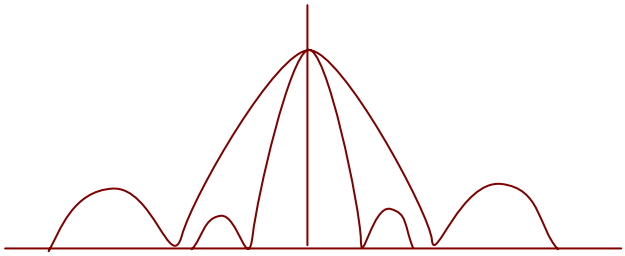
# EC 622 Statistical Signal Processing



**P. K. Bora**

**Department of Electronics & Communication Engineering**

**INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI**

# EC 622 Statistical Signal Processing Syllabus

1. Review of random variables: distribution and density functions, moments, independent, uncorrelated and orthogonal random variables; Vector-space representation of Random variables, Schwarz Inequality Orthogonality principle in estimation, Central Limit theorem, Random process, stationary process, autocorrelation and autocovariance functions, Spectral representation of random signals, Wiener Khinchin theorem, Properties of power spectral density, Gaussian Process and White noise process

2. Linear System with random input, Spectral factorization theorem and its importance, innovation process and whitening filter

3. Random signal modelling: MA(q), AR(p) , ARMA(p,q) models

4. Parameter Estimation Theory: Principle of estimation and applications, Properties of estimates, unbiased and consistent estimators, MVUE, CR bound, Efficient estimators; Criteria of estimation: the methods of maximum likelihood and its properties ; Baysean estimation: Mean Square error and MMSE, Mean Absolute error, Hit and Miss cost function and MAP estimation

5. Estimation of signal in presence of White Gaussian Noise (WGN)

Linear Minimum Mean-Square Error (LMMSE) Filtering: Wiener Hoff Equation
FIR Wiener filter, Causal IIR Wiener filter, Noncausal IIR Wiener filter
Linear Prediction of Signals, Forward and Backward Predictions, Levinson Durbin Algorithm, Lattice filter realization of prediction error filters

6. Adaptive Filtering: Principle and Application, Steepest Descent Algorithm Convergence characteristics; LMS algorithm, convergence, excess mean square error
Leaky LMS algorithm; Application of Adaptive filters ;RLS algorithm, derivation, Matrix inversion Lemma, Intialization, tracking of nonstationarity

7. Kalman filtering: Principle and application, Scalar Kalman filter, Vector Kalman filter

8. Spectral analysis: Estimated autocorrelation function, periodogram, Averaging the periodogram (Bartlett Method), Welch modification, Blackman and Tukey method of smoothing periodogram, Parametric method, AR(p) spectral estimation and detection of Harmonic signals, MUSIC algorithm.

## Acknowledgement

I take this opportunity to thank Prof. A. Mahanta who inspired me to take the course *Statistical Signal Processing*. I am also thankful to my other faculty colleagues of the ECE department for their constant support. I acknowledge the help of my students, particularly Mr. Diganta Gogoi and Mr. Gaurav Gupta for their help in preparation of the handouts. My appreciation goes to Mr. Sanjib Das who painstakingly edited the final manuscript and prepared the power-point presentations for the lectures. I acknowledge the help of Mr. L.N. Sharma and Mr. Nabajyoti Dutta for word-processing a part of the manuscript. Finally I acknowledge QIP, IIT Guwhati for the financial support for this work.

# SECTION – I

## REVIEW OF
## RANDOM VARIABLES & RANDOM PROCESS

# Table of Contents

# SECTION – I

## REVIEW OF
## RANDOM VARIABLES & RANDOM PROCESS

# CHAPTER - 1: REVIEW OF RANDOM VARIABLES

## 1.1 Introduction

- Mathematically a random variable is neither random nor a variable
- It is a mapping from sample space into the real-line ( "real-valued" random variable) or the complex plane ( "complex-valued " random variable) .

Suppose we have a probability space $\{S, \Im, P\}$ .

Let $X : S \to \Re$ be a function mapping the sample space $S$ into the real line such that For each $s \in S$, there exists a unique $X(s) \in \Re$. Then $X$ is called a random variable.

Thus a random variable associates the points in the sample space with real numbers.

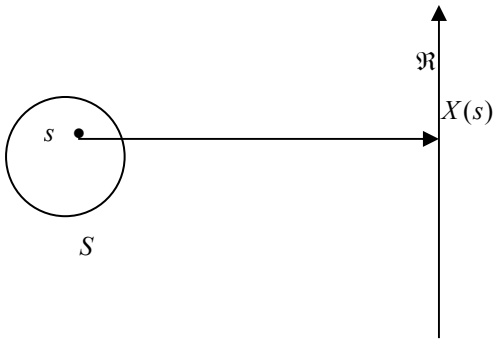



Figure Random Variable

**Notations:**

- Random variables are represented by upper-case letters.
- Values of a random variable are denoted by lower case letters
- $Y = y$ means that $y$ is the value of a random variable $X$.

**Example 1:** Consider the example of tossing a fair coin twice. The sample space is S= {HH, HT, TH, TT} and all four outcomes are equally likely. Then we can define a random variable $X$ as follows

| Sample Point | Value of the random Variable $X = x$ | $P\{X = x\}$ |
|---|---|---|
| HH | 0 | $\frac{1}{4}$ |
| HT | 1 | $\frac{1}{4}$ |
| TH | 2 | $\frac{1}{4}$ |
| TT | 3 | $\frac{1}{4}$ |

**Example 2:** Consider the sample space associated with the single toss of a fair die. The sample space is given by $S = \{1,2,3,4,5,6\}$. If we define the random variable $X$ that associates a real number equal to the number in the face of the die, then $X = \{1,2,3,4,5,6\}$

## 1.2 Discrete and Continuous Random Variables

- A random variable $X$ is called discrete if there exists a countable sequence of distinct real number $x_i$ such that $\sum_i P_m(x_i) = 1$. $P_m(x_i)$ is called the ***probability mass function***. The random variable defined in Example 1 is a discrete random variable.
- A continuous random variable $X$ can take any value from a continuous interval
- A random variable may also de mixed type. In this case the RV takes continuous values, but at each finite number of points there is a finite probability.

## 1.3 Probability Distribution Function

We can define an event $\{X \le x\} = \{s / X(s) \le x, \text{s} \in \text{S}\}$

The probability

$F_X(x) = P\{X \le x\}$ is called the probability distribution function.

