Imperial College London

Probability and Stochastic Processes

Lecture 1: Random Variables

Dr. Cong Ling

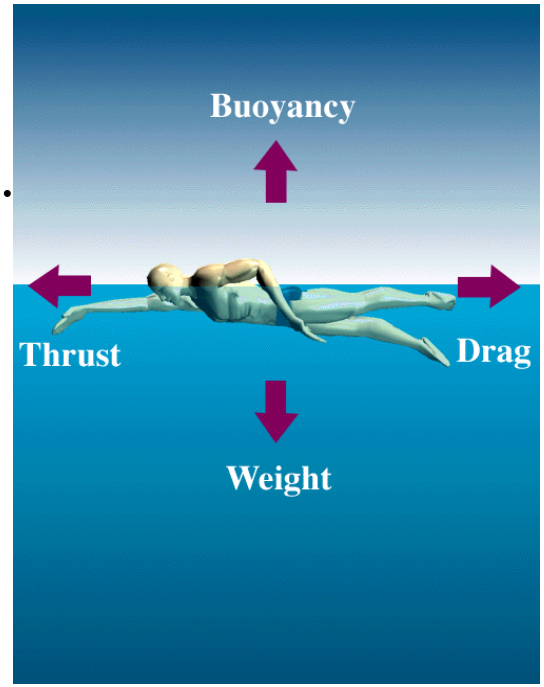
Department of Electrical and Electronic Engineering

Course Information

- Lecturer: Dr. Cong Ling
  - Office: Room 815, EE Building
  - Email: c.ling@imperial.ac.uk
- Teaching Assistant: Mr. Edmund Dable-Heath
  - Office: Room 805, EE Building
  - Email: e.dable-heath18@imperial.ac.uk
- Course notes are available in Blackboard
- Desirable knowledge: Elementary probability
- Grading
  - 3-hour exam
  - Coursework accounts for 15% (deadline: **end of term**)
- Non-assessed problem sheets (for problem classes)

About the Classes

- Lectures: remote asynchronous mode
  - You're expected to watch the recordings in Teams.
  - Essential for the problem classes (synchronous, timetabled/Tuesday, also in Teams).
- Our responsibility is to facilitate you to learn. You have to make the effort.
- Spend time reviewing lecture notes afterwards.
- If you have a question on the lecture material after a class, then
  - Look up a book! Be resourceful.
  - Try to work it out yourself.
  - Ask during classes or by email.



Lectures

Probability

1. Random variables
2. Joint distributions
3. Sequences of random variables
4. Parameter estimation

Stochastic processes

5. Stochastic processes
6. Power spectrum
7. Mean-square estimation
8. Markov chains
9. Continuous-time processes
10. Martingales

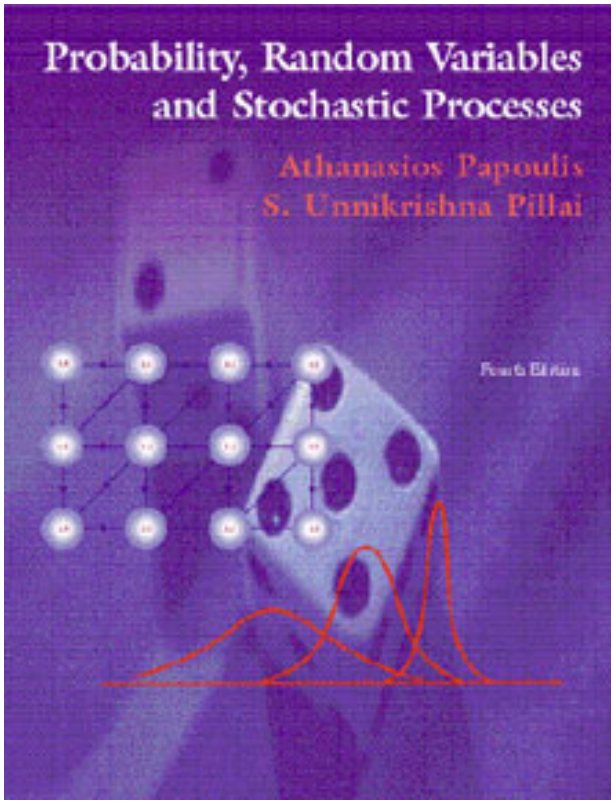
1 lecture = 2 hours

Please watch 1 lecture/week

Why This Course?

- This is a foundational course for many areas of EEE (e.g., automatic control, statistical signal processing, machine learning, network traffic/queuing theory, digital communications, information theory and coding) and beyond (e.g., finance, quantum).
- This is intended to be a medium-level course on probability and random processes, biased towards EEE applications (but not a survey of applications).
- Main objectives:
  - To develop the main ideas of probability theory in a systematic way;
  - To study randomly-varying functions of time, known as stochastic processes or random processes;
  - To demonstrate how to set up probabilistic models for engineering problems.

Textbook



- A. Papoulis and S. Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 4th edition (old editions are ok)
- Book website including slides, problem hints etc.: http://www.mhhe.com/papoulis/
- These lecture notes (except the later lectures) are largely adapted from the slides on the book website
- Other books:
  - Grimmett & Stirzaker, *Probability and Random Processes*, Oxford
  - Stark & Woods, *Probability, Random Processes, and Estimation Theory for Engineers*, Prentice Hall
  - Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, Addison-Wesley

PROBABILITY THEORY

Basics

Probability theory deals with the study of random phenomena, which under repeated experiments yield different outcomes that have certain underlying patterns about them. The notion of an experiment assumes a set of repeatable conditions that allow any number of identical repetitions. When an experiment is performed under these conditions, certain elementary events ξ_i occur in different but *uncertain* ways. We can assign nonnegative number as the probability $P(\xi_i)$, of the event ξ_i in various ways:

Laplace's Classical Definition: The probability of an event A is defined a-priori without actual experimentation as

$$P(A) = \frac{\text{Number of outcomes favorable to } A}{\text{Total number of possible outcomes}},$$

provided all these outcomes are *equally likely*.

Relative Frequency Definition: The probability of an event A is defined as

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

where n_A is the number of occurrences of A and n is the total number of trials.

Axioms of Probability

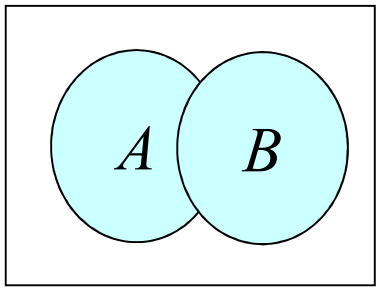
The axiomatic approach to probability, due to Kolmogorov, developed through a set of axioms is generally recognized as superior to the above definitions, as it provides a solid foundation for complicated applications.

For any event A, we assign a number $P(A)$, called the probability of the event A. This number satisfies the following three conditions that act the axioms of probability.

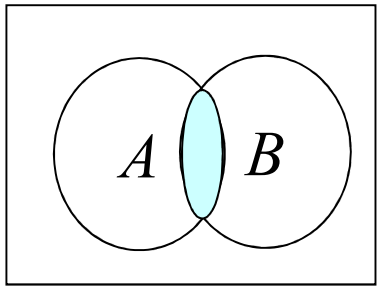
(i) $P(A) \geq 0$ (Probability is a nonnegative number)

(ii) $P(\Omega) = 1$ (Probability of the whole set is unity)

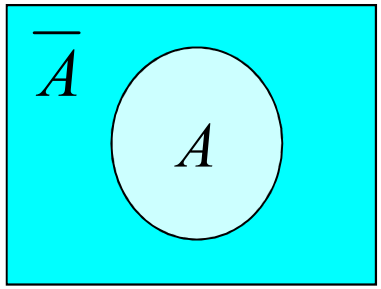
(iii) If $A \cap B = \varphi$, then $P(A \cup B) = P(A) + P(B)$.



$A \cup B$



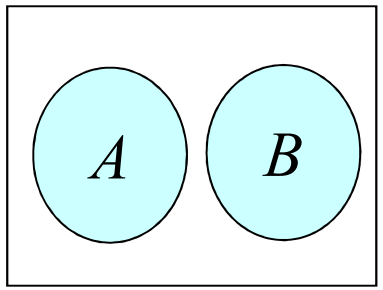
$A \cap B$



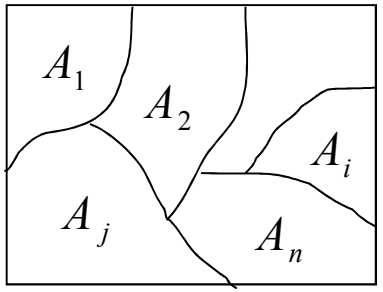
$\overline{A}$

- If $A \cap B = \phi$, the empty set, then A and B are said to be mutually exclusive (M.E).
- A partition of Ω is a collection of mutually exclusive subsets of Ω such that their union is Ω.

$$A_i \cap A_j = \phi, \quad \text{and} \quad \bigcup_{i=1} A_i = \Omega.$$



$A \cap B = \phi$



Suppose A and B are *not* mutually exclusive (M.E.). How does one compute $P(A \cup B) = ?$

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

where AB is a shorthand for $A \cap B$. The probability of more complicated events must follow from this framework by deduction.

Historical perspectives:

Origin of probability theory: Fermat, Pascal, Huygens

Later: Bernoulli, De Moivre, Bayes, Laplace, Gauss, Poisson

Russian school: Chebyshev, Markov, Lyapunov, Kolmogorov

Conditional Probability and Independence

In N independent trials, suppose N_A, N_B, N_{AB} denote the number of times events A, B and AB occur respectively. According to the frequency interpretation of probability, for large N

$$P(A) \approx \frac{N_A}{N}, \quad P(B) \approx \frac{N_B}{N}, \quad P(AB) \approx \frac{N_{AB}}{N}.$$

Among the N_A occurrences of A, only N_{AB} of them are also found among the N_B occurrences of B. Thus the ratio

$$\frac{N_{AB}}{N_B} = \frac{N_{AB}/N}{N_B/N} = \frac{P(AB)}{P(B)}$$

is a measure of "the event A given that B has already occurred". We denote this conditional probability by

$P(A|B)$ = Probability of "the event A given that B has occurred".

We define

$$P(A \mid B) = \frac{P(AB)}{P(B)},$$

provided $P(B) \neq 0.$

With the notion of conditional probability, next we introduce the notion of "independence" of events.

Independence: A and B are said to be independent events, if

$$P(AB) = P(A) \cdot P(B).$$

Notice that the above definition is a probabilistic statement, *not* a set theoretic notion such as mutually exclusiveness.

Suppose A and B are independent, then

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

Thus if A and B are independent, the event that B has occurred does not shed any more light into the event A. It makes no difference to A whether B has occurred or not.

In general (i.e., A and B are not necessarily independent),

$$P(AB) = P(A \mid B)P(B).$$

Similarly,

$$P(AB) = P(B \mid A)P(A).$$

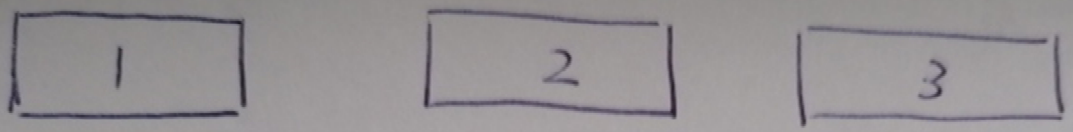
We get

$$P(A \mid B) = \frac{P(B \mid A)}{P(B)} \cdot P(A)$$

This is known as Bayes' theorem.

Example: Three-envelope puzzle.

PUZZLE



A: index of envelope containing film check

B: you bet #1, I open #3 (no check)

$P(A=1) = P(A=2) = P(A=3) = \frac{1}{3}$

$P(B|A=1) = \frac{1}{2}$

$P(B|A=2) = 1$

$P(B|A=3) = 0$

$P(B) = \sum P(B|A=i)\,P(A=i)$

$= \frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3}$

$= \frac{1}{2}$

$P(A=1|B) = \frac{P(B|A=1)}{P(B)} P(A=1) = \frac{\frac{1}{2}}{\frac{1}{2}} \times \frac{1}{3} = \frac{1}{3}$

$P(A=2|B) = \frac{1}{\frac{1}{2}} \times \frac{1}{3} = \frac{2}{3}$

Random Variables

A random variable is a variable whose value is subject to variations due to randomness. As opposed to other mathematical variables, a random variable conceptually does not have a single, fixed value; rather, it can take on a set of possible different values, each with an associated probability.

Let X be a random variable (r.v). All r.v's will be written in capital letters. Denote

$$P\{ X \le x \} = F_X(x) \ge 0.$$

The role of the subscript X is only to identify the actual r.v. $F_X(x)$ is said to the Probability Distribution Function (PDF) associated with the r.v X.

Distribution Function: Note that a distribution function $F_X(x)$ is nondecreasing, right-continuous and satisfies

(i) $F_X(+\infty) = 1, \quad F_X(-\infty) = 0,$

(ii) if $x_1 < x_2$, then $F_X(x_1) \le F_X(x_2),$

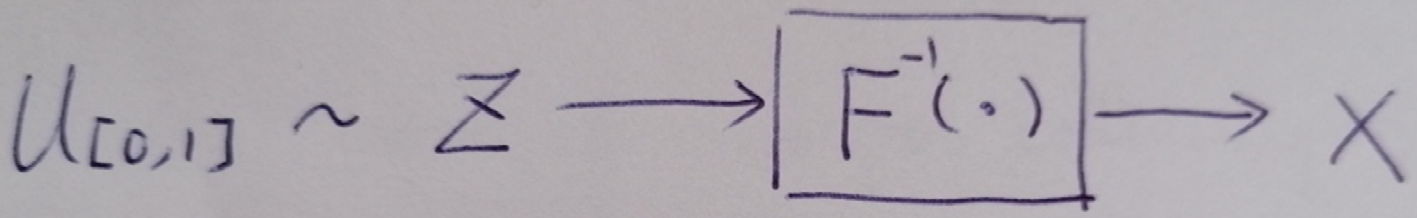
(iv) if $F_X(x_0) = 0$ for some x_0, then $F_X(x) = 0, \quad x \le x_0.$

(v) $P\{ X > x \} = 1 - F_X(x).$

(vi) $P\{ x_1 < X \le x_2 \} = F_X(x_2) - F_X(x_1), \quad x_2 > x_1.$

Application: sample from a given distribution.

SAMPLING



$$F_X(x) = P(X \leq x) = P(F^{-1}(Z) \leq x)$$

$$= P(Z \leq F(x))$$

$$= F(x)$$

• X is said to be a **continuous-type r.v.** if its distribution function $F_X(x)$ is continuous. In that case for all x, and we get

$$P\{X = x\} = 0.$$

• If $F_X(x)$ is constant except for a finite number of jump discontinuities (piece-wise constant; step-type), then X is said to be a **discrete-type r.v.** If x_i is such a discontinuity point, then

$$p_i = P\{X = x_i\}.$$

Probability density function (p.d.f)

The derivative of the distribution function $F_X(x)$ is called the probability density function $f_X(x)$ of the r.v X. Thus

$$f_X(x) \stackrel{\Delta}{=} \frac{dF_X(x)}{dx}.$$

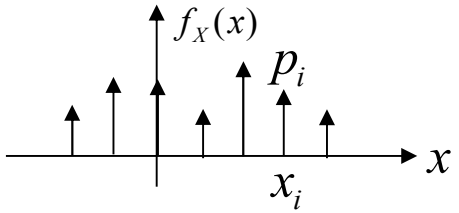
Since

$$\frac{dF_X(x)}{dx} = \lim_{\Delta x \to 0} \frac{F_X(x+\Delta x) - F_X(x)}{\Delta x} \geq 0,$$

from the monotone-nondecreasing nature of $F_X(x)$, it follows that $f_X(x) \geq 0$ for all x. $f_X(x)$ will be a continuous function, if X is a continuous type r.v.

However, if X is a discrete type r.v, then its p.d.f has the general form

$$f_X(x) = \sum_i p_i \delta(x - x_i),$$



where x_i represent the jump-discontinuity points in $F_X(x)$. As Figure shows $f_X(x)$ represents a collection of positive discrete masses, and it is known as the probability mass function (p.m.f) in the discrete case. We also obtain by integration

$$F_X(x) = \int_{-\infty}^{x} f_x(u)du.$$

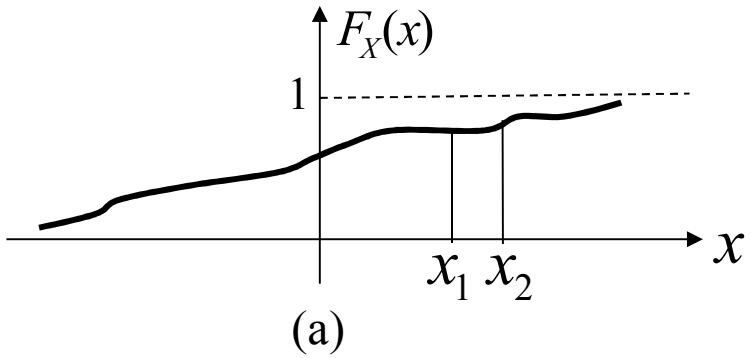
Since $F_X(+\infty) = 1,$ it yields

$$\int_{-\infty}^{+\infty} f_x(x)dx = 1,$$

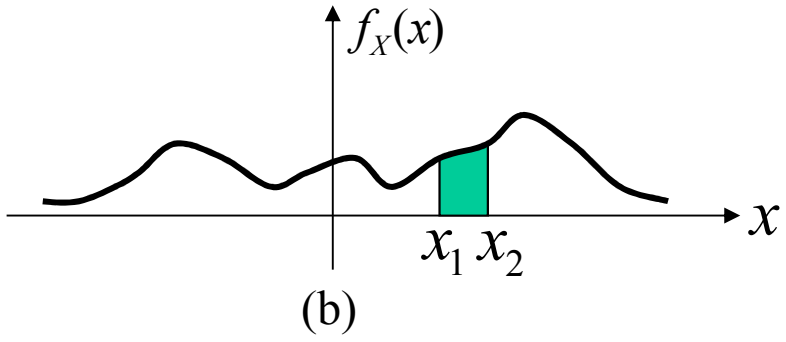
which justifies its name as the density function. Further, we also get

$$P\{x_1 < X \le x_2\} = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x)dx.$$

Thus the area under $f_X(x)$ in the interval (x_1, x_2) represents the probability in that interval.



(a)



(b)

Often, r.vs are referred by their specific density functions - both in the continuous and discrete cases - and in what follows we shall list a number of them in each category.

Continuous-type random variables

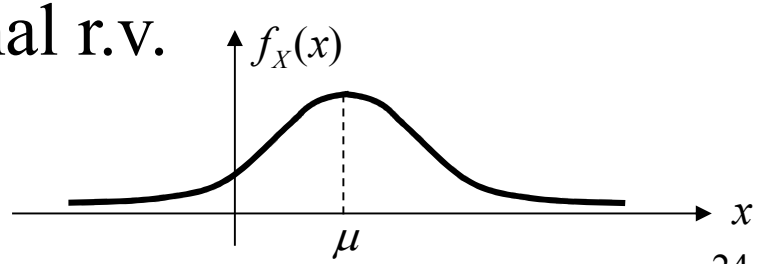
1. Normal (Gaussian): X is said to be normal or Gaussian r.v, if

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}.$$

This is a bell shaped curve, symmetric around the mean μ, and its distribution function is given by

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2} dy \triangleq G\left(\frac{x-\mu}{\sigma}\right),$$

where $G(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$ is often tabulated. Since $f_X(x)$ depends on mean μ and variance σ^2, the notation $X \sim N(\mu, \sigma^2)$ will be used to represent a normal r.v.

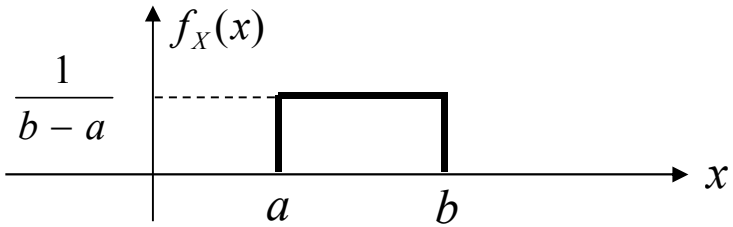


2. Uniform: $X \sim U(a,b)$, $a<b$, if (left figure)

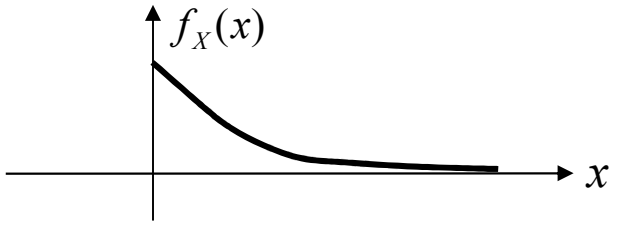
$$f_X(x)=\begin{cases}\dfrac{1}{b-a}, & a\le x\le b,\\ 0, & \text{otherwise.}\end{cases}$$

3. Exponential: $X\sim \varepsilon(\lambda)$ if (right figure)

$$f_X(x)=\begin{cases}\dfrac{1}{\lambda}e^{-x/\lambda}, & x\ge 0,\\ 0, & \text{otherwise.}\end{cases}$$



Uniform



Exponential

Discrete-type random variables

1. Bernoulli: X takes the values $(0,1)$, and

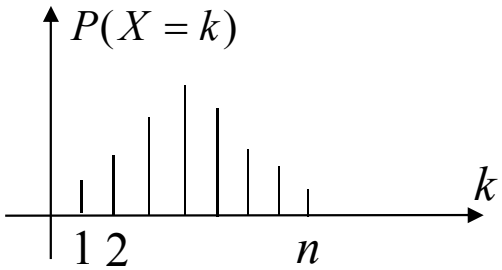
$$P(X=0)=q, \quad P(X=1)=p.$$

2. Binomial: $X \sim B(n,p)$, if

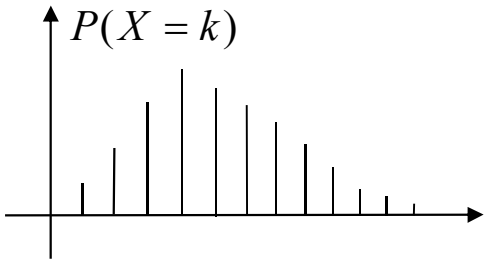
$$P(X=k)=\binom{n}{k}p^k q^{n-k}, \quad k=0,1,2,\cdots,n.$$

3. Poisson: $X \sim P(\lambda)$, if

$$P(X=k)=e^{-\lambda}\frac{\lambda^k}{k!}, \quad k=0,1,2,\cdots,\infty.$$



Binomial



Poission

Functions of a Random Variable

Let X be a r.v and suppose $g(x)$ is a function of the variable x. Define

$$Y = g(X).$$

What is its PDF $F_Y(y)$, pdf $f_Y(y)$?

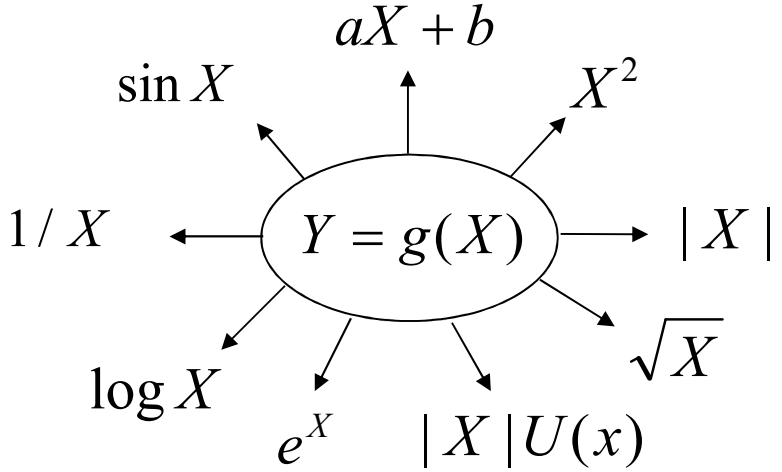
How to compute the mean, variance, and more generally, moments of X?

It will be helpful to use the characteristic function of X: Fourier transform of the pdf.

We start from the fact

$$F_Y(y) = P(Y \le y) = P\big(g(X) \le y\big).$$

Thus the distribution function as well of the density function of Y can be determined in terms of that of X. To obtain the distribution function of Y, we must determine the set on the x-axis such that $X \le g^{-1}(y)$ for every given y, and the probability of that set. We shall consider some of the following functions to illustrate the method (see Text, Chap. 5 for all technical details).



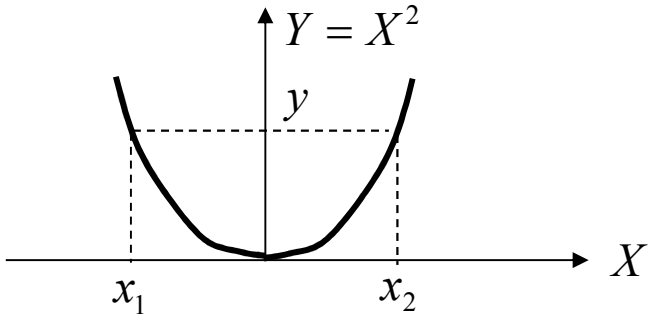
Example 1: $Y = X^2$.

$$F_Y(y) = P(Y \le y) = P(X^2 \le y).$$

If $y < 0$, then the event $\{X^2 \le y\} = \phi,$ and hence

$$F_Y(y) = 0, \quad y < 0.$$

For $y > 0$, from figure, the event $\{Y \le y\} = \{X^2 \le y\}$ is equivalent to $\{x_1 < X \le x_2\}$.



Hence

$$F_Y(y) = P\left(x_1 < X \le x_2\right) = F_X(x_2) - F_X(x_1)$$
$$= F_X(\sqrt{y}) - F_X(-\sqrt{y}), \quad y > 0.$$

By direct differentiation, we get

$$f_Y(y) = \begin{cases} \dfrac{1}{2\sqrt{y}}\left(f_X(\sqrt{y}) + f_X(-\sqrt{y})\right), & y > 0, \\ 0, & \text{otherwise}. \end{cases}$$

If $f_X(x)$ represents an even function, then this reduces to

$$f_Y(y) = \frac{1}{\sqrt{y}} f_X\left(\sqrt{y}\right), \quad y > 0.$$

In particular if $X \sim N(0,1)$, so that

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

and substituting this, we obtain the p.d.f of $Y = X^2$ to be

$$f_Y(y) = \frac{1}{\sqrt{2\pi y}} e^{-y/2} \qquad y > 0.$$

We notice that it represents a Chi-square r.v with $n = 1$.

Thus, if X is a Gaussian r.v with $\mu = 0,$ then $Y = X^2$ represents a Chi-square r.v with one degree of freedom ($n = 1$).

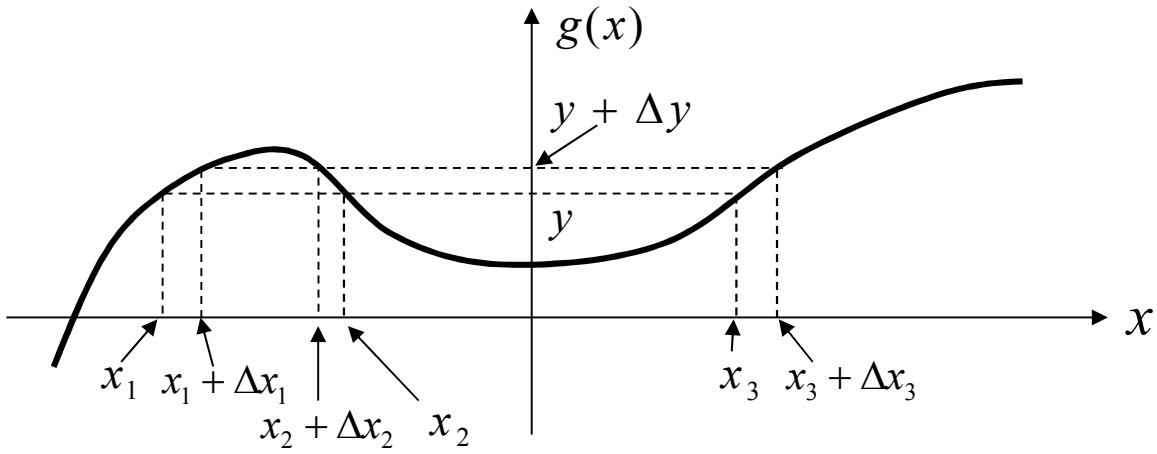
Note: As a general approach, given $Y = g(X)$, first sketch the graph $y = g(x)$, and determine the range space of y. Suppose $a < y < b$ is the range space of $y = g(x)$. Then clearly for $y < a$, $F_Y(y) = 0$, and for $y > b$, $F_Y(y) = 1$, so that $F_Y(y)$ can be nonzero only in $a < y < b$. Next, determine whether there are discontinuities in the range space of y. If so evaluate $P(Y = y_i)$ at these discontinuities. In the continuous region of y, use the basic approach

$$F_Y(y) = P\big(g(X) \le y\big)$$

and determine appropriate events in terms of the r.v X for every y. Finally, we must have $F_Y(y)$ for $-\infty < y < +\infty$, and obtain

$$f_Y(y) = \frac{dF_Y(y)}{dy} \quad \text{in} \quad a < y < b.$$

However, if $Y = g(X)$ is a continuous function, it is easy to establish a direct procedure to obtain $f_Y(y)$. A continuous function $g(x)$ with $g'(x)$ nonzero at all but a finite number of points, has only a finite number of maxima and minima, and it eventually becomes monotonic as $|x| \to \infty$. Consider a specific y on the y-axis, and a positive increment Δy as shown in the following figure



$f_Y(y)$ for $Y = g(X)$, where $g(\cdot)$ is of continuous type.

We can write

$$P\{y < Y \le y + \Delta y\} = \int_{y}^{y+\Delta y} f_Y(u)du \approx f_Y(y) \cdot \Delta y.$$

But the event $\{y < Y \le y + \Delta y\}$ can be expressed in terms of X as well. To see this, referring back to figure, we notice that the equation $y = g(x)$ has three solutions x_1, x_2, x_3 (for the specific y chosen there). As a result when $\{y < Y \le y + \Delta y\}$, the r.v X could be in any one of the three mutually exclusive intervals

$$\{x_1 < X \le x_1 + \Delta x_1\},\ \{x_2 + \Delta x_2 < X \le x_2\} \text{ or } \{x_3 < X \le x_3 + \Delta x_3\}.$$

Hence the probability of the event is the sum of the probability of the above three events, i.e.,

$$P\{y < Y \le y + \Delta y\} = P\{x_1 < X \le x_1 + \Delta x_1\} + P\{x_2 + \Delta x_2 < X \le x_2\} + P\{x_3 < X \le x_3 + \Delta x_3\}.$$

For small $\Delta y, \Delta x_i$, making use of the approximation, we get

$$f_Y(y)\Delta y = f_X(x_1)\Delta x_1 + f_X(x_2)(-\Delta x_2) + f_X(x_3)\Delta x_3.$$

In this case, $\Delta x_1 > 0,\ \Delta x_2 < 0$ and $\Delta x_3 > 0,$ so that it can be rewritten as

$$f_Y(y) = \sum_i f_X(x_i)\frac{|\Delta x_i|}{\Delta y} = \sum_i \frac{1}{|\Delta y / \Delta x_i|} f_X(x_i)$$

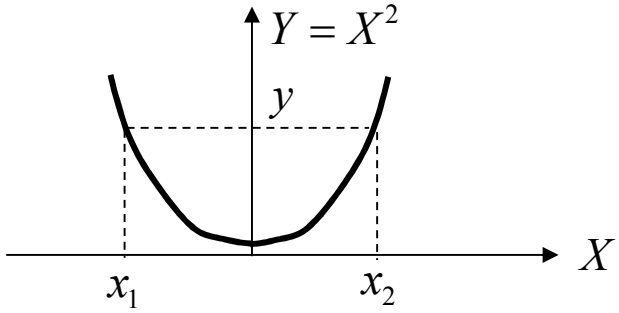
and as $\Delta y \to 0,$ it can be expressed as

$$f_Y(y) = \sum_i \frac{1}{|dy/dx|_{x_i}} f_X(x_i) = \sum_i \frac{1}{|g'(x_i)|} f_X(x_i).$$

The summation index *i* depends on *y*, and for every *y* the equation $y = g(x_i)$ must be solved to obtain the total number of solutions at every *y*, and the actual solutions $x_1, x_2, \cdots$ all in terms of *y*.

For example, if $Y = X^2$, then for all $y > 0$, $x_1 = -\sqrt{y}$ and $x_2 = +\sqrt{y}$ represent the two solutions for each y. Notice that the solutions x_i are all in terms of y so that the right side is only a function of y. Referring back to the example $Y = X^2$ (Example 1) here for each $y > 0$, there are two solutions given by $x_1 = -\sqrt{y}$ and $x_2 = +\sqrt{y}$. ($f_Y(y) = 0$ for $y < 0$). Moreover

$$\frac{dy}{dx} = 2x \quad \text{so that} \quad \left|\frac{dy}{dx}\right|_{x=x_i} = 2\sqrt{y}$$



and we get

$$f_Y(y) = \begin{cases} \dfrac{1}{2\sqrt{y}}\left(f_X(\sqrt{y}) + f_X(-\sqrt{y})\right), & y > 0, \\ 0, & \text{otherwise}, \end{cases}$$

which agrees with Example 1.

Mean, Variance, Moments and Characteristic Functions

Quite often it is desirable to characterize the r.v in terms of its average behavior by some parameters. In this context, we will introduce two parameters - mean and variance - that are universally used to represent the overall properties of the r.v and its p.d.f.

More generally, we will introduce moments, which are extension of mean and variance.

The characteristic function encodes the information about all the moments.

Mean or the **Expected Value** of a r.v X is defined as

$$\eta_X = \overline{X} = E(X) = \int_{-\infty}^{+\infty} x\, f_X(x)dx.$$

If X is a discrete-type r.v, then we get

$$\eta_X = \overline{X} = E(X) = \int x \sum_i p_i \delta(x - x_i)dx = \sum_i x_i p_i \underbrace{\int \delta(x - x_i)dx}_{1}$$

$$= \sum_i x_i p_i = \sum_i x_i\, P(X = x_i).$$

Mean represents the average (mean) value of the r.v in a very large number of trials. For example if $X \sim U(a,b)$, then ,

$$E(X) = \int_a^b \frac{x}{b-a}dx = \frac{1}{b-a}\frac{x^2}{2}\bigg|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}$$

is the midpoint of the interval (a,b).

On the other hand if X is exponential with parameter λ , then

$$E(X) = \int_0^\infty \frac{x}{\lambda} e^{-x/\lambda} dx = \lambda \int_0^\infty y e^{-y} dy = \lambda,$$

implying that the parameter λ represents the mean value of the exponential r.v.

Similarly if X is Poisson with parameter λ , we get

$$E(X) = \sum_{k=0}^{\infty} kP(X = k) = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!}$$

$$= e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} = \lambda e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

Thus the parameter λ also represents the mean of the Poisson r.v.

