilon}(1-\frac{\tau}{T})C_X(\tau)d\tau + \frac{1}{T}\int_{T_\epsilon}^T(1-\frac{\tau}{T})\epsilon d\tau \le \epsilon$.

- **Ergodic**
  - A stationary random process $(X_n : n \in \mathbb{Z})$ is defined to be ergodic if in any of the three senses (**a.s., m.s., or p.**) (function h which is bounded and Borel measurable on $\mathbb{R}^k$.)

$$\boxed{\lim_{n\to\infty}\sum_{j=1}^n h(X_j,..,X_{j+k-1}) = E[h(X_1,..,X_k)], \quad \forall k, \forall h}$$

  - Importance of Ergodicity
    - If X_n is ergodic, then all of its **finite dimensional distributions** are determined as **time averages**.
  - `e.g.` consider ergodic process X_k and function $h = (X_{k-1}, X_k) = \begin{cases} 1 & X_{k-1} > 0 \ge X_k \\ 0 & otherwise \end{cases}$ #tounderstand compute $P(X_1 > 0 \ge X_2)$ =? #todo
  - `e.g.` Two ergodic random process
    1. $\{X_k\}$ i.i.d.
    2. $\{X_t\}$: stationary Gaussian random process with $\lim_{\tau\to\infty} C_X(\tau) = 0$
- **WSS process through LTI system**
  - **Joint Wide Sense Stationary** (**J-WSS**). if both the following condition holds:
    1. $\{X_t\}$ and $\{Y_t\}$ are both WSS.
    2. cross correlation function $R_{XY}(t_1,t_2) := E[X(t_1)Y(t_2)]$ depends on t_1 and t_2 **only via their difference**.
  - Theorem. Let $\{X_t\}$ be a **WSS** process which is passed a LTI system with impulse response h. The output process $\{Y_t\}$ and $\{X_t\}$ are J-WSS.

> `proof` #tounderstand
>
> $$\begin{aligned} m_Y(t) &= E[Y_t] = E\left[\int_{-\infty}^\infty h(t-\tau)X(\tau)d\tau\right] \\ &= \int_{-\infty}^\infty h(t-\tau)\underbrace{E[X(\tau)]}_{m_X(\tau)=c}d\tau = c\int_{-\infty}^\infty h(\tau)d\tau \quad \to \text{indep. of t} \\ R_{XY}(t_1,t_2) &= E\left[X(t_1)\int_{-\infty}^\infty h(\tau)X(t_2-\tau)d\tau\right] = \int_{-\infty}^\infty h(\tau)E[X(t_1)X(t_2-\tau)]d\tau \\ &= \int_{-\infty}^\infty h(\tau)R_X(t_1-t_2+\tau)d\tau = (\underbrace{\tilde{h}}_{\tilde{h}(x)=h(-x)} * R_X)(t_1-t_2) =: R_{XY}(t_1-t_2) \\ R_Y(t_1,t_2) &= E[Y_{t_1}Y_{t_2}] = E\left[\left(\int_{-\infty}^\infty h(\tau)X(t_1=\tau)d\tau\right)Y(t_2)\right] \\ &= \int_{-\infty}^\infty h(\tau)E[X(t_1-\tau)Y(t_2)]d\tau = \int_{-\infty}^\infty h(\tau)R_{XY}(t_1-\tau,t_2)d\tau \\ &= \int_{-\infty}^\infty h(\tau)R_{XY}(t_1-t_2-\tau)d\tau = (h * R_{XY})(t_1-t_2) =: R_Y(t_1-t_2) \end{aligned}$$

- **Linear time invariant (LTI) systems**. (linear; time-invariant; convolution;)
  - Suppose when input is $e^{j\omega t}$, $e^{j\omega t} \to \boxed{LTI} \to y(t)$
    - $e^{j\omega(t-\tau)} \to \boxed{LTI} \to y(t-\tau)$.
    - $e^{j\omega(t-\tau)} = e^{-j\omega\tau}e^{j\omega t} \to \boxed{LTI} \to e^{-i\omega\tau}y(t)$
    - $\implies y(t)e^{-j\omega t} = y(0), y(t) = y(0)e^{j\omega t}$