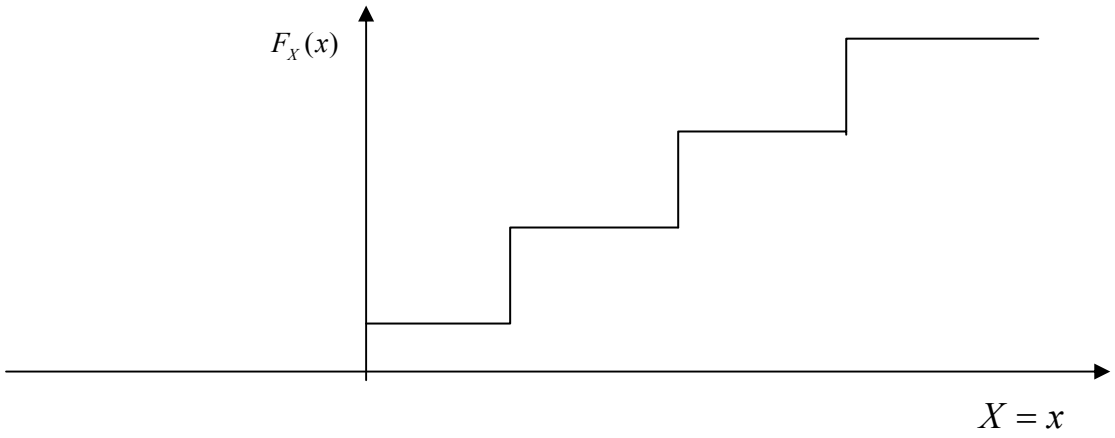
Given $F_X(x)$, we can determine the probability of any event involving values of the random variable $X$.

- $F_X(x)$ is a non-decreasing function of $X$.
- $F_X(x)$ is right continuous

  => $F_X(x)$ approaches to its value from right.
- $F_X(-\infty) = 0$
- $F_X(\infty) = 1$
- $P\{x_1 < X \le x\} = F_X(x) - F_X(x_1)$

**Example 3:** Consider the random variable defined in Example 1. The distribution function $F_X(x)$ is as given below:

| **Value of the random Variable** $X = x$ | $F_X(x)$ |
|---|---|
| $x < 0$ | $0$ |
| $0 \le x < 1$ | $\frac{1}{4}$ |
| $1 \le x < 2$ | $\frac{1}{2}$ |
| $2 \le x < 3$ | $\frac{3}{4}$ |
| $x \ge 3$ | $1$ |




## 1.4 Probability Density Function

If $F_X(x)$ is differentiable $f_X(x) = \frac{d}{dx} F_X(x)$ is called the probability density function and has the following properties.

- $f_X(x)$ is a non- negative function
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$
- $P(x_1 < X \le x_2) = \int_{-x_1}^{x_2} f_X(x)dx$

**Remark:** Using the Dirac delta function we can define the density function for a discrete random variables.

## 1.5 Joint random variable

$X$ and $Y$ are two random variables defined on the same sample space $S$. $P\{X \le x, Y \le y\}$ is called the joint distribution function and denoted by $F_{X,Y}(x,y)$.

Given $F_{X,Y}(x,y), -\infty < x < \infty, -\infty < y < \infty,$ we have a complete description of the random variables $X$ and $Y$.

- $P\{0 < X \le x,\ 0 < Y \le y\} = F_{X,Y}(x,y) - F_{X,Y}(x,0) - F_{X,Y}(0,y) + F_{X,Y}(x,y)$
- $F_X(x) = F_{XY}(x,+\infty)$.

  To prove this

$$(X \le x) = (X \le x) \cap (Y \le +\infty)$$
$$\therefore\ F_X(x) = P(X \le x) = P(X \le x, Y \le \infty) = F_{XY}(x,+\infty)$$
$$F_X(x) = P(X \le x) = P(X \le x, Y \le \infty) = F_{XY}(x,+\infty)$$

Similarly $F_Y(y) = F_{XY}(\infty, y)$.

- Given $F_{X,Y}(x,y), -\infty < x < \infty, -\infty < y < \infty,$ each of $F_X(x)$ and $F_Y(y)$ is called a marginal distribution function.

We can define joint probability density function $f_{X,Y}(x,y)$ of the random variables $X$ and $Y$ by

$f_{X,Y}(x,y) = \dfrac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$, provided it exists

- $f_{X,Y}(x,y)$ is always a positive quantity.
- $F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(x,y)\,dx\,dy$

## 1.6 Marginal density functions

$$\begin{aligned} f_X(x) &= \tfrac{d}{dx} F_X(x) \\ &= \tfrac{d}{dx} F_X(x,\infty) \\ &= \tfrac{d}{dx} \int_{-\infty}^{x} \left( \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \right) dx \\ &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \end{aligned}$$

and $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$

## 1.7 Conditional density function

$f_{Y/X}(y/X=x) = f_{Y/X}(y/x)$ is called conditional density of $Y$ given $X$.

Let us define the conditional distribution function.

We cannot define the conditional distribution function for the continuous random variables $X$ and $Y$ by the relation

$$F_{Y/X}(y/x) = P(Y \le y / X = x)$$
$$= \frac{P(Y \le y, X = x)}{P(X = x)}$$

as both the numerator and the denominator are zero for the above expression.

The conditional distribution function is defined in the limiting sense as follows:

$$F_{Y/X}(y/x) = lim_{\Delta x \to 0} P(Y \le y / x < X \le x + \Delta x)$$
$$= lim_{\Delta x \to 0} \frac{P(Y \le y, x < X \le x + \Delta x)}{P(x < X \le x + \Delta x)}$$
$$= lim_{\Delta x \to 0} \frac{\int_{\infty}^{y} f_{X,Y}(x,u)\Delta x du}{f_X(x)\Delta x}$$
$$= \frac{\int_{\infty}^{y} f_{X,Y}(x,u) du}{f_X(x)}$$

The conditional density is defined in the limiting sense as follows

$$f_{Y/X}(y/X=x) = \lim_{\Delta y \to 0}(F_{Y/X}(y+\Delta y/X=x) - F_{Y/X}(y/X=x))/\Delta y$$
$$= \lim_{\Delta y \to 0, \Delta x \to 0}(F_{Y/X}(y+\Delta y/x<X\le x+\Delta x) - F_{Y/X}(y/x<X\le x+\Delta x))/\Delta y \qquad (1)$$