Given $X \sim f_X(x)$, suppose $Y = g(X)$ defines a new r.v with p.d.f $f_Y(y)$. Then from the previous discussion, the new r.v Y has a mean given by

$$\mu_Y = E(Y) = \int_{-\infty}^{+\infty} y\, f_Y(y) dy.$$

By change of variables, we get the useful formula

$$E(Y) = E\big(g(X)\big) = \int_{-\infty}^{+\infty} y\, f_Y(y) dy = \int_{-\infty}^{+\infty} g(x) f_X(x) dx.$$

In the discrete case, it reduces to

$$E(Y) = \sum_i g(x_i) P(X = x_i).$$

We can use it to determine the mean of $Y = X^2$, where X is a Poisson r.v.

$$E\left(X^2\right)=\sum_{k=0}^{\infty}k^2P(X=k)=\sum_{k=0}^{\infty}k^2e^{-\lambda}\frac{\lambda^k}{k!}=e^{-\lambda}\sum_{k=1}^{\infty}k^2\frac{\lambda^k}{k!}$$

$$=e^{-\lambda}\sum_{k=1}^{\infty}k\frac{\lambda^k}{(k-1)!}=e^{-\lambda}\sum_{i=0}^{\infty}(i+1)\frac{\lambda^{i+1}}{i!}$$

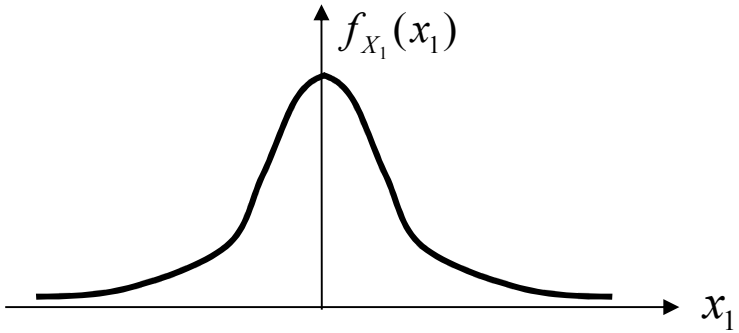
$$=\lambda e^{-\lambda}\left(\sum_{i=0}^{\infty}i\frac{\lambda^i}{i!}+\sum_{i=0}^{\infty}\frac{\lambda^i}{i!}\right)=\lambda e^{-\lambda}\left(\sum_{i=1}^{\infty}i\frac{\lambda^i}{i!}+e^{\lambda}\right)$$

$$=\lambda e^{-\lambda}\left(\sum_{i=1}^{\infty}\frac{\lambda^i}{(i-1)!}+e^{\lambda}\right)=\lambda e^{-\lambda}\left(\sum_{m=0}^{\infty}\frac{\lambda^{m+1}}{m!}+e^{\lambda}\right)$$

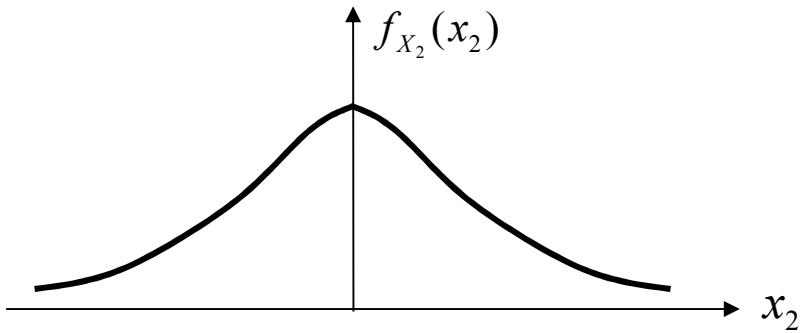
$$=\lambda e^{-\lambda}\left(\lambda e^{\lambda}+e^{\lambda}\right)=\lambda^2+\lambda.$$

In general, $E\left(X^k\right)$ is known as the kth moment of r.v X. Thus if $X\sim P(\lambda)$, its second moment is given by $\lambda^2+\lambda$.

Mean alone will not be able to truly represent the p.d.f of any r.v. To illustrate this, consider the following scenario: Consider two Gaussian r.vs $X_1 \sim N(0,1)$ and $X_2 \sim N(0,10)$. Both of them have the same mean $\mu = 0$. However, as the figure shows, their p.d.fs are quite different. One is more concentrated around the mean, whereas the other one (X_2) has a wider spread. Clearly, we need at least an additional parameter to measure this spread around the mean!



(a) $\sigma^2 = 1$



(b) $\sigma^2 = 10$

For a r.v X with mean μ, $X-\mu$ represents the deviation of the r.v from its mean. Since this deviation can be either positive or negative, consider the quantity $(X-\mu)^2$, and its average value $E[(X-\mu)^2]$ represents the average mean square deviation of X around its mean. Define

$$\sigma_X^2 \stackrel{\Delta}{=} E[(X-\mu)^2] > 0.$$

With $g(X)=(X-\mu)^2$ we get

$$\sigma_X^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f_X(x)dx > 0.$$

σ_X^2 is known as the variance of the r.v X, and its square root $\sigma_X = \sqrt{E(X-\mu)^2}$ is known as the standard deviation of X. Note that the standard deviation represents the root mean square spread of the r.v X around its mean μ.

Using the linearity of the integrals, we get

$$Var(X) = \sigma_X^2 = \int_{-\infty}^{+\infty} \left(x^2 - 2x\mu + \mu^2\right) f_X(x)dx$$

$$= \int_{-\infty}^{+\infty} x^2 f_X(x)dx - 2\mu \int_{-\infty}^{+\infty} x\, f_X(x)dx + \mu^2$$

$$= E\left(X^2\right) - \mu^2 = E\left(X^2\right) - \left[E(X)\right]^2 = \overline{X^2} - \overline{X}^2.$$

Alternatively, we can use this formula to compute σ_X^2.

Thus , for example, returning back to the Poisson r.v, we get

$$\sigma_X^2 = \overline{X^2} - \overline{X}^2 = \left(\lambda^2 + \lambda\right) - \lambda^2 = \lambda.$$

Thus for a Poisson r.v, mean and variance are both equal to its parameter λ.

Moments: As remarked earlier, in general

$$m_n = E(X^n), \quad n \geq 1$$

are known as the moments of the r.v X, and

$$\mu_n = E[(X - \mu)^n]$$

are known as the **central moments** of X. Clearly, the mean $\mu = m_1$, and the variance $\sigma^2 = \mu_2$.

In general, the quantities

$$E[(X - a)^n]$$

are known as the generalized moments of X about a, and

$$E[|X|^n]$$

are known as the absolute moments of X.

For example, if $X \sim N(0, \sigma^2)$, then it can be shown that

$$E(X^n) = \begin{cases} 0, & n \text{ odd}, \\ 1 \cdot 3 \cdots (n-1)\sigma^n, & n \text{ even}. \end{cases}$$

It is often a tedious procedure to directly compute the moments, and in this context, the notion of the characteristic function can be quite helpful.

Characteristic Function

The characteristic function of a r.v X is defined as*

$$\Phi_X(\omega) = E\left(e^{jX\omega}\right) = \int_{-\infty}^{+\infty} e^{jx\omega} f_X(x) dx.$$

Thus $\Phi_X(0) = 1$, and $\left|\Phi_X(\omega)\right| \le 1$ for all ω.

*Note that this is slightly different from the ordinary Fourier transform $F(\omega) = \int_{-\infty}^{\infty} f(x) e^{-j\omega x} dx$. Cf. the moment generating function $\Phi_X(s) = E\left(e^{Xs}\right)$.

Characteristic function of the Gaussian r.v: if $X \sim N(\mu, \sigma^2)$, then

$$\Phi_X(\omega) = \int_{-\infty}^{+\infty} e^{j\omega x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} dx \quad (\text{Let } x - \mu = y)$$

$$= e^{j\mu\omega} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} e^{j\omega y} e^{-y^2/2\sigma^2} dy = e^{j\mu\omega} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} e^{-y/2\sigma^2 (y - j2\sigma^2\omega)} dy$$

(Let $y - j\sigma^2\omega = u$ so that $y = u + j\sigma^2\omega$)

$$= e^{j\mu\omega} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} e^{-(u + j\sigma^2\omega)(u - j\sigma^2\omega)/2\sigma^2} du$$

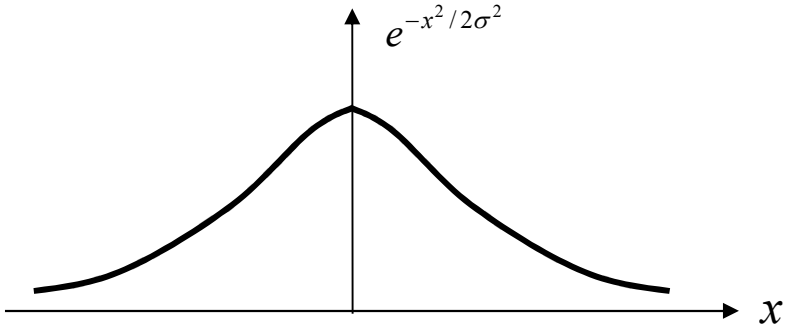
$$= e^{j\mu\omega} e^{-\sigma^2\omega^2/2} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} e^{-u^2/2\sigma^2} du = e^{(j\mu\omega - \sigma^2\omega^2/2)}.$$

Thus if $X \sim N(0, \sigma^2)$, then

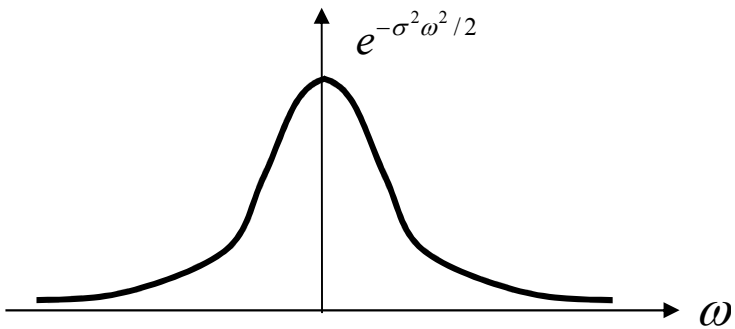
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2},$$

and

$$\Phi_X(\omega) = e^{-\sigma^2\omega^2/2}.$$



(a)



(b)

Notice that the characteristic function of a Gaussian r.v itself has the "Gaussian" bell shape. From the figure, the reverse roles of σ^2 in $f_X(x)$ and $\Phi_X(\omega)$ are noteworthy $(\sigma^2$ vs $\frac{1}{\sigma^2})$.

To illustrate the usefulness of the characteristic function of a r.v in computing its moments, first it is necessary to derive the relationship between them. Towards this, from definition

$$\Phi_X(\omega) = E\left(e^{jX\omega}\right) = E\left[\sum_{k=0}^{\infty} \frac{(j\omega X)^k}{k!}\right] = \sum_{k=0}^{\infty} j^k \frac{E(X^k)}{k!}\omega^k$$

$$= 1 + jE(X)\omega + j^2 \frac{E(X^2)}{2!}\omega^2 + \cdots + j^k \frac{E(X^k)}{k!}\omega^k + \cdots.$$

Taking the first derivative with respect to ω, and letting it to be equal to zero, we get

$$\left.\frac{\partial \Phi_X(\omega)}{\partial \omega}\right|_{\omega=0} = jE(X) \quad \text{or} \quad E(X) = \frac{1}{j}\left.\frac{\partial \Phi_X(\omega)}{\partial \omega}\right|_{\omega=0}.$$

Similarly, the second derivative gives

$$E(X^2) = \frac{1}{j^2}\left.\frac{\partial^2 \Phi_X(\omega)}{\partial \omega^2}\right|_{\omega=0},$$

and repeating this procedure k times, we obtain the kth moment of X to be

$$E(X^k) = \frac{1}{j^k} \frac{\partial^k \Phi_X(\omega)}{\partial \omega^k}\bigg|_{\omega=0}, \quad k \geq 1.$$

We can use these formulas to compute the mean, variance and other higher order moments of any r.v X.

Example: Moments of a Gaussian r.v

$$\phi_x(w) = e^{-\sigma^2 w^2/2}$$

$$E(x^k) = \frac{1}{j^k} \frac{\partial^k \phi_x(w)}{\partial w^k}\bigg|_{w=0}$$

$$\frac{\partial \phi_x(w)}{\partial w} = -\sigma^2 w \cdot e^{-\sigma^2 w^2/2}$$

$$E(x) = 0$$

$$\frac{\partial^2 \phi_x(w)}{\partial w} = -\sigma^2 e^{-\sigma^2 w^2/2} + \sigma^4 w^2 e^{-\sigma^2 w^2/2}$$

$$E(x^2) = \sigma^2$$

$$\frac{\partial^3 \phi_x(w)}{\partial w} = \sigma^4 w e^{-\sigma^2 w^2/2} + 2\sigma^4 w e^{-\sigma^2 w^2/2} - \sigma^6 w^3 e^{-\sigma^2 w^2/2}$$

$$E(x^3) = 0$$

$$\frac{\partial^4 \phi_x(w)}{\partial w} = \sigma^4 e^{-\sigma^2 w^2/2} + 2\sigma^4 e^{-\sigma^2 w^2/2} + \cdots$$

$$E(x^4) = 3\sigma^4$$

Imperial College
London

EE4-10: Probability and Stochastic Processes

Lecture 2: Joint Distributions

Two Random Variables

In many experiments, the observations are expressible not as a single quantity, but as a family of quantities. For example to record the height and weight of each person in a community or the number of people and the total income in a family, we need two numbers.

Let X and Y denote two random variables (r.v). Then

$$P\left(x_1 < X(\xi) \le x_2\right) = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x)dx,$$

and

$$P\left(y_1 < Y(\xi) \le y_2\right) = F_Y(y_2) - F_Y(y_1) = \int_{y_1}^{y_2} f_Y(y)dy.$$

What about the probability that the pair of r.vs (X,Y) belongs to an arbitrary region D? In other words, how does one estimate, for example, $P\left[(x_1 < X \le x_2) \cap (y_1 < Y \le y_2)\right] = ?$ Towards this, we define the joint probability distribution function of X and Y to be

$$\begin{aligned} F_{XY}(x,y) &= P\left[(X(\xi) \le x) \cap (Y(\xi) \le y)\right] \\ &= P(X \le x, Y \le y) \ge 0, \end{aligned}$$

where x and y are arbitrary real numbers.

By definition, the joint p.d.f of X and Y is given by

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \, \partial y}.$$

the probability that (X,Y) belongs to an arbitrary region D

$$P\big((X,Y)\in D\big)=\int\int_{(x,y)\in D} f_{XY}(x,y)dxdy.$$

Marginal Statistics

In the context of several r.vs, the statistics of each individual ones are called marginal statistics. Thus $F_X(x)$ is the marginal probability distribution function of X, and $f_X(x)$ is the marginal p.d.f of X. It is interesting to note that all marginals can be obtained from the joint p.d.f. In fact

$$F_X(x)=F_{XY}(x,+\infty),\quad F_Y(y)=F_{XY}(+\infty,y).$$

Also

$$f_X(x)=\int_{-\infty}^{+\infty} f_{XY}(x,y)dy,\quad f_Y(y)=\int_{-\infty}^{+\infty} f_{XY}(x,y)dx.$$

If X and Y are discrete r.vs, then $p_{ij} \stackrel{\Delta}{=} P(X = x_i, Y = y_j)$ represents their joint p.d.f, and their respective marginal p.d.fs are given by

$$P(X = x_i) = \sum_j P(X = x_i, Y = y_j) = \sum_j p_{ij}$$

and

$$P(Y = y_j) = \sum_i P(X = x_i, Y = y_j) = \sum_i p_{ij}$$

Assuming that $P(X = x_i, Y = y_j)$ is written out in the form of a rectangular array, to obtain $P(X = x_i)$, one need to add up all entries in the i-th row.

It used to be a practice for insurance companies routinely to scribble out these sum values in the left and top margins, thus suggesting the name marginal densities!

	$\sum_i p_{ij}$					
	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	p_{1n}
	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$	p_{2n}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\sum_j p_{ij}$	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	p_{in}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	p_{m1}	p_{m2}	$\cdots$	p_{mj}	$\cdots$	p_{mn}

The joint P.D.F and/or the joint p.d.f represent complete information about the r.vs, and their marginal p.d.fs can be evaluated from the joint p.d.f. However, given marginals, (most often) it will not be possible to compute the joint p.d.f. Consider the following example:

Example: X and Y are said to be jointly normal (Gaussian) distributed, if their joint p.d.f has the following form:

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{\frac{-1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)},$$

$$-\infty < x < +\infty,\ -\infty < y < +\infty,\ |\rho| < 1.$$

By direct integration, it can be shown that

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dy = \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-(x-\mu_X)^2/2\sigma_X^2} \sim N(\mu_X, \sigma_X^2),$$

and similarly

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dx = \frac{1}{\sqrt{2\pi\sigma_Y^2}} e^{-(y-\mu_Y)^2/2\sigma_Y^2} \sim N(\mu_Y, \sigma_Y^2),$$

Following the above notation, we will denote the joint r.v. as $N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$. Clearly, knowing the marginals alone doesn't tell us everything about the parameter ρ.

As we show below, the only situation where the marginal p.d.fs can be used to recover the joint p.d.f is when the random variables are statistically independent.

Independence of r.vs

Definition: The random variables X and Y are said to be statistically independent if

$$F_{XY}(x,y) = F_X(x)F_Y(y)$$

or equivalently, if X and Y are independent, then we must have

$$f_{XY}(x,y) = f_X(x)f_Y(y).$$

If X and Y are discrete-type r.vs then their independence implies

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j) \quad \text{for all} \quad i, j.$$

Example: Birthday paradox

In a class of about 20 students, there are two students with the same birthday, with high probability.

The probability that there are two (or more) students with the same birthday

$$p(n) = 1 - 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{n-1}{365}\right)$$

Using the Taylor series approximation $e^{-x} \approx 1 - x$,

$$p(n) \approx 1 - 1 \times e^{-\frac{1}{365}} \times e^{-\frac{2}{365}} \times \cdots \times e^{-\frac{n-1}{365}}$$

$$= 1 - e^{-\frac{1+2+\cdots+(n-1)}{365}} = 1 - e^{-\frac{n(n-1)}{730}}$$

Let $p(n) = 0.5$, we obtain

$$e^{-\frac{n(n-1)}{730}} = 0.5$$

which gives $n = 23$.

In reality, birthdays are not uniform, so n can be smaller.

Example: Given

$$f_{XY}(x,y)=\begin{cases} xy^2e^{-y}, & 0<y<\infty, \quad 0<x<1, \\ 0, & \text{otherwise.} \end{cases}$$

Determine whether X and Y are independent.

Solution:

$$f_X(x)=\int_0^{+\infty} f_{XY}(x,y)dy = x\int_0^{\infty} y^2e^{-y}dy$$

$$= x\left(-y^2e^{-y}\Big|_0^{\infty} + 2\int_0^{\infty} ye^{-y}dy \right) = 2x, \quad 0<x<1.$$

Similarly

$$f_Y(y)=\int_0^{1} f_{XY}(x,y)dx = \frac{y^2}{2}e^{-y}, \quad 0<y<\infty.$$

In this case

$$f_{XY}(x,y)=f_X(x)f_Y(y),$$

and hence X and Y are independent random variables.

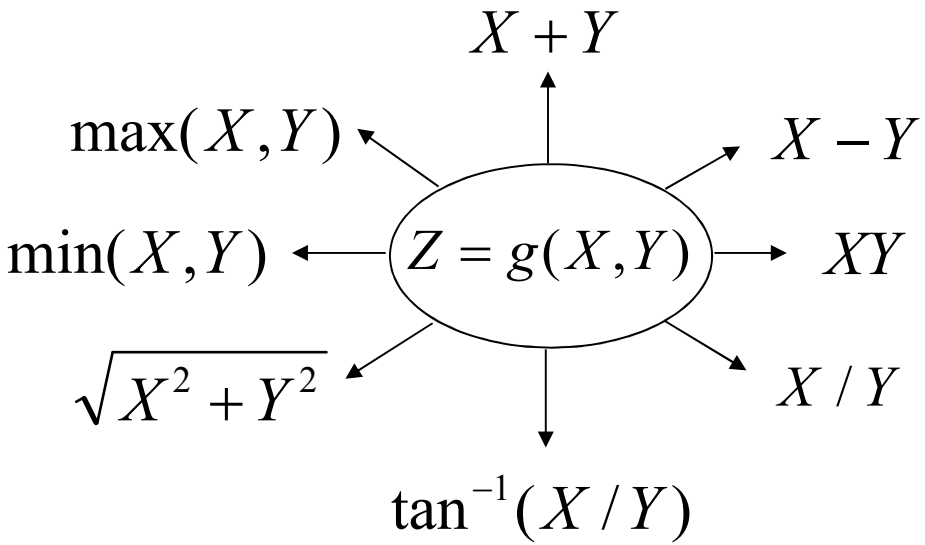
Function of Two Random Variables

Given two random variables X and Y and a function $g(x,y)$, we form a new random variable Z as

$$Z = g(X,Y).$$

Given the joint p.d.f $f_{XY}(x,y)$, how does one obtain $f_Z(z)$, the p.d.f of Z ? Problems of this type are of interest from a practical standpoint. For example, a receiver output signal usually consists of the desired signal buried in noise, and the above formulation in that case reduces to $Z = X + Y$.

It is important to know the statistics of the incoming signal for proper receiver design. In this context, problems of the following type are analyzed in Text, Chap. 6.



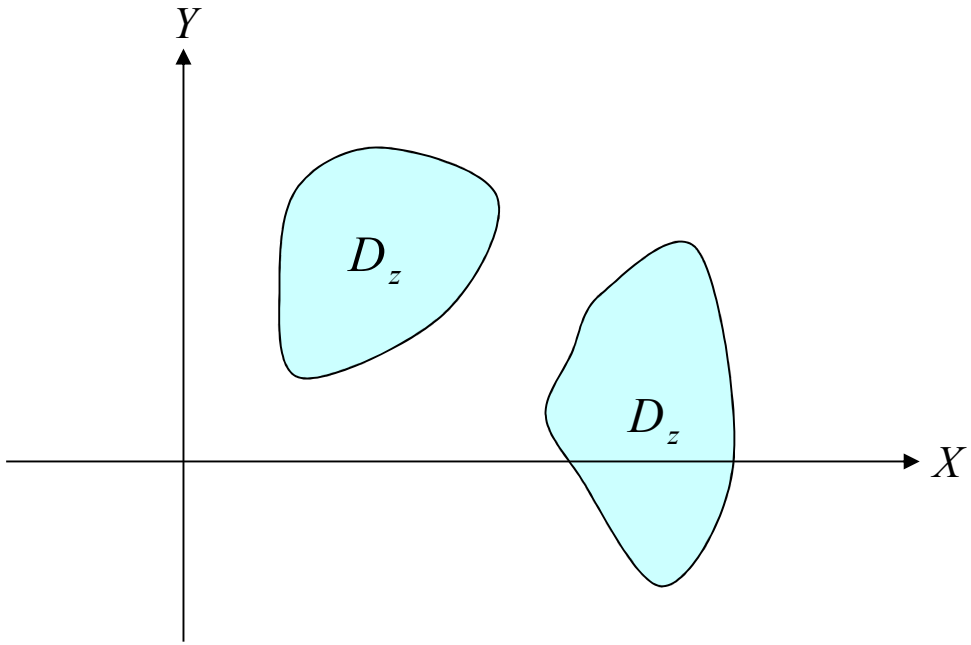
We start with

$$F_Z(z) = P(Z \le z) = P(g(X,Y) \le z) = P[(X,Y) \in D_z]$$

$$= \int\int_{x,y \in D_z} f_{XY}(x,y)\,dxdy,$$

where D_z in the XY plane represents the region such that $g(x, y) \le z$ is satisfied. Note that D_z need not be simply connected (see the following figure). To determine $F_Z(z)$ it is enough to find the region D_z for every z, and then evaluate the integral there.

We shall illustrate this method through the example $X + Y$.

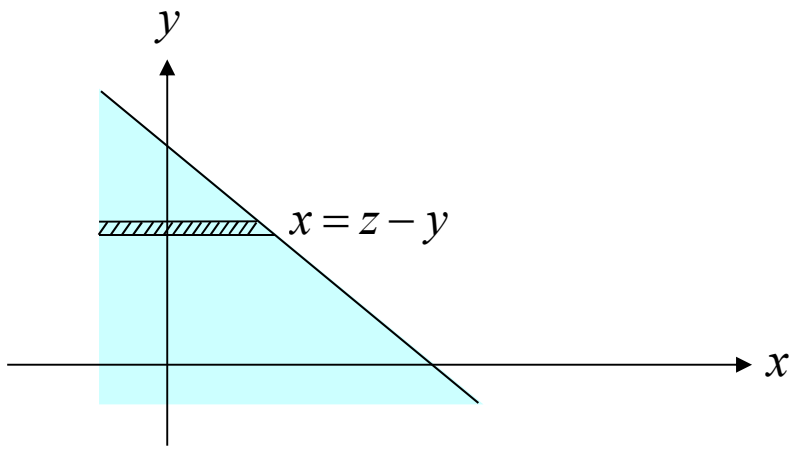


Example: $Z = X + Y$. Find $f_Z(z)$.

Solution:

$$F_Z(z) = P(X + Y \le z) = \int_{y=-\infty}^{+\infty} \int_{x=-\infty}^{z-y} f_{XY}(x, y)dxdy,$$

since the region D_z of the xy plane where $x + y \le z$ is the shaded area in the figure to the left of the line $x + y = z$. Integrating over the horizontal strip along the x-axis first (inner integral) followed by sliding that strip along the y-axis from $-\infty$ to $+\infty$ (outer integral) we cover the entire shaded area.



We can find $f_Z(z)$ by differentiating $F_Z(z)$ directly. In this context, it is useful to recall the differentiation rule due to Leibnitz. Suppose

$$H(z) = \int_{a(z)}^{b(z)} h(x,z)dx.$$

Then

$$\frac{dH(z)}{dz} = \frac{db(z)}{dz} h\big(b(z),z\big) - \frac{da(z)}{dz} h\big(a(z),z\big) + \int_{a(z)}^{b(z)} \frac{\partial h(x,z)}{\partial z} dx.$$

Using this we get

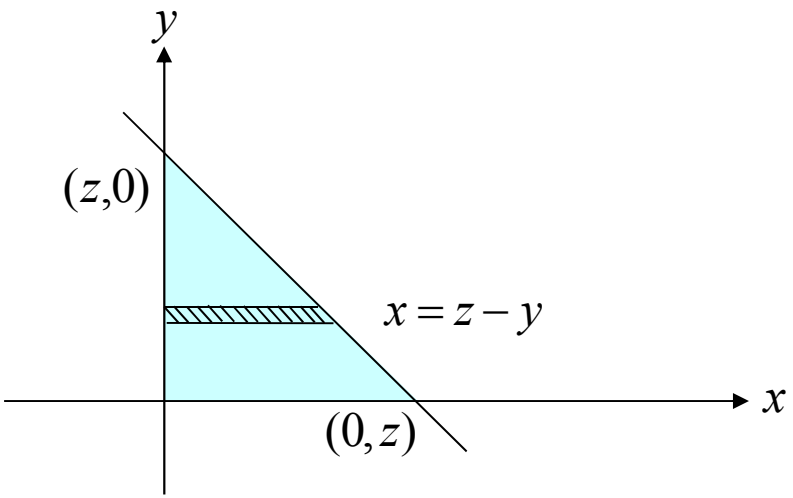
$$f_Z(z) = \frac{dF_Z(z)}{dz} = \int_{-\infty}^{+\infty} \left(\frac{\partial}{\partial z} \int_{-\infty}^{z-y} f_{XY}(x,y)dx \right) dy = \int_{-\infty}^{+\infty} f_{XY}(z-y,y)dy.$$

If X and Y are independent, then $f_{XY}(x,y) = f_X(x) f_Y(y)$, and inserting it into the p.d.f of Z, we get

$$f_Z(z) = \int_{y=-\infty}^{+\infty} f_X(z-y) f_Y(y)dy.$$

The above integral is the standard **convolution** of the functions $f_X(z)$ and $f_Y(z)$ expressed two different ways. We thus reach the following conclusion: **If two r.vs are independent, then the density of their sum equals the convolution of their density functions.**

As a special case, suppose that $f_X(x) = 0$ for $x < 0$ and $f_Y(y) = 0$ for $y < 0,$ then we can make use of the next figure to determine the new limits for D_z.



Two Functions of Two Random Variables

In the spirit of the previous section, let us look at an immediate generalization: Suppose X and Y are two random variables with joint p.d.f $f_{XY}(x,y)$. Given two functions $g(x,y)$ and $h(x,y)$, define the new random variables

$$Z = g(X,Y)$$
$$W = h(X,Y).$$

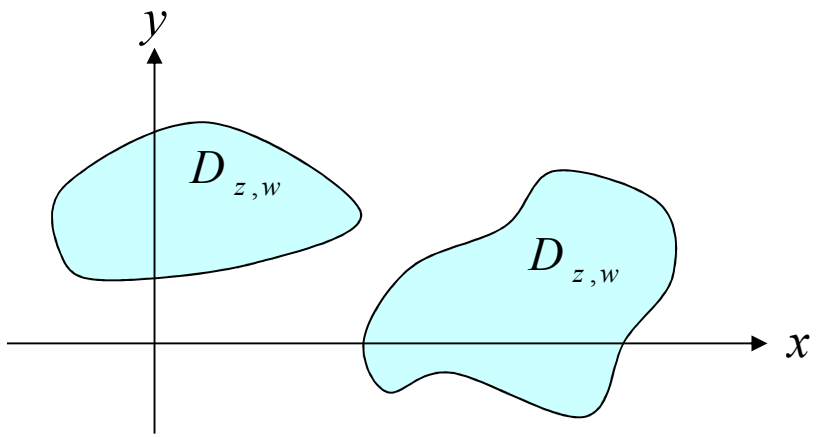
How does one determine their joint p.d.f $f_{ZW}(z,w)$? Obviously with $f_{ZW}(z,w)$ in hand, the marginal p.d.fs $f_Z(z)$ and $f_W(w)$ can be easily determined.

The procedure is the same as that for one function.

In fact for given z and w,

$$F_{ZW}(z,w) = P\big(Z(\xi) \le z, W(\xi) \le w\big) = P\big(g(X,Y) \le z, h(X,Y) \le w\big)$$
$$= P\big((X,Y) \in D_{z,w}\big) = \int\int_{(x,y)\in D_{z,w}} f_{XY}(x,y)\,dxdy,$$

where $D_{z,w}$ is the region in the xy plane such that the inequalities $g(x,y) \le z$ and $h(x,y) \le w$ are simultaneously satisfied.



If $g(x,y)$ and $h(x,y)$ are continuous and differentiable functions, then it is possible to develop a formula to obtain the joint p.d.f $f_{ZW}(z,w)$ directly.

Towards this, consider the equations

$$g(x,y) = z, \qquad h(x,y) = w.$$

For a given point (z,w), this equation can have many solutions. Let us say they are

$$(x_1, y_1),\ (x_2, y_2),\ \cdots,\ (x_n, y_n)$$

Then the p.d.f is given by

$$f_{ZW}(z,w) = \sum_i \frac{1}{|J(x_i,y_i)|} f_{XY}(x_i,y_i),$$

where $J(x_i, y_i)$ represents the Jacobian of the transformation given by (det denotes matrix determinant)

$$J(x_i,y_i) = \det \begin{pmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \\ \frac{\partial h}{\partial x} & \frac{\partial h}{\partial y} \end{pmatrix}_{x=x_i, y=y_i}$$

Next we shall illustrate the usefulness of the formula through an example:

Example: Suppose X and Y are zero mean independent Gaussian r.vs with common variance σ^2.

Define $Z = \sqrt{X^2 + Y^2}$, $W = \tan^{-1}(Y/X)$, where $|w| \leq \pi/2$.

Obtain $f_{ZW}(z, w)$.

Solution: Here

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}.$$

Since

$$z = g(x, y) = \sqrt{x^2 + y^2}; w = h(x, y) = \tan^{-1}(y/x), \quad |w| \leq \pi/2,$$

if (x_1, y_1) is a solution pair so is $(-x_1, -y_1)$.

We can compute

$$J(x,y)=\det\begin{pmatrix}\dfrac{x}{\sqrt{x^2+y^2}} & \dfrac{y}{\sqrt{x^2+y^2}}\\ \dfrac{-y}{x^2+y^2} & \dfrac{x}{x^2+y^2}\end{pmatrix}=\frac{1}{\sqrt{x^2+y^2}}=\frac{1}{z}.$$

Thus, we get

$$f_{ZW}(z,w)=z\big(f_{XY}(x_1,y_1)+f_{XY}(x_2,y_2)\big)$$

$$=\frac{z}{\pi\sigma^2}e^{-z^2/2\sigma^2},\qquad 0<z<\infty,\quad |w|<\frac{\pi}{2}.$$

Thus

$$f_Z(z)=\int_{-\pi/2}^{\pi/2}f_{ZW}(z,w)dw=\frac{z}{\sigma^2}e^{-z^2/2\sigma^2},\qquad 0<z<\infty,$$

which represents a Rayleigh r.v with parameter σ^2, and

$$f_W(w)=\int_0^{\infty}f_{ZW}(z,w)dz=\frac{1}{\pi},\quad |w|<\frac{\pi}{2},$$

which represents a uniform r.v in the interval $(-\pi/2, \pi/2)$. Moreover by direct computation

$$f_{ZW}(z,w) = f_Z(z) \cdot f_W(w)$$

implying that Z and W are independent. We summarize these results in the following statement: If X and Y are zero mean independent Gaussian random variables with common variance, then $\sqrt{X^2+Y^2}$ has a Rayleigh distribution and $\tan^{-1}(Y/X)$ has a uniform distribution. Moreover these two derived r.vs are statistically independent.

Alternatively, with X and Y as independent zero mean r.vs, $X + jY$ represents a complex Gaussian r.v. But $X + jY = Ze^{jW}$, it follows that the magnitude and phase of a complex Gaussian r.v are independent with Rayleigh and uniform distributions $(U \sim (-\pi, \pi))$ respectively.

Covariance: Given any two r.vs X and Y, define

$$Cov(X,Y) = E\big[(X-\mu_X)(Y-\mu_Y)\big].$$

By expanding and simplifying the right side, we also get

$$Cov(X,Y) = E(XY) - \mu_X\mu_Y = E(XY) - E(X)E(Y)$$

$$= \overline{XY} - \overline{X}\,\overline{Y}.$$

It is easy to see that $\big|Cov(X,Y)\big| \le \sqrt{Var(X)Var(Y)}$.

We may define the normalized parameter

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}, \quad -1 \le \rho_{XY} \le 1,$$

Or $Cov(X,Y) = \rho_{XY}\sigma_X\sigma_Y$ and it represents the **correlation coefficient** between X and Y.

Uncorrelated r.vs: If $\rho_{XY} = 0,$ then X and Y are said to be uncorrelated r.vs. If X and Y are uncorrelated, then

$$E(XY) = E(X)E(Y).$$

Orthogonality: X and Y are said to be orthogonal if

$$E(XY) = 0.$$

Thus, if either X or Y has zero mean, then orthogonality implies uncorrelatedness also and vice-versa.

Suppose X and Y are independent r.vs. Then

$$E(XY) = E(X)E(Y),$$

and we conclude that the random variables are uncorrelated. Thus independence implies uncorrelatedness.

Naturally, if two random variables are statistically independent, then there cannot be any correlation between them $(\rho_{XY} = 0)$. However, the converse is in general not true. As the next example shows, random variables can be uncorrelated without being independent.

Example : Let $X \sim U(0,1),\ Y \sim U(0,1)$. Suppose X and Y are independent. Define $Z = X + Y,\ W = X - Y$. Show that Z and W are dependent, but uncorrelated r.vs.