- In general, $y(0)$ may depend on ω, $y(t)=H(\omega)e^{j\omega t}$
- Fourier series and Fourier transforms
  - **Fourier series**. $g(t)=\sum_{n=-\infty}^{\infty}c_n e^{j\frac{2\pi}{T}nt}$. $\leftrightarrow c_m=\frac{1}{T}\int_{-T/2}^{T/2}g(t)e^{-j\frac{2\pi mt}{T}}dt$. (frequency: $\frac{2\pi n}{T}^{(Radians/sec)}$, $\frac{n}{T}^{(Hz)}$)
  - **Fourier transform**. $G(\omega)=\int_{-\infty}^{\infty}g(t)e^{-j\omega t}dt.\leftrightarrow g(t)=\int_{-\infty}^{\infty}G(\omega)\frac{e^{j\omega t}}{2\pi}d\omega.$
  - With LTI system:
    - $x(t)=\int\frac{X(\omega)}{2\pi}\boxed{e^{j\omega t}}d\omega\xrightarrow{LTI}y(t)=\int\frac{X(\omega)}{2\pi}\boxed{H(\omega)e^{j\omega t}}d\omega=\int\frac{X(\omega)H(\omega)}{2\pi}e^{j\omega t}d\omega$
    - $Y(\omega)=H(\omega)X(\omega)$. $H(\omega)$: transfer function.
    - Convolution. $y(t)=\int h(t-\tau)x(\tau)d\tau$.
- **Energy** spectral density. $|X(\omega)|^2$.
  - energy of X in the frequency band $[a,b]$ is: $\|y(t)\|^2=\int_{-\infty}^{\infty}|\mathbb{I}_{[a,b]}(\omega)|^2|X(\omega)|^2\frac{d\omega}{2\pi}=\int_a^b|X(\omega)|^2\frac{d\omega}{2\pi}$.
  - The energy of a waveform $x(t)$: $\int_{-\infty}^{\infty}|x(t)|^2dt$
- **Power** in a process #toreview
  - Periodic signals with finite average power: $\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}|x(t)|^2dt<\infty$.
  - Consider **WSS random process** $X=(X_t:t\in\mathbb{R})$

$$\begin{aligned}E[P_X]&=E\left[\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}X(t)^2dt\right]=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}E[X(t)^2]dt\\&=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}R_X(0)dt=R_X(0)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\underbrace{S_X(\omega)}_{\substack{\text{power spectral}\\\text{density}}}d\omega\\R_X(t)&=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{j\omega t}S_X(\omega)d\omega\longrightarrow E[|X_t|^2]=R_X(0)=\frac{1}{2\pi}\int_{-\infty}^{\infty}S_X(\omega)d\omega\end{aligned}$$

  - $E[|X_t|^2]$ is the power of X so $S_X(\omega)$ is the power spectral density.
  - $S_{YX}(\omega)=H(\omega)S_X(\omega)$. $\quad S_Y(\omega)=|H(\omega)|^2S_X(\omega)$.

> `example` Suppose X is WSS and Y is a moving average of X , with averaging window duration T for some $T>0$.
>
> $$\begin{aligned}&y(t)=\frac{1}{T}\int_{t-T}^{t}x(s)ds,\quad h(\tau)=\begin{cases}1/T & 0\le\tau\le T\\0 & else\end{cases}\\&H(\omega)=e^{-\frac{j\omega T}{2}}\operatorname{sinc}(\omega T/2).\quad or\quad H(2\pi f)=e^{-j\pi fT}\operatorname{sinc}(fT).\end{aligned}$$
>
> So the power density is $S_Y(2\pi f)=S_X(2\pi f)|\operatorname{sinc}(ft)|^2$.