Because $(X=x) = \lim_{\Delta x \to 0}(x < X \le x + \Delta x)$

The right hand side in equation (1) is

$$\lim_{\Delta y \to 0, \Delta x \to 0}(F_{Y/X}(y+\Delta y/x<X<x+\Delta x) - F_{Y/X}(y/x<X<x+\Delta x))/\Delta y$$
$$= \lim_{\Delta y \to 0, \Delta x \to 0}(P(y<Y\le y+\Delta y/x<X\le x+\Delta x))/\Delta y$$
$$= \lim_{\Delta y \to 0, \Delta x \to 0}(P(y<Y\le y+\Delta y, x<X\le x+\Delta x))/P(x<X\le x+\Delta x)\Delta y$$
$$= \lim_{\Delta y \to 0, \Delta x \to 0} f_{X,Y}(x,y)\Delta x\Delta y / f_X(x)\Delta x\Delta y$$
$$= f_{X,Y}(x,y)/f_X(x)$$

$$\boxed{\therefore f_{Y/X}(x/y) = f_{X,Y}(x,y)/f_X(x)} \qquad (2)$$

Similarly, we have

$$\boxed{\therefore f_{X/Y}(x/y) = f_{X,Y}(x,y)/f_Y(y)} \qquad (3)$$

From (2) and (3) we get Baye's rule

$$\therefore f_{X/Y}(x/y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$
$$= \frac{f_X(x) f_{Y/X}(y/x)}{f_Y(y)}$$
$$= \frac{f_{X,Y}(x,y)}{\int_{-\infty}^{\infty} f_{X,Y}(x,y)dx}$$
$$= \frac{f_{Y/X}(y/x) f_X(x)}{\int_{-\infty}^{\infty} f_X(u) f_{Y/X}(y/x)du} \qquad (4)$$

**Given the joint density function we can find out the conditional density function.**

**Example 4:**

For random variables X and Y, the joint probability density function is given by

$$f_{X,Y}(x,y) = \frac{1+xy}{4} \quad |x| \le 1, \ |y| \le 1$$
$$= 0 \quad \text{otherwise}$$

Find the marginal density $f_X(x), f_Y(y)$ and $f_{Y/X}(y/x)$. Are $X$ and $Y$ independent?

$$f_X(x) = \int_{-1}^{1} \frac{1+xy}{4} dy = \frac{1}{2}$$

Similarly

$$f_Y(y) = \frac{1}{2} \qquad -1 \le y \le 1$$

and

$$f_{Y/X}(y/x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{1+xy}{4}, \ |x| \le 1, \ |y| \le 1$$
$$= 0 \quad \text{otherwise}$$

$\therefore X$ and $Y$ are not independent

## 1.8 Baye's Rule for mixed random variables

Let $X$ be a discrete random variable with probability mass function $P_X(x)$ and $Y$ be a continuous random variable. In practical problem we may have to estimate $X$ from observed $Y$. Then

$$P_{X/Y}(x/y) = \lim_{\Delta y \to 0} P_{X/Y}(x/y < Y \le y + \Delta y)\infty$$

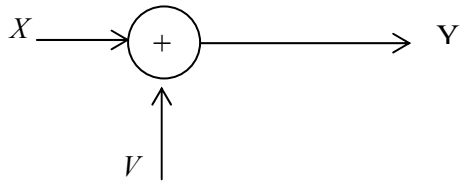
$$= \lim_{\Delta y \to 0} \frac{P_{X,Y}(x, y < Y \le y + \Delta y)}{P_Y(y < Y \le y + \Delta y)}$$

$$= \lim_{\Delta y \to 0} \frac{P_X(x) f_{Y/X}(y/x)\Delta y}{f_Y(y)\Delta y}$$

$$= \frac{P_X(x) f_{Y/X}(y/x)}{f_Y(y)}$$

$$== \frac{P_X(x) f_{Y/X}(y/x)}{\sum_x P_X(x) f_{Y/X}(y/x)}$$

**Example 5:**




$X$ is a binary random variable with

$$X = \begin{cases} 1 \text{ with probability } \frac{1}{2} \\ -1 \text{ with probability } \frac{1}{2} \end{cases}$$

$V$ is the Gaussian noise with mean 0 and variance $\sigma^2$.

Then

$$P_{X/Y}(x = 1/y) = \frac{P_X(x) f_{Y/X}(y/x)}{\sum_x P_X(x) f_{Y/X}(y/x)}$$

$$= \frac{e^{-(y-1)^2/2\sigma^2}}{e^{-(y-1)^2/2\sigma^2} + (e^{-(y+1)^2/2\sigma^2}}$$

## 1.9 Independent Random Variable

Let $X$ and $Y$ be two random variables characterised by the joint density function

$$F_{X,Y}(x, y) = P\{X \le x, Y \le y\}$$

and $f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y)$

Then $X$ and $Y$ are independent if $f_{X/Y}(x/y) = f_X(x) \ \ \forall x \in \Re$

and equivalently

$f_{X,Y}(x, y) = f_X(x) f_Y(y)$ , where $f_X(x)$ and $f_Y(y)$ are called the marginal density functions.

## 1.10 Moments of Random Variables

- Expectation provides a description of the random variable in terms of a few parameters instead of specifying the entire distribution function or the density function
- It is far easier to estimate the expectation of a R.V. from data than to estimate its distribution

**First Moment or mean**

The mean $\mu_X$ of a random variable $X$ is defined by

$$\mu_X = EX = \sum x_i P(x_i) \text{ for a discrete random variable } X$$
$$= \int_{-\infty}^{\infty} x f_X(x)dx \text{ for a continuous random variable } X$$

For any piecewise continuous function $y = g(x)$, the expectation of the R.V. $Y = g(X)$ is given by $EY = Eg(X) = \int_{-\infty}^{-\infty} g(x) f_x(x)dx$

**Second moment**

$$EX^2 = \int_{-\infty}^{\infty} x^2 f_X(x)dx$$

**Variance**

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)dx$$

- Variance is a central moment and measure of dispersion of the random variable about the mean.
- $\sigma_x$ is called the ***standard deviation***.
-

For two random variables $X$ and $Y$ the joint expectation is defined as

$$E(XY) = \mu_{X,Y} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y)dxdy$$

The correlation between random variables $X$ and $Y$, measured by the covariance, is given by

$$Cov(X,Y) = \sigma_{XY} = E(X-\mu_X)(Y-\mu_Y)$$
$$= E(XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y)$$
$$= E(XY) - \mu_X\mu_Y$$

The ratio $$\rho = \frac{E(X-\mu_X)(Y-\mu_Y)}{\sqrt{E(X-\mu_X)^2 E(Y-\mu_Y)^2}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

is called the ***correlation coefficient***. The correlation coefficient measures how much two random variables are similar.