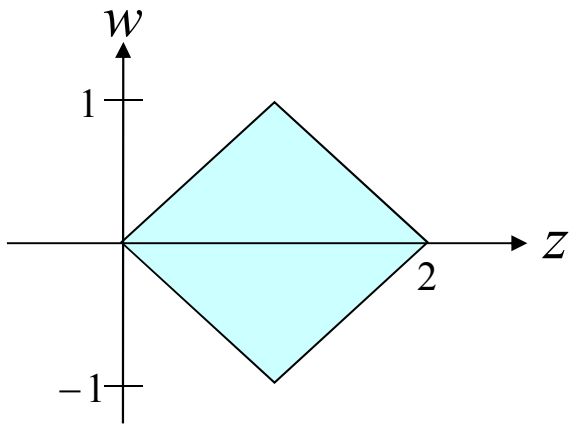
Solution: $z = x + y,\quad w = x - y$ gives the only solution set to be

$$x = \frac{z + w}{2},\quad y = \frac{z - w}{2}.$$

Moreover $0 < z < 2,\ -1 < w < 1,\ z + w \le 2,\ z - w \le 2,\ z > |w|$ and $|J(z,w)| = 1/2.$

Thus (see the shaded region in the figure)

$$f_{ZW}(z,w)=\begin{cases}1/2, & 0<z<2,\ -1<w<1,\ z+w\le 2,\ z-w\le 2,\ |w|<z,\\ 0, & \text{otherwise,}\end{cases}$$



and hence

$$f_Z(z)=\int f_{ZW}(z,w)dw=\begin{cases}\displaystyle\int_{-z}^{z}\frac{1}{2}dw=z, & 0<z<1,\\ \displaystyle\int_{z-2}^{2-z}\frac{1}{2}dw=2-z, & 1<z<2,\end{cases}$$

or by direct computation ($Z=X+Y$)

$$f_Z(z) = f_X(z) \otimes f_Y(z) = \begin{cases} z, & 0 < z < 1, \\ 2 - z, & 1 < z < 2, \\ 0, & \text{otherwise}, \end{cases}$$

and

$$f_W(w) = \int f_{ZW}(z,w)dz = \int_{|w|}^{2-|w|} \frac{1}{2} dz = \begin{cases} 1 - |w|, & -1 < w < 1, \\ 0, & \text{otherwise}. \end{cases}$$

Clearly $f_{ZW}(z,w) \neq f_Z(z) f_W(w)$. Thus Z and W are not independent. However

$$E(ZW) = E\left[(X+Y)(X-Y)\right] = E(X^2) - E(Y^2) = 0,$$

and

$$E(W) = E(X-Y) = 0,$$

and hence

$$Cov(Z,W) = E(ZW) - E(Z)E(W) = 0$$

implying that Z and W are uncorrelated random variables.

Joint Moments:

$$E[X^k Y^m] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x^k y^m f_{XY}(x,y)dx\,dy,$$

represents the joint moment of order (k,m) for X and Y.

Following the one random variable case, we can define the joint characteristic function between two random variables which will turn out to be useful for moment calculations.

Joint characteristic functions:

The joint characteristic function between X and Y is defined as

$$\Phi_{XY}(u,v) = E\left(e^{j(Xu+Yv)}\right) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{j(Xu+Yv)} f_{XY}(x,y)dxdy.$$

Note that $\left|\Phi_{XY}(u,v)\right| \le \Phi_{XY}(0,0) = 1.$

It is easy to show that

$$E(XY) = \frac{1}{j^2} \left. \frac{\partial^2 \Phi_{XY}(u,v)}{\partial u \partial v} \right|_{u=0,v=0}.$$

If X and Y are independent r.vs, then we obtain

$$\Phi_{XY}(u,v) = E(e^{juX})E(e^{jvY}) = \Phi_X(u)\Phi_Y(v).$$

Also

$$\Phi_X(u) = \Phi_{XY}(u,0), \quad \Phi_Y(v) = \Phi_{XY}(0,v).$$

More on Gaussian r.vs :

If X and Y are jointly Gaussian $N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, we obtain the joint characteristic function of two jointly Gaussian r.vs to be

$$\Phi_{XY}(u,v) = E(e^{j(Xu+Yv)}) = e^{j(\mu_X u+\mu_Y v)-\frac{1}{2}(\sigma_X^2 u^2+2\rho\sigma_X\sigma_Y uv+\sigma_Y^2 v^2)}.$$

This can be used to make various conclusions. Letting $v = 0$ we get

$$\Phi_X(u) = \Phi_{XY}(u,0) = e^{j\mu_X u-\frac{1}{2}\sigma_X^2 u^2},$$

and it agrees with the characteristic function of X.

By direct computation, it is easy to show that for two jointly Gaussian random variables

$$Cov(X,Y) = \rho\,\sigma_X\sigma_Y.$$

Hence, ρ in $N(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)$ represents the actual correlation coefficient of the two jointly Gaussian r.vs. Notice that $\rho = 0$ implies

$$f_{XY}(X,Y) = f_X(x) f_Y(y).$$

Thus if X and Y are jointly Gaussian, uncorrelatedness does imply independence between the two random variables. Gaussian case is the only exception where the two concepts imply each other.

Example : Let X and Y be jointly Gaussian r.vs with parameters $N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$. Define $Z = aX + bY$. Determine $f_Z(z)$.

Solution: In this case we can make use of characteristic function to solve this problem.

$$\Phi_Z(u) = E(e^{jZu}) = E(e^{j(aX+bY)u}) = E(e^{jauX+jbuY})$$
$$= \Phi_{XY}(au, bu).$$

With u and v replaced by au and bu respectively we get

$$\Phi_Z(u) = e^{j(a\mu_X + b\mu_Y)u - \frac{1}{2}(a^2\sigma_X^2 + 2\rho ab\sigma_X\sigma_Y + b^2\sigma_Y^2)u^2} = e^{j\mu_Z u - \frac{1}{2}\sigma_Z^2 u^2},$$

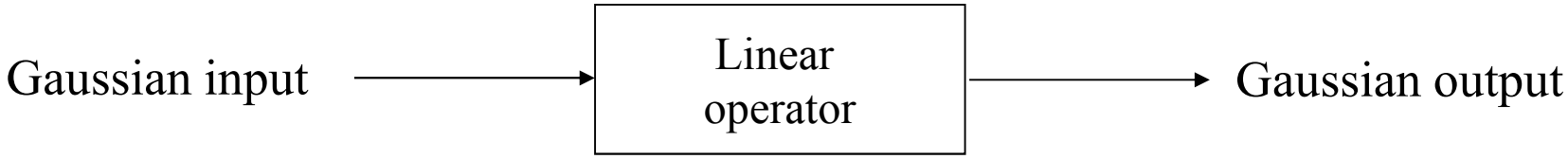
where

$$\mu_Z \stackrel{\Delta}{=} a\mu_X + b\mu_Y,$$
$$\sigma_Z^2 \stackrel{\Delta}{=} a^2\sigma_X^2 + 2\rho ab\sigma_X\sigma_Y + b^2\sigma_Y^2.$$

Notice that it has the same form as the characteristic function of a Gaussian r.v, and hence we conclude that $Z = aX + bY$ is also Gaussian with mean and variance as shown above.

From the previous example, we conclude that any linear combination of jointly Gaussian r.vs generate a Gaussian r.v.

In other words, linearity preserves Gaussianity. We can use the characteristic function relation to conclude an even more general result. That is, the output of any linear system with a Gaussian input is also Gaussian.



This result is very useful in the analysis of linear systems. For example, it implies that the noise at the output of a linear filter is also Gaussian if the input noise is Gaussian.

Joint characteristic functions are useful in determining the p.d.f of linear combinations of r.vs. For example, with X and Y as independent Poisson r.vs with parameters λ_1 and λ_2 respectively, let

$$Z = X + Y.$$

Then

$$\Phi_Z(u) = \Phi_X(u)\Phi_Y(u).$$

But

$$\Phi_X(u) = e^{\lambda_1(e^{ju}-1)}, \quad \Phi_Y(u) = e^{\lambda_2(e^{ju}-1)}$$

so that

$$\Phi_Z(u) = e^{(\lambda_1+\lambda_2)(e^{ju}-1)} \quad \sim \quad P(\lambda_1+\lambda_2)$$

i.e., sum of independent Poisson r.vs is also a Poisson random variable.

Conditional p.d.f

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

If the r.vs X and Y are independent, then $f_{XY}(x, y) = f_X(x) f_Y(y)$ and it reduces to

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

implying that the conditional p.d.fs coincide with their unconditional p.d.fs. This makes sense, since if X and Y are independent r.vs, information about Y shouldn't be of any help in updating our knowledge about X.

In the case of discrete-type r.vs

$$P\big(X = x_i \mid Y = y_j\big) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}.$$

Extension to multiple r.v's

- Joint PDF

$$F_{X_1X_2\ldots X_n}(x_1,x_2,\ldots x_n) \equiv P(X_1 \le x_1, X_2 \le x_2, \ldots X_n \le x_n)$$

- Joint pdf

$$f_{X_1X_2\ldots X_n}(x_1,x_2,\ldots x_n) \equiv \frac{\partial^n F_{X_1X_2\ldots X_n}(x_1,x_2,\ldots x_n)}{\partial x_1 \partial x_2 \ldots \partial x_n}$$

- **Independent**

$$F_{X_1X_2\ldots X_n}(x_1,x_2,\ldots x_n) = F_{X_1}(x_1)F_{X_2}(x_2)\ldots F_{X_n}(x_n)$$
$$f_{X_1X_2\ldots X_n}(x_1,x_2,\ldots x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\ldots f_{X_n}(x_n)$$

- **i.i.d.** (independent, identically distributed)
  - The random variables are independent and have the same distribution.
  - Example: outcomes from repeatedly flipping a coin.

Multidimensional Gaussian distribution

The most widely used continuous joint distribution is the multivariate normal (or Gaussian). Multivariate normal pdf of the vector $X = [X_1, X_2, \ldots, X_n]$ is completely determined by mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:

$$f_X(\mathbf{x}) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]}{(2\pi)^{n/2} \left|\det(\boldsymbol{\Sigma})\right|^{1/2}}$$

where $\det(\Sigma)$ represents determinant of Σ, whose (i, j)th entry is given by the covariance between X_i and X_j .

Imperial College
London

EE4-10: Probability and Stochastic Processes

Lecture 3: Sequences of Random Variables

Outline

In this lecture, we are concerned with the asymptotical properties of a sequence of random variables

$$X_1, X_2, \cdots, X_n, \cdots$$

In particular:

- Concentration inequalities
- Law of large numbers
  - Weak law of large numbers
  - Strong law of large numbers (non-examinable)
- Central limit theorem

Concentration inequalities

We begin with concentration inequalities which provide bounds on the probability that a random variable deviates from some value (e.g. its expectation).

Markov Inequality: for nonnegative r.v. X,

$$P(X \geq a) \leq \frac{E(X)}{a}, \qquad a > 0$$

Proof:

$$E(X) = \int_0^a x f_X(x)dx + \int_a^\infty x f_X(x)dx$$

$$\geq \int_a^\infty x f_X(x)dx \geq \int_a^\infty a f_X(x)dx = aP(X \geq a)$$

Markov inequality is not tight because only the mean is utilized

Generalized Markov Inequality: for any strictly increasing and non-negative function g

$$P(X \geq a) = P(g(X) \geq g(a)) \leq \frac{E(g(X))}{g(a)}, \qquad a > 0$$

Proof: Let $Y = g(X)$ and then proceed as in previous proof.

Example: $P(|X| \geq a) \leq \frac{E(|X|^m)}{a^m}, \qquad a > 0, \quad m = 1,2,\cdots$

Chebyshev Inequality: is a special case of generalized Markov's inequality when $g(X) = X^2$

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}, \qquad a > 0$$

Chebyshev inequality is tighter than Markov inequality as the former uses the variance

Interestingly, to compute the above probability bound the knowledge of $f_X(x)$ is not necessary. We only need σ^2, the variance of the r.v. In particular with $\varepsilon = k\sigma$ we obtain

$$P\big(|X-\mu| \geq k\sigma\big) \leq \frac{1}{k^2}.$$

Thus with $k=3,$ we get the probability of X being outside the 3σ interval around its mean to be 0.111 for any r.v. Compare with Markov inequality $(\mu = 0, \sigma = 1)$

$$P(|X| \geq a) \leq \frac{E(|X|)}{a} = \frac{\sqrt{2/\pi}\sigma}{3\sigma} = 0.265$$

Obviously this cannot be a tight bound as it includes all r.vs. For example, in the case of a Gaussian r.v, $P\big(|X| \geq 3\sigma\big) = 0.0027.$ Markov and Chebyshev inequalities always overtimate the exact probability.

Chernoff Bound: gives exponentially decreasing bounds on tail distributions.

$$P(X > a) \leq \min_{\lambda>0} e^{-\lambda a} E[e^{\lambda X}] = \min_{\lambda>0} e^{-\lambda a} \Phi_X(\lambda), \qquad a > 0$$

where $\Phi_X(s) = \int_{-\infty}^{\infty} f_X(x) e^{sx} dx$ is the moment generating function of X

Proof: This is a special case of the generalized Markov inequality for $g(X) = e^{\lambda X}$, $\lambda > 0$

Open Question: For the bound to be tight, we have to choose appropriate value for λ.

Example: Chernoff bound for normal distribution.

Consider $X \sim N(0, \sigma^2)$ i.e., normal distribution

Obtain the Chernoff bound on $P(X > a)$ for $a > 0$:

$$P(X > a) \le e^{-a\lambda} e^{\sigma^2 \lambda^2 / 2} = e^{-a\lambda + \sigma^2 \lambda^2 / 2}$$

The RHS is minimized when $\lambda = a / \sigma^2$

Thus, $P(X > a) \le e^{-a^2 / 2\sigma^2}$.

Applying it to our problem, $P(|X| > 3\sigma) \le 2 \cdot e^{-9/2} = 0.022$

This is commonly used to bound the Q – function

$$Q(x) = \int_{t=x}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

Chernoff bound: $Q(x) \le e^{-x^2/2}$

Bernstein Inequality is a more precise formulation of the classical Chebyshev inequality. It gives exponential bounds on the probability that the sum of random variables deviates from its mean.

Let $X_1, ..., X_n$ be independent and identically distributed (iid) random variables with mean μ, variance σ^2, and $|X_i - \mu| < M$, then for the sample average

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

we have

$$P\left(|\overline{X}_n - \mu| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2(\sigma^2 + tM/3)}\right), \qquad t > 0$$

Hoeffding Inequality:

If $X_1, ..., X_n$ are independent and bounded; that is, assume $X_i \in [a_i, b_i]$. Then, for the sample average

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

we have

$$P\left(| \overline{X}_n - E(\overline{X}_n) | \geq t \right) \leq 2 \exp\left(- \frac{2n^2 t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right), \qquad t > 0$$

Note that, unlike Bernstein's inequality, this bound does not depend on the variance.

The Law of Large Numbers (LLN)

Bernoulli proved the weak law of large numbers (WLLN) around 1700 which was published posthumously in 1713 in his treatise *Ars Conjectandi.* Poisson generalized Bernoulli's theorem around 1800, and in 1866 Chebychev discovered the method bearing his name. Later on one of his students, Markov observed that Chebychev's reasoning can be used to extend Bernoulli's theorem to dependent random variables as well.

In 1909 the French mathematician Emile Borel proved a deeper theorem known as the strong law of large numbers (SLLN) that further generalizes Bernoulli's theorem. In 1926 Kolmogorov derived conditions that were necessary and sufficient for a set of mutually independent random variables to obey the law of large numbers.

Let $X_1, X_2 ..., X_n$ be a sequence of independent and identically distributed (iid) random variables, each having a mean $E(X_i) = \mu$ and standard deviation σ. Define a new variable, the "sample average"

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

The law of large numbers states that the sample average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.

There are two forms of the Law of Large Numbers, namely, the Strong Law and the Weak Law. These forms do not describe different laws but instead refer to different ways of describing the mode of convergence. Both versions of the law state that – with virtual certainty – the sample average converges to the expected value. The strong form implies the weak.

Weak Law of Large Numbers

Suppose $X_1, X_2 \ldots, X_n$ are i.i.d. with finite $E(X_i) = \mu$.

Then, $\bar{X}_n$ converges to μ in probability.

Strong Law of Large Numbers

Suppose $X_1, X_2 \ldots, X_n$ are i.i.d. with finite $E(X_i) = \mu$.

Then, $\bar{X}_n$ converges to μ almost surely.

Convergence in probability

A sequence $\{X_n\}$ of random variables converges in probability towards X if for all $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(|X_n - X| > \varepsilon\right) = 0$$

Almost sure convergence

A sequence $\{X_n\}$ converges *almost surely* or *almost everywhere* or *with probability 1* or *strongly* towards X if

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1$$

Almost sure convergence implies convergence in probability, but the converse is not true. For a counterexample, see http://math.stackexchange.com/questions/233861

Proof of the weak law

Using the properties of expectations it is easy to prove that

$$E(\bar{X}_n) = \frac{E(X_1)+E(X_2)+\cdots+E(X_n)}{n} = \mu$$

Using the properties of the variances we can also prove that

$$var(\bar{X}_n) = var\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)$$

$$= \frac{1}{n^2} var(X_1 + X_2 + \cdots + X_n)$$

$$= \frac{1}{n^2}[var(X_1) + var(X_2) + \cdots + var(X_n)] = \frac{\sigma^2}{n}$$

Then, using the Chebyshev inequality, one can find

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{var(\bar{X}_n)=}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \quad \text{as} \quad n \to \infty$$

where ε is any (but typically small) positive number. Let us discuss this result. The left side contains the probability that the deviation between the sample average, $\bar{X}_n$ and the mean value of the individual random variable, μ, exceeds ε. The right hand side shows that this probability decreases as $1/n$ for large n. Thus, one can say that " $\bar{X}_n$ converges to μ in probability".

The LLN has found a wide range of applications. In particular, **Shannon's information theory is based on the weak law**.

The proof of the strong law is more complicated, and is not required in this course.

Example: $2n$ red cards and $2n$ black cards (all distinct) are shuffled together to form a single deck, and then split into half. What is the probability that each half will contain n red and n black cards?

Solution: From a deck of $4n$ cards, $2n$ cards can be chosen in $\binom{4n}{2n}$ different ways. To determine the number of favorable draws of n red and n black cards in each half, consider the unique draw consisting of $2n$ red cards and $2n$ black cards in each half. Among those $2n$ red cards, n of them can be chosen in $\binom{2n}{n}$ different ways; similarly for each such draw there are $\binom{2n}{n}$ ways of choosing n black cards. Thus the total number of favorable draws containing n red and n black cards in each half are $\binom{2n}{n}\binom{2n}{n}$ among a total of $\binom{4n}{2n}$ draws. This gives the desired probability p_n to be

$$p_n = \frac{\binom{2n}{n}\binom{2n}{n}}{\binom{4n}{2n}} = \frac{((2n)!)^4}{(4n)!(n!)^4}.$$

For large n, using Stirling's formula we get

$$p_n \approx \frac{[\sqrt{2\pi(2n)}\,(2n)^{2n}e^{-2n}]^4}{\sqrt{2\pi(4n)}\,(4n)^{4n}e^{-4n}[\sqrt{2\pi\,n}\;\,n^n e^{-n}]^4} = \sqrt{\frac{2}{\pi\,n}}$$

For a full deck of 52 cards, we have $n = 13,$ which gives

$$p_n \approx 0.221$$

and for a partial deck of 20 cards (that contains 10 red and 10 black cards), we have $n = 5$ and $p_n \approx 0.3568.$

One summer afternoon, 20 cards (containing 10 red and 10 black cards) were given to a 5 year old child. The child split that partial deck into two equal halves and the outcome was declared a success if each half contained exactly 5 red and 5 black cards. With adult supervision (in terms of shuffling) the experiment was repeated 100 times that very same afternoon. The results are tabulated below in Table 1, and the relative frequency vs the number of trials plot in Fig 1 shows the convergence of *k /n* to *p*.

Table 1

Expt	Number of successes	Expt	Number of successes	Expt	Number of successes	Expt	Number of successes	Expt	Number of successes
1	0	21	8	41	14	61	23	81	29
2	0	22	8	42	14	62	23	82	29
3	1	23	8	43	14	63	23	83	30
4	1	24	8	44	14	64	24	84	30
5	2	25	8	45	15	65	25	85	30
6	2	26	8	46	16	66	25	86	31
7	3	27	9	47	17	67	25	87	31
8	4	28	10	48	17	68	25	88	32
9	5	29	10	49	17	69	26	89	32
10	5	30	10	50	18	70	26	90	32
11	5	31	10	51	19	71	26	91	33
12	5	32	10	52	20	72	26	92	33
13	5	33	10	53	20	73	26	93	33
14	5	34	10	54	21	74	26	94	34
15	6	35	11	55	21	75	27	95	34
16	6	36	12	56	22	76	27	96	34
17	6	37	12	57	22	77	28	97	34
18	7	38	13	58	22	78	29	98	34
19	7	39	14	59	22	79	29	99	34
20	8	40	14	60	22	80	29	100	35

The figure below shows results of an experiment of 100 trials.

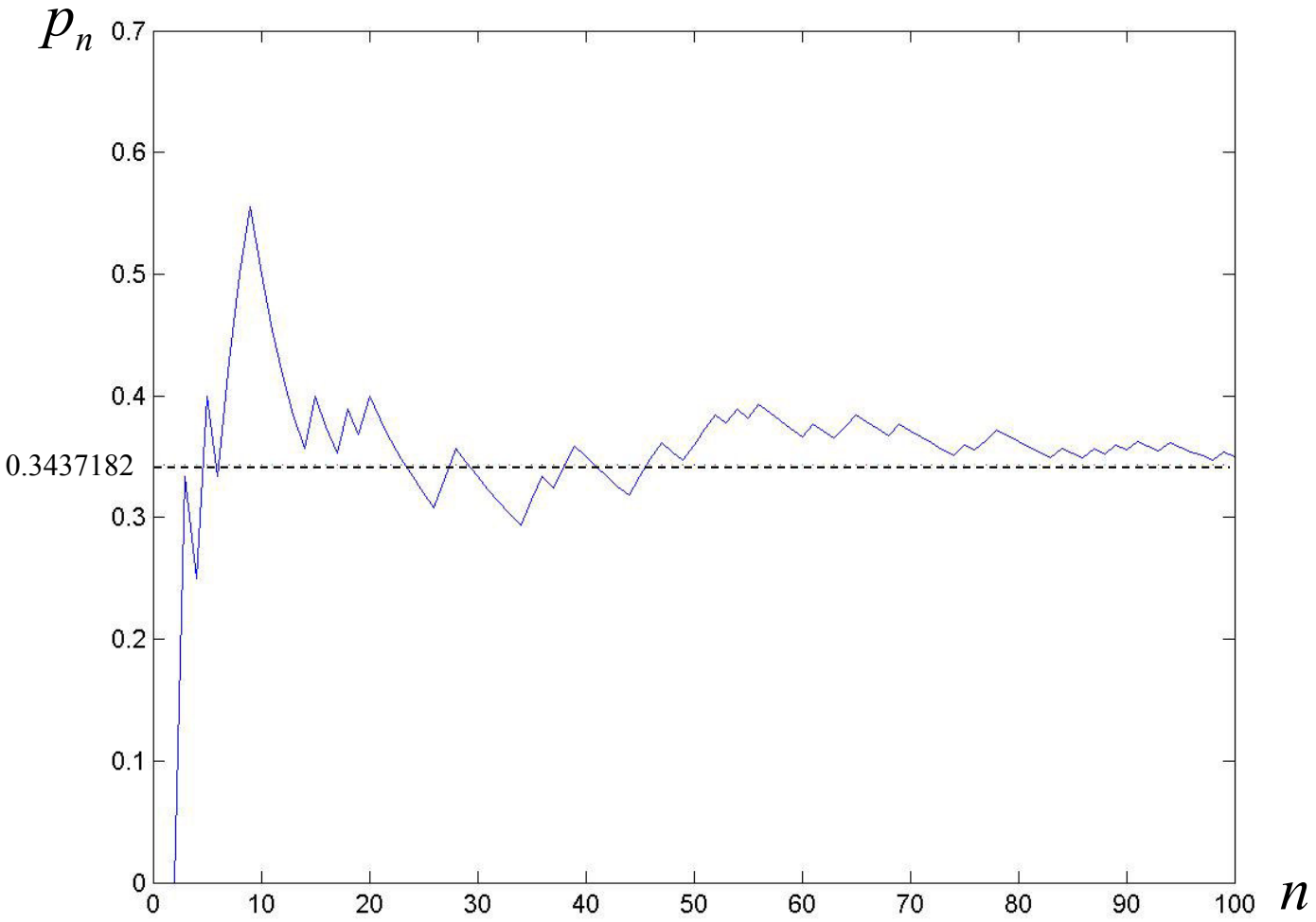


Fig 1

Stirling's Formula: What is it?

Stirling's formula gives an accurate approximation for $n!$ as follows:

$$n! \sim \sqrt{2\pi}\, n^{n+\frac{1}{2}} e^{-n} \qquad (*)$$

in the sense that the ratio of the two sides in (*) is near to one; i.e., their relative error is small, or the percentage error decreases steadily as n increases. The approximation is remarkably accurate even for small n. Thus $1! = 1$ is approximated as $\sqrt{2\pi}/e \approx 0.922$ and $3! = 6$ is approximated as 5.836.

Sometimes, a rough estimation $n! \sim (n/e)^n$ is used.

Central Limit Theorem

The Law of Large Numbers indicated that the sample average converges to the population average. Now our concern are the probability distributions. Suppose that we know a distribution for i.i.d. X_i. We wonder how the distribution of the average of X_i is related to the individual distributions.

Central Limit Theorem: Suppose $X_1, X_2, \cdots, X_n$ are a set of zero mean independent, identically distributed (i.i.d) random variables with some common distribution.

Let σ^2 represent their common variance. Consider their scaled sum

$$Y = \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}}.$$

Then asymptotically (as $n \to \infty$)

$$Y \to N(0, \sigma^2).$$

Proof: Although the theorem is true under even more general conditions, we shall prove it here under the independence assumption. Since

$$E(X_i) = 0,$$

we have

$$Var(X_i) = E(X_i^2) = \sigma^2.$$

Consider

$$\Phi_Y(\omega) = E(e^{jY\omega}) = E\left(e^{j(X_1+X_2+\cdots+X_n)\omega/\sqrt{n}} \right) = \prod_{i=1}^{n} E(e^{jX_i\omega/\sqrt{n}})$$

$$= \prod_{i=1}^{n} \Phi_{X_i}(\omega/\sqrt{n})$$

where we have made use of the independence of the r.vs $X_1, X_2, \cdots, X_n$. But

$$E(e^{jX_i\omega/\sqrt{n}}) = E\left(1 + \frac{jX_i\omega}{\sqrt{n}} + \frac{j^2X_i^2\omega^2}{2!n} + \frac{j^3X_i^3\omega^3}{3!n^{3/2}} + \cdots \right) = 1 - \frac{\sigma^2\omega^2}{2n} + o\left(\frac{1}{n}\right)$$

Substituting it into the characteristic function, we obtain

$$\Phi_Y(\omega) = \left[1 - \frac{\sigma^2\omega^2}{2n} + o\left(\frac{1}{n}\right) \right]^n$$

and

$$\lim_{n\to\infty} \Phi_Y(\omega) = e^{-\sigma^2\omega^2/2}$$

since

$$\lim_{n\to\infty}\left(1-\frac{x}{n}\right)^{n} \to e^{-x}.$$

[Note that $o(1/n)$ terms decay faster than $1/n$.]
But it represents the characteristic function of a zero mean normal r.v with variance σ^2 and the theorem follows.

The central limit theorem states that a large sum of independent random variables each with finite variance tends to behave like a normal random variable. Thus the individual p.d.fs become unimportant to analyze the collective sum behavior. If we model the noise phenomenon as the sum of a large number of independent random variables (eg: electron motion in resistor components), then this theorem allows us to conclude that noise behaves like a Gaussian r.v.

It may be remarked that the finite variance assumption is necessary for the theorem to hold good. To prove its importance, consider the r.vs to be Cauchy distributed, and let

$$Y = \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}}.$$

where each $X_i \sim C(\alpha).$ Then since

$$\Phi_{X_i}(\omega) = e^{-\alpha|\omega|},$$

substituting this, we get

$$\Phi_Y(\omega) = \prod_{i=1}^{n} \Phi_{X_i}(\omega/\sqrt{n}) = \left(e^{-\alpha|\omega|/\sqrt{n}}\right)^n \sim C(\alpha\sqrt{n})$$

which shows that Y is still Cauchy with parameter $\alpha\sqrt{n}$. In other words, central limit theorem doesn't hold good for a set of Cauchy r.vs as their variances are undefined.

Imperial College
London

EE4-10: Probability and Stochastic Processes

Lecture 4: Parameter Estimation

Principles of Parameter Estimation

The purpose of this lecture is to illustrate the usefulness of the various concepts introduced and studied in earlier lectures to practical problems of interest. In this context, consider the problem of estimating an unknown parameter of interest from a few of its noisy observations. For example, determining the daily temperature in a city, or the coefficient of a fading wireless channel, are problems that fall into this category.

Observations (measurement) are made on data that contain the desired nonrandom parameter θ and undesired noise. Thus, for example,

Observation = signal (desired part) + noise,

or, the i th observation can be represented as

$$X_i = \theta + n_i, \quad i = 1,2,\cdots,n.$$

Here θ represents the unknown nonrandom desired parameter, and $n_i, \ i = 1,2,\cdots,n$ represent random variables that may be dependent or independent from observation to observation. Given n observations $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$, the estimation problem is to obtain the "best" estimator for the unknown parameter θ in terms of these observations.

Let us denote by $\hat{\theta}(X)$ the estimator for θ. Obviously $\hat{\theta}(X)$ is a function of only the observations. "Best estimator" in what sense? Various optimization strategies can be used to define the term "best".

Ideal solution would be when the estimate $\hat{\theta}(X)$ coincides with the unknown θ. This of course may not be possible, and almost always any estimate will result in an error given by

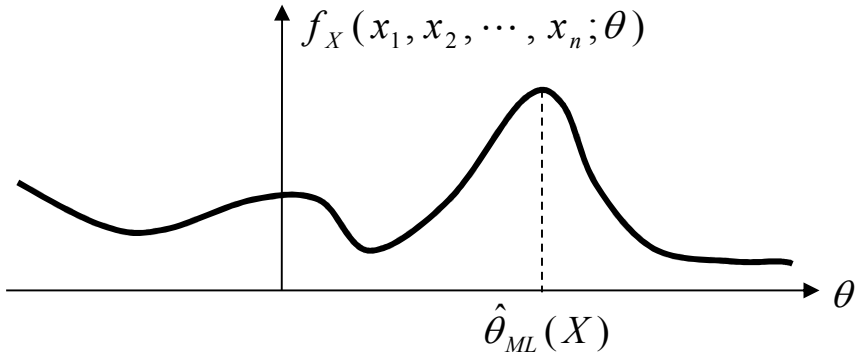
$$e = \hat{\theta}(X) - \theta.$$

One strategy would be to select the estimator $\hat{\theta}(X)$ so as to minimize some function of this error - such as - minimization of the mean square error (MMSE), or minimization of the absolute value of the error etc.

A more fundamental approach is that of the **principle of Maximum Likelihood** (ML).

The underlying assumption in any estimation problem is

that the available data $X_1, X_2, \cdots, X_n$ has something to do with the unknown parameter θ. More precisely, we assume that the joint p.d.f of $X_1, X_2, \cdots, X_n$ given by $f_X(x_1, x_2, \cdots, x_n; \theta)$ depends on θ. The method of maximum likelihood assumes that the given sample data set is representative of the population $f_X(x_1, x_2, \cdots, x_n; \theta)$, and chooses that value for θ that most likely caused the observed data to occur, i.e., once observations $x_1, x_2, \cdots, x_n$ are given, $f_X(x_1, x_2, \cdots, x_n; \theta)$ is a function of θ alone, and the value of θ that maximizes the above p.d.f is the most likely value for θ, and it is chosen as the ML estimate $\hat{\theta}_{ML}(X)$ for θ.



Given $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n,$ the joint p.d.f $f_X(x_1, x_2, \cdots, x_n; \theta)$ represents the likelihood function, and the ML estimate can be determined either from the likelihood equation

$$\hat{\theta}_{ML} = \arg\max_{\theta} \; f_X(x_1, x_2, \cdots, x_n; \theta)$$

or using the log-likelihood function (arg max in the above formula represents the argument that maximizes f_X)

$$L(x_1, x_2, \cdots, x_n; \theta) \stackrel{\Delta}{=} \log f_X(x_1, x_2, \cdots, x_n; \theta).$$

If $L(x_1, x_2, \cdots, x_n; \theta)$ is differentiable and a maximum $\hat{\theta}_{ML}$ exists, then that must satisfy the equation

$$\left. \frac{\partial \log f_X(x_1, x_2, \cdots, x_n; \theta)}{\partial \theta} \right|_{\theta = \hat{\theta}_{ML}} = 0.$$

We will illustrate the above procedure through several examples:

Example 1: Let $X_i = \theta + w_i, \quad i = 1 \to n,$ represent n observations where θ is the unknown parameter of interest, and $w_i, \quad i = 1 \to n,$ are zero mean independent normal r.vs with common variance σ^2. Determine the ML estimate for θ.

Solution: Since w_i are independent r.vs and θ is an unknown constant, we have X_i s are independent normal random variables. Thus the likelihood function takes the form

$$f_X(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta).$$

Moreover, each X_i is Gaussian with mean θ and variance σ^2 (Why?). Thus

$$f_{X_i}(x_i; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i - \theta)^2 / 2\sigma^2}.$$

Thus we get the likelihood function to be

$$f_X(x_1, x_2, \cdots, x_n; \theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n} (x_i - \theta)^2 / 2\sigma^2}.$$

It is easier to work with the log-likelihood function $L(X;\theta)$ in this case:

$$L(X;\theta) = \ln f_X(x_1, x_2, \cdots, x_n; \theta) = -\frac{n}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2},$$

and taking derivative with respect to θ, we get

$$\left.\frac{\partial \ln f_X(x_1, x_2, \cdots, x_n; \theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{ML}} = 2\sum_{i=1}^{n} \left.\frac{(x_i - \theta)}{2\sigma^2}\right|_{\theta=\hat{\theta}_{ML}} = 0,$$

or

$$\hat{\theta}_{ML}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i. \qquad (1)$$

Thus (1) represents the ML estimate for θ, which happens to be a linear estimator (linear function of the data) in this case.

Notice that the estimator is a r.v. Taking its expected value, we get

$$E[\hat{\theta}_{ML}(x)] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \theta,$$

i.e., the expected value of the estimator does not differ from the desired parameter, and hence there is no bias between the two. Such estimators are known as **unbiased estimators**. Thus (1) represents an unbiased estimator for θ. Moreover the variance of the estimator is given by

$$Var(\hat{\theta}_{ML}) = E[(\hat{\theta}_{ML} - \theta)^2] = \frac{1}{n^2} E\left\{\left(\sum_{i=1}^{n}(X_i - \theta)\right)^2\right\}$$

$$= \frac{1}{n^2}\left\{\sum_{i=1}^{n} E(X_i - \theta)^2 + \sum_{i=1}^{n}\sum_{j=1,\ i\neq j}^{n} E(X_i - \theta)(X_j - \theta)\right\}.$$

The later terms are zeros since X_i and X_j are independent r.vs.

Then

$$Var(\hat{\theta}_{ML}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Thus

$$Var(\hat{\theta}_{ML}) \to 0 \quad \text{as} \quad n \to \infty, \tag{2}$$

another desired property. We say such estimators (that satisfy (2)) are **consistent estimators**.

Next two examples show that ML estimator can be highly nonlinear.

Example 2: Let $X_1, X_2, \cdots, X_n$ be independent, identically distributed uniform random variables in the interval $(0, \theta)$ with common p.d.f

$$f_{X_i}(x_i;\theta) = \frac{1}{\theta}, \quad 0 < x_i < \theta,$$

where θ is an unknown parameter. Find the ML estimate for θ.