Weiner filter: Linear MMSE estimation in random process

- Input signal is characterized by a random process $X(t)$. The signal $X(t)$ goes through a channel that modifies $X(t)$ and adds noise. We observe the noisy output $Y(t)$ of the channel.
  - **Linear estimate**: $\hat{X}(t)=\int h(\tau)Y(t-\tau)d\tau$. ("LTI system")
  - **objective**: square error loss: $\to\min_{h(\cdot)}E\left[(X(t)-\hat{X}(t))^2\right]$.
  - **Assumption**: The signal $X(t)$ and the observation $Y(t)$ are jointly WSS with **known autocorrelation functions** $R_X(\tau)$, $R_Y(\tau)$, respectively, and **cross correlation function** $R_{XY}(\tau)$.
- **Orthogonality theorem** `cf.` LMMSE, for random variables.
  - Linear estimator with impulse response hpt q is optimal if and only if $E[(X(t)-\hat{X}(t))Y(s)]=0$ for every t and s i.e., the estimation error is orthogonal to every sample of the observation.
  - Application of the theorem to obtain the Weiner filter.

$$\begin{aligned}0&=E[(X(t)-\hat{X}(t))Y(s)]=E[X(t)Y(s)]-E[\hat{X}(t)Y(s)]\\&=E[X(t)Y(s)]-\int_{-\infty}^{\infty}\boxed{h(\tau)E[Y(t-\tau)Y(s)]}d\tau\\\Longrightarrow&R_{XY}(t-s)=h\otimes R_Y(t-s)\implies H(\omega)S_Y(\omega)=S_{XY}(\omega)\\\Longrightarrow&\boxed{H(\omega)=\frac{S_{XY}(\omega)}{S_Y(\omega)}}\end{aligned}$$

  This can be interpreted as separate LMMSE estimation of frequency component $X(\omega)$ from the frequency component $Y(\omega)$. #tounderstand

> `proof` of Orthogonality Theorem.
> Suppose that the impulse response $h(\cdot)$ satisfies the orthogonality relation. Consider another arbitrary estimator with impulse response $\tilde{h}(t)$, and let $\tilde{X}(t)$ be the corresponding linear estimate of $X(t)$. Then we have
>
> $$\begin{aligned}&E[(X(t)-\tilde{X}(t))^2]\\=&E\left[\left\{X(t)-\hat{X}(t)+\hat{X}(t)-\tilde{X}\right\}^2\right]\\=&E\left[\left(X(t)-\hat{X}(t)\right)^2\right]+E\left[\left(\hat{X}(t)-\tilde{X}(t)\right)^2\right]+2E\left[(X(t)-\hat{X}(t))\times(\hat{X}(t)-\tilde{X}(t))\right]\end{aligned}$$

$$
\begin{aligned}
&E\left[(X(t)-\hat{X}(t)) \times (\hat{X}(t)-\tilde{X}(t))\right] \\
=&E\left[(X(t)-\hat{X}(t)) \times \left[\int h(\tau)Y(t-\tau)d\tau - \int \tilde{h}(\tau)Y(t-\tau)d\tau\right]\right] \\
=&\int h(\tau)E[(x(t)-\hat{X}(t))Y(t-\tau)]d\tau - \int \tilde{h}(\tau)E[(X(t)-\hat{X}(t))Y(t-\tau)]d\tau \\
=&0
\end{aligned}
$$

We can conclude the proof by observing, $E[(X(t)-\tilde{X}(t))^2] = E[(X(t)-\hat{X}(t))^2] + E[(\hat{X}(t)-\tilde{X}(t))^2] \geq E[(X(t)-\hat{X}(t))^2]$.

- **Weiner filter**.
  - $\boxed{H(\omega) = \frac{S_{XY}(\omega)}{S_Y(\omega)}}$
  - the minimum mean square error. $E[|X(t)-\hat{X}(t)|] = E[X^2(t)] - E[\hat{X}^2(t)] = R_X(0) - R_{\hat{X}}(0)$.

`example` Find the best linear estimate of $X(t)$ given observation $Y(t) = X(t) + N(t)$ Assume $X(t)$ and $N(t)$ are jointly WSS with mean zero. Suppose $X(t)$ and $N(t)$ have known autocorrelation functions and suppose that $R_{XN}(t) = 0$, i.e. X and N are uncorrelated.