## 1.11 Uncorrelated random variables

Random variables $X$ and $Y$ are uncorrelated if covariance

$Cov(X,Y) = 0$

Two random variables may be dependent, but still they may be uncorrelated. If there exists correlation between two random variables, one may be represented as a linear regression of the others. We will discuss this point in the next section.

## 1.12 Linear prediction of $Y$ from $X$

$\hat{Y} = aX + b$ Regression

Prediction error $Y - \hat{Y}$

Mean square prediction error

$E(Y - \hat{Y})^2 = E(Y - aX - b)^2$

For minimising the error will give optimal values of $a$ and $b$. Corresponding to the optimal solutions for $a$ and $b$, we have

$\frac{\partial}{\partial a} E(Y - aX - b)^2 = 0$

$\frac{\partial}{\partial b} E(Y - aX - b)^2 = 0$

Solving for $a$ and $b$, $\hat{Y} - \mu_Y = \frac{1}{\sigma_X^2}\sigma_{X,Y}(x - \mu_X)$

so that $\hat{Y} - \mu_Y = \rho_{X,Y}\frac{\sigma_y}{\sigma_x}(x - \mu_X)$, where $\boxed{\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}}$ is the **correlation coefficient**.

If $\rho_{X,Y} = 0$ then $X$ and $Y$ are uncorrelated.

$\Rightarrow \hat{Y} - \mu_Y = 0$

$\Rightarrow \hat{Y} = \mu_Y$ is the best prediction.

Note that independence => Uncorrelatedness. But uncorrelated generally does not imply independence (except for jointly Gaussian random variables).

**Example 6:**

$Y = X^2$ and $f_X(x)$ is uniformly distributed between (1,-1).

$X$ and $Y$ are dependent, but they are uncorrelated.

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

Because $$= EXY = EX^3 = 0$$

$$= EXEY \quad (\because \mathrm{EX} = 0)$$

In fact for any zero- mean symmetric distribution of X, $X$ and $X^2$ are uncorrelated.

## 1.13 Vector space Interpretation of Random Variables

The set of all random variables defined on a **sample space** form a vector space with respect to addition and scalar multiplication. This is very easy to verify.

## 1.14 Linear Independence

Consider the sequence of random variables $X_1, X_2, ....X_N$.

If $c_1X_1 + c_2X_2 + .... + c_NX_N = 0$ implies that

$c_1 = c_2 = .... = c_N = 0,$ then $X_1, X_2, ....X_N$. are linearly independent.

## 1.15 Statistical Independence

$X_1, X_2, ....X_N$ are statistically independent if

$$f_{X_1, X_2, ....X_N}(x_1, x_2, ....x_N) = f_{X_1}(x_1) f_{X_2}(x_2) .... f_{X_N}(x_N)$$

Statistical independence in the case of zero mean random variables also implies linear independence

## 1.16 Inner Product

If $x$ and $y$ are real vectors in a vector space $V$ defined over the field $\mathbb{R}$, the inner product $< x, y >$ is a scalar such that

$\forall x, y, z \in V$ and $a \in \mathbb{R}$

1. $< x, y > = < y, x >$
2. $< x, x > = \|x\|^2 \geq 0$
3. $< x + y, z > = < x, z > + < y, z >$
4. $< ax, y > = a < x, y >$

In the case of RVs, inner product between $X$ and $Y$ is defined as

$$< X, Y > = EXY = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \, f_{X,Y}(x, y) dy \, dx.$$

Magnitude / Norm of a vector

$$\|x\|^2 = < x, x >$$

So, for R.V.

$$\|X\|^2 = EX^2 = \int_{-\infty}^{\infty} x^2 f_X(x) dx$$

- The set of RVs along with the inner product defined through the joint expectation operation and the corresponding norm defines a Hilbert Space.

## 1.17 Schwary Inequality

For any two vectors $x$ and $y$ belonging to a Hilbert space $V$

$$|< x, y >| \le \|x\| \, \|y\|$$

For RV $X$ and $Y$

$$E^2(XY) \le EX^2 EY^2$$

**Proof:**

Consider the random variable $Z = aX + Y$

$$E(aX + Y)^2 \ge 0$$
$$\Rightarrow a^2 EX^2 + EY^2 + 2aEXY \ge 0$$

Non-negatively of the left-hand side => its minimum also must be nonnegative.

For the minimum value,

$$\frac{dEZ^2}{da} = 0 => a = -\frac{EXY}{EX^2}$$

so the corresponding minimum is $\frac{E^2XY}{EX^2} + EY^2 - 2\frac{E^2XY}{EX^2}$

Minimum is nonnegative =>

$$EY^2 - \frac{E^2XY}{EX^2} \ge 0$$
$$=> \boxed{E^2XY < EX^2 EY^2}$$
$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_{Xx}\sigma_X} = \frac{E(X-\mu_X)(Y-\mu_Y)}{\sqrt{E(X-\mu_X)^2 E(Y-\mu_Y)^2}}$$

From schwarz inequality

$$\|\rho(X,Y)\| \le 1$$

## 1.18 Orthogonal Random Variables

Recall the definition of orthogonality. Two vectors $x$ and $y$ are called orthogonal if $< x, y > = 0$

Similarly two random variables $X$ and $Y$ are called orthogonal if $EXY = 0$

If each of $X$ and $Y$ is zero-mean

$Cov(X,Y) = EXY$

Therefore, if $EXY = 0$ then $Cov(XY) = 0$ for this case.

For zero-mean random variables,

Orthogonality ⇔ uncorrelatedness

## 1.19 Orthogonality Principle

$X$ is a random variable which is not observable. $Y$ is another observable random variable which is statistically dependent on $X$. Given a value of $Y$ what is the best guess for $X$? (Estimation problem).

Let the best estimate be $\hat{X}(Y)$. Then $E(X-\hat{X}(Y))^2$ is a minimum with respect to $\hat{X}(Y)$.