Solution: The likelihood function in this case is given by

$$f_X(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n; \theta) = \frac{1}{\theta^n}, \quad 0 < x_i \le \theta, \quad i = 1 \to n$$

$$= \frac{1}{\theta^n}, \quad 0 \le \max(x_1, x_2, \cdots, x_n) \le \theta.$$

The likelihood function in this case is maximized by the minimum value of θ, and since $\theta \ge \max(X_1, X_2, \cdots, X_n)$, we get

$$\hat{\theta}_{ML}(X) = \max(X_1, X_2, \cdots, X_n) \tag{3}$$

to be the ML estimate for θ. Notice that (3) represents a nonlinear function of the observations. To determine whether (3) represents an unbiased estimate for θ, we need to evaluate its mean. To accomplish that in this case, it is easier to determine its p.d.f and proceed directly. Let

$$Z = \max(X_1, X_2, \cdots, X_n)$$

with X_i s uniform. Then

$$F_Z(z) = P[\max(X_1, X_2, \cdots, X_n) \le z] = P(X_1 \le z, X_2 \le z, \cdots, X_n \le z)$$

$$= \prod_{i=1}^{n} P(X_i \le z) = \prod_{i=1}^{n} F_{X_i}(z) = \left(\frac{z}{\theta}\right)^n, \quad 0 < z < \theta,$$

so that

$$f_Z(z) = \begin{cases} \dfrac{nz^{n-1}}{\theta^n}, & 0 < z < \theta, \\ 0, & \text{otherwise} \quad . \end{cases}$$

Thus, we get

$$E[\hat{\theta}_{ML}(X)] = E(Z) = \int_0^\theta z\, f_Z(z) dz = \frac{n}{\theta^n} \int_0^\theta z^n dz = \frac{n}{n+1} \frac{\theta^{n+1}}{\theta^n} = \frac{\theta}{(1+1/n)} \quad . \qquad (4)$$

In this case $E[\hat{\theta}_{ML}(X)] \neq \theta,$ and hence the ML estimator is not an unbiased estimator for θ. However, from (4) as $n \to \infty$

$$\lim_{n\to\infty} E[\hat{\theta}_{ML}(X)] = \lim_{n\to\infty} \frac{\theta}{(1+1/n)} = \theta,$$

i.e., the ML estimator is an **asymptotically unbiased estimator**. We also get

$$E(Z^2) = \int_0^\theta z^2 f_Z(z)dz = \frac{n}{\theta^n}\int_0^\theta z^{n+1}dz = \frac{n\theta^2}{n+2}$$

so that

$$Var[\hat{\theta}_{ML}(X)] = E(Z^2) - [E(Z)]^2 = \frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2} = \frac{n\theta^2}{(n+1)^2(n+2)}.$$

Once again $Var[\hat{\theta}_{ML}(X)] \to 0$ as $n \to \infty$, implying that the estimator in (3) is a **consistent estimator**.

Example 3: Let $X_1, X_2, \cdots, X_n$ be i.i.d Gamma random variables with unknown parameters α and β. Determine the ML estimator for α and β.

Solution: Here $x_i \geq 0,$ and

$$f_X(x_1, x_2, \cdots, x_n; \alpha, \beta) = \frac{\beta^{n\alpha}}{(\Gamma(\alpha))^n} \prod_{i=1}^{n} x_i^{\alpha-1} \; e^{-\beta \sum_{i=1}^{n} x_i}.$$

This gives the log-likelihood function to be

$$L(x_1, x_2, \cdots, x_n; \alpha, \beta) = \log f_X(x_1, x_2, \cdots, x_n; \alpha, \beta)$$

$$= n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1)\left(\sum_{i=1}^{n} \log x_i \right) - \beta \sum_{i=1}^{n} x_i.$$

Differentiating L with respect to α and β we get

$$\frac{\partial L}{\partial \alpha} = n \log \beta - \frac{n}{\Gamma(\alpha)} \Gamma'(\alpha) + \sum_{i=1}^{n} \log x_i \Big|_{\alpha, \beta = \hat{\alpha}, \hat{\beta}} = 0, \tag{5}$$

$$\frac{\partial L}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i \Big|_{\alpha, \beta = \hat{\alpha}, \hat{\beta}} = 0.$$

Thus

$$\hat{\beta}_{ML}(X) = \frac{\hat{\alpha}_{ML}}{\frac{1}{n} \sum_{i=1}^{n} x_i}, \tag{6}$$

and substituting (6) into (5), it gives

$$\log \hat{\alpha}_{ML} - \frac{\Gamma'(\hat{\alpha}_{ML})}{\Gamma(\hat{\alpha}_{ML})} = \log\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - \frac{1}{n}\sum_{i=1}^{n} \log(x_i). \qquad (7)$$

Notice that (7) is highly nonlinear in $\hat{\alpha}_{ML}$.

In general the (log)-likelihood function can have more than one solution, or no solutions at all. Further, the (log)-likelihood function may not be even differentiable, or it can be extremely complicated to solve explicitly
(see example 3, equation (7)).

Best Unbiased Estimator:

Referring back to example 1, we have seen that (1) represents an unbiased estimator for θ with certain variance. It is possible that, for a given *n*, there may be other

unbiased estimators to this problem with even lower variances. If such is indeed the case, those estimators will be naturally preferable compared to (1). In a given scenario, is it possible to determine the lowest possible value for the variance of *any* unbiased estimator?

Cramer - Rao Bound: Variance of any unbiased estimator based on observations $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$ for θ must satisfy the lower bound

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)} \tag{8}$$

where $I(\theta)$ is called the **Fisher information**

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ln f_X(x_1,\cdots,x_n;\theta)\right)^2\right] = -E\left[\frac{\partial^2}{\partial\theta^2}\ln f_X(x_1,\cdots,x_n;\theta)\right]$$

To see why the two expressions are equal, we abbreviate $f_X(x_1,\cdots,x_n;\theta)$ by f, and note that

$$\frac{\partial^2 \ln f}{\partial\theta^2} = \frac{\partial}{\partial\theta}\left(\frac{\frac{\partial f}{\partial\theta}}{f}\right) = \frac{\frac{\partial^2 f}{\partial\theta^2}\cdot f - \left(\frac{\partial f}{\partial\theta}\right)^2}{f^2} = \frac{\frac{\partial^2 f}{\partial\theta^2}}{f} - \left(\frac{\partial \ln f}{\partial\theta}\right)^2$$

and that

$$E\left[\frac{\frac{\partial^2 f}{\partial\theta^2}}{f}\right] = \int \frac{\frac{\partial^2 f}{\partial\theta^2}}{f}\cdot f dx = \int \frac{\partial^2 f}{\partial\theta^2} dx = \frac{\partial^2}{\partial\theta^2}\int f dx = \frac{\partial^2}{\partial\theta^2} 1 = 0$$

Thus,

$$E\left[\frac{\partial^2 \ln f}{\partial\theta^2}\right] = -E\left[\left(\frac{\partial \ln f}{\partial\theta}\right)^2\right].$$

This important result states that the right side of (8) acts as a lower bound on the variance of *all* unbiased estimator for θ, provided their joint p.d.f satisfies certain regularity restrictions.

Naturally any unbiased estimator whose variance coincides with that in (8), must be the best. There are no better solutions! Such estimates are known as **efficient estimators**.

Example 1': Let us examine whether (1) represents an efficient estimator. Towards this we note

$$\left(\frac{\partial \ln f_X(x_1,\cdots,x_n;\theta)}{\partial\theta}\right)^2 = \frac{1}{\sigma^4}\left(\sum_{i=1}^{n}(X_i-\theta)\right)^2;$$

and

$$E\left(\frac{\partial \ln f_X(x_1,\cdots,x_n;\theta)}{\partial\theta}\right)^2 = \frac{1}{\sigma^4}\left\{\sum_{i=1}^{n}E[(X_i-\theta)^2]+\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}E[(X_i-\theta)(X_j-\theta)]\right\}$$

$$= \frac{1}{\sigma^4}\sum_{i=1}^{n}\sigma^2 = \frac{n}{\sigma^2},$$

and substituting this into the first form on the right side of (8), we obtain the Cramer - Rao lower bound for this problem to be

$$\frac{\sigma^2}{n}. \qquad (9)$$

But the variance of the ML estimator is the same as (9), implying that (1) indeed represents an efficient estimator in this case, the best of all possibilities!

It is possible that in certain cases there are no unbiased estimators that are efficient. In that case, the best estimator will be an unbiased estimator with the lowest possible variance.
How does one find such an unbiased estimator?
Fortunately Rao-Blackwell theorem (page 335-337, Text) gives a complete answer to this problem.
Cramer-Rao bound can be extended to multiparameter case as well (see page 343-345,Text).

So far, we discussed nonrandom parameters that are unknown. What if the parameter of interest is a r.v with a-priori p.d.f $f_\theta(\theta)$? How does one obtain a good estimate for θ based on the observations $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$?

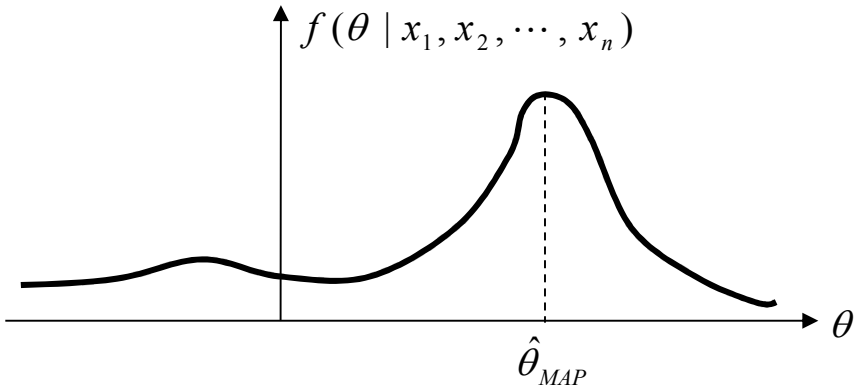
One technique is to use the observations to compute its a-posteriori probability density function $f_{\theta|X}(\theta \mid x_1, x_2, \cdots, x_n)$. Of course, we can use the Bayes' theorem to obtain this a-posteriori p.d.f. This gives

$$f_{\theta|X}(\theta \mid x_1, x_2, \cdots, x_n) = \frac{f_{X|\theta}(x_1, x_2, \cdots, x_n \mid \theta) f_\theta(\theta)}{f_X(x_1, x_2, \cdots, x_n)}. \qquad (10)$$

Notice that (10) is only a function of θ, since $x_1, x_2, \cdots, x_n$ represent given observations. Once again, we can look for

the most probable value of θ suggested by the above a-posteriori p.d.f. Naturally, the most likely value for θ is that corresponding to the maximum of the a-posteriori p.d.f (see Fig.). This estimator - maximum of the a-posteriori p.d.f is known as the **MAP estimator** for θ.

It is possible to use other optimality criteria as well. Of course, that should be the subject matter of another course!



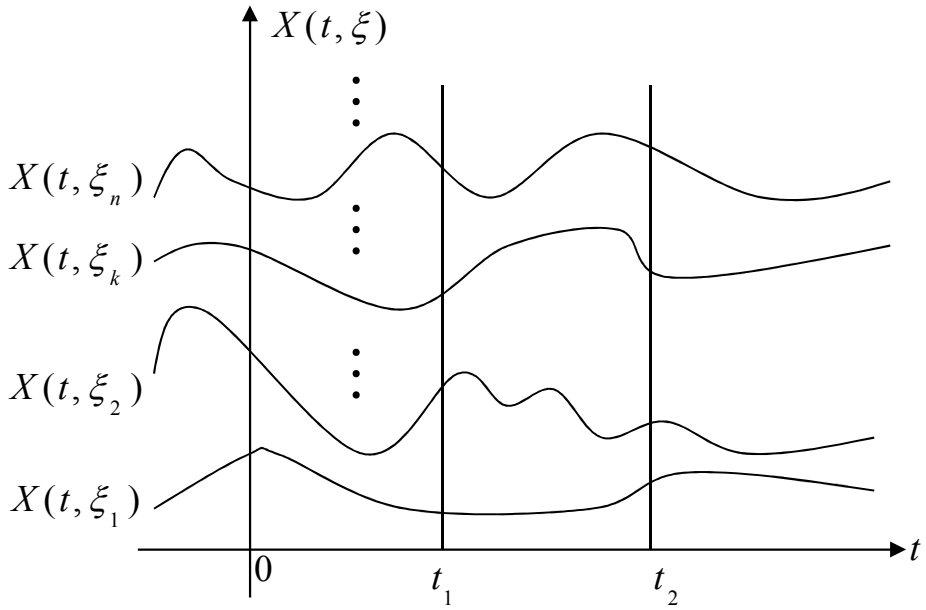
Imperial College
London

EE4-10: Probability and Stochastic Processes

Lecture 5: Stochastic Processes

Introduction

Let ξ denote the random outcome of an experiment. To every such outcome suppose a waveform $X(t,\xi)$ is assigned. The collection of such waveforms form a stochastic process. The set of $\{\xi_k\}$ and the time index t can be continuous or discrete (countably infinite or finite) as well.



For fixed $\xi_i \in S$ (the set of all experimental outcomes), $X(t,\xi)$ is a specific time function. For fixed t,

$$X_1 = X(t_1,\xi_i)$$

is a random variable. The ensemble of all such realizations $X(t,\xi)$ over time represents the stochastic

process $X(t)$. (see Figure). For example

$$X(t) = a\cos(\omega_0 t + \varphi),$$

where φ is a uniformly distributed random variable in $(0, 2\pi)$, represents a stochastic process. Stochastic processes are everywhere: Brownian motion, stock market fluctuations, various queuing systems all represent stochastic phenomena.

If $X(t)$ is a stochastic process, then for fixed t, $X(t)$ represents a random variable. Its distribution function is given by

$$F_X(x,t) = P\{X(t) \le x\}$$

Notice that $F_X(x,t)$ depends on t, since for a different t, we obtain a different random variable. Further

$$f_X(x,t) \triangleq \frac{dF_X(x,t)}{dx}$$

represents the first-order probability density function of the process $X(t)$.

For $t = t_1$ and $t = t_2$, $X(t)$ represents two different random variables $X_1 = X(t_1)$ and $X_2 = X(t_2)$ respectively. Their joint distribution is given by

$$F_X(x_1, x_2, t_1, t_2) = P\{X(t_1) \le x_1, X(t_2) \le x_2\}$$

and

$$f_X(x_1, x_2, t_1, t_2) \triangleq \frac{\partial^2 F_X(x_1, x_2, t_1, t_2)}{\partial x_1 \, \partial x_2}$$

represents the second-order density function of the process $X(t)$. Similarly $f_X(x_1, x_2, \cdots x_n, \; t_1, t_2 \cdots, t_n)$ represents the n^{th} order density function of the process $X(t)$. Complete specification of the stochastic process $X(t)$ requires the knowledge of $f_X(x_1, x_2, \cdots x_n, \; t_1, t_2 \cdots, t_n)$ for all $t_i, \quad i = 1, 2, \cdots, n$ and for all n. (an almost impossible task in reality).

Mean of a Stochastic Process:

$$\mu(t) \triangleq E\{X(t)\} = \int_{-\infty}^{+\infty} x f_X(x,t)dx$$

represents the mean value of a process $X(t)$. In general, the mean of a process can depend on the time index t.

Autocorrelation function of a process $X(t)$ is defined as

$$R_{XX}(t_1,t_2) \triangleq E\{X(t_1)X^*(t_2)\} = \int\int x_1 x_2^* f_X(x_1,x_2,t_1,t_2)dx_1 dx_2$$

where * denotes the complex conjugate, and it represents the interrelationship between the random variables $X_1 = X(t_1)$ and $X_2 = X(t_2)$ generated from the process $X(t)$.

Properties:

1. $R_{XX}(t_1,t_2) = R_{XX}^*(t_2,t_1) = [E\{X(t_2)X^*(t_1)\}]^*$

2. $R_{XX}(t,t) = E\{|X(t)|^2\} > 0.$ (Average instantaneous power)

3. $R_{XX}(t_1, t_2)$ represents a nonnegative definite function, i.e., for *any* set of constants $\{a_i\}_{i=1}^{n}$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j^* R_{XX}(t_i, t_j) \geq 0.$$

The inequality follows by noticing that $E\{|Y|^2\} \geq 0$ for $Y = \sum_{i=1}^{n} a_i X(t_i)$.
The function

$$C_{XX}(t_1, t_2) = R_{XX}(t_1, t_2) - \mu_X(t_1)\mu_X^*(t_2)$$

represents the **autocovariance** function of the process $X(t)$.

Example

$$X(t) = a\cos(\omega_0 t + \varphi), \qquad \varphi \sim U(0,2\pi).$$

This gives

$$\begin{aligned}\mu_X(t) &= E\{X(t)\} = aE\{\cos(\omega_0 t + \varphi)\} \\ &= a\cos\omega_0 t\, E\{\cos\varphi\} - a\sin\omega_0 t\, E\{\sin\varphi\} = 0,\end{aligned}$$

since $E\{\cos\varphi\} = \frac{1}{2\pi}\int_0^{2\pi}\cos\varphi\, d\varphi = 0 = E\{\sin\varphi\}.$

Similarly

$$\begin{aligned}R_{XX}(t_1,t_2) &= a^2 E\{\cos(\omega_0 t_1 + \varphi)\cos(\omega_0 t_2 + \varphi)\} \\ &= \frac{a^2}{2}E\{\cos\omega_0(t_1 - t_2) + \cos(\omega_0(t_1 + t_2) + 2\varphi)\} \\ &= \frac{a^2}{2}\cos\omega_0(t_1 - t_2).\end{aligned}$$

Stationary Stochastic Processes

Stationary processes exhibit statistical properties that are invariant to shift in the time index. Thus, for example, second-order stationarity implies that the statistical properties of the pairs $\{X(t_1), X(t_2)\}$ and $\{X(t_1+c), X(t_2+c)\}$ are the same for *any* c. Similarly first-order stationarity implies that the statistical properties of $X(t_i)$ and $X(t_i+c)$ are the same for any c.

In strict terms, the statistical properties are governed by the joint probability density function. Hence a process is n^{th}-order **Strict-Sense Stationary (S.S.S)** if

$$f_X(x_1, x_2, \cdots x_n,\ t_1, t_2 \cdots, t_n) \equiv f_X(x_1, x_2, \cdots x_n,\ t_1 + c, t_2 + c \cdots, t_n + c)$$

for *any* c, where the left side represents the joint density function of the random variables $X_1 = X(t_1),\ X_2 = X(t_2),\ \cdots,\ X_n = X(t_n)$ and the right side corresponds to the joint density function of the random variables $X_1' = X(t_1 + c),\ X_2' = X(t_2 + c),\ \cdots,\ X_n' = X(t_n + c)$.

A process $X(t)$ is said to be **strict-sense stationary** if the above condition is true for all $t_i,\ \ i = 1, 2, \cdots, n,\ \ n = 1, 2, \cdots$ and *any* c.

For a **first-order strict sense stationary process**,
we have

$$f_X(x,t) \equiv f_X(x,t+c)$$

for any c. In particular $c = -t$ gives

$$f_X(x,t) = f_X(x)$$

i.e., the first-order density of $X(t)$ is independent of t. In that case

$$E[X(t)] = \int_{-\infty}^{+\infty} x f(x) dx = \mu, \quad \text{a constant.} \qquad (1)$$

Similarly, for a **second-order strict-sense stationary process**
we have

$$f_X(x_1, x_2,\ t_1, t_2) \equiv f_X(x_1, x_2,\ t_1 + c, t_2 + c)$$

for any c. For $c = -t_2$ we get

$$f_X(x_1, x_2,\ t_1, t_2) \equiv f_X(x_1, x_2,\ t_1 - t_2)$$

i.e., the second order density function of a strict sense stationary process depends only on the difference of the time indices $t_1 - t_2 = \tau$. In that case the autocorrelation function is given by

$$\begin{aligned} R_{XX}(t_1, t_2) &\triangleq E\{X(t_1)X^*(t_2)\} \\ &= \int\int x_1 x_2^* f_X(x_1, x_2, \tau = t_1 - t_2) dx_1 dx_2 \\ &= R_{XX}(t_1 - t_2) \triangleq R_{XX}(\tau) = R_{XX}^*(-\tau), \end{aligned} \qquad (2)$$

i.e., the autocorrelation function of a second order strict-sense stationary process depends only on the difference of the time indices $\tau = t_1 - t_2$.

Notice that (1) and (2) are consequences of the stochastic process being first and second-order strict sense stationary.

On the other hand, the basic conditions for the first and second order stationarity –involving the p.d.f– are usually difficult to verify.

In that case, we often resort to a looser definition of stationarity, known as **Wide-Sense Stationarity (W.S.S),** by making use of

(1) and (2) as the necessary conditions. Thus, a process $X(t)$ is said to be **Wide-Sense Stationary** if

(i) $E\{X(t)\} = \mu$

and

(ii) $E\{X(t_1)X^*(t_2)\} = R_{XX}(t_1 - t_2)$,

i.e., for wide-sense stationary processes, the mean is a constant and the autocorrelation function depends only on the difference between the time indices. Notice that (i)-(ii) do not say anything about the nature of the probability density functions, and instead deal with the average behavior of the process. Since (i)-(ii) follow from the p.d.f, strict-sense stationarity always implies wide-sense stationarity. However, the converse is *not true* in general, the only exception being the Gaussian process.

This follows, since if $X(t)$ is a Gaussian process, then by definition $X_1 = X(t_1), X_2 = X(t_2), \cdots, X_n = X(t_n)$ are jointly Gaussian random variables for any $t_1, t_2 \cdots, t_n$ whose joint characteristic function is given by

$$\phi_{\underline{X}}(\omega_1,\omega_2,\cdots,\omega_n) = e^{j\sum_{k=1}^{n}\mu(t_k)\omega_k - \sum_{l,k}^{n}\sum C_{XX}(t_i,t_k)\omega_i\omega_k/2}$$

where $C_{XX}(t_i,t_k)$ is the autocovariance function. If $X(t)$ is wide-sense stationary, then we get

$$\phi_{\underline{X}}(\omega_1,\omega_2,\cdots,\omega_n) = e^{j\sum_{k=1}^{n}\mu\omega_k - \frac{1}{2}\sum_{1=1}^{n}\sum_{k=1}^{n}C_{XX}(t_i-t_k)\omega_i\omega_k}$$

and hence if the set of time indices are shifted by a constant c to generate a new set of jointly Gaussian random variables $X_1' = X(t_1+c)$, $X_2' = X(t_2+c),\cdots,X_n' = X(t_n+c)$ then their joint characteristic function is identical. Thus the set of random variables $\{X_i\}_{i=1}^{n}$ and $\{X_i'\}_{i=1}^{n}$ have the same joint probability distribution for all n and all c, establishing the strict sense stationarity of Gaussian processes from its wide-sense stationarity. To summarize

If *X*(*t*) is a Gaussian process, then

wide-sense stationarity $\Rightarrow$ strict-sense stationarity.

We also note that the process $X(t) = a\cos(\omega_0 t+\varphi)$, is wide-sense stationary, but not strict-sense stationary.

Ergodic Processes

A wide-sense stationary random process is **ergodic in the mean** if the time-average of $X(t)$ converges to the ensemble average:

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}X(t)dt = E[X(t)] = \mu$$

A wide-sense stationary random process is **ergodic in the autocorrelation** if the time-autocorrelation function converges to the ensemble-average autocorrelation:

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}X(t)X^{*}(t-\tau)dt = E[X(t)X^{*}(t-\tau)] = R_{XX}(\tau)$$

A random process is **ergodic** (in the wide sense) when it is ergodic in the mean and in its autocorrelation function.

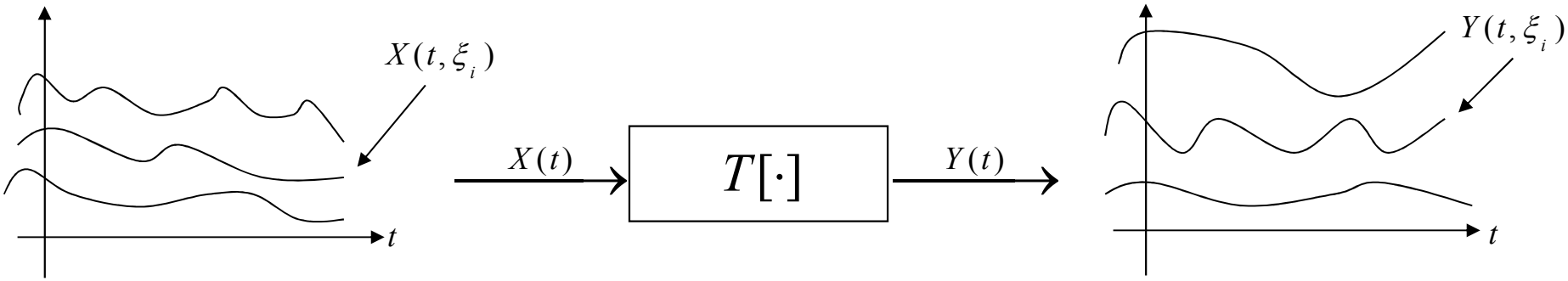
An important example of an ergodic processes is the stationary Gaussian process with continuous spectrum.

The implication of ergodicity is that it is possible to determine the statistical behavior by examining only one typical sample function, since almost every member of the ensemble shows the same statistical behavior as the whole ensemble.

For example, in electrical engineering, the mean corresponds to the DC component of a signal, while $R_{XX}(0)$ corresponds to the power. For ergodic processes, both could be measured in the lab using a signal over a sufficiently long interval.

Systems with Stochastic Inputs

A deterministic system[1] transforms each input waveform $X(t,\xi_i)$ into an output waveform $Y(t,\xi_i) = T[X(t,\xi_i)]$ by operating only on the time variable t. Thus a set of realizations at the input corresponding to a process $X(t)$ generates a new set of realizations $\{Y(t,\xi)\}$ at the output associated with a new process $Y(t)$.



Our goal is to study the output process statistics in terms of the input process statistics and the system function.

---

[1]A stochastic system on the other hand operates on both the variables t and ξ.

Deterministic Systems

Memoryless Systems

$$Y(t) = g[X(t)]$$

Systems with Memory

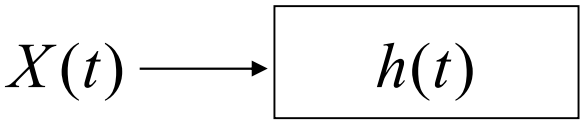
Time-varying systems

Time-Invariant systems

Linear systems

$$Y(t) = L[X(t)]$$

Linear-Time Invariant (LTI) systems



LTI system

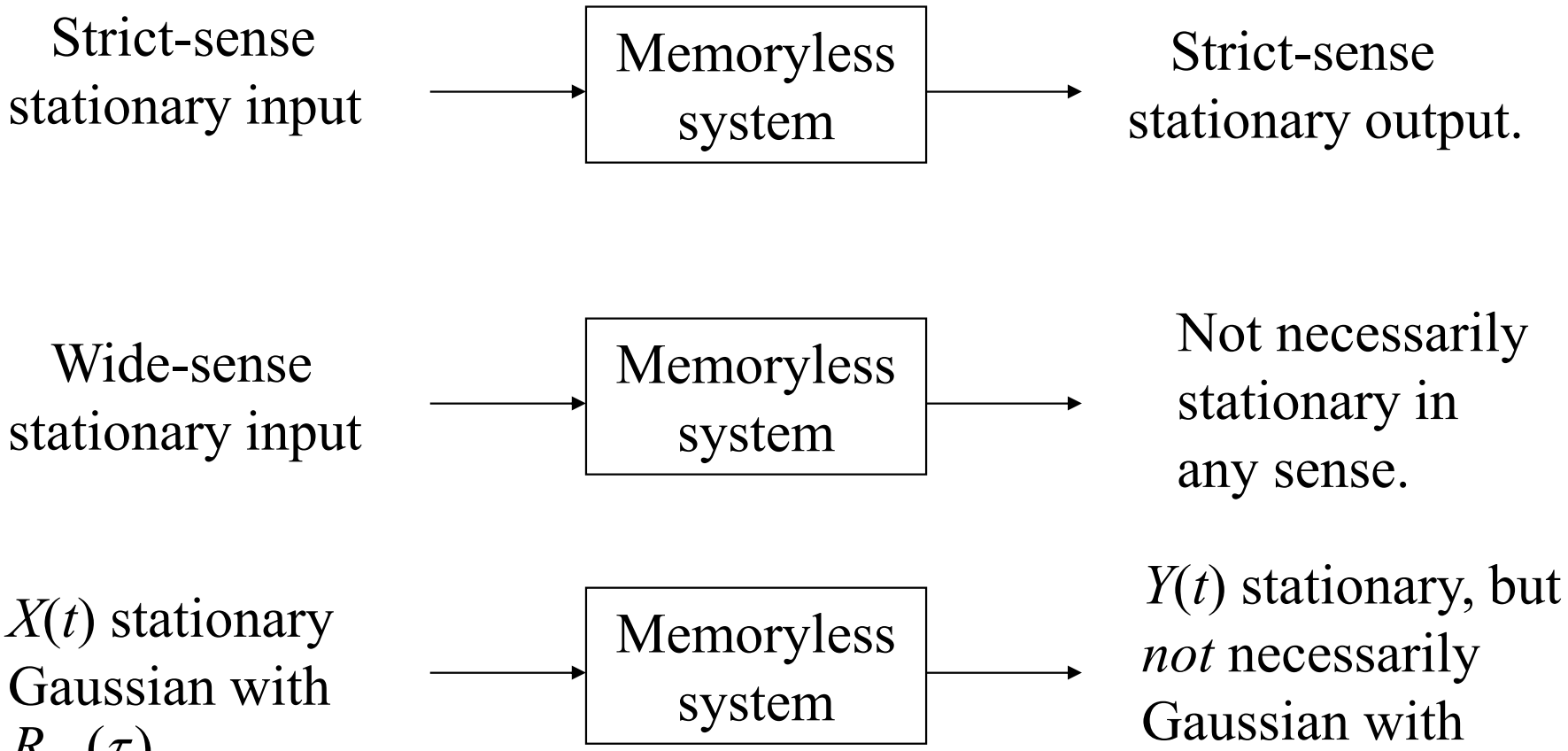
$$Y(t) = \int_{-\infty}^{+\infty} h(t-\tau)X(\tau)d\tau$$

$$= \int_{-\infty}^{+\infty} h(\tau)X(t-\tau)d\tau.$$

Memoryless Systems:

The output $Y(t)$ in this case depends only on the present value of the input $X(t)$. i.e.,

$$Y(t) = g\{X(t)\}$$



Linear Systems: $L[\cdot]$ represents a linear system if

$$L\{a_1X(t_1)+a_2X(t_2)\}=a_1L\{X(t_1)\}+a_2L\{X(t_2)\}. \qquad (3)$$

Let

$$Y(t)=L\{X(t)\}$$

represent the output of a linear system.

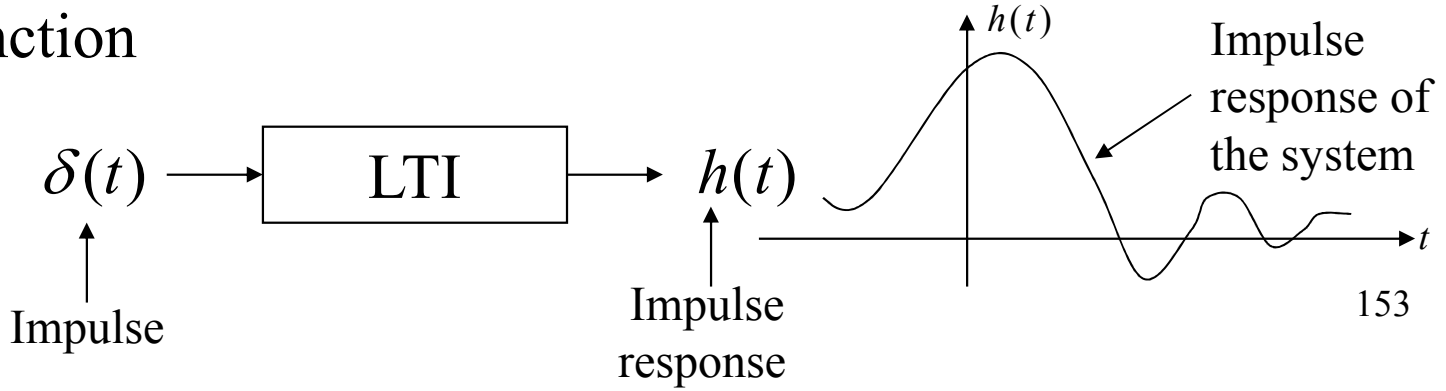
Time-Invariant System: $L[\cdot]$ represents a time-invariant system if

$$Y(t)=L\{X(t)\}\Rightarrow L\{X(t-t_0)\}=Y(t-t_0) \qquad (4)$$

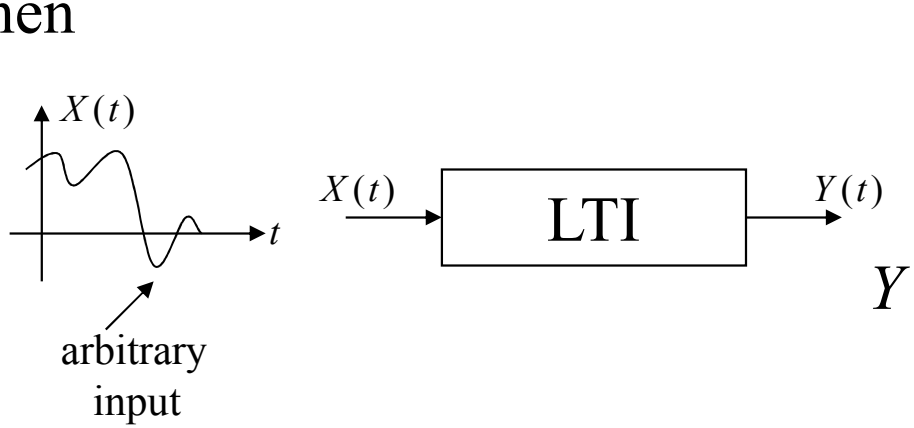
i.e., shift in the input results in the same shift in the output also.

If $L[\cdot]$ satisfies both (3) and (4), then it corresponds to a linear time-invariant (LTI) system.

LTI systems can be uniquely represented in terms of their output to a delta function



then



$$Y(t) = \int_{-\infty}^{+\infty} h(t-\tau)X(\tau)d\tau$$

$$= \int_{-\infty}^{+\infty} h(\tau)X(t-\tau)d\tau$$

It follows by expressing $X(t)$ as

$$X(t) = \int_{-\infty}^{+\infty} X(\tau)\delta(t-\tau)d\tau$$

Thus

$$Y(t) = L\{X(t)\} = L\{\int_{-\infty}^{+\infty} X(\tau)\delta(t-\tau)d\tau\}$$

$$= \int_{-\infty}^{+\infty} L\{X(\tau)\delta(t-\tau)d\tau\}$$ ← By Linearity

$$= \int_{-\infty}^{+\infty} X(\tau)L\{\delta(t-\tau)\}d\tau$$ ← By Time-invariance

$$= \int_{-\infty}^{+\infty} X(\tau)h(t-\tau)d\tau = \int_{-\infty}^{+\infty} h(\tau)X(t-\tau)d\tau.$$

Output Statistics: The mean of the output process is given by

$$\mu_Y(t) = E\{Y(t)\} = \int_{-\infty}^{+\infty} E\{X(\tau)h(t-\tau)d\tau\}$$

$$= \int_{-\infty}^{+\infty} \mu_X(\tau)h(t-\tau)d\tau = \mu_X(t) * h(t).$$

Similarly the cross-correlation function between the input and output processes is given by

$$R_{XY}(t_1,t_2) = E\{X(t_1)Y^*(t_2)\}$$

$$= E\{X(t_1)\int_{-\infty}^{+\infty} X^*(t_2-\alpha)h^*(\alpha)d\alpha\}$$

$$= \int_{-\infty}^{+\infty} E\{X(t_1)X^*(t_2-\alpha)\}h^*(\alpha)d\alpha$$

$$= \int_{-\infty}^{+\infty} R_{XX}(t_1,t_2-\alpha)h^*(\alpha)d\alpha$$

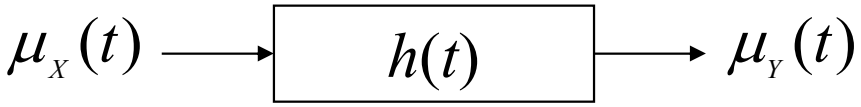
$$= R_{XX}(t_1,t_2) * h^*(t_2).$$

Finally the output autocorrelation function is given by

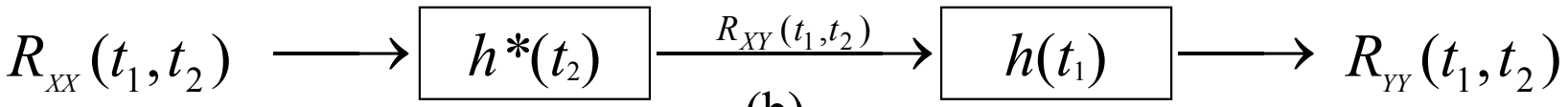
$$
\begin{aligned}
R_{YY}(t_1,t_2) &= E\{Y(t_1)Y^*(t_2)\} \\
&= E\{\int_{-\infty}^{+\infty} X(t_1-\beta)h(\beta)d\beta\, Y^*(t_2)\} \\
&= \int_{-\infty}^{+\infty} E\{X(t_1-\beta)Y^*(t_2)\}h(\beta)d\beta \\
&= \int_{-\infty}^{+\infty} R_{XY}(t_1-\beta,t_2)h(\beta)d\beta \\
&= R_{XY}(t_1,t_2) * h(t_1),
\end{aligned}
$$

or

$$
R_{YY}(t_1,t_2) = R_{XX}(t_1,t_2) * h^*(t_2) * h(t_1).
$$



(a)



(b)

In particular if $X(t)$ is wide-sense stationary, then we have $\mu_X(t) = \mu_X$ so that from

$$\mu_Y(t) = \mu_X \int_{-\infty}^{+\infty} h(\tau)d\tau = \mu_X c, \quad \text{a constant.}$$

Also $R_{XX}(t_1, t_2) = R_{XX}(t_1 - t_2)$ so that R_{XY} reduces to

$$\begin{aligned} R_{XY}(t_1, t_2) &= \int_{-\infty}^{+\infty} R_{XX}(t_1 - t_2 + \alpha)h^*(\alpha)d\alpha \\ &= R_{XX}(\tau) * h^*(-\tau) \stackrel{\Delta}{=} R_{XY}(\tau), \quad \tau = t_1 - t_2. \end{aligned}$$

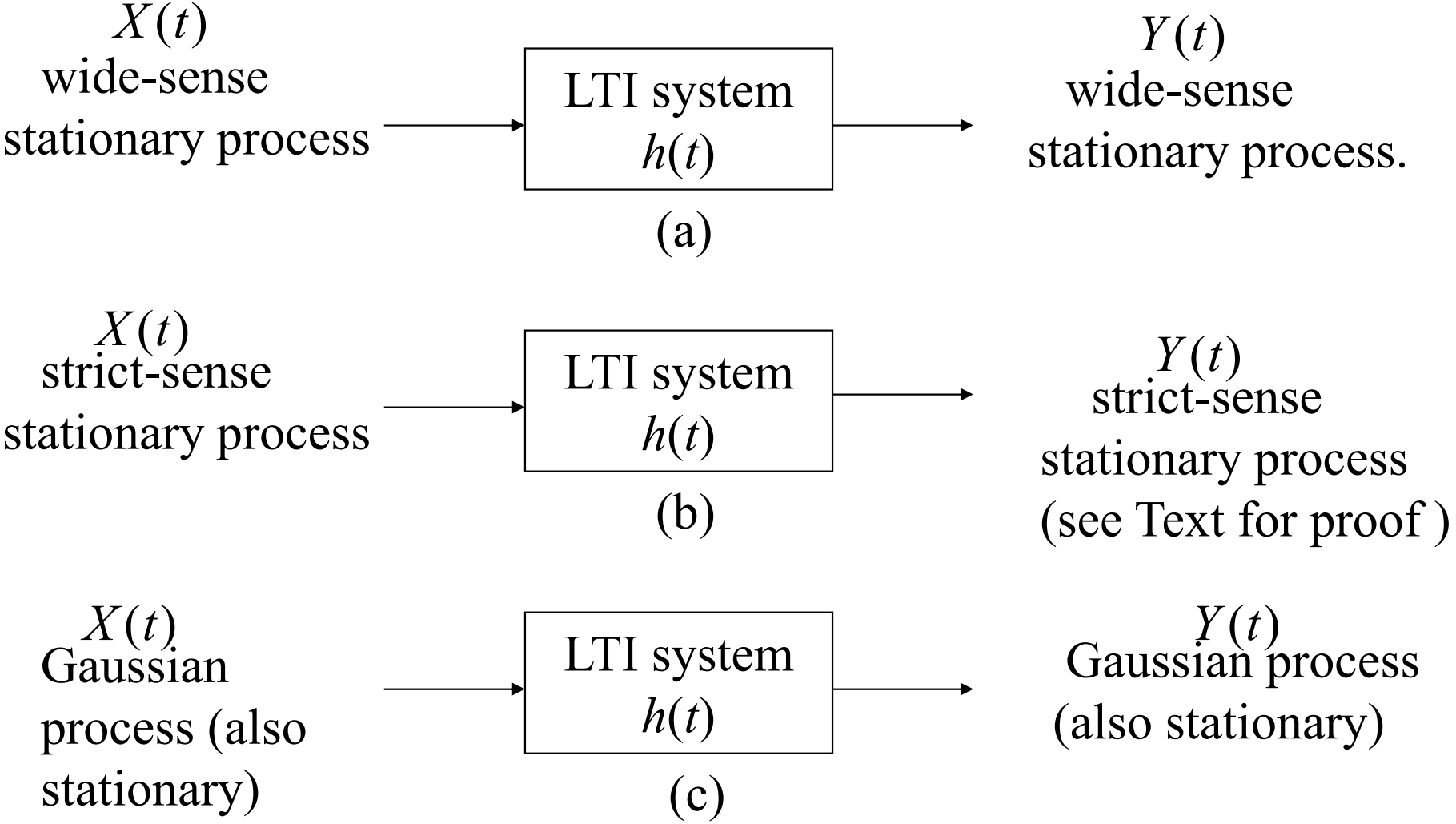
Thus $X(t)$ and $Y(t)$ are jointly w.s.s. Further, the output autocorrelation simplifies to

$$\begin{aligned} R_{YY}(t_1, t_2) &= \int_{-\infty}^{+\infty} R_{XY}(t_1 - \beta - t_2)h(\beta)d\beta, \quad \tau = t_1 - t_2 \\ &= R_{XY}(\tau) * h(\tau) = R_{YY}(\tau). \end{aligned}$$

We also obtain

$$R_{YY}(\tau) = R_{XX}(\tau) * h^*(-\tau) * h(\tau).$$

Thus, the output process is also wide-sense stationary.
This gives rise to the following representation



White Noise Process

$W(t)$ is said to be a white noise process if

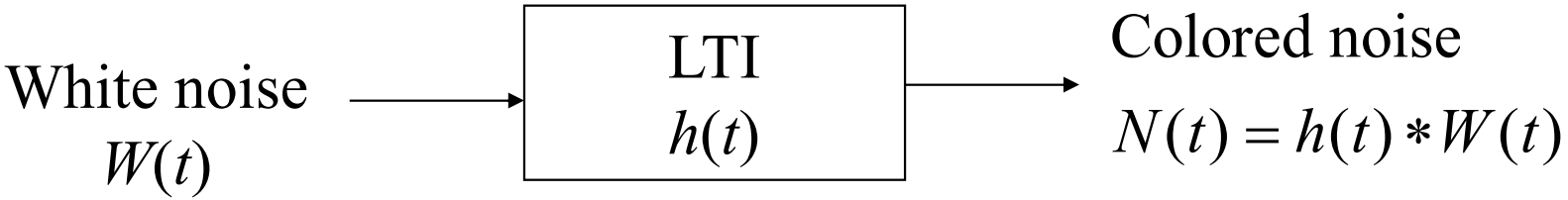
$$R_{WW}(t_1, t_2) = q(t_1)\delta(t_1 - t_2),$$

i.e., $E[W(t_1)\, W^*(t_2)] = 0$ unless $t_1 = t_2$.

$W(t)$ is said to be wide-sense stationary (w.s.s) white noise if $E[W(t)]$ = constant, and

$$R_{WW}(t_1, t_2) = q\delta(t_1 - t_2) = q\delta(\tau).$$

If $W(t)$ is also a Gaussian process (white Gaussian process), then all of its samples are independent random variables (why?).



For w.s.s. white noise input $W(t)$, we have

$$E[N(t)] = \mu_W \int_{-\infty}^{+\infty} h(\tau)d\tau, \quad \text{a constant}$$

and

$$\begin{aligned} R_{nn}(\tau) &= q\delta(\tau) * h^*(-\tau) * h(\tau) \\ &= qh^*(-\tau) * h(\tau) = q\rho(\tau) \end{aligned}$$

where

$$\rho(\tau) = h(\tau) * h^*(-\tau) = \int_{-\infty}^{+\infty} h(\alpha)h^*(\alpha - \tau)d\alpha.$$

Thus the output of a white noise process through an LTI system represents a (colored) noise process.

Note: White noise need not be Gaussian.

"White" and "Gaussian" are two different concepts!

Discrete Time Stochastic Processes

A discrete time stochastic process $X_n = X(nT)$ is a sequence of random variables. The mean, autocorrelation and auto-covariance functions of a discrete-time process are gives by

$$\mu_n = E\{X(nT)\}$$

$$R(n_1, n_2) = E\{X(n_1T)X^*(n_2T)\}$$

and

$$C(n_1, n_2) = R(n_1, n_2) - \mu_{n_1}\mu_{n_2}^*$$

respectively. As before strict sense stationarity and wide-sense stationarity definitions apply here also.
For example, $X(nT)$ is wide sense stationary if

$$E\{X(nT)\} = \mu, \quad \text{a constant}$$

and

$$E[X\{(k+n)T\}X^*\{(k)T\}] = R(n) = r_n \stackrel{\Delta}{=} r_{-n}^* \qquad (5)$$

i.e., $R(n_1, n_2) = R(n_1 - n_2) = R^*(n_2 - n_1)$. The positive-definite property of the autocorrelation sequence can be expressed in terms of certain Hermitian-Toeplitz matrices as follows:

Theorem: A sequence $\{r_n\}_{-\infty}^{+\infty}$ forms an autocorrelation sequence of a wide sense stationary stochastic process if and only if every Hermitian-Toeplitz matrix T_n given by

$$T_n = \begin{pmatrix} r_0 & r_1 & r_2 & \cdots & r_n \\ r_1^* & r_0 & r_1 & \cdots & r_{n-1} \\ & & \vdots & & \\ r_n^* & r_{n-1}^* & \cdots & r_1^* & r_0 \end{pmatrix} = T_n^*$$

is non-negative (positive) definite for $n = 0,\ 1,\ 2, \cdots, \infty$.

Proof: Let $\underline{a} = [a_0, a_1, \cdots, a_n]^T$ represent an arbitrary constant vector. Then

$$\underline{a}^* T_n \underline{a} = \sum_{i=0}^{n} \sum_{k=0}^{n} a_i a_k^* r_{k-i}$$

Using (5), it reduces to

$$\underline{a}^{*} T_n \underline{a} = \sum_{i=0}^{n}\sum_{k=0}^{n} a_i a_k^{*} E\{X(kT)X^{*}(iT)\} = E\left\{\left|\sum_{k=0}^{n} a_k^{*} X(kT)\right|^2\right\} \geq 0.$$

Thus, if $X(nT)$ is a wide sense stationary stochastic process then T_n is a non-negative definite matrix for every $n = 0, 1, 2, \cdots, \infty$. Similarly the converse also follows from this argument.

If $X(nT)$ represents a wide-sense stationary input to a discrete-time system $\{h(nT)\}$, and $Y(nT)$ the system output, then as before the cross correlation function satisfies

$$R_{XY}(n) = R_{XX}(n) * h^{*}(-n)$$

and the output autocorrelation function is given by

$$R_{YY}(n) = R_{XY}(n) * h(n)$$

or

$$R_{YY}(n) = R_{XX}(n) * h^{*}(-n) * h(n).$$

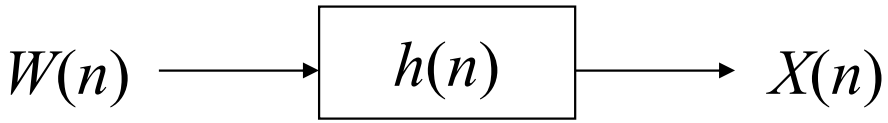
Thus wide-sense stationarity from input to output is preserved for discrete-time systems also.

Auto Regressive Moving Average (ARMA) Processes

Consider an input – output representation

$$X(n) = -\sum_{k=1}^{p} a_k X(n-k) + \sum_{k=0}^{q} b_k W(n-k), \qquad (6)$$

where $X(n)$ may be considered as the output of a system $\{h(n)\}$ driven by the input $W(n)$.



Z – transform of (6) gives

$$X(z)\sum_{k=0}^{p} a_k z^{-k} = W(z)\sum_{k=0}^{q} b_k z^{-k}, \quad a_0 \equiv 1$$

or

$$H(z) = \sum_{k=0}^{\infty} h(k) z^{-k} = \frac{X(z)}{W(z)} = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2} + \cdots + b_q z^{-q}}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_p z^{-p}} \triangleq \frac{B(z)}{A(z)}$$

represents the transfer function of the associated system response $\{h(n)\}$ so that

$$X(n) = \sum_{k=0}^{\infty} h(n-k)W(k).$$

Notice that the transfer function $H(z)$ is rational with p poles and q zeros that determine the model order of the underlying system. From (6), the output undergoes regression over p of its previous values and at the same time a moving average based on $W(n), W(n-1), \cdots, W(n-q)$ of the input over $(q + 1)$ values is added to it, thus generating an **Auto Regressive Moving Average** (ARMA (p, q)) process $X(n)$. Generally the input $\{W(n)\}$ represents a sequence of uncorrelated random variables of zero mean and constant variance σ_W^2 so that

$$R_{WW}(n) = \sigma_W^2 \delta(n).$$

If in addition, $\{W(n)\}$ is normally distributed then the output $\{X(n)\}$ also represents a strict-sense stationary normal process.

If $q = 0$, then (6) represents an AR(p) process (all-pole process), and if $p = 0$, then (6) represents an MA(q)

process (all-zero process). Next, we shall discuss the AR(1) process through explicit calculations.

AR(1) process: An AR(1) process has the form

$$X(n) = aX(n-1) + W(n)$$

and the corresponding system transfer

$$H(z) = \frac{1}{1 - az^{-1}} = \sum_{n=0}^{\infty} a^n z^{-n}$$

provided $|a| < 1$. Thus

$$h(n) = a^n, \qquad |a| < 1$$

represents the impulse response of an AR(1) stable system.

We get the output autocorrelation sequence of an AR(1) process to be

$$R_{XX}(n) = \sigma_W^2 \delta(n) * \{a^{-n}\} * \{a^n\} = \sigma_W^2 \sum_{k=0}^{\infty} a^{|n|+k} a^k = \sigma_W^2 \frac{a^{|n|}}{1 - a^2}$$

The normalized (in terms of $R_{XX}(0)$) output autocorrelation sequence is given by

$$\rho_X(n) = \frac{R_{XX}(n)}{R_{XX}(0)} = a^{|n|}, \quad |n| \geq 0.$$

Application of convolution and Fourier transform:

Integer multiplication with $O(n \log n)$ complexity, where n is the number of digits.

Imperial College
London

EE4-10: Probability and Stochastic Processes

Lecture 6: Power Spectrum

Reminder of Fourier Transforms

Fourier transform in angular frequency (rad/s)

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt$$

Inverse transform

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{j\omega t} d\omega$$

Fourier transform in ordinary frequency (Hz)

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi f t} dt$$

Inverse transform

$$x(t) = \int_{-\infty}^{\infty} X(f) e^{j2\pi f t} df$$

We will use the first definition. Students should be aware that in engineering applications (e.g., telecommunications) where frequency is measured in Hz, the second definition may be preferred.

Power Spectrum

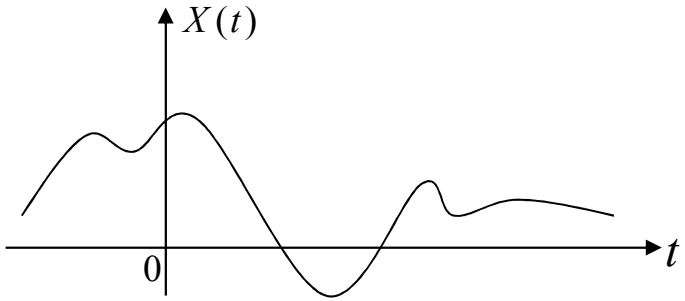
For a deterministic signal x(t), the spectrum is well defined: If $X(\omega)$ represents its Fourier transform, i.e., if

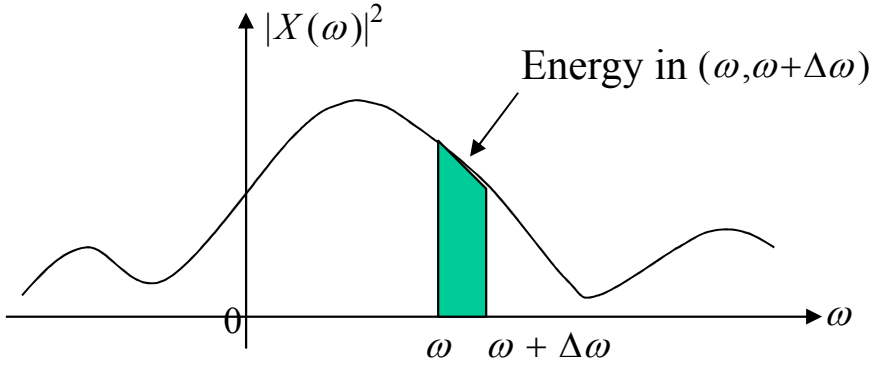
$$X(\omega) = \int_{-\infty}^{+\infty} x(t) e^{-j\omega t} dt,$$

then $|X(\omega)|^2$ represents its energy spectrum. This follows from Parseval's theorem since the signal energy is given by

$$\int_{-\infty}^{+\infty} x^2(t) dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |X(\omega)|^2 \, d\omega = E.$$

Thus $|X(\omega)|^2 \, \Delta\omega$ represents the signal energy in the band $(\omega, \omega + \Delta\omega)$ (see Figure).





However for stochastic processes, a direct application of this formula generates a sequence of random variables for every ω. Moreover, for a stochastic process, $E\{|X(t)|^2\}$ represents the ensemble average power (instantaneous energy) at the instant t.

To obtain the spectral distribution of power versus frequency for stochastic processes, it is best to avoid infinite intervals to begin with, and start with a finite interval $(-T,\ T)$. Formally, partial Fourier transform of a process $X(t)$ based on $(-T,\ T)$ is given by

$$X_T(\omega) = \int_{-T}^{T} X(t) e^{-j\omega t} dt$$

so that

$$\frac{|X_T(\omega)|^2}{2T} = \frac{1}{2T}\left|\int_{-T}^{T} X(t) e^{-j\omega t} dt\right|^2$$

represents the power distribution associated with that realization based on $(-T,\ T)$. Notice that this represents a random variable for every ω, and its ensemble average gives, the average power distribution based on $(-T,\ T)$. Thus power density

$$P_T(\omega) = E\left\{\frac{|X_T(\omega)|^2}{2T}\right\} = \frac{1}{2T}\int_{-T}^{T}\int_{-T}^{T} E\{X(t_1)X^*(t_2)\}e^{-j\omega(t_1-t_2)}dt_1dt_2$$

$$= \frac{1}{2T}\int_{-T}^{T}\int_{-T}^{T} R_{XX}(t_1,t_2)e^{-j\omega(t_1-t_2)}dt_1dt_2$$

represents the power distribution of $X(t)$ based on $(-T,\ T)$. For wide sense stationary (w.s.s) processes, further simplification is possible, i.e., if $X(t)$ is assumed to be w.s.s, then $R_{XX}(t_1,t_2) = R_{XX}(t_1-t_2)$ and we have

$$P_T(\omega) = \frac{1}{2T}\int_{-T}^{T}\int_{-T}^{T} R_{XX}(t_1-t_2)e^{-j\omega(t_1-t_2)}dt_1dt_2. \qquad (*)$$

Let $\tau = t_1 - t_2$, we get

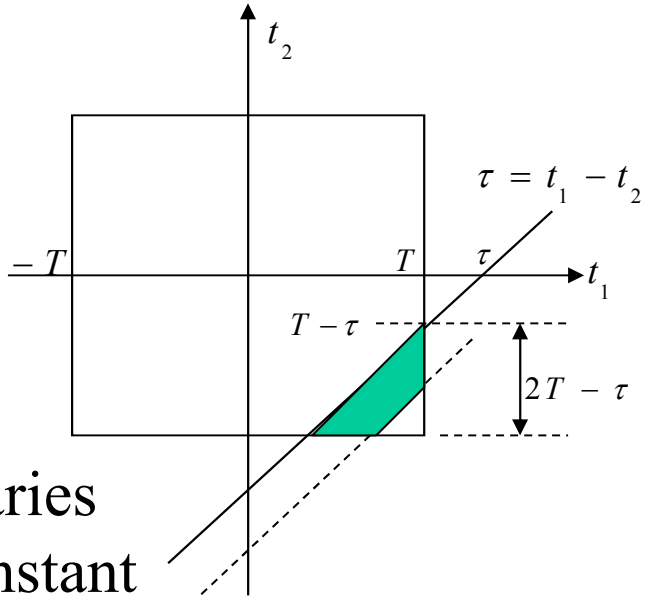
$$P_T(\omega) = \frac{1}{2T}\int_{-2T}^{2T} R_{XX}(\tau)e^{-j\omega\tau}(2T-|\tau|)d\tau \qquad (**)$$

$$= \int_{-2T}^{2T} R_{XX}(\tau)e^{-j\omega\tau}(1-\tfrac{|\tau|}{2T})d\tau \ \geq\ 0$$

to be the power distribution of the w.s.s. process $X(t)$ based on $(-T,\ T)$.

Derivation of Eq. (**)

The step from (*) to (**) in the last slide is a bit tricky. Details are given here.



As t_1, t_2 varies from $-T$ to $+T$, $\tau = t_1 - t_2$ varies from $-2T$ to $+2T$. Moreover $R_{XX}(\tau)$ is a constant over the shaded region in the figure, whose area is given by

$$\frac{1}{2}(2T-\tau)^2 - \frac{1}{2}(2T-\tau-d\tau)^2 = (2T-\tau)d\tau \qquad (\tau > 0)$$

and hence the integral (*) reduces to (**).

Finally letting $T \to \infty$, we obtain

$$S_{XX}(\omega) = \lim_{T\to\infty} P_T(\omega) = \int_{-\infty}^{+\infty} R_{XX}(\tau) e^{-j\omega\tau} d\tau \geq 0$$

to be the ***power spectral density*** of the w.s.s process $X(t)$. Notice that

$$R_{XX}(\tau) \quad \xleftrightarrow{\text{F·T}} \quad S_{XX}(\omega) \geq 0.$$

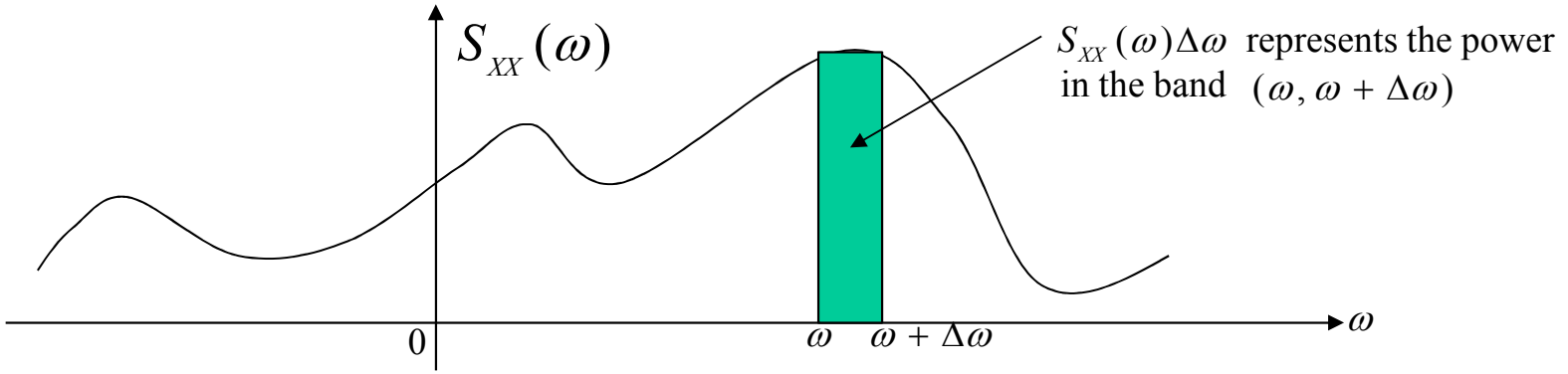
i.e., the autocorrelation function and the power spectrum of a w.s.s Process form a Fourier transform pair, a relation known as the **Wiener-Khinchin Theorem**. The inverse formula gives

$$R_{XX}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} S_{XX}(\omega) e^{j\omega\tau} d\omega$$

and in particular for $\tau = 0,$ we get

$$\frac{1}{2\pi}\int_{-\infty}^{+\infty} S_{XX}(\omega) d\omega = R_{XX}(0) = E\{|X(t)|^2\} = P, \quad \textit{the total power.}$$

Thus, the area under $S_{XX}(\omega)$ represents the total power of the process $X(t)$, and hence $S_{XX}(\omega)$ truly represents the power spectrum (see the following figure).



The nonnegative-definiteness property of the autocorrelation function translates into the "nonnegative" property for its Fourier transform (power spectrum), since

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j^* R_{XX}(t_i - t_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j^* \frac{1}{2\pi}\int_{-\infty}^{+\infty} S_{XX}(\omega) e^{j\omega(t_i - t_j)} d\omega$$

$$= \frac{1}{2\pi}\int_{-\infty}^{+\infty} S_{XX}(\omega)\left|\sum_{i=1}^{n} a_i e^{j\omega t_i}\right|^2 d\omega \geq 0.$$

From the above formula, it follows that

$$R_{XX}(\tau)\ \textit{nonnegative-definite} \quad \Leftrightarrow \quad S_{XX}(\omega) \geq 0.$$

If $X(t)$ is a real w.s.s process, then $R_{XX}(\tau) = R_{XX}(-\tau)$ so that

$$S_{XX}(\omega) = \int_{-\infty}^{+\infty} R_{XX}(\tau) e^{-j\omega\tau} d\tau$$

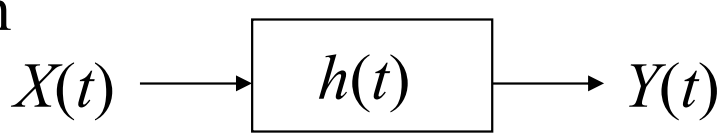
$$= \int_{-\infty}^{+\infty} R_{XX}(\tau) \cos\omega\tau d\tau$$

$$= 2\int_{0}^{\infty} R_{XX}(\tau) \cos\omega\tau d\tau = S_{XX}(-\omega) \geq 0$$

so that the power spectrum is an even function, (in addition to being real and nonnegative).

Power Spectra and Linear Systems

If a w.s.s process $X(t)$ with autocorrelation function $R_{XX}(\tau) \leftrightarrow S_{XX}(\tau) \geq 0$ is applied to a linear system with impulse response $h(t)$, the cross correlation function $R_{XY}(\tau)$ and the output autocorrelation function $R_{YY}(\tau)$ are given by



$$R_{XY}(\tau) = R_{XX}(\tau) * h^*(-\tau), \quad R_{YY}(\tau) = R_{XX}(\tau) * h^*(-\tau) * h(\tau).$$

But if

$$f(t) \leftrightarrow F(\omega), \quad g(t) \leftrightarrow G(\omega)$$

Then

$$f(t) * g(t) \leftrightarrow F(\omega)G(\omega)$$

since

$$\mathcal{F}\{f(t) * g(t)\} = \int_{-\infty}^{+\infty} f(t) * g(t) e^{-j\omega t} dt$$

$$\mathcal{F}\{f(t)*g(t)\}=\int_{-\infty}^{+\infty}\left\{\int_{-\infty}^{+\infty}f(\tau)g(t-\tau)d\tau\right\}e^{-j\omega t}dt$$

$$=\int_{-\infty}^{+\infty}f(\tau)e^{-j\omega\tau}d\tau\int_{-\infty}^{+\infty}g(t-\tau)e^{-j\omega(t-\tau)}d(t-\tau)$$

$$=F(\omega)G(\omega).$$

Using these relations we get

$$S_{XY}(\omega)=\mathcal{F}\{R_{XX}(\tau)*h^{*}(-\tau)\}=S_{XX}(\omega)H^{*}(\omega)$$

since

$$\int_{-\infty}^{+\infty}h^{*}(-\tau)e^{-j\omega\tau}d\tau=\left(\int_{-\infty}^{+\infty}h(t)e^{-j\omega t}dt\right)^{*}=H^{*}(\omega),$$

where

$$H(\omega)=\int_{-\infty}^{+\infty}h(t)e^{-j\omega t}dt$$

represents the transfer function of the system, and

$$S_{YY}(\omega)=\mathcal{F}\{R_{YY}(\tau)\}=S_{XY}(\omega)H(\omega)$$

$$=S_{XX}(\omega)\,|\,H(\omega)\,|^{2}\ .$$

Thus, the cross spectrum need not be real or nonnegative;
However the output power spectrum is real and nonnegative and is related to the input spectrum and the system transfer function.

W.S.S White Noise Process: If $W(t)$ is a w.s.s white noise process, then

$$R_{WW}(\tau) = q\delta(\tau) \quad \Rightarrow \quad S_{WW}(\omega) = q.$$

Thus the spectrum of a white noise process is flat, thus justifying its name. Notice that a white noise process is unrealizable since its total power is indeterminate.

Now, suppose the input to an unknown system is a white noise process, then the output spectrum is given by

$$S_{YY}(\omega) = q\,|H(\omega)|^2$$

Notice that the output spectrum captures the system transfer function characteristics entirely, and for rational systems one may determine the pole/zero locations of the underlying system.

Example 1: A w.s.s white noise process $W(t)$ is passed through a low pass filter (LPF) with bandwidth $B/2$. Find the autocorrelation function of the output process.

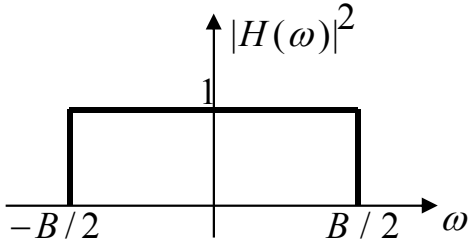
Solution: Let $X(t)$ represent the output of the LPF. Then

$$S_{XX}(\omega) = q\,|H(\omega)|^2 = \begin{cases} q, & |\omega| \le B/2 \\ 0, & |\omega| > B/2 \end{cases} .$$

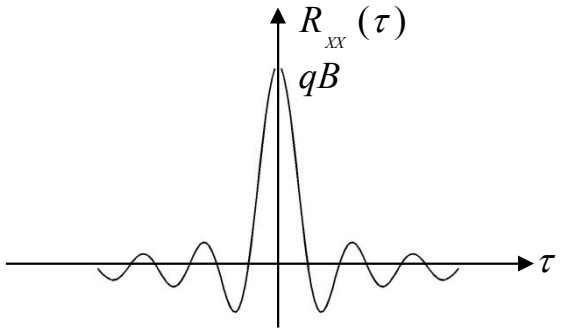
Inverse transform of $S_{XX}(\omega)$ gives the output autocorrelation function to be

$$R_{XX}(\tau) = \frac{1}{2\pi}\int_{-B/2}^{B/2} S_{XX}(\omega)e^{j\omega\tau}d\omega = \frac{1}{2\pi}q\int_{-B/2}^{B/2} e^{j\omega\tau}d\omega$$

$$= \frac{1}{2\pi}qB\,\frac{\sin(B\tau/2)}{(B\tau/2)} = \frac{1}{2\pi}qB\,\mathrm{sinc}(B\tau/2)$$



(a) LPF



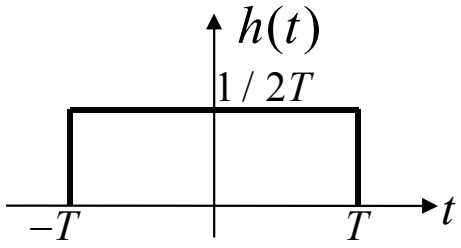
(b)

Example 2: Let

$$Y(t) = \frac{1}{2T}\int_{t-T}^{t+T} X(\tau)d\tau$$

represent a "smoothing" operation using a moving window on the input process $X(t)$. Find the spectrum of the output $Y(t)$ in term of that of $X(t)$.

Solution: If we define an LTI system with impulse response $h(t)$ as in the figure, then in term of $h(t)$, $Y(t)$ reduces to



$$Y(t) = \int_{-\infty}^{+\infty} h(t-\tau)X(\tau)d\tau = h(t) * X(t)$$

so that

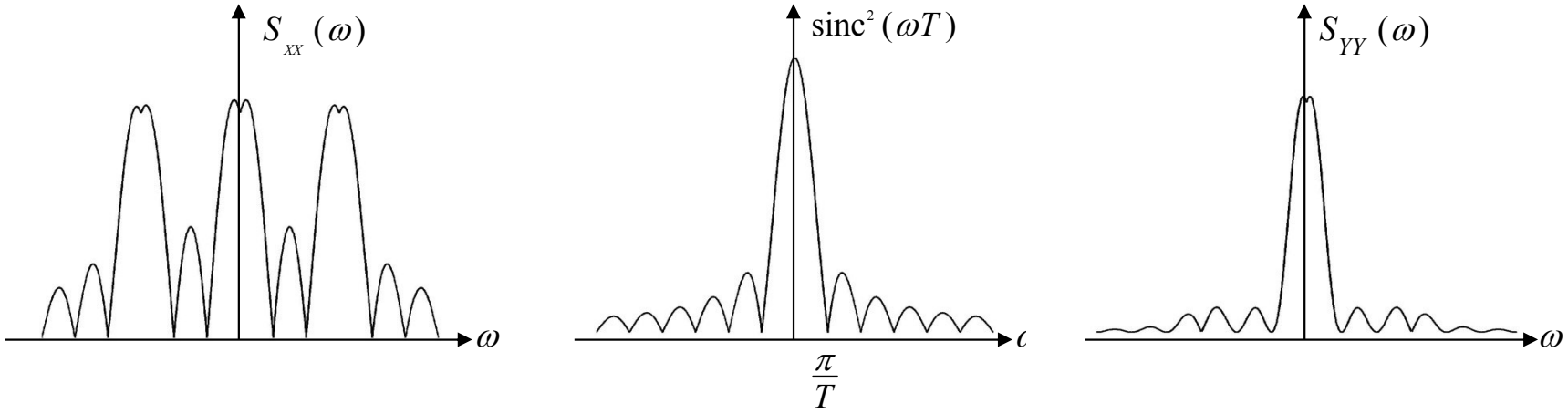
$$S_{YY}(\omega) = S_{XX}(\omega)\,|\,H(\omega)\,|^2 .$$

Here

$$H(\omega) = \int_{-T}^{+T} \tfrac{1}{2T} e^{-j\omega t} dt = \mathrm{sinc}(\omega T)$$

so that

$$S_{YY}(\omega) = S_{XX}(\omega)\,\mathrm{sinc}^2(\omega T).$$



Notice that the effect of the smoothing operation is to suppress the high frequency components in the input (beyond π / T), and the equivalent linear system acts as a low-pass filter (continuous-time moving average) with bandwidth $2\pi / T$ in this case.