And the corresponding estimation principle is called *minimum mean square error principle*. For finding the minimum, we have

$$\frac{\partial}{\partial \hat{X}} E(X-\hat{X}(Y))^2 = 0$$

$$\Rightarrow \frac{\partial}{\partial \hat{X}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\hat{X}(y))^2 f_{X,Y}(x,y)dydx = 0$$

$$\Rightarrow \frac{\partial}{\partial \hat{X}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\hat{X}(y))^2 f_Y(y) f_{X/Y}(x)dydx = 0$$

$$\Rightarrow \frac{\partial}{\partial \hat{X}} \int_{-\infty}^{\infty} f_Y(y)\left(\int_{-\infty}^{\infty} (x-\hat{X}(y))^2 f_{X/Y}(x)dx\right)dy = 0$$

Since $f_Y(y)$ in the above equation is always positive, therefore the minimization is equivalent to

$$\frac{\partial}{\partial \hat{X}} \int_{-\infty}^{\infty} (x-\hat{X}(y))^2 f_{X/Y}(x)dx = 0$$

$$\text{Or } 2\int_{-\infty}^{\infty} (x-\hat{X}(y)) f_{X/Y}(x)dx = 0$$

$$\Rightarrow \int_{-\infty}^{\infty} \hat{X}(y) f_{X/Y}(x)dx = \int_{-\infty}^{\infty} x f_{X/Y}(x)dx$$

$$\Rightarrow \hat{X}(y) = E(X/Y)$$

Thus, the minimum *mean-square error* estimation involves conditional expectation which is difficult to obtain numerically.

Let us consider a simpler version of the problem. We assume that $\hat{X}(y) = ay$ and the estimation problem is to find the optimal value for $a$. Thus we have the *linear minimum mean-square error* criterion which minimizes $E(X-aY)^2$.

$$\frac{d}{da} E(X-aY)^2 = 0$$

$$\Rightarrow E\frac{d}{da}(X-aY)^2 = 0$$

$$\Rightarrow E(X-aY)Y = 0$$

$$\Rightarrow EeY = 0$$

where $e$ is the estimation error.

The above result shows that for the *linear minimum mean-square error* criterion, estimation error is orthogonal to data. This result helps us in deriving optimal filters to estimate a random signal buried in noise.

The mean and variance also give some quantitative information about the bounds of RVs. Following inequalities are extremely useful in many practical problems.

## 1.20 Chebysev Inequality

Suppose $X$ is a parameter of a manufactured item with known mean $\mu_X$ and variance $\sigma_X^2$. The quality control department rejects the item if the absolute deviation of $X$ from $\mu_X$ is greater than $2\sigma_X$. What fraction of the manufacturing item does the quality control department reject? Can you roughly guess it?

The standard deviation gives us an intuitive idea how the random variable is distributed about the mean. This idea is more precisely expressed in the remarkable *Chebysev Inequality* stated below. For a random variable $X$ with mean $\mu_X$ and variance $\sigma_X^2$

$$P\{|X-\mu_X| \geq \varepsilon\} \leq \frac{\sigma_X^2}{\varepsilon^2}$$

**Proof:**

$$\begin{aligned}
\sigma_x^2 &= \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)dx \\
&\geq \int_{|X-\mu_X|\geq\varepsilon} (x-\mu_X)^2 f_X(x)dx \\
&\geq \int_{|X-\mu_X|\geq\varepsilon} \varepsilon^2 f_X(x)dx \\
&= \varepsilon^2 P\{|X-\mu_X| \geq \varepsilon\}
\end{aligned}$$

$$\therefore \quad P\{|X-\mu_X| \geq \varepsilon\} \leq \frac{\sigma_X^2}{\varepsilon^2}$$

## 1.21 Markov Inequality

For a random variable $X$ which take only nonnegative values

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

$$\begin{aligned}
E(X) &= \int_0^{\infty} x f_X(x)dx \\
&\geq \int_a^{\infty} x f_X(x)dx \\
&\geq \int_a^{\infty} a f_X(x)dx \\
&= aP\{X \geq a\}
\end{aligned}$$

$\therefore \quad P\{X \geq a\} \leq \dfrac{E(X)}{a}$

**Result:** $P\{(X-k)^2 \geq a\} \leq \dfrac{E(X-k)^2}{a}$

## 1.22 Convergence of a sequence of random variables

Let $X_1, X_2, ..., X_n$ be a sequence $n$ independent and identically distributed random variables. Suppose we want to estimate the mean of the random variable on the basis of the observed data by means of the relation

$$\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{N} X_i$$

How closely does $\hat{\mu}_X$ represent $\mu_X$ as $n$ is increased? How do we measure the closeness between $\hat{\mu}_X$ and $\mu_X$ ?

Notice that $\hat{\mu}_X$ is a random variable. What do we mean by the statement $\hat{\mu}_X$ converges to $\mu_X$ ?

Consider a deterministic sequence $x_1, x_2, ....x_n....$ The sequence converges to a limit $x$ if correspond to any $\varepsilon > 0$ we can find a positive integer $m$ such that $|x - x_n| < \varepsilon$ for $n > m.$

Convergence of a random sequence $X_1, X_2, ....X_n....$ cannot be defined as above.

A sequence of random variables is said to converge everywhere to $X$ if

$|X(\xi) - X_n(\xi)| \to 0$ for $n > m$ and $\forall \xi.$

## 1.23 Almost sure (a.s.) convergence or convergence with probability 1

For the random sequence $X_1, X_2, ....X_n....$

$\{X_n \to X\}$ this is an event.

If $P\{s \mid X_n(s) \to X(s)\} = 1 \quad as \quad n \to \infty,$
$P\{s \mid X_n(s) - X(s)| < \varepsilon \text{ for } n \geq m\} = 1 \quad as \quad m \to \infty,$

then the sequence is said to converge to $X$ almost sure or with probability 1.

One important application is the ***Strong Law of Large Numbers***:

If $X_1, X_2, ....X_n....$ are iid random variables, then $\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu_X$ with probability 1as $n \to \infty.$

## 1.24 Convergence in mean square sense

If $E(X_n - X)^2 \to 0 \quad as \quad n \to \infty$, we say that the sequence converges to $X$ in mean square (M.S).

**Example 7:**

If $X_1, X_2, ....X_n ....$ are iid random variables, then

$\frac{1}{n}\sum_{i=1}^{N} X_i \to \mu_X$ in the mean square 1as $n \to \infty$.