Discrete – Time Processes

For discrete-time w.s.s stochastic processes $X(nT)$ with autocorrelation sequence $\{r_k\}_{-\infty}^{+\infty}$, (proceeding as above) or formally defining a continuous time process $X(t) = \sum_n X(nT)\delta(t - nT)$, we get the corresponding autocorrelation function to be

$$R_{XX}(\tau) = \sum_{k=-\infty}^{+\infty} r_k \delta(\tau - kT).$$

Its Fourier transform is given by

$$S_{XX}(\omega) = \sum_{k=-\infty}^{+\infty} r_k e^{-jk\omega T} \geq 0,$$

and it defines the power spectrum of the discrete-time process $X(nT)$. Clearly,

$$S_{XX}(\omega) = S_{XX}(\omega + 2\pi / T)$$

so that $S_{XX}(\omega)$ is a periodic function with period

$$2B = \frac{2\pi}{T}.$$

This gives the inverse relation

$$r_k = \frac{1}{2B}\int_{-B}^{B} S_{XX}(\omega)e^{jk\omega T}\,d\omega$$

and

$$r_0 = E\{|\,X(nT)\,|^2\} = \frac{1}{2B}\int_{-B}^{B} S_{XX}(\omega)\,d\omega$$

represents the total power of the discrete-time process $X(nT)$. The input-output relations for discrete-time system $h(nT)$ translate into

$$S_{XY}(\omega) = S_{XX}(\omega)H^*(e^{j\omega})$$

and

$$S_{YY}(\omega) = S_{XX}(\omega)\,|\,H(e^{j\omega})\,|^2$$

where

$$H(e^{j\omega}) = \sum_{n=-\infty}^{+\infty} h(nT)\,e^{-j\omega nT}$$

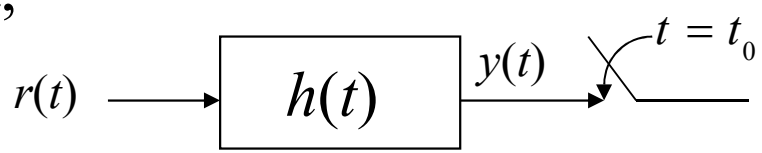
represents the discrete-time system transfer function.

Matched Filter

Let $r(t)$ represent a deterministic signal $s(t)$ corrupted by noise. Thus

$$r(t) = s(t) + w(t), \qquad 0 < t < t_0$$

where $r(t)$ represents the observed data, and it is passed through a receiver with impulse response $h(t)$. The output $y(t)$ is given by



Matched Filter

$$y(t) \triangleq y_s(t) + n(t)$$

where

$$y_s(t) = s(t) * h(t), \qquad n(t) = w(t) * h(t),$$

and it can be used to make a decision about the presence of absence of $s(t)$ in $r(t)$. Towards this, one approach is to require that the receiver output signal to noise ratio $(SNR)_0$ at time instant t_0 be maximized. Notice that

$$(SNR)_0 \triangleq \frac{\text{Output signal power at } t = t_0}{\text{Average output noise power}} = \frac{|y_s(t_0)|^2}{E\{|n(t)|^2\}}$$

$$= \frac{|y_s(t_0)|^2}{\frac{1}{2\pi}\int_{-\infty}^{+\infty} S_{nn}(\omega)d\omega} = \frac{\left|\frac{1}{2\pi}\int_{-\infty}^{+\infty} S(\omega)H(\omega)e^{j\omega t_0}d\omega\right|^2}{\frac{1}{2\pi}\int_{-\infty}^{+\infty} S_{WW}(\omega)|H(\omega)|^2\, d\omega} \qquad (1)$$

represents the output SNR, and the problem is to maximize $(\text{SNR})_0$ by optimally choosing the receiver filter $H(\omega)$.

Optimum Receiver for White Noise Input: The simplest input noise model assumes $w(t)$ to be white noise with spectral density N_0, so that (1) simplifies to

$$(SNR)_0 = \frac{\left|\int_{-\infty}^{+\infty} S(\omega)H(\omega)e^{j\omega t_0}d\omega\right|^2}{2\pi N_0\int_{-\infty}^{+\infty}|H(\omega)|^2\, d\omega} \qquad (2)$$

and a direct application of Cauchy-Schwarz' inequality in (2) gives

$$(SNR)_0 \le \frac{1}{2\pi N_0}\int_{-\infty}^{+\infty} |S(\omega)|^2\, d\omega = \frac{\int_0^{+\infty} s(t)^2\, dt}{N_0} = \frac{E_s}{N_0} \qquad (3)$$

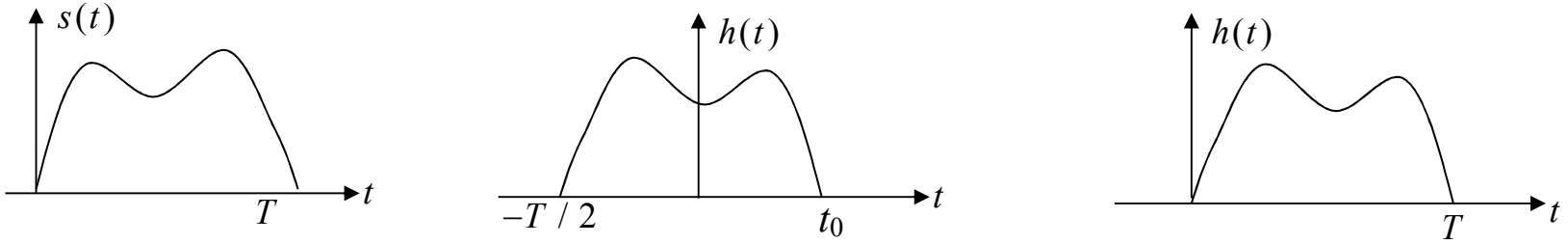
and equality in (3) is guaranteed if and only if

$$H(\omega) = S^*(\omega)e^{-j\omega t_0} \qquad (4)$$

or

$$h(t) = s(t_0 - t). \qquad (5)$$

From (3), the optimum receiver that maximizes the output SNR at $t = t_0$ is given by (4)-(5). Notice that (5) need not be causal, and the corresponding SNR is given by (3).



(a) (b) $t_0=T/2$ (c) $t_0=T$

The above figure shows the optimum $h(t)$ for two different values of t_0. In Fig (b), the receiver is noncausal, whereas in Fig (c) the receiver represents a causal waveform.

If the receiver is causal, the optimum causal receiver can be shown to be

$$h_{opt}(t) = s(t_0 - t)u(t)$$

and the corresponding maximum $(SNR)_0$ in that case is given by

$$(SNR_0) = \frac{1}{N_0}\int_0^{t_0} s^2(t)dt$$

Imperial College
London

EE4-10: Probability and Stochastic Processes

Lecture 7: Mean-Square Estimation

Mean Square Estimation

Given some information that is related to an unknown quantity of interest, the problem is to obtain a good estimate for the unknown in terms of the observed data.

Suppose $X_1, X_2, \cdots, X_n$ represent a sequence of random variables about whom one set of observations are available, and Y represents an unknown random variable. The problem is to obtain a good estimate for Y in terms of the observations $X_1, X_2, \cdots, X_n$.

Let

$$\hat{Y} = \varphi(X_1, X_2, \cdots, X_n) = \varphi(\underline{X})$$

represent such an estimate for Y.

Note that $\varphi(\cdot)$ can be a linear or a nonlinear function of the observation $X_1, X_2, \cdots, X_n$. Clearly

$$\varepsilon(\underline{X}) = Y - \hat{Y} = Y - \varphi(\underline{X})$$

represents the error in the above estimate, and $|\varepsilon|^2$ the square of

the error. Since ε is a random variable, $E\{|\varepsilon|^2\}$ represents the mean square error. One strategy to obtain a good estimator would be to minimize the mean square error by varying over all possible forms of $\varphi(\cdot)$, and this procedure gives rise to the **Minimization of the Mean Square Error (MMSE)** criterion for estimation. Thus under MMSE criterion, the estimator $\varphi(\cdot)$ is chosen such that the mean square error $E\{|\varepsilon|^2\}$ is at its minimum.

Next we show that the conditional mean of Y given $\underline{X}$ is the best estimator in the above sense.

Theorem 1: Under MMSE criterion, the best estimator for the unknown Y in terms of $X_1, X_2, \cdots, X_n$ is given by the conditional mean of Y gives $\underline{X}$. Thus

$$\hat{Y} = \varphi(\underline{X}) = E\{Y \mid \underline{X}\}.$$

Proof : Let $\hat{Y} = \varphi(\underline{X})$ represent an estimate of Y in terms of $\underline{X} = (X_1, X_2, \cdots, X_n)$. Then the error $\varepsilon = Y - \hat{Y}$, and the mean square error is given by

$$\sigma_\varepsilon^2 = E\{|\varepsilon|^2\} = E\{|Y - \hat{Y}|^2\} = E\{|Y - \varphi(\underline{X})|^2\}$$

Since

$$E[z] = E_X[E_z\{z \mid \underline{X}\}]$$

we can rewrite the mean square error as

$$\sigma_\varepsilon^2 = E\{\underbrace{|Y - \varphi(\underline{X})|^2}_{z}\} = E_{\underline{X}}[E_Y\{\underbrace{|Y - \varphi(\underline{X})|^2}_{z} \big| \underline{X}\}]$$

where the inner expectation is with respect to Y, and the outer one is with respect to $\underline{X}$.

Thus

$$\sigma_\varepsilon^2 = E[E\{|Y - \varphi(\underline{X})|^2 \, \big| \underline{X}\}]$$

$$= \int_{-\infty}^{+\infty} E\{|Y - \varphi(\underline{X})|^2 \, \big| \underline{X}\} f_X(\underline{X}) dx.$$

To obtain the best estimator φ, we need to minimize σ_ε^2 with respect to φ. Since $f_{\underline{X}}(\underline{X}) \geq 0$, $E\{|Y - \varphi(\underline{X})|^2 \, \big| \underline{X}\} \geq 0$, and the variable φ appears only in the integrand term, minimization of the mean square error σ_ε^2 with respect to φ is equivalent to minimization of $E\{|Y - \varphi(\underline{X})|^2 \, \big| \underline{X}\}$ with respect to φ.

Since $\underline{X}$ is fixed at some value, $\varphi(\underline{X})$ is no longer random, and hence minimization of $E\{|Y-\varphi(\underline{X})|^2 \,|\, \underline{X}\}$ is equivalent to

$$\frac{\partial}{\partial \varphi} E\{|Y-\varphi(\underline{X})|^2 \,|\, \underline{X}\} = 0.$$

This gives

$$E\{Y-\varphi(\underline{X}) \,|\, \underline{X}\} = 0$$

or

$$E\{Y \,|\, \underline{X}\} - E\{\varphi(\underline{X}) \,|\, \underline{X}\} = 0.$$

But

$$E\{\varphi(\underline{X}) \,|\, \underline{X}\} = \varphi(\underline{X}),$$

since when $\underline{X} = \underline{x}, \varphi(\underline{X})$ is a fixed number $\varphi(\underline{x})$.

We get the desired estimator to be

$$\hat{Y} = \varphi(\underline{X}) = E\{Y \mid \underline{X}\} = E\{Y \mid X_1, X_2, \cdots, X_n\}.$$

Thus the conditional mean of Y given $X_1, X_2, \cdots, X_n$ represents the best estimator for Y that minimizes the mean square error.

The minimum value of the mean square error is given by

$$\sigma^2_{\min} = E\{|Y - E(Y \mid \underline{X})|^2\} = E[\underbrace{E\{|Y - E(Y \mid \underline{X})|^2 \,|\, \underline{X}\}}_{\text{var}(Y|\underline{X})}]$$

$$= E\{\text{var}(Y \mid \underline{X})\} \geq 0.$$

As an example, suppose $Y = X^3$ is the unknown. Then the best MMSE estimator is given by

$$\hat{Y} = E\{Y \mid X\} = E\{X^3 \mid X\} = X^3.$$

Clearly if $Y = X^3$, then indeed $\hat{Y} = X^3$ is the best estimator for Y

in terms of X. Thus the best estimator can be nonlinear.

Next, we will consider a less trivial example.

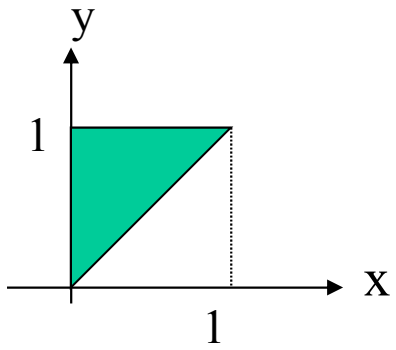
Example : Let

$$f_{X,Y}(x,y) = \begin{cases} kxy, & 0 < x < y < 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $k > 0$ is a suitable normalization constant. To determine the best estimate for Y in terms of X, we need $f_{Y|X}(y \mid x)$.

$$f_X(x) = \int_x^1 f_{X,Y}(x,y)dy = \int_x^1 kxydy$$

$$= \frac{kxy^2}{2}\Big|_x^1 = \frac{kx(1-x^2)}{2}, \quad 0 < x < 1.$$



Thus

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{kxy}{kx(1-x^2)/2} = \frac{2y}{1-x^2}; \quad 0 < x < y < 1.$$

Hence the best MMSE estimator is given by

$$\hat{Y} = \varphi(X) = E\{Y \mid \underline{X}\} = \int_x^1 y\, f_{Y|X}(y \mid x)dy$$

$$= \int_x^1 y \frac{2y}{1-x^2} dy = \frac{2}{1-x^2}\int_x^1 y^2 dy$$

$$= \frac{2}{3}\frac{y^3}{1-x^2}\Big|_x^1 = \frac{2}{3}\frac{1-x^3}{1-x^2} = \frac{2}{3}\frac{(1+x+x^2)}{1+x}.$$

Once again the best estimator is nonlinear. In general the best estimator $E\{Y \mid \underline{X}\}$ is difficult to evaluate, and hence next we will examine the special subclass of best linear estimators.

Linear MMSE Estimator

In this case the estimator $\hat{Y}$ is a linear function of the observations $X_1, X_2, \cdots, X_n$. Thus

$$\hat{Y}_l = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n = \sum_{i=1}^{n} a_i X_i.$$

where $a_1, a_2, \cdots, a_n$ are unknown quantities to be determined. The mean square error is given by $(\varepsilon = Y - \hat{Y}_l)$

$$E\{|\varepsilon|^2\} = E\{|Y - \hat{Y}_l|^2\} = E\{|Y - \sum a_i X_i|^2\}$$

and under the MMSE criterion $a_1, a_2, \cdots, a_n$ should be chosen so that the mean square error $E\{|\varepsilon|^2\}$ is at its minimum possible value. Let σ_n^2 represent that minimum possible value. Then

$$\sigma_n^2 = \min_{a_1, a_2, \cdots, a_n} E\{|Y - \sum_{i=1}^{n} a_i X_i|^2\}.$$

To minimize it, we can equate

$$\frac{\partial}{\partial a_k} E\{|\varepsilon|^2\} = 0, \quad k = 1,2,\cdots,n.$$

This gives

$$\frac{\partial}{\partial a_k} E\{|\varepsilon|^2\} = E\left\{\frac{\partial |\varepsilon|^2}{\partial a_k}\right\} = 2E\left[\varepsilon\left\{\frac{\partial \varepsilon}{\partial a_k}\right\}^*\right] = 0.$$

But

$$\frac{\partial \varepsilon}{\partial a_k} = \frac{\partial (Y - \sum_{i=1}^{n} a_i X_i)}{\partial a_k} = \frac{\partial Y}{\partial a_k} - \frac{\partial (\sum_{i=1}^{n} a_i X_i)}{\partial a_k} = -X_k .$$

Thus, we get

$$\frac{\partial E\{|\varepsilon|^2\}}{\partial a_k} = -2E\{\varepsilon X_k^*\} = 0,$$

or the best linear estimator must satisfy

$$E\{\varepsilon X_k^*\} = 0, \quad k = 1,2,\cdots,n. \tag{1}$$

Notice that in (1), ε represents the estimation error $(Y - \sum_{i=1}^{n} a_i X_i)$, and X_k, $k = 1 \to n$ represents the data. Thus from (1), the error ε is orthogonal to the data X_k, $k = 1 \to n$ for the best linear estimator. This is the **orthogonality principle**.

In other words, in the linear estimator, the unknown constants $a_1, a_2, \cdots, a_n$ must be selected such that the error

$\varepsilon = Y - \sum_{i=1}^{n} a_i X_i$ is orthogonal to every data $X_1, X_2, \cdots, X_n$ for the best linear estimator that minimizes the mean square error.

Interestingly a general form of the orthogonality principle holds good in the case of nonlinear estimators also.

Nonlinear Orthogonality Rule: Let $h(\underline{X})$ represent *any* functional form of the data and $E\{Y \mid \underline{X}\}$ the best estimator for Y given $\underline{X}$. With $\varepsilon = Y - E\{Y \mid \underline{X}\}$ we shall show that

$$E\{\varepsilon h(\underline{X})\} = 0,$$

implying that

$$\varepsilon = Y - E\{Y \mid \underline{X}\} \quad \perp \quad h(\underline{X}).$$

This follows since

$$\begin{aligned} E\{\varepsilon h(\underline{X})\} &= E\{(Y - E[Y \mid \underline{X}])h(\underline{X})\} \\ &= E\{Yh(\underline{X})\} - E\{E[Y \mid \underline{X}]h(\underline{X})\} \\ &= E\{Yh(\underline{X})\} - E\{E[Yh(\underline{X}) \mid \underline{X}]\} \\ &= E\{Yh(\underline{X})\} - E\{Yh(\underline{X})\} = 0. \end{aligned}$$

Thus in the nonlinear version of the orthogonality rule the error is orthogonal to *any* functional form of the data.

The orthogonality principle can be used to obtain the unknowns $a_1, a_2, \cdots, a_n$ in the linear case.

For example suppose $n = 2$, and we need to estimate Y in terms of X_1 and X_2 linearly. Thus

$$\hat{Y}_l = a_1 X_1 + a_2 X_2$$

Now, the orthogonality rule gives

$$E\{\varepsilon \, \mathrm{X}_1^*\} = E\{(Y - a_1 X_1 - a_2 X_2) X_1^*\} = 0$$

$$E\{\varepsilon \, \mathrm{X}_2^*\} = E\{(Y - a_1 X_1 - a_2 X_2) X_2^*\} = 0$$

Thus

$$E\{|\, X_1 \,|^2\} a_1 + E\{X_2 X_1^*\} a_2 = E\{Y X_1^*\}$$

$$E\{X_1 X_2^*\} a_1 + E\{|\, X_2 \,|^2\} a_2 = E\{Y X_2^*\}$$

or

$$\begin{pmatrix} E\{|X_1|^2\} & E\{X_2X_1^*\} \\ E\{X_1X_2^*\} & E\{|X_2|^2\} \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} E\{YX_1^*\} \\ E\{YX_2^*\} \end{pmatrix}$$

This equation can be solved to obtain a_1 and a_2 in terms of the cross-correlations.

The minimum value of the mean square error σ_n^2 is given by

$$\sigma_n^2 = \min_{a_1,a_2,\cdots,a_n} E\{|\varepsilon|^2\}$$

$$= \min_{a_1,a_2,\cdots,a_n} E\{\varepsilon\varepsilon^*\} = \min_{a_1,a_2,\cdots,a_n} E\{\varepsilon(Y - \sum_{i=1}^{n} a_i X_i)^*\}$$

$$= \min_{a_1,a_2,\cdots,a_n} E\{\varepsilon\ Y^*\} - \min_{a_1,a_2,\cdots,a_n} \sum_{i=1}^{n} a_i E\{\varepsilon\ X_i^*\}.$$

But the second term in the above equation is zero, since the error is orthogonal to the data X_i, where $a_1, a_2, \cdots, a_n$ are chosen to be optimum. Thus the minimum value of the mean square error is given by

$$\sigma_n^2 = E\{\varepsilon\, Y^*\} = E\{(Y - \sum_{i=1}^{n} a_i X_i) Y^*\}$$

$$= E\{|\, Y\,|^2\} - \sum_{i=1}^{n} a_i E\{X_i Y^*\}$$

where $a_1, a_2, \cdots, a_n$ are the optimum values.

Since the linear estimate is only a special case of the general estimator $\varphi(\underline{X})$, the best linear estimator cannot be superior to the best nonlinear estimator $E\{Y \mid \underline{X}\}$. Often the best linear estimator will be inferior to the best estimator.

This raises the following question. Are there situations in which the best estimator also turns out to be linear? In those situations it is enough to use linear MMSE and obtain the best linear estimators, since they also represent the best global estimators. Such is the case if Y and $X_1, X_2, \cdots, X_n$ are distributed as jointly Gaussian.

We summarize this in the next theorem and prove that result.

Theorem 2: If $X_1, X_2, \cdots, X_n$ and Y are jointly Gaussian zero

mean random variables, then the best estimate for Y in terms of $X_1, X_2, \cdots, X_n$ is always linear.

Proof : Let

$$\hat{Y} = \varphi(X_1, X_2, \cdots, X_n) = E\{Y \mid \underline{X}\}$$

represent the best (possibly nonlinear) estimate of Y, and

$$\hat{Y}_l = \sum_{i=1}^{n} a_i X_i$$

the best linear estimate of Y. Then

$$\varepsilon = Y - \hat{Y}_l = Y - \sum_{i=1}^{n} a_i X_i$$

is orthogonal to the data $X_k, \quad k = 1 \to n.$ Thus

$$E\{\varepsilon\, X_k^*\} = 0, \quad k = 1 \to n.$$

Also,

$$E\{\varepsilon\} = E\{Y\} - \sum_{i=1}^{n} a_i E\{X_i\} = 0.$$

Using these equations, we get

$$E\{\varepsilon\, X_k^*\} = E\{\varepsilon\}E\{X_k^*\} = 0, \quad k = 1 \to n.$$

It is clear that we obtain that ε and X_k are zero mean uncorrelated random variables for $k = 1 \to n.$ But ε itself represents a Gaussian random variable, since it represents a linear combination of a set of jointly Gaussian random variables. Thus ε and $\underline{X}$ are jointly Gaussian and uncorrelated random variables. As a result, ε and $\underline{X}$ are independent random variables. Thus from their independence

$$E\{\varepsilon \,|\, \underline{X}\} = E\{\varepsilon\}.$$

But, $E\{\varepsilon\} = 0,$ and hence

$$E\{\varepsilon \,|\, \underline{X}\} = 0.$$

Finally, we get

$$E\{\varepsilon \,|\, \underline{X}\} = E\{Y - \sum_{i=1}^{n} a_i X_i \,|\, \underline{X}\} = 0$$

or

$$E\{Y \mid \underline{X}\} = E\{\sum_{i=1}^{n} a_i X_i \mid \underline{X}\} = \sum_{i=1}^{n} a_i X_i = Y_l.$$

$E\{Y \mid \underline{X}\} = \varphi(\underline{x})$ represents the best possible estimator,
and $\sum_{i=1}^{n} a_i X_i$ represents the best linear estimator.
Thus the best linear estimator is also the best possible overall estimator in the Gaussian case.

Next we turn our attention to prediction problems using linear estimators.

Linear Prediction

Suppose $X_1, X_2, \cdots, X_n$ are known and X_{n+1} is unknown. Thus $Y = X_{n+1}$, and this represents a one-step prediction problem. If the unknown is X_{n+k}, then it represents a *k*-step ahead prediction problem. Returning back to the one-step predictor, let $\hat{X}_{n+1}$ represent the best linear predictor. Then

$$\hat{X}_{n+1} = \sum_{i=1}^{n} a_i X_i,$$

where the error

$$\varepsilon_n = X_{n+1} - \hat{X}_{n+1} = X_{n+1} - \sum_{i=1}^{n} a_i X_i$$

is orthogonal to the data, i.e.,

$$E\{\varepsilon_n \; X_k^*\} = 0, \qquad k = 1,..., n.$$

Hence, we get

$$E\{\varepsilon_n \; X_k^*\} = E\{X_{n+1} X_k^*\} - \sum_{i=1}^{n} a_i E\{X_i X_k^*\} = 0, \qquad k = 1,..., n. \quad (2)$$

Suppose X_i represents the sample of a wide sense stationary

stochastic process $X(t)$ so that

$$E\{X_i X_k^*\} = R(i-k) = r_{i-k} = r_{k-i}^*$$

Thus (2) becomes

$$E\{\varepsilon_n X_k^*\} = r_{n+1-k} - \sum_{i=1}^{n} a_i r_{i-k} = 0, \quad k = 1,\ldots,n. \tag{3}$$

Expanding (3) for $k = 1,2,\cdots,n,$ we get the following set of linear equations

$$\begin{aligned}
&a_1 r_0 + a_2 r_1 + a_3 r_2 + \cdots + a_n r_{n-1} = r_n && \leftarrow k = 1 \\
&a_1 r_1^* + a_2 r_0 + a_3 r_1 + \cdots + a_n r_{n-2} = r_{n-1} && \leftarrow k = 2 \\
&\qquad\qquad\qquad\qquad \vdots \\
&a_1 r_{n-1}^* + a_2 r_{n-2}^* + a_3 r_{n-3}^* + \cdots + a_n r_0 = r_1 && \leftarrow k = n. \qquad (4)
\end{aligned}$$

Then, the minimum mean square error is given by

$$\sigma_n^2 = E\{|\varepsilon|^2\} = E\{\varepsilon_n\, Y^*\} = E\{\varepsilon_n\, X_{n+1}^*\}$$

$$= E\{(X_{n+1} - \sum_{i=1}^{n} a_i X_i) X_{n+1}^*\} \qquad (5)$$

$$= r_0 - \sum_{i=1}^{n} a_i r_{n+1-i}^*$$

The n equations in (4) can be represented as

$$\begin{pmatrix} r_0 & r_1 & r_2 & \cdots & r_{n-1} \\ r_1^* & r_0 & r_1 & \cdots & r_{n-2} \\ r_2^* & r_1^* & r_0 & \cdots & r_{n-3} \\ & & \vdots & & \\ r_{n-1}^* & r_{n-2}^* & \cdots & r_1^* & r_0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} r_n \\ r_{n\text{-}1} \\ r_{n\text{-}2} \\ \vdots \\ r_1 \end{pmatrix}. \qquad (6)$$

Eq. (6) can be written compactly as

$$\mathbf{Ra} = \mathbf{r}$$

with matrix **R**, vectors **a** and **r** obviously defined. This equation is known as the **Wiener-Hopf equation**. Notice that **R** is Hermitian Toeplitz and positive definite. Thus, the prediction coefficients vector **a** is given by

$$\mathbf{a} = \mathbf{R}^{-1}\mathbf{r} \tag{7}$$

Substituting (7) back into (5), we obtain another expression of the mean-square error:

$$\sigma_n^2 = r_0 - \mathbf{r}^*\mathbf{R}^{-1}\mathbf{r} \tag{8}$$

We have assumed the autocorrelation function is known.
What if it is unknown? We may use adaptive (learning) methods, e.g., least mean-square (LMS) algorithm, Kalman filter etc.

Imperial College
London

EE4-10: Probability and Stochastic Processes

Lecture 8: Markov Chains

Markov Chains

A discrete-time stochastic process $X = \{X_n, n = 0, 1, 2, \ldots\}$ is called a *Markov chain* provided that

$$P\{X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0\}$$
$$= P(X_{n+1} = j \mid X_n = i\}$$

for all states $i_0, i_1, \ldots, i_{n-1}, i, j$ and all $n \geq 0$.

- X_n is called the *state* of the process at time *n*.
- The *state space E* is the set of all possible values that the random variables X_n can assume.

We restrict ourselves to Markov chains such that the conditional probabilities

$$P_{ij} = P\{X_{n+1} = j \mid X_n = i\}$$

are independent of *n*, and for which $E \subset \{0, 1, 2, \ldots\}$ (i.e., *E* is finite or countable). Such a Markov chain is called *homogeneous*.

Since

- probabilities are non-negative, and
- the process must make a transition into some state at each time, then

$$P_{ij} \geq 0 \quad for\ i, j \in E; \qquad \sum_{j \in E} P_{ij} = 1 \quad for\ i \in E.$$

We can arrange the probabilities P_{ij} into a square matrix $\{P_{ij}\}$ called the *transition matrix*.

Let P be a square matrix of entries defined for all $i, j \in E$.

Then P is a *Markov matrix* (or stochastic matrix) provided that

$$\text{for each } i, j \in E, \ P_{ij} \geq 0$$

$$\text{for each } i \in E, \ \sum_{j \in E} P_{ij} = 1.$$

Obviously the transition matrix for a Markov chain is a Markov matrix.

Example:

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}$$

This Markov matrix corresponds to a 3-state Markov chain.

Comment: The continuous-time version is often referred to as the **Markov process**.

Markov chains have many applications in modeling real-world processes, for example, stock markets, Internet, queuing theory, modeling of language etc.

Chapman-Kolmogorov Equations

It follows from the definition of a Markov chain that, for example,

$$
\begin{aligned}
&P\{X_9 = j, X_8 = k \mid X_7 = i\} \\
&\quad = P\{X_9 = j \mid X_8 = k, X_7 = i\} P\{X_8 = k \mid X_7 = i\} \\
&\quad = P\{X_9 = j \mid X_8 = k\} P\{X_8 = k \mid X_7 = i\} \\
&\quad = P_{ik} P_{kj}
\end{aligned}
$$

Hence we have the following:

Theorem 1. For any $n = 0, 1, 2, \ldots$, $m = 1, 2, 3, \ldots$, and $i_0, \ldots, \quad i_m \in E$,

$$
\begin{aligned}
&P\{X_{n+1} = i_1, \ldots, X_{n+m} = i_m \mid X_n = i_0\} \\
&\quad = P_{i_0 i_1} \cdot P_{i_1 i_2} \cdot \ldots \cdot P_{i_{m-1} i_m} .
\end{aligned}
$$

Now note that

$$P\{X_{n+2} = j \mid X_n = i\}$$

$$= \sum_{k \in E} P\{X_{n+2} = j \mid X_{n+1} = k, X_n = i\} P\{X_{n+1} = k \mid X_n = i\}$$

$$= \sum_{k \in E} P\{X_{n+2} = j \mid X_{n+1} = k\} P\{X_{n+1} = k \mid X_n = i\}$$

$$= \sum_{k \in E} P_{ik} P_{kj}$$

$$= P^{2}_{ij} \equiv (P^2)_{ij}$$

Example:

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix} \quad \rightarrow \quad P^2 = P \cdot P = \begin{pmatrix} \frac{7}{16} & \frac{3}{16} & \frac{3}{8} \\ \frac{3}{8} & \frac{1}{4} & \frac{3}{8} \\ \frac{3}{8} & \frac{3}{16} & \frac{7}{16} \end{pmatrix}$$

Hence, $$P\{X_{n+2} = 2 \mid X_n = 3\} = P^2_{32} = \frac{3}{16}.$$

In general, for all $n, m \geq 0$ and $i, j \in E$,

$$P\{X_{n+m} = j \mid X_n = i\} = P_{ij}^m \equiv (P^m)_{ij}.$$

This leads to the *Chapman-Kolmogorov equations*:

$$P_{ij}^{n+m} = \sum_{k \in E} P_{ik}^n P_{kj}^m \qquad \text{for all } n, m \geq 0, \text{ all } i, j \in E.$$

Also note that if $\pi^{(0)}$ is the probability distribution of the states at time 0, i.e.,

$$\pi_i^{(0)} = P\{X_0 = i\}.$$

Then the probability distribution at time n is

$$\pi^{(n)} = \pi^{(0)} P^n.$$

Classification of States

Fix initial state $X_0 = j$, and let T_j be the time of the first return to state j (**recurrent time**).

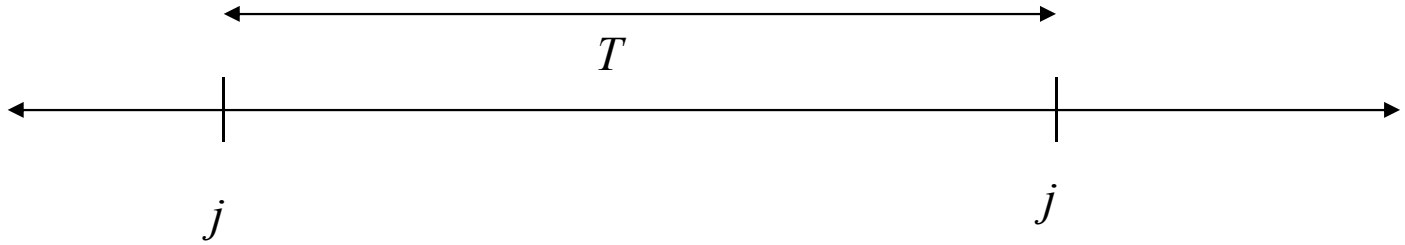
- State j is called *recurrent* if

$$P\{T_j < \infty\} = 1.$$

- A recurrent state j is call *positive* if

$$E[T_j] < \infty$$

Otherwise, it is called *null*.