We have to show that $\lim_{n\to\infty} E(\frac{1}{n}\sum_{i=1}^{N} X_i - \mu_X)^2 = 0$

Now,

$$E(\frac{1}{n}\sum_{i=1}^{N} X_i - \mu_X)^2 = E(\frac{1}{n}(\sum_{i=1}^{N}(X_i - \mu_X))^2$$

$$= \frac{1}{n^2}\sum_{i=1}^{N} E(X_i - \mu_X)^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} E(X_i - \mu_X)(X_j - \mu_X)$$

$$= \frac{n\sigma_X^2}{n^2} + 0 \text{ ( Because of independence)}$$

$$= \frac{\sigma_X^2}{n}$$

$$\therefore \lim_{n\to\infty} E(\frac{1}{n}\sum_{i=1}^{N} X_i - \mu_X)^2 = 0$$

## 1.25 Convergence in probability

$P\{|X_n - X| > \varepsilon\}$ is a sequence of probability. $X_n$ is said to convergent to $X$ in probability if this sequence of probability is convergent that is

$P\{|X_n - X| > \varepsilon\} \to 0 \quad as \quad n \to \infty.$

If a sequence is convergent in mean, then it is convergent in probability also, because

$P\{|X_n - X|^2 > \varepsilon^2\} \le E(X_n - X)^2 / \varepsilon^2$ (Markov Inequality)

We have

$P\{|X_n - X| > \varepsilon\} \le E(X_n - X)^2 / \varepsilon^2$

If $E(X_n - X)^2 \to 0 \quad as \quad n \to \infty,$ (mean square convergent) then

$P\{|X_n - X| > \varepsilon\} \to 0 \quad as \quad n \to \infty.$

**Example 8:**

Suppose $\{X_n\}$ be a sequence of random variables with

$$P(X_n = 1\} = 1 - \frac{1}{n}$$

and

$$P(X_n = -1\} = \frac{1}{n}$$

Clearly

$$P\{|X_n - 1| > \varepsilon\} = P\{X_n = -1\} = \frac{1}{n} \to 0$$

$as\ n \to \infty.$

Therefore $\{X_n\} \xrightarrow{P} \{X = 0\}$

## 1.26 Convergence in distribution

The sequence $X_1, X_2, ....X_n ....$ is said to converge to $X$ in distribution if

$F_{X_n}(x) \to F_X(x) \qquad as \quad n \to \infty.$

Here the two distribution functions eventually coincide.

## 1.27 Central Limit Theorem

Consider independent and identically distributed random variables $X_1, X_2, ....X_n$.

Let $Y = X_1 + X_2 + ...X_n$

Then $\mu_Y = \mu_{X_1} + \mu_{X_2} + ...\mu_{X_n}$

And $\sigma_Y^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 + ... + \sigma_{X_n}^2$

The central limit theorem states that under very general conditions $Y$ converges to $N(\mu_Y, \sigma_Y^2)$ as $n \to \infty.$ The conditions are:

1. The random variables $X_1, X_2, ..., X_n$ are independent with same mean and variance, but not identically distributed.
2. The random variables $X_1, X_2, ..., X_n$ are independent with different mean and same variance and not identically distributed.

## 1.28 Jointly Gaussian Random variables

Two random variables $X$ and $Y$ are called jointly Gaussian if their joint density function is

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

where $A = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{X,Y}^2}}$

**Properties:**

(1) If $X$ and $Y$ are jointly Gaussian, then for any constants $a$ and $b$, then the random variable

$Z$, given by $Z = aX + bY$ is Gaussian with mean $\mu_Z = a\mu_X + b\mu_Y$ and variance

$\sigma_Z{}^2 = a^2\sigma_X{}^2 + b^2\sigma_Y{}^2 + 2ab\sigma_X\sigma_Y\rho_{X,Y}$

(2) If two jointly Gaussian RVs are uncorrelated, $\rho_{X,Y} = 0$ then they are statistically independent.

$f_{X,Y}(x,y) = f_X(x)f_Y(y)$ in this case.

(3) If $f_{X,Y}(x,y)$ is a jointly Gaussian distribution, then the marginal densities $f_X(x)$ and $f_Y(y)$ are also Gaussian.

(4) If $X$ and $Y$ are joint by Gaussian random variables then the optimum nonlinear estimator $\hat{X}$ of $X$ that minimizes the mean square error $\xi = E\{[X - \hat{X}]^2\}$ is a linear estimator $\hat{X} = aY$

# CHAPTER - 2 : REVIEW OF RANDOM PROCESS

## 2.1 Introduction

Recall that a random variable maps each sample point in the sample space to a point in the real line. A random process maps each sample point to a waveform.

- A random process can be defined as an indexed family of random variables $\{X(t),\ t \in T\}$ where $T$ is an index set which may be discrete or continuous usually denoting time.
- The random process is defined on a common probability space $\{S, \Im, P\}$.
- A random process is a function of the sample point $\xi$ and index variable $t$ and may be written as $X(t,\xi)$.
- For a fixed $t(=t_0)$, $X(t_0,\xi)$ is a random variable.
- For a fixed $\xi(=\xi_0)$, $X(t,\xi_0)$ is a single realization of the random process and is a deterministic function.
- When both $t$ and $\xi$ are varying we have the random process $X(t,\xi)$.

The random process $X(t,\xi)$ is normally denoted by $X(t)$.

We can define a discrete random process $X[n]$ on discrete points of time. Such a random process is more important in practical implementations.

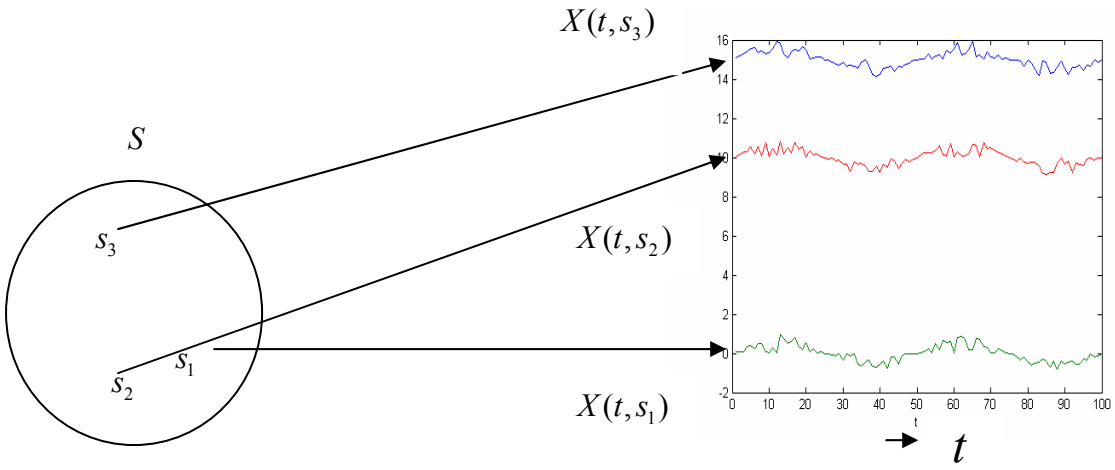



Figure Random Process

## 2.2 How to describe a random process?

To describe $X(t)$ we have to use joint density function of the random variables at different $t$.

For any positive integer $n$, $X(t_1), X(t_2), .....X(t_n)$ represents $n$ jointly distributed random variables. Thus a random process can be described by the joint distribution function $F_{X(t_1),X(t_2).....X(t_n)}(x_1, x_2.....x_n) = F(x_1, x_2.....x_n, t_1, t_2.....t_n), \forall n \in N$ and $\forall t_n \in T$

Otherwise we can determine all the possible moments of the process.

$E(X(t)) = \mu_x(t)$ = mean of the random process at $t$.

$R_X(t_1, t_2) = E(X(t_1)X(t_2))$ = autocorrelation function at $t_1, t_2$

$R_X(t_1, t_2, t_3) = E(X(t_1), X(t_2), X(t_3))$ = Triple correlation function at $t_1, t_2, t_3$, etc.

We can also define the auto-covariance function $C_X(t_1, t_2)$ of $X(t)$ given by

$$\begin{aligned} C_X(t_1, t_2) &= E(X(t_1) - \mu_X(t_1))(X(t_2) - \mu_X(t_2)) \\ &= R_X(t_1, t_2) - \mu_X(t_1)\mu_X(t_2) \end{aligned}$$

**Example 1:**

(a) Gaussian Random Process

For any positive integer $n$, $X(t_1), X(t_2), .....X(t_n)$ represent $n$ jointly random variables. These $n$ random variables define a random vector $\mathbf{X} = [X(t_1), X(t_2), .....X(t_n)]'$. The process $X(t)$ is called Gaussian if the random vector $[X(t_1), X(t_2), .....X(t_n)]'$ is jointly Gaussian with the joint density function given by

$$f_{X(t_1),X(t_2)...X(t_n)}(x_1, x_2, ...x_n) = \frac{e^{-\frac{1}{2}\mathbf{X}'\mathbf{C}_X^{-1}\mathbf{X}}}{\left(\sqrt{2\pi}\right)^n \sqrt{\det(\mathbf{C}_\mathbf{X})}}$$ where $\mathbf{C}_\mathbf{X} = E(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})'$

and $\boldsymbol{\mu}_\mathbf{X} = E(\mathbf{X}) = \left[E(X_1), E(X_2).....E(X_n)\right]'$.

(b) Bernouli Random Process

(c) A sinusoid with a random phase.

## 2.3 Stationary Random Process

A random process $X(t)$ is called strict-sense stationary if its probability structure is invariant with time. In terms of the joint distribution function

$$F_{X(t_1),X(t_2).....X(t_n)}(x_1,x_2.....x_n) = F_{X(t_1+t_0),X(t_2+t_0).....X(t_n+t_0)}(x_1,x_2.....x_n) \ \forall n \in N \text{ and } \forall t_0, t_n \in T$$

For $n = 1$,

$$F_{X(t_1)}(x_1) = F_{X(t_1+t_0)}(x_1) \ \forall t_0 \in T$$

Let us assume $t_0 = -t_1$

$$F_{X(t_1)}(x_1) = F_{X(0)}(x_1)$$

$$\Rightarrow EX(t_1) = EX(0) = \mu_X(0) = \text{constant}$$

**For** $n = 2$,

$$F_{X(t_1),X(t_2)}(x_1,x_2.) = F_{X(t_1+t_0),X(t_2+t_0)}(x_1,x_2)$$

Put $t_0 = -t_2$

$$F_{X(t_1),X(t_2)}(x_1,x_2) = F_{X(t_1-t_2),X(0)}(x_1,x_2)$$

$$\Rightarrow R_X(t_1,t_2) = R_X(t_1-t_2)$$

A random process $X(t)$ is called **wide sense stationary process (WSS)** if

$\mu_X(t) = \text{constant}$

$R_X(t_1,t_2) = R_X(t_1-t_2)$ is a function of time lag.

For a Gaussian random process, WSS implies strict sense stationarity, because this process is completely described by the mean and the autocorrelation functions.

The autocorrelation function $R_X(\tau) = EX(t+\tau)X(t)$ is a crucial quantity for a WSS process.

- $R_X(0) = EX^2(t)$ is the mean-square value of the process.
- $R_X(-\tau) = R_X(\tau)$for real process (for a complex process *X(t)*, $R_X(-\tau) = R_X^*(\tau)$
- $|R_X(\tau)| \le R_X(0)$ which follows from the Schwartz inequality

$$\begin{aligned} R_X^2(\tau) &= \{EX(t)X(t+\tau)\}^2 \\ &= |< X(t), X(t+\tau) >|^2 \\ &\le \|X(t)\|^2 \|X(t+\tau)\|^2 \\ &= EX^2(t)EX^2(t+\tau) \\ &= R_X^2(0)R_X^2(0) \end{aligned}$$

$$\therefore |R_X(\tau)| \le R_X(0)$$

- $R_X(\tau)$ is a positive semi-definite function in the sense that for any positive integer $n$ and real $a_j, a_j$, $\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j R_X(t_i, t_j) \geq 0$
- If $X(t)$ is periodic (in the mean square sense or any other sense like with probability 1), then $R_X(\tau)$ is also periodic.

For a discrete random sequence, we can define the autocorrelation sequence similarly.

- If $R_X(\tau)$ drops quickly , then the signal samples are less correlated which in turn means that the signal has lot of changes with respect to time. Such a signal has high frequency components. If $R_X(\tau)$ drops slowly, the signal samples are highly correlated and such a signal has less high frequency components.
- $R_X(\tau)$ is directly related to the frequency domain representation of WSS process.