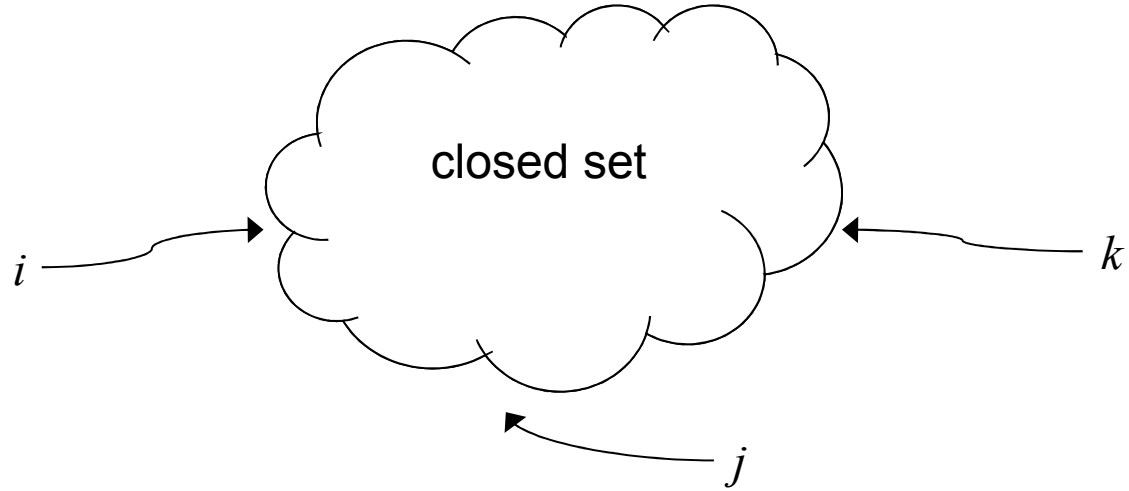


- A recurrent state *j* is said to be *periodic* with period *d* if $P_{jj}^{n} = 0$ whenever *n* is not divisible by *d*, and *d* is that largest integer with this property. A state with period 1 is said to be *aperiodic*.
- Recurrent positive, aperiodic states are called *ergodic*. A Markov chain is called *ergodic* if all of its states are ergodic.
- State *j* is called *transient* if

$$P\{T_j < \infty\} < 1.$$

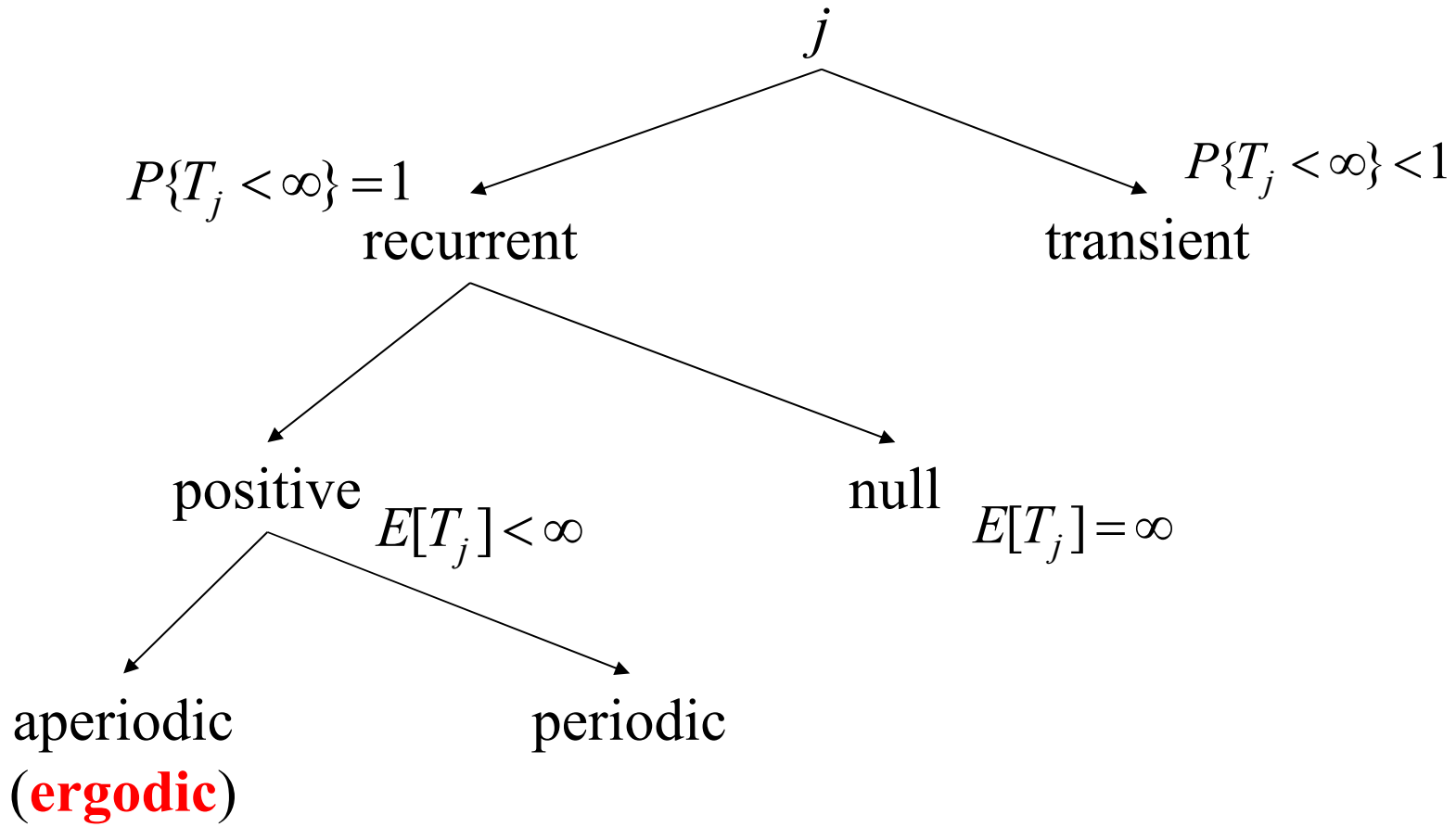
Namely, there is a positive probability of never returning to that state.

- A state j is *accessible* from state i if there exists $n \geq 0$ such that $P_{ij}^{n} > 0$.
- A set of states is said to be *closed* if no state outside it is accessible from any state in it.
- A state forming a closed set by itself is called an *absorbing* state.



- A closed set is *irreducible* if no proper subset of it is closed.
- A Markov chain is called *irreducible* if its only closed set is the set of all states.

Classification of States



Theorem 2. In an irreducible Markov chain, either all states are transient, all states are recurrent null, or all states are recurrent positive (i.e., all states are of the same type).

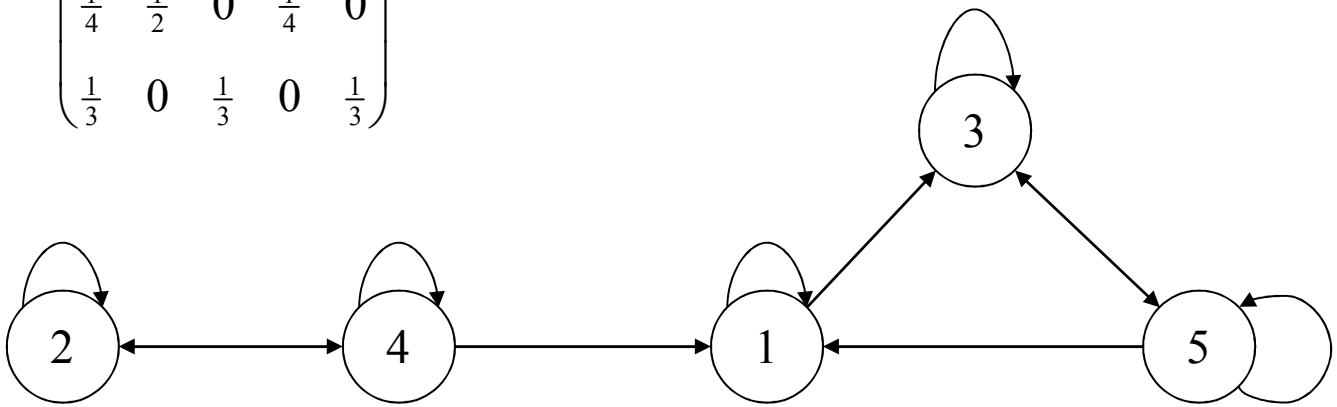
Theorem 3. In a finite-state Markov chain, all recurrent states are positive, and it is impossible that all states are transient. If the Markov chain is also irreducible, then it has no transient states.

Example: Consider a Markov chain with state space {1, 2, 3, 4, 5} and transition matrix P as shown in the right. What is the classification of each state?

$$P = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & 0 & \frac{3}{4} & 0 \\ 0 & 0 & \frac{1}{3} & 0 & \frac{2}{3} \\ \frac{1}{4} & \frac{1}{2} & 0 & \frac{1}{4} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \end{pmatrix}$$

It is useful to draw a graph with the states as vertices and a directed edge from i to j if $P(i, j) > 0$.

$$P = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & 0 & \frac{3}{4} & 0 \\ 0 & 0 & \frac{1}{3} & 0 & \frac{2}{3} \\ \frac{1}{4} & \frac{1}{2} & 0 & \frac{1}{4} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \end{pmatrix}$$



Note that:

- $\{1, 3, 5\}$ and $\{1, 2, 3, 4, 5\}$ are closed.
- $\{1, 3, 5\}$ is irreducible.

Limiting Probabilities

Theorem 5. In an irreducible aperiodic homogenous Markov chain,

$$\pi = \lim_{n\to\infty} \pi^{(n)} = \lim_{n\to\infty} \pi^{(0)} P^n$$

always exists and is independent of the initial state probability distribution $\pi^{(0)}$.

Moreover, either:

a. all states are transient or all states are recurrent null, in which cases $\pi_j = 0$ for all j (only possible for infinite chains), or

b. all states are recurrent positive (i.e., **ergodic**), in which case $\pi_j > 0$ for all j and the π_j are uniquely determined by the following equations:

$$\pi_j = \sum_{i\in E} \pi_i P_{ij}, \quad \sum_{j\in E} \pi_j = 1.$$

This can also be written as

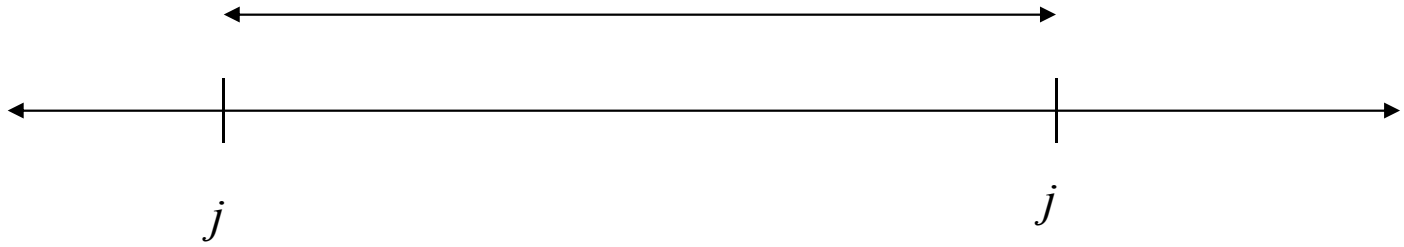
$$\pi = \pi P, \pi 1 = 1.$$

Furthermore, if m_j is the expected time between two returns to j, then

$$\pi_j = \frac{1}{m_j}.$$

In other words, the limiting probability of being in state j is equal to the rate at which j is visited.

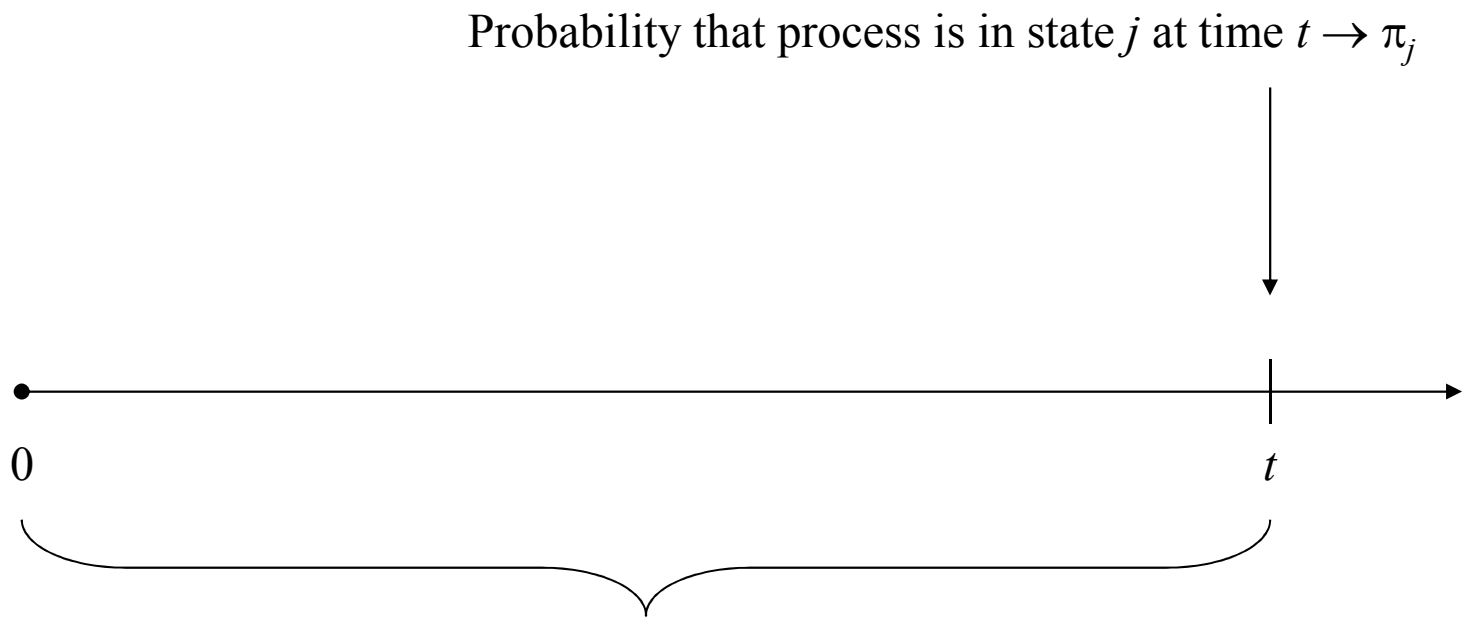
$$m_j = \frac{1}{\pi_j}$$



Another Look at Ergodicity

It follows that for recurrent positive states, the limiting probabilities have two interpretations:

- the limiting distribution of the state at time t
- the long-run proportion of time that the process spends in each state.



Ergodic Theorem

If a Markov chain $X = \{X_n, n = 0, 1, 2, \ldots\}$ is ergodic, then for any bounded function f we have

$$\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \rightarrow E[f(X)]$$

where the expectation on right-hand side is with respect to the stationary distribution π of the process.

In particular, we have

$$\frac{1}{n}\sum_{i=0}^{n-1} X_i \rightarrow E[X]$$

which generalizes the standard law of large numbers where X_i's are required to be independent.

This means that, for ergodic Markov chains, the time average can be replaced by ensemble average. This property is useful in practice.

Computing Limiting Probabilities

If π is a solution of $\pi = \pi P$, then any constant c, $c\pi$ is also a solution. In solving $\pi = \pi P$, $\pi \mathbf{1}=1$, it is best to solve $\pi = \pi P$ first and then normalize the resulting solution to satisfy the second condition.

In the finite state space case, the equations of $\pi = \pi P$ are linearly dependent; therefore one can throw one of the equations out of consideration.

Example:

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}$$

$\pi = \pi P, \quad \pi 1 = 1$ implies that $\pi = \left(\frac{2}{5}, \frac{1}{5}, \frac{2}{5}\right).$

Compare $\pi = \left(\frac{2}{5}, \frac{1}{5}, \frac{2}{5}\right)$ with

$$P^3 = \begin{pmatrix} \frac{13}{32} & \frac{13}{64} & \frac{25}{64} \\ \frac{13}{32} & \frac{3}{16} & \frac{13}{32} \\ \frac{25}{64} & \frac{13}{64} & \frac{13}{32} \end{pmatrix} \qquad P^5 = \begin{pmatrix} \frac{205}{512} & \frac{205}{1024} & \frac{409}{1024} \\ \frac{205}{512} & \frac{51}{256} & \frac{205}{512} \\ \frac{409}{1024} & \frac{205}{1024} & \frac{205}{512} \end{pmatrix}$$

Observation. In a finite-state irreducible, aperiodic Markov chain, the rows of P^n converge to π.

Perron-Frobenius Theorem

For finite chains, the convergence to the stationary distribution is exponentially fast. This follows from the Perron-Frobenious theorem.

A Markov chain is irreducible and aperiodic if and only if its transition matrix P is **primitive**, namely, there is some integer such that P^n is a *positive matrix*, i.e., its entries are all strictly positive

$$(P^n)_{ij} > 0$$

If P is primitive, then, the maximum eigenvalue of P is $\lambda_1 = 1$, while all other eigenvalues $\lambda_2, \ldots, \lambda_k$ are strictly less than unity in magnitude; further, the eigenvector associated with $\lambda_1 = 1$ has all positive entries. In fact, this (left) eigenvector corresponds to the stationary distribution $\pi = \pi P$.

Assuming P is diagonalizable, we have the eigen-decomposition

$$P = U\Sigma U^T$$

where U is an orthonormal matrix whose rows are composed of the eigenvectors $\{\mathbf{u}_1,\mathbf{u}_2,\ldots, \mathbf{u}_k\}$, and $\Sigma = \text{diag}\{\lambda_1,\lambda_2,\ldots, \lambda_k\}$. Starting from any initial distribution π_0, the distribution at step n is given by

$$\pi_n = \pi_0 P^n = \pi_0\, U\Sigma^n U^T$$

Let $\mathbf{a} = \pi_0 U$. We have

$$\pi_n = a_1\lambda_1^n\mathbf{u}_1 + a_2\lambda_2^n\mathbf{u}_2 + \cdots + a_k\lambda_k^n\mathbf{u}_k$$
$$\rightarrow a_1\lambda_1^n\mathbf{u}_1 = a_1\mathbf{u}_1$$

since $|\lambda_i| < 1$ for $i \geq 2$. After normalization, we have $\pi_n \rightarrow \pi$, since π corresponds to $\mathbf{u}_1$ (upon normalization).

Random Walk With Left Barrier

Let X be a Markov chain with state space $E = \{0, 1, 2, \ldots\}$, and transition matrix

$$P = \begin{pmatrix} 0 & 1 & & & & \\ q & 0 & p & & & 0 \\ & q & 0 & p & & \\ & & \cdot & \cdot & \cdot & \\ 0 & & & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot \end{pmatrix}$$

where $0 < p < 1, q = 1 - p.$

All states can be reached from each other, so the Markov chain is irreducible. Consequently, all states are transient, all are recurrent positive, or all are recurrent null.

To identify the states, let us assume that a limiting π probability distribution exists, that is, there exists a vector satisfying

$$\pi = \pi P.$$

Any solution is of the form

$$\pi_j = \frac{1}{q}\left(\frac{p}{q}\right)^{j-1} \pi_o, \quad j = 1, 2, \ldots$$

for some constant $\pi_0 \geq 0$.

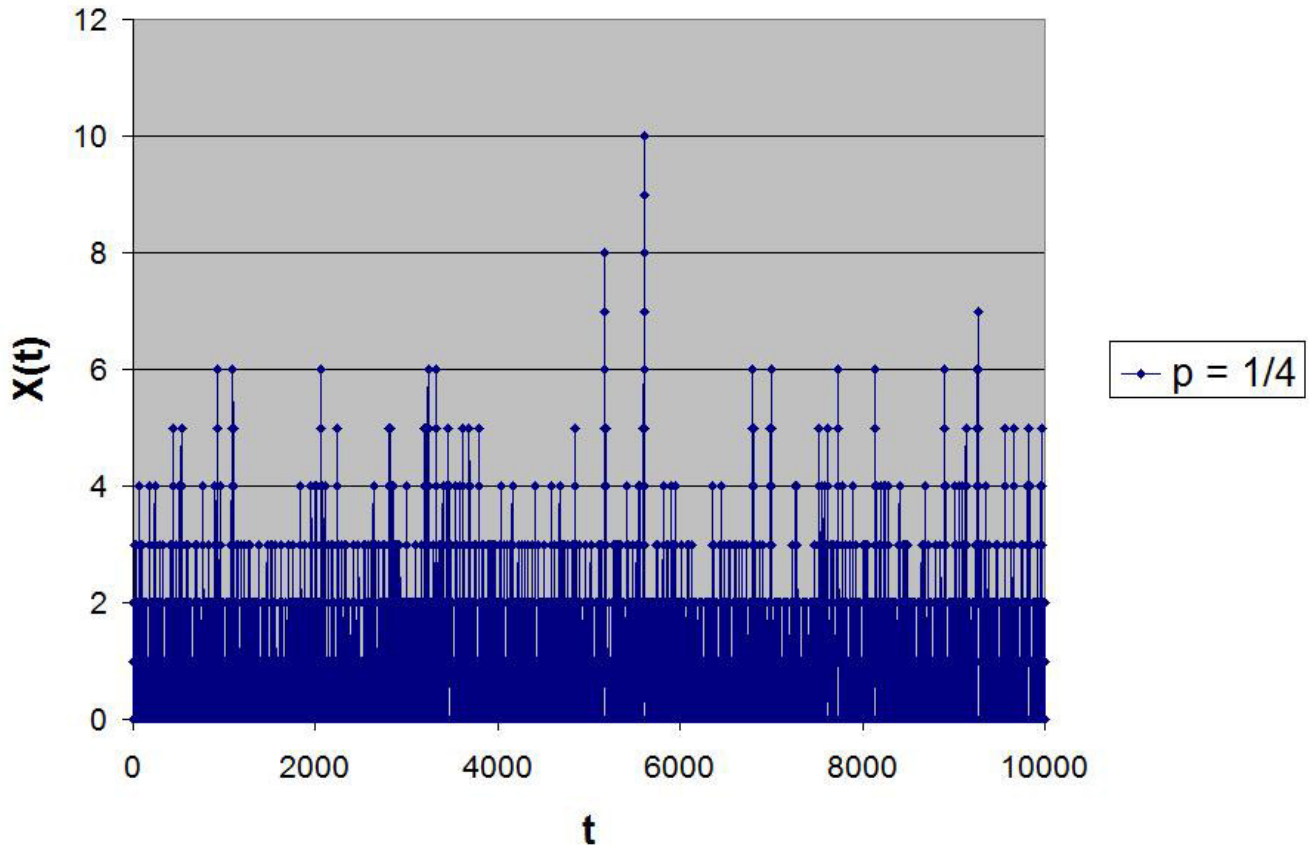
If $p < q$, then $p/q < 1$ and

$$\sum_{j=0}^{\infty} \pi_j = \frac{2q}{q-p}\pi_0, \quad \pi_0 = \frac{1}{2}\left(1 - \frac{p}{q}\right)$$

and

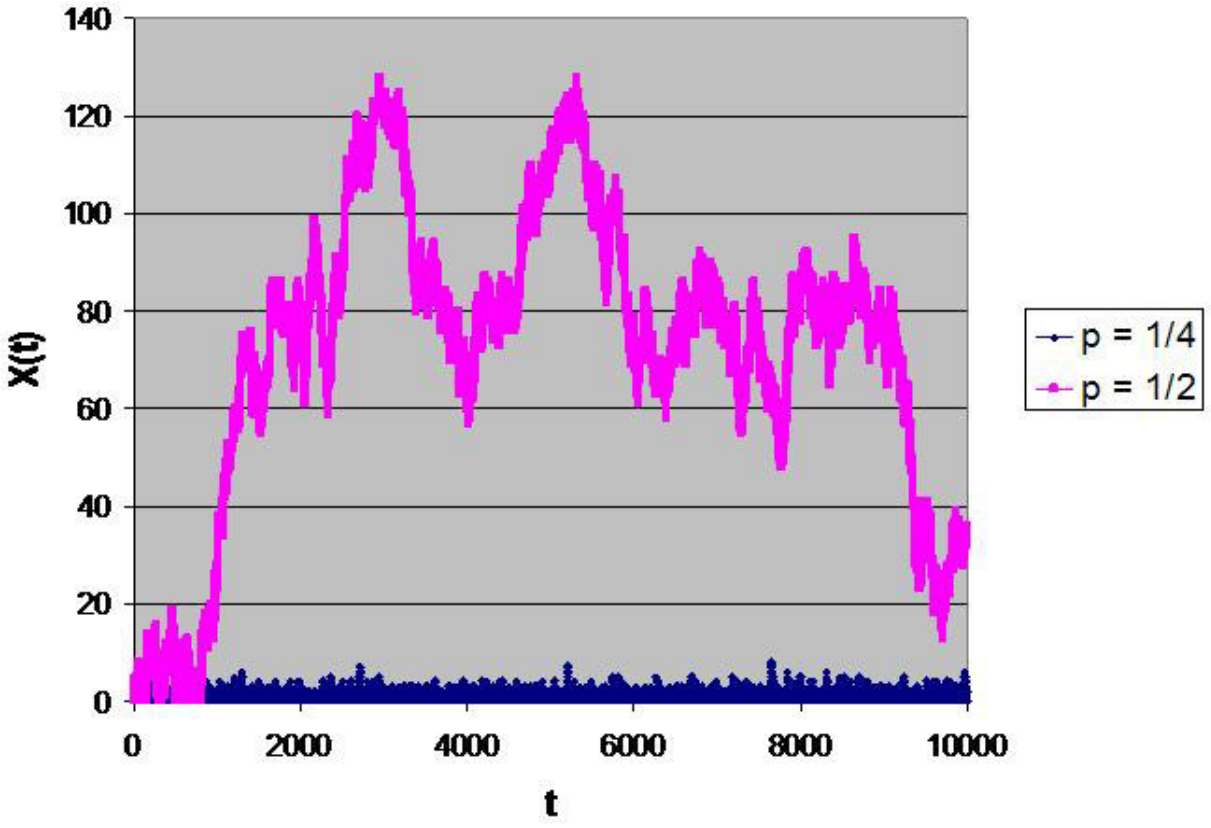
$$\pi_j = \begin{cases} \frac{1}{2}\left(1-\frac{p}{q}\right) & \text{if } j = 0 \\ \frac{1}{2q}\left(1-\frac{p}{q}\right)\left(\frac{p}{q}\right)^{j-1} & \text{if } j \geq 1 \end{cases}$$

In this case, all states are recurrent positive.

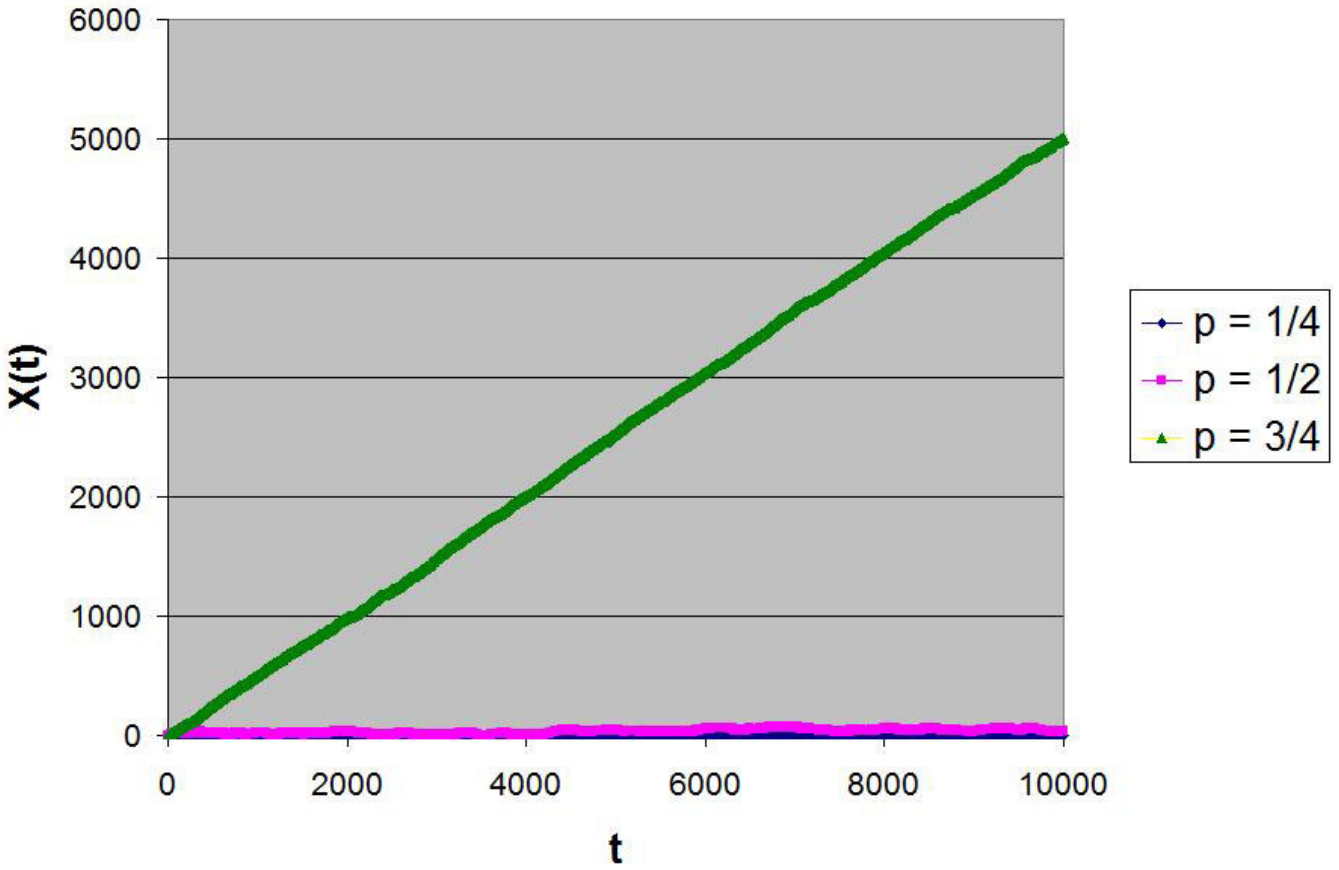


If $p \geq q$, then $\pi_j = 0$ for all j. In fact, the states are recurrent null if $p = q$, and transient if $p > q$.

p = 1/2



p = 3/4


6000
5000
4000
3000
2000
1000
0
X(t)
0
2000
4000
6000
8000
10000
t
p = 1/4
p = 1/2
p = 3/4

Counterexample: The Gambler's Ruin

Let X be a Markov chain with state space $E = \{0, 1, 2, \ldots, N\}$, and transition matrix

$$P = \begin{pmatrix} 1 & & & & & 0 \\ q & 0 & p & & & \\ & q & 0 & p & & \\ & & \cdot & \cdot & \cdot & \\ & & & q & 0 & p \\ & 0 & & & & 1 \end{pmatrix}$$

where $0 < p < 1, q = 1 - p$. This is also called a random walk with absorbing barriers (states 0 and N).

This Markov chain models a gamble where the gambler wins with probability p and loses with probability q at each step. The gamble ends if it reaches either of the absorbing barriers. In particular, reaching state 0 corresponds to the gambler's ruin.

Here, the transition matrix P is not primitive, hence there exists no stationary distribution. In fact, any probability distribution of the form $q = (x, 0, \ldots, 0, 1 - x)$ assigning positive weights to the absorbing states satisfies the equation $q = qP$.

If the gambler's initial capital is i, his ruin probability P_i is the probability of reaching 0, starting in state i. We have

$$P_i = \begin{cases} \dfrac{1 - \left(p/q\right)^{N-i}}{1 - \left(p/q\right)^{N}}, & \text{if } p \neq \dfrac{1}{2} \\ \dfrac{N-i}{N}, & \text{if } p = \dfrac{1}{2} \end{cases}$$

It is possible to derive P_i from the difference equation (exercise)

$$P_i = pP_{i+1} + qP_{i-1}$$

with initial conditions $P_0 = 1$ and $P_N = 0$. However, we will take another approach and defer the derivation to the last lecture.

Discussion

Now, suppose a player has wealth a, and the adversary has wealth b. If both players are of equal skill, that is, $p = q = ½$, then the ruin probability

$$P_a = \frac{b}{a+b}$$

Thus, it is unwise to play against someone with a large fortune, since $P_a \to 1$ if $b \gg a$.

Needless to say if the adversary is skillful ($p < q$) and wealthy ($b \to \infty$). In this case, the gambler's ruin is certain in the long run:

$$P_a \to 1$$

Casinos and lotteries work on this principle. They keep a slight advantage to themselves and posses large capitals. Interestingly, the same principle underlies the operations of more respectable institutions such as insurance companies.

Application: MCMC

Monte Carlo methods for counting, integration, signal processing, machine learning etc. requires random sampling, i.e., to sample points from "very large" sample set at random with distribution π. However, this can be difficult to implement due to the large set size.

Markov chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. The state of the chain after a large number of steps is then used as a sample of the desired distribution.

An example of MCMC is top-in-at-random for card shuffling, which eventually leads to uniform distribution.

Imperial College
London

EE4-10: Probability and Stochastic Processes

Lecture 9: Continuous-Time Processes

The Markov chains we studied have discrete time and discrete state spaces. In general, random processes come in many types. For example, they may run in discrete or continuous time, and their state spaces may also be discrete or continuous.

In this lecture, we are concerned with Markov processes with continuous time and/or state spaces. Markov processes have many applications in queuing theory. In particular, we will look at

- Poisson process with continuous time and discrete state space.
- Wiener process with continuous time and continuous state space.

Our emphasis will be the Poisson process due to its usefulness in modelling network traffic.

Markov Processes

A continuous-time Markov chain $\{X(t) : t \geq 0\}$ is often called a Markov process. The process $\{X(t)\}$ satisfies the Markov property if

$$P(X(t_n) = j | X(t_1) = i_1, \ldots, X(t_{n-1}) = i_{n-1}) = P(X(t_n) = j | X(t_{n-1}) = i_{n-1})$$

for all $j, i_1, \ldots, i_{n-1}$ and any sequence $t_1 < t_2 < \cdots < t_n$ of times.

The evolution of Markov processes can be described in much the same way as that for Markov chains.

However, the theory of continuous state-space Markov processes (and also Markov chains) is more difficult.

Poisson Processes

Definition: $X(t) = n(0, t)$ represents a **Poisson process** if

(i) the number of arrivals $n(t_1, t_2)$ in an interval (t_1, t_2) of length $t = t_2 - t_1$ is a Poisson random variable with parameter λt.

Thus

$$P\{n(t_1, t_2) = k\} = e^{-\lambda t} \frac{(\lambda t)^k}{k!}, \quad k = 0, 1, 2, \cdots, \quad t = t_2 - t_1$$

and

(ii) If the intervals (t_1, t_2) and (t_3, t_4) are nonoverlapping, then the random variables $n(t_1, t_2)$ and $n(t_3, t_4)$ are independent.

Parameter λ is referred to as *intensity* of the Poisson process.

Examples:

- Goals scored in a football match.
- Requests for individual documents on a web server.

•Requests for telephone calls at a switchboard.
•Particle emissions due to radioactive decay.
•Outbreak of war.
•Shot noise: photons landing on a photodiode.

Since $n(0, t) \sim P(\lambda t)$, we have

$$E[X(t)] = E[n(0,t)] = \lambda t$$

and

$$E[X^2(t)] = E[n^2(0,t)] = \lambda t + \lambda^2 t^2.$$

To determine the autocorrelation function $R_{XX}(t_1, t_2)$, let $t_2 > t_1$, then from (ii) above $n(0, t_1)$ and $n(t_1, t_2)$ are independent Poisson random variables with parameters λt_1 and $\lambda(t_2 - t_1)$ respectively. Thus

$$E[n(0,t_1)n(t_1,t_2)] = E[n(0,t_1)]E[n(t_1,t_2)] = \lambda^2 t_1(t_2 - t_1).$$

But

$$n(t_1, t_2) = n(0, t_2) - n(0, t_1) = X(t_2) - X(t_1)$$

and hence the left side can be rewritten as

$$E[X(t_1)\{X(t_2) - X(t_1)\}] = R_{XX}(t_1, t_2) - E[X^2(t_1)].$$

All together, we obtain

$$\begin{aligned} R_{XX}(t_1, t_2) &= \lambda^2 t_1 (t_2 - t_1) + E[X^2(t_1)] \\ &= \lambda t_1 + \lambda^2 t_1\, t_2, \quad t_2 \geq t_1 . \end{aligned}$$

Similarly

$$R_{XX}(t_1, t_2) = \lambda t_2 + \lambda^2 t_1\, t_2, \quad t_2 < t_1 .$$

Thus

$$R_{XX}(t_1, t_2) = \lambda^2 t_1\, t_2 + \lambda \min(t_1, t_2).$$

notice that the Poisson process $X(t)$ *does not* represent a wide sense stationary process.

Sum of Poisson Processes:

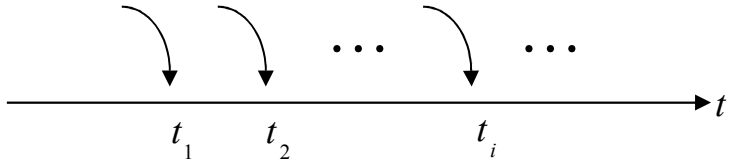
If $X_1(t)$ and $X_2(t)$ represent two independent Poisson processes, then their sum $X_1(t) + X_2(t)$ is also a Poisson process with parameter $(\lambda_1 + \lambda_2)t$.

Random selection of Poisson Points:

Let $t_1, t_2, \cdots, t_i, \cdots$ represent random arrival points associated with a Poisson process $X(t)$ with parameter λt,
and associated with
each arrival point,
define an independent
Bernoulli random
variable N_i, where

$$P(N_i = 1) = p,$$

$$P(N_i = 0) = q = 1 - p.$$



Define the processes

$$Y(t) = \sum_{i=1}^{X(t)} N_i \quad ; \quad Z(t) = \sum_{i=1}^{X(t)} (1 - N_i) = X(t) - Y(t)$$

we claim that both $Y(t)$ and $Z(t)$ are independent Poisson processes with parameters λpt and λqt respectively.

Proof:

$$P\{Y(t) = k\} = \sum_{n=k}^{\infty} P\{Y(t) = k \mid X(t) = n\} P\{X(t) = n)\}.$$

But given $X(t) = n$, we have $Y(t) = \sum_{i=1}^{n} N_i \sim B(n, p)$ so that

$$P\{Y(t) = k \mid X(t) = n\} = \binom{n}{k} p^k q^{n-k}, \quad 0 \le k \le n,$$

and

$$P\{X(t) = n\} = e^{-\lambda t} \frac{(\lambda t)^n}{n!}.$$

We get

$$P\{Y(t)=k\} = e^{-\lambda t}\sum_{n=k}^{\infty}\frac{n!}{(n-k)!k!}p^k q^{n-k}\,\frac{(\lambda t)^n}{n!} = \frac{p^k e^{-\lambda t}}{k!}\,(\lambda t)^k \underbrace{\sum_{n=k}^{\infty}\frac{(q\lambda t)^{n-k}}{(n-k)!}}_{e^{q\lambda t}}$$

$$= (\lambda pt)^k\,\frac{e^{-(1-q)\lambda t}}{k!} = e^{-\lambda pt}\,\frac{(\lambda pt)^k}{k!}, \qquad k = 0,\,1,\,2,\,\cdots$$

$$\sim \quad P(\lambda pt).$$

Similary, $Z(t) \sim P(\lambda qt)$.

More generally,

$$P\{Y(t)=k, Z(t)=m\} = P\{Y(t)=k, X(t)-Y(t)=m\}$$

$$= P\{Y(t)=k, X(t)=k+m\}$$

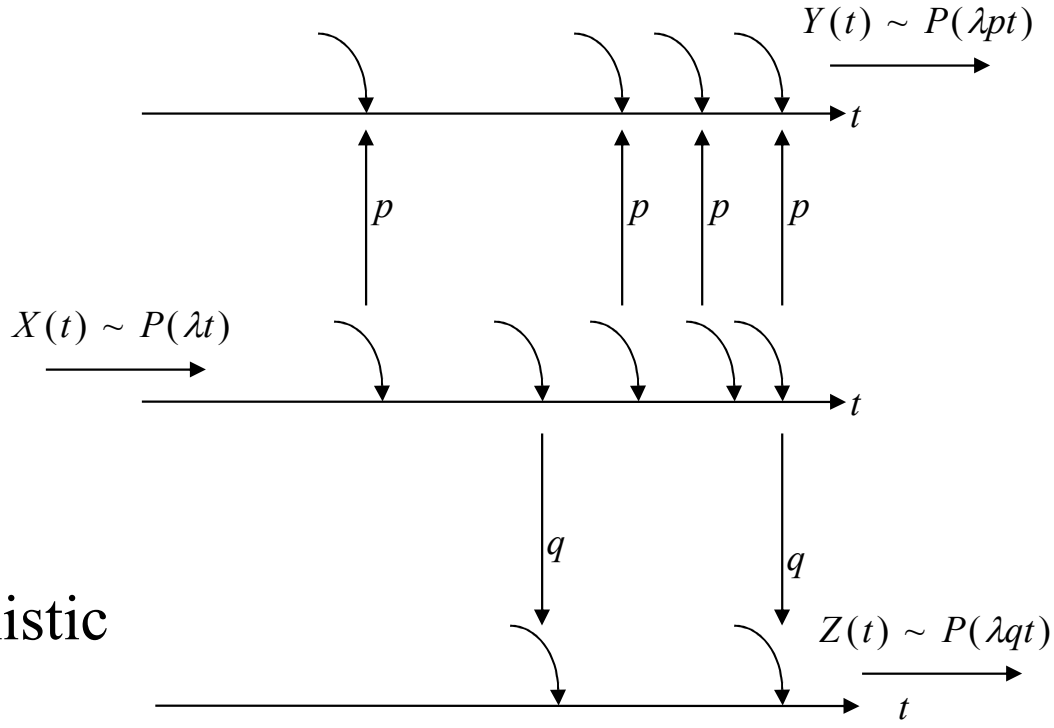
$$= P\{Y(t)=k \mid X(t)=k+m\}P\{X(t)=k+m\}$$

$$= \binom{k+m}{k}p^k q^m \cdot e^{-\lambda t}\,\frac{(\lambda t)^{k+m}}{(k+m)!} = \underbrace{e^{-\lambda pt}\,\frac{(\lambda pt)^k}{k!}}_{P(Y(t)=k)}\quad \underbrace{e^{-\lambda qt}\,\frac{(\lambda qt)^m}{m!}}_{P(Z(t)=m)}$$

$$= P\{Y(t)=k\}P\{Z(t)=m\},$$

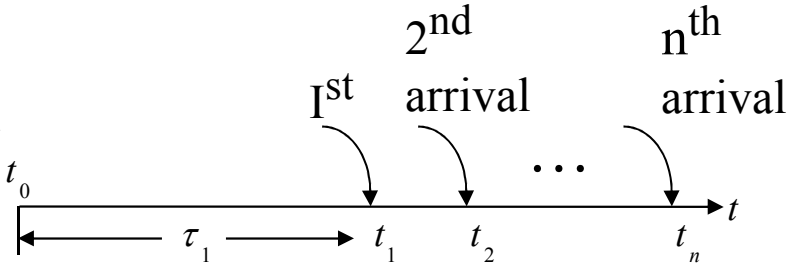
which completes the proof.

Notice that $Y(t)$ and $Z(t)$ are generated as a result of random Bernoulli selections from the original Poisson process $X(t)$ (see the figure), where each arrival gets tossed over to either $Y(t)$ with probability p or to $Z(t)$ with probability q. Each such sub-arrival stream is also a Poisson process. Thus random selection of Poisson points preserve the Poisson nature of the resulting processes. However, deterministic selection from a Poisson process destroys the Poisson property for the resulting processes.



Inter-arrival Distribution for Poisson Processes

Let τ_1 denote the time interval (delay) to the first arrival from *any* fixed point t_0. To determine the probability distribution of the random variable τ_1, we argue as follows: Observe that the event "$\tau_1 > t$" is the same as "$n(t_0, t_0+t) = 0$", or the complement event "$\tau_1 \le t$" is the same as the event "$n(t_0, t_0+t) > 0$".



Hence the distribution function of τ_1 is given by

$$F_{\tau_1}(t) \triangleq P\{\tau_1 \le t\} = P\{X(t) > 0\} = P\{n(t_0, t_0 + t) > 0\}$$

$$= 1 - P\{n(t_0, t_0 + t) = 0\} = 1 - e^{-\lambda t}$$

and hence its derivative gives the probability density function

$$f_{\tau_1}(t) = \frac{dF_{\tau_1}(t)}{dt} = \lambda e^{-\lambda t}, \qquad t \ge 0$$

i.e., τ_1 is an exponential random variable with parameter λ so that $E(\tau_1) = 1/\lambda$.

Similarly, let t_n represent the n^{th} random arrival point for a Poisson process. Then

$$F_{t_n}(t) \triangleq P\{t_n \le t\} = P\{X(t) \ge n\}$$

$$= 1 - P\{X(t) < n\} = 1 - \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

and hence

$$f_{t_n}(t) = \frac{dF_{t_n}(t)}{dt} = -\sum_{k=1}^{n-1} \frac{\lambda(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t} + \sum_{k=0}^{n-1} \frac{\lambda(\lambda t)^k}{k!} e^{-\lambda t}$$

$$= \frac{\lambda^n t^{n-1}}{(n-1)!} e^{-\lambda t}, \qquad t \ge 0$$

which represents an Erlang-*n* density function. i.e., the waiting time to the n^{th} Poisson arrival instant has an Erlang-*n* distribution.

Moreover

$$t_n = \sum_{i=1}^{n} \tau_i$$

where τ_i is the random inter-arrival duration between the $(i-1)^{th}$ and i^{th} events. Notice that τ_i s are independent, identically distributed random variables. Hence using their characteristic functions, it follows that all inter-arrival durations of a Poisson process are independent exponential random variables with common parameter λ. i.e.,

$$f_{\tau_i}(t) = \lambda e^{-\lambda t}, \quad t \geq 0.$$

Alternatively, we have τ_1 is an exponential random variable. By repeating that argument after shifting t_0 to the new point t_1 in the preceding figure, we conclude that τ_2 is an exponential random variable. Thus the sequence $\tau_1, \tau_2, \cdots, \tau_n, \cdots$ are independent exponential random variables with common p.d.f.

Thus if we systematically tag every m^{th} outcome of a Poisson process $X(t)$ with parameter λt to generate a new process $e(t)$, then the inter-arrival time between any two events of $e(t)$ is an Erlang-m random variable.

Notice that

$$E[e(t)] = m / \lambda.$$

The inter-arrival time of $e(t)$ in that case represents an Erlang-m random variable, and $e(t)$ is an Erlang-m process.

In summary, if Poisson arrivals are randomly redirected to form new queues, then each such queue generates a new Poisson process. However if the arrivals are systematically redirected (1^{st} arrival to 1^{st} counter, 2^{nd} arrival to 2^{nd} counter, …, m^{th} to m^{th} , $(m+1)^{st}$ arrival to 1^{st} counter, …), then the new subqueues form Erlang-m processes.

Gaussian Process

A continuous-time process X is called a Gaussian process if each finite-dimensional vector $\{X(t_1), X(t_2), \ldots, X(t_n)\}$ has the multivariate normal distribution $N(\boldsymbol{\mu}(\boldsymbol{t}), \boldsymbol{\Sigma}(\boldsymbol{t}))$ for some mean vector $\boldsymbol{\mu}(\boldsymbol{t})$ and some covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{t})$. A Gaussian process is stationary if and only if $E[X(t)]$ is constant for all t and $\boldsymbol{\Sigma}(\boldsymbol{t}) = \boldsymbol{\Sigma}(\boldsymbol{t} + c)$ for all $\boldsymbol{t}$ and c.

The Gaussian process is widely applied to model noise in communication and signal processing systems.

The Wiener process is a non-stationary Gaussian process.

Wiener Processes

The random walk has two interesting and basic properties:

- Time-homogeneity: X_m and $X_{m+n} - X_n$ have the same distribution;
- Independent increments: the increments $X_{n_i} - X_{m_i}$ are independent whenever the intervals $(m_i, n_i]$ are disjoint.

The Wiener process can be seen as the continuous analogue of the random walk♣. Formally, a Wiener process $W = \{W(t): t \geq 0\}$ is Gaussian process, starting from $W(0) = 0$, such that

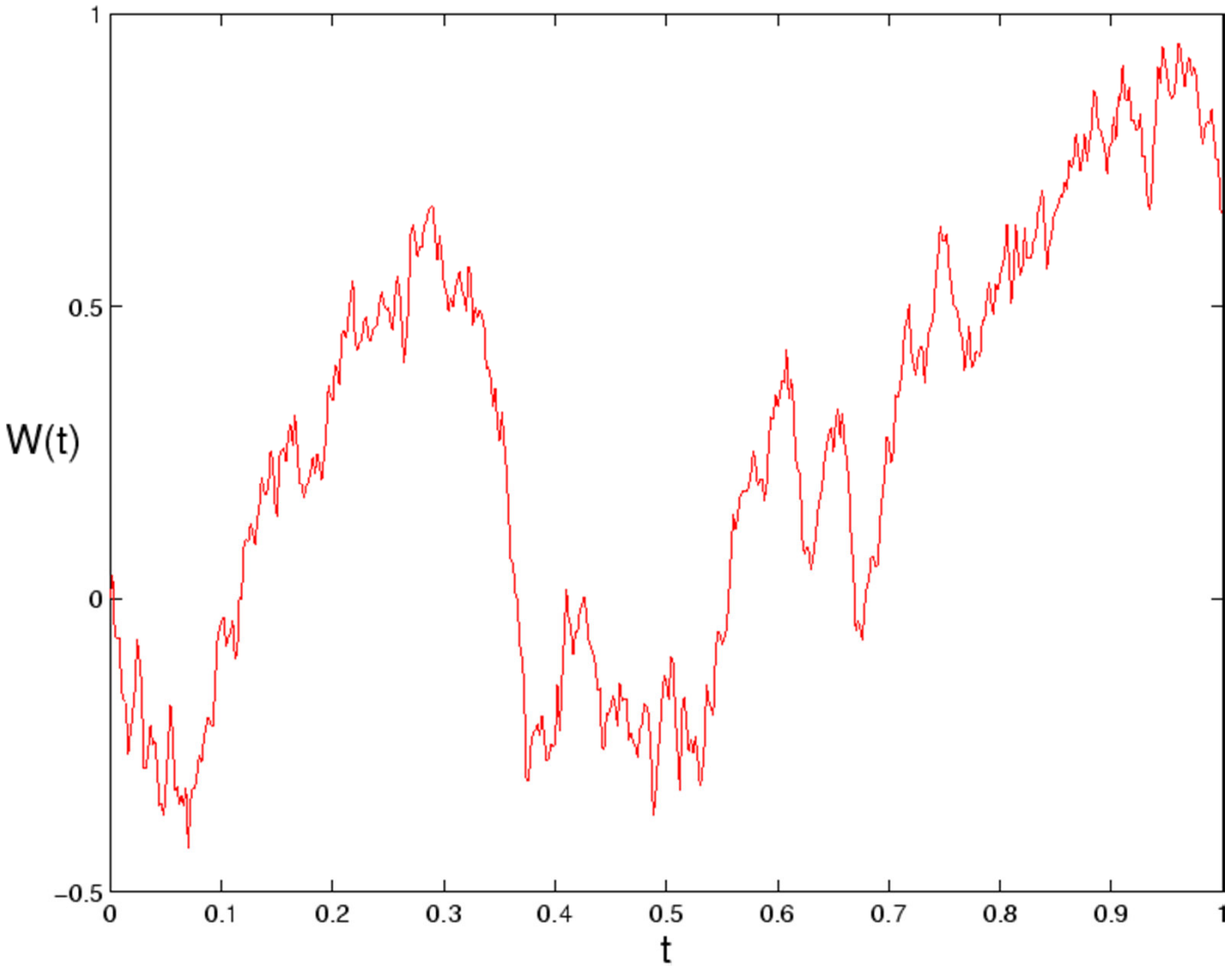
a) W has independent increments,

b) $W(s + t) - W(s)$ is distributed as $N(0, \sigma^2 t)$ for all $s, t \geq 0$,

c) The sample paths of W are continuous.

♣ Note that these properties are also shared by the Poisson process.

The Wiener process can be used to model the Brownian motion, diffusion phenomena, stock prices, etc. The following figure shows a typical realization of a Wiener process.



Obviously, the expectation is zero

$$E[W(t)] = 0$$

and the variance is

$$Var[W(t)] = \sigma^2 t.$$

The autocovariance function of W is given by

$$\begin{aligned} C(s,t) &= E\Big[\big(W(s) - W(0)\big)\big(W(t) - W(0)\big)\Big] \\ &= E\Big[\big(W(s) - W(0)\big)^2 \\ &\quad + \big(W(s) - W(0)\big)\big(W(t) - W(s)\big)\Big] \\ &= \sigma^2 s + 0 \qquad\qquad \text{if} \quad 0 \le s \le t \end{aligned}$$

Thus,

$$C(s,t) = \sigma^2 \min\{s,t\} \quad \text{for all } s,t \ge 0$$

Therefore, the Wiener process is not stationary.

Imperial College
London

EE4-10: Probability and Stochastic Processes

Lecture 10: Martingales

Martingales refer to a specific class of stochastic processes that maintain a form of "stability" in an overall sense. Let $\{X_i, i \geq 0\}$ refer to a discrete time stochastic process. If n refers to the present instant, then in any realization the random variables $X_0, X_1, \cdots, X_n$ are known, and the future values $X_{n+1}, X_{n+2}, \cdots$ are unknown. The process is "stable" in the sense that conditioned on the available information (past and present), no change is expected on the average for the future values, and hence the conditional expectation of the immediate future value is the same as that of the present value. Thus, if

$$E\{X_{n+1} \mid X_n, X_{n-1}, \cdots, X_1, X_0\} = X_n \qquad (1)$$

for all n, then the sequence $\{X_n\}$ represents a **Martingale**.

Example: Gambler's fortune. Historically *martingales* refer to the "doubling the stake" strategy in gambling where the gambler doubles the bet on every loss till the almost sure win occurs eventually at which point the entire loss is recovered by the wager together with a modest profit.
If the game is fair, then the gambler's fortune X_n satisfies the definition (1).

Suppose the gambler wins for the first time at the Nth play.
N is a random variable with distribution

$$P(N = n) = 2^{-n}$$

So $P(N < \infty) = 1$.

However, by this time, he will have lost an amount L with mean

$$E[L] = \sum_{n=1}^{\infty} 2^{-n}(1 + 2 + \cdots + 2^{n-2}) = \infty$$

This means that the gambler must be prepared to lose a huge amount of money! And so must be the owner of the casino.

Example: Markov chain. If $\{X_n\}$ refers to a Markov chain, then as we have seen, with

$$p_{ij} = P\{X_{n+1} = j \mid X_n = i\},$$

Eq. (1) reduces to the simpler expression

$$\sum_j j\, p_{ij} = i. \qquad (2)$$

For finite chains of size N, interestingly, Eq. (2) reads

$$P\,x_2 = x_2, \qquad x_2 = [1,\ 2,\ 3,\ \cdots,\ N]^T$$

implying that x_2 is a right-eigenvector of the $N \times N$ transition probability matrix $P = (p_{ij})$ associated with the eigenvalue 1. However, the "all one" vector $x_1 = [1,\ 1,\ 1,\ \cdots,\ 1]^T$ is always an eigenvector for any P corresponding to the unit eigenvalue [see Eq. (15-179), Text], and from Perron's theorem and the discussion there [Theorem 15-8, Text] it follows that, for finite Markov chains that are also martingales, P *cannot* be a *primitive* matrix, and the corresponding chains are in fact *not irreducible*. Hence every finite state martingale has at least two closed sets embedded in it. (The closed sets in the two martingales in Example 15-13, Text correspond to two absorbing states.)

Example (DeMoivre's Martingale): The gambler's ruin problem gives rise to various martingales.
From there, if S_n refers to player A's cumulative capital at stage n, (note that $S_0 = \$\, a$), then as DeMoivre observed

$$Y_n = \left(\frac{q}{p}\right)^{S_n}$$

generates a martingale. This follows since

$$S_{n+1} = S_n + Z_{n+1}$$

where the instantaneous gain or loss given by Z_{n+1} obeys

$$P\{Z_{n+1} = 1\} = p, \qquad P\{Z_{n+1} = -1\} = q,$$

and hence

$$E\{Y_{n+1} \mid Y_n, Y_{n-1}, \cdots, Y_0\} = E\{\left(\frac{q}{p}\right)^{S_{n+1}} \mid S_n, S_{n-1}, \cdots, S_0\}$$

$$= E\{\left(\frac{q}{p}\right)^{S_n + Z_{n+1}} \mid S_n\},$$

since $\{S_n\}$ generates a Markov chain.

Thus

$$E\{Y_{n+1} \mid Y_n, Y_{n-1}, \cdots, Y_0\} = \left(\frac{q}{p}\right)^{S_n} \left(\frac{q}{p} \cdot p + \left(\frac{q}{p}\right)^{-1} \cdot q\right) = \left(\frac{q}{p}\right)^{S_n} \quad = Y_n$$

i.e., Y_n defines a martingale!

Martingales have excellent convergence properties in the long run. To start with, from (1) for any *given* n, taking expectations on both sides we get

$$E\{X_{n+1}\} = E\{X_n\} = E\{X_0\}. \qquad (3)$$

Observe that, as stated, (3) is true only when n is *known* or n is a *given* number. Under what conditions is this result true if we replace n by a *random time* T? i.e., if T is a random variable, then when is

$$E\{X_T\} \stackrel{?}{=} E\{X_0\}.$$

The answer turns out to be that T has to be a *stopping time*. What is a **stopping time**?

A stochastic process may be known to assume a particular value, but the time at which it happens is in general unpredictable or random. In other words, the nature of the outcome is fixed but the timing is random. When that outcome actually occurs, the time instant corresponds to a *stopping time*. Consider a gambler starting with $a and let T refer to the time instant at which his capital becomes zero. The random variable T represents a stopping time (Time to go home for the gambler!). It corresponds to the gambler's ruin.

Recall that in a Poisson process the occurrences of the first, second, $\cdots$ arrivals correspond to stopping times $T_1, T_2, \cdots$. Stopping times refer to those random instants at which there is sufficient information to decide whether or not a specific condition is satisfied.

Stopping Time: The random variable T is a stopping time for the process $X(t)$, if for all $t \geq 0$, the event $\{T \leq t\}$ is a function of the values $\{X(\tau) \mid \tau > 0, \tau \leq t\}$ of the process up to t, i.e., it should be possible to decide whether T has occurred or not by the time t, knowing *only* the value of the process $X(t)$ up to that time t. Thus the Poisson arrival times T_1 and T_2 referred above are stopping times.

A key result in martingales states that so long as

T is a stopping time (under some additional mild restrictions)

$$E\{X_T\} = E\{X_0\}. \qquad (10)$$

See [Grimmett & Stirzaker, Chap. 12.5] for a proof.
Notice that (10) generalizes (3) to certain random time instants (stopping times) as well. Eq. (10) is an extremely useful tool in analyzing martingales. We shall illustrate its usefulness by rederiving the gambler's ruin probability in last lecture.

We make use of DeMoivre's martingale Y_n in the gambler's ruin problem. Let T refer to the random instant at which the game ends; i.e., the instant at which either player A loses all his wealth and P_a is the associated probability of ruin for player A, or player A gains all wealth $\$(a + b) = \$N$ with probability $(1 - P_a)$. In that case, T is a stopping time and hence from (10), we get

$$E\{Y_T\} = E\{Y_0\} = \left(\frac{q}{p}\right)^a$$

since player A starts with \$a (recall the definition of Y_n). But

$$E\{Y_T\} = \left(\frac{q}{p}\right)^0 P_a + \left(\frac{q}{p}\right)^{a+b} (1 - P_a)$$

$$= P_a + \left(\frac{q}{p}\right)^{a+b} (1 - P_a).$$

Equating the two equations and simplifying we get

$$P_a = \frac{1 - \left(\frac{p}{q}\right)^b}{1 - \left(\frac{p}{q}\right)^{a+b}}$$

This equation can be used to derive other useful probabilities and advantageous plays as well.

As the following result shows, martingales do not fluctuate wildly. There is in fact only a small probability that a large deviation for a martingale from its initial value will occur.

Generalization

Generalized definition of martingales. A sequence $\{S_n : n \geq 1\}$ with finite mean is a martingale with respect to the sequence $\{X_n : n \geq 1\}$ if for all $n \geq 1$

$$E[S_{n+1}|X_1, X_2, \cdots, X_n] = S_n$$

The extra generality is useful in the following way: A specific sequence $\{X_n\}$ may itself not be a martingale. However, it is often possible to find some function ϕ such that $\{S_n = \phi(X_n) : n \geq 1\}$ is a martingale.

We call $\{S_n : n \geq 1\}$ a **submartingale** if

$$E[S_{n+1}|X_1, X_2, \cdots, X_n] \geq S_n$$

or a **supermartingale** if

$$E[S_{n+1}|X_1, X_2, \cdots, X_n] \leq S_n$$

For most casinos, the gambler's fortune in actually a supermartingale.

Note that $\{S_n\}$ is a martingale if it is both a submartingale and a supermatingale. Also, $\{S_n\}$ is a submartingale if and only if $\{-S_n\}$ is a supermartingale.

Doob decomposition. A submartingale $\{S_n\}$ with finite means may be expressed in the form

$$S_n = M_n + Y_n$$

where $\{M_n\}$ is a martingale, and $\{Y_n\}$ is an increasing predictable sequence.

Example: Submartingale. Let X_1, X_2, ... be independent random variables with zero means and finite variances. The partial sum $S_n = \sum_{i=1}^{n} X_i$ forms a martingale (exercise). Define

$$T_n = S_n^2 = \left(\sum_{i=1}^{n} X_i\right)^2$$

Then

$$E[T_{n+1}|X_1, \ldots, X_n] = E[S_n^2 + 2S_n X_{n+1} + X_{n+1}^2 \big| X_1, \ldots, X_n]$$
$$= T_n + E[X_{n+1}^2] \geq T_n$$

Thus, $\{T_n\}$ is not a martingale, but a submartingale.

Example: Random walk. A particle jumps either one step to the right or one step to the left, with corresponding probabilities p and $q = 1 - p$. Assuming the usual independence, the position of the particle after n steps $S_n = \sum_{i=1}^{n} X_i$ satisfies

$$E[S_{n+1}|X_1, \ldots, X_n] = S_n + (p - q)$$

Thus, $\{S_n\}$ is a martingale only if $p = q = ½$.

If $p > q$, it is a submartingale; If $p < q$, it is a supermartingale.

Martingale convergenc theorem

Let $\{S_n\}$ be a submartingale with finite means. There exists a random variable S_∞ such that

$$S_n \to S_\infty \text{ almost surely}$$

as $n \to \infty$.

It follows of course a supermatingale also converges almost surely. See [Grimmett & Stirzaker, Chap. 12] for a proof of this theorem.

This theorem is of immense value and has many applications in communications and signal processing.

Example: Doob's martingale. Let Y have finite second meoment, and let $X_1, X_2, \ldots$ be a sequence of random variables. Define

$$Y_n = E[Y|X_1, X_2, \cdots, X_n]$$

which is recognized as the MMSE estimate of Y, given X_1, $X_2, \ldots X_n$. We claim $\{Y_n\}$ is a martingale with respect to $\{X_n\}$.

To see this, we note that

$$E[Y_{n+1}|X_1, \ldots, X_n] = E[E[Y|X_1, \ldots, X_n, X_{n+1}]|X_1, \ldots, X_n]$$
$$= E[Y|X_1, \ldots, X_n] = Y_n$$

where we applied the tower property

$$E[E[Y|\mathbf{X}_1, \mathbf{X}_2]|\mathbf{X}_1] = E[Y|\mathbf{X}_1].$$

By the convergence theorem, Y_n converges to a random variable Y_∞ almost surely.

Example: Likelihood ratios. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with density function f. Suppose that it is known that f is either p or q. The problem is to decide which of the two is the true density. A common approach is to calculate the likelihood ratio

$$Y_n = \frac{p(X_1)p(X_2)\cdots P(X_n)}{q(X_1)q(X_2)\cdots q(X_n)}$$

and to adopt the strategy:

Decide p if $Y_n \geq a$; decide q if $Y_n < a$

where a is some threshold. $(a > 0)$

If $f = q$, then

$$E[Y_{n+1}\,|X_1, X_2, \cdots, X_n] = Y_n E\left[\frac{p(X_{n+1})}{q(X_{n+1})}\right] =$$

$$Y_n \int_{-\infty}^{\infty} \frac{p(x)}{q(x)} q(x)dx = Y_n$$

Further,

$$E[|Y_n|]$$

$$= \int \frac{p(x_1)p(x_2)\cdots P(x_n)}{q(x_1)q(x_2)\cdots q(x_n)} q(x_1)q(x_2)\cdots q(x_n)dx_1\cdots dx_n = 1$$

It follows that $\{Y_n\}$ is a martingale. By the convergence theorem, the limit $Y_\infty = \lim_{n\to\infty} Y_n$ exists almost surely. We may calculate the limit explicitly as follows: Define the log likelihood ratio

$$\log Y_n = \sum_{i=1}^{n} \log \frac{p(X_i)}{q(X_i)}$$

By Jensen's inequality,

$$E\left[\log \frac{p(X_i)}{q(X_i)}\right] < \log\left[E\left[\frac{p(X_i)}{q(X_i)}\right]\right] = 0$$

By the strong law of large numbers, $n^{-1}\log Y_n$ converge to a negative number, say $b \in (-\infty, 0)$ almost surely. This implies that

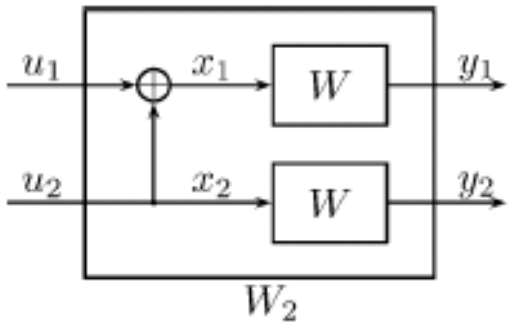
$$Y_n \to Y_\infty = 0 \text{ almost surely.}$$

The fact that $Y_n \to 0$ almost surely tell us that $Y_n < a$ for all large n, and hence the decision strategy gives the correct answer.

Example: Polar codes. Polar codes are capacity-achieving error-correction codes that are deployed in 5G. Consider a binary-input discrete memoryless channel W with mutual information $I(W)$ ($I(W) \in [0, 1]$, which means the amount of information that can be sent through the channel). The construction of a polar code is simple.

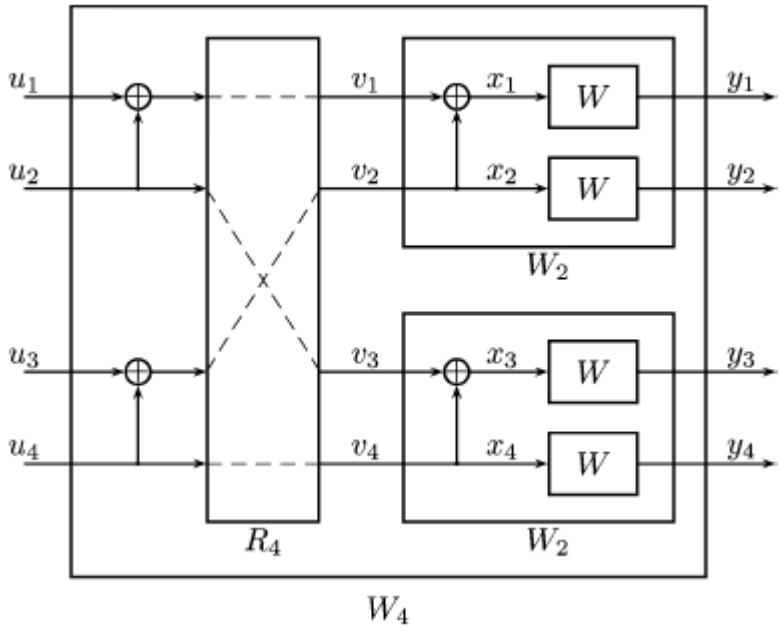
Suppose the codeword length is $N = 2^n$. Let $\mathbf{F}_N = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{\otimes n}$ where $\otimes n$ denotes the n-fold Kronecker product. For example, $\mathbf{F}_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$, $\mathbf{F}_4 = \begin{bmatrix} \mathbf{F}_2 & 0 \\ \mathbf{F}_2 & \mathbf{F}_2 \end{bmatrix}$ and so on. Let $\mathbf{u}$ be the N-length input to the encoder, and let $\mathbf{x} = \mathbf{u}\mathbf{F}_N$ be the codeword. Each element of $\mathbf{x}$ is sent through the channel W, and vector $\mathbf{y}$ is received.

Consider the 'bit channel' $W_N^{(i)}$ from the i-th element u_i of U to the output. Using the martingale theory, we can prove that mutual information $I(W_N^{(i)})$ tends to either 0 or 1, as $n \to \infty$. This is known as the "**polarization**" phenomenon. Then, encoding becomes trivial: put information bits in those u_i's for which $I(W_N^{(i)}) \approx 1$, and put zeros in the others. To prove polarization, we use the recursive structure of a polar code. Firstly, note that the single-step transform gives births to two 'bit channels', while preserving mutual information

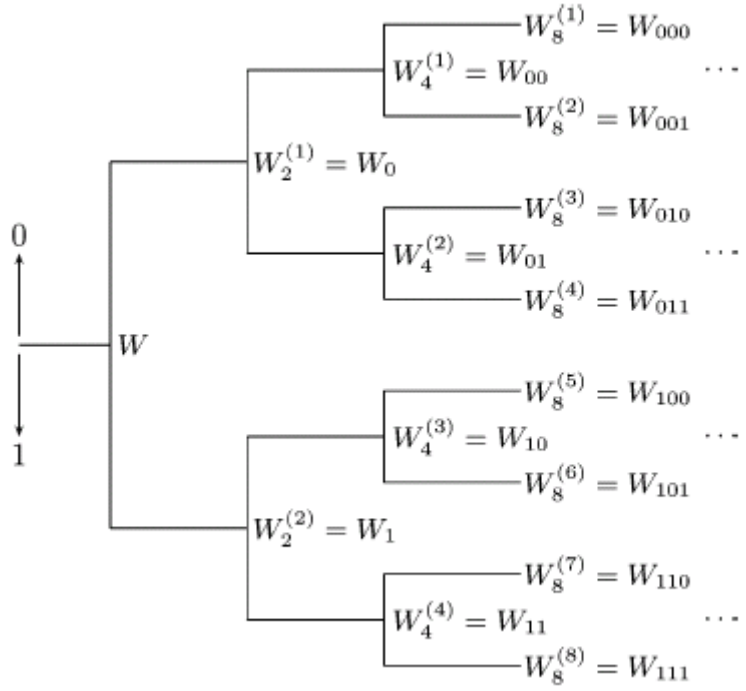


$$I\left(W_2^{(1)}\right) + I\left(W_2^{(2)}\right) = 2I(W)$$
$$I\left(W_2^{(1)}\right) \leq I\left(W_2^{(2)}\right) \tag{11}$$

Then, $W_2^{(1)}$ gives birth to $W_4^{(1)}$ and $W_4^{(2)}$, and so on.



Repeating this, we obtain N 'bit channels' at the n-th step. More conveniently, this process can be described as a binary tree. Note how the 'bit channels' $W_{b_1 b_2 \cdots b_n}$ are labelled in the tree.



Now pick a 'bit channel' uniformly at random on the n-th level of the tree, which is equivalent to a random traverse on the tree, namely, at each step the r.v b_i takes the value of 0 or 1 with equal probability. We claim mutual information I_n at the n-th step is a martingale. This is because

$$E[I_{n+1}|b_1,\ldots,b_n] = \frac{1}{2}I\left(W_{b_1 b_2 \cdots b_n 0}\right) + \frac{1}{2}I\left(W_{b_1 b_2 \cdots b_n 1}\right)$$
$$= I\left(W_{b_1 b_2 \cdots b_n}\right) = I_n$$

again due to the information-preserving property of the transform. By the martingale convergence theorem, I_n converges to a random variable I_∞ such that $E[I_\infty] = E[I_0] = I_0 = I(W)$. In fact, the limit $I_\infty = 0$ or 1 is a binary random variable (these are the fixed points of transform (11)). So the portion of almost prefect bit channels is $I(W)$, meaning that the (symmetric) capacity is achieved.

For details, see

E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Transactions on Information Theory,*, vol. 55, no. 7, pp. 3051–3073, 2009.