The following figure illustrates the above concepts

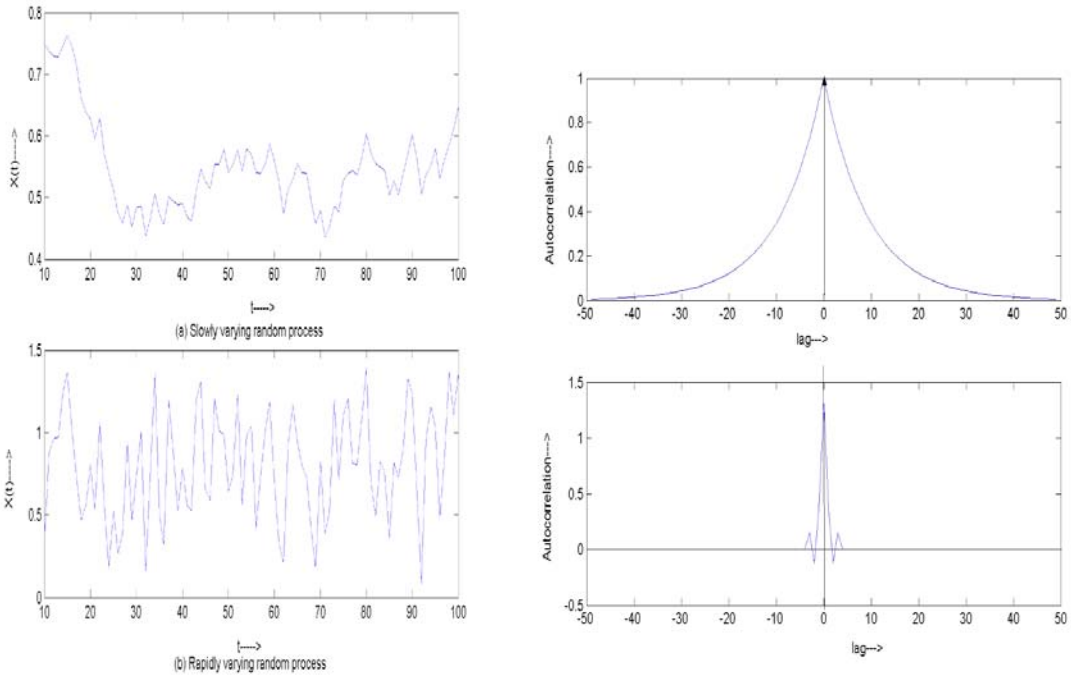



**Figure** Frequency Interpretation of Random process: for slowly varying random process Autocorrelation decays slowly

## 2.4 Spectral Representation of a Random Process

*How to have the frequency-domain representation of a random process?*

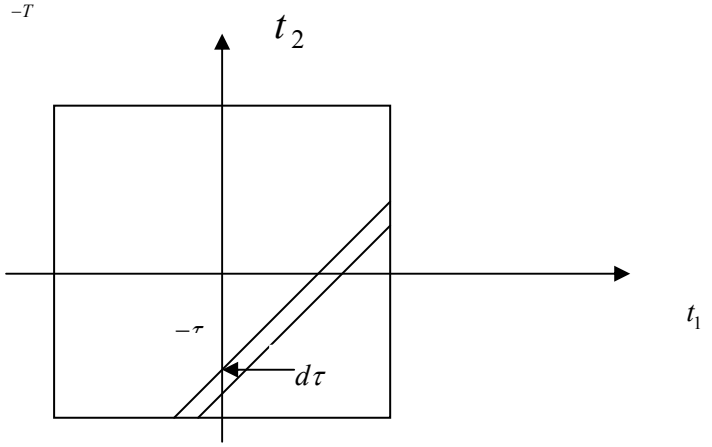
- Wiener (1930) and Khinchin (1934) independently discovered the spectral representation of a random process. Einstein (1914) also used the concept.
- Autocorrelation function and power spectral density forms a Fourier transform pair

Lets define

$$X_T(t) = X(t) \quad -T < t < T$$
$$= 0 \qquad \text{otherwise}$$

as $t \to \infty, X_T(t)$ will represent the random process $X(t)$.

Define $X_T(w) = \int_{-T}^{T} X_T(t) e^{-jwt}\, dt$ in mean square sense.




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

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

$$= \frac{1}{2T}\int_{-2T}^{2T} R_X(\tau)e^{-j\omega\tau}(2T-|\tau|)d\tau$$

Substituting $t_1 - t_2 = \tau$ so that $t_2 = t_1 - \tau$ is a line, we get

$$E\frac{X_T(\omega)X_T^{*}(\omega)}{2T} = \int_{-2T}^{2T} R_x(\tau)e^{-j\omega\tau}(1-\frac{|\tau|}{2T})d\tau$$

If $R_X(\tau)$ is integrable then as $T \to \infty$,

$$\lim_{T\to\infty}\frac{E|X_T(\omega)|^2}{2T} = \int_{-\infty}^{\infty} R_X(\tau)e^{-j\omega\tau}d\tau$$

$\frac{E|X_T(\omega)|^2}{2T}$ = contribution to average power at freq $\omega$ and is called the power spectral density.

Thus $$S_X(\omega) = \int_{-\infty}^{\infty} R_x(\tau) e^{-j\omega\tau} d\tau$$

and $$R_X(\tau) = \int_{-\infty}^{\infty} S_X(\omega) e^{j\omega\tau} dw$$

**Properties**

- $EX^2(t) = R_X(0) = \int_{-\infty}^{\infty} S_X(\omega) dw$ = average power of the process.
- The average power in the band $(w_1, w2)$ is $\int_{w_1}^{w_2} S_X(w) dw$
- $R_X(\tau)$ is real and even $\Rightarrow S_X(\omega)$ is real, even.
- From the definition $S_X(w) = \lim_{T\to\infty} \frac{E|X_T(\omega)|^2}{2T}$ is always positive.
- $h_X(w) = \frac{S_x(\omega)}{EX^2(t)}$ = normalised power spectral density and has properties of PDF, (always +ve and area=1).

## 2.5 Cross-correlation & Cross power Spectral Density

Consider two real random processes *X(t)* and *Y(t).*

Joint stationarity of *X(t)* and *Y(t)* implies that the joint densities are invariant with shift of time.

The ***cross-correlation function*** $R_{X,Y}(\tau)$ for a jointly wss processes *X(t)* and *Y(t)* is defined as

$$R_{X,Y}(\tau) = E\, X(t+\tau)Y(t)$$

so that $$R_{YX}(\tau) = E\, Y(t+\tau)X(t)$$
$$= E\, X(t)Y(t+\tau)$$
$$= R_{X,Y}(-\tau)$$

$$\therefore \quad R_{YX}(\tau) = R_{X,Y}(-\tau)$$

Cross power spectral density

$$S_{X,Y}(w) = \int_{-\infty}^{\infty} R_{X,Y}(\tau) e^{-jw\tau} d\tau$$

For real processes *X(t)* and *Y(t)*

$$S_{X,Y}(w) = S^*_{Y,X}(w)$$

The Wiener-Khinchin theorem is also valid for discrete-time random processes.

If we define $R_X\ [m]\ =\ E\,X[n\ +\ m]\,X[n]$

Then corresponding PSD is given by

$$S_X(w) = \sum_{m=-\infty}^{\infty} R_x[m]e^{-j\omega m} \qquad -\pi \le w \le \pi$$

$$\text{or } S_X(f) = \sum_{m=-\infty}^{\infty} R_x[m]e^{-j2\pi m} \qquad -1 \le f \le 1$$

$$\therefore R_X[m] = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_X(w)e^{j\omega m}dw$$

For a discrete sequence the generalized PSD is defined in the $z-$domain as follows

$$S_X(z) = \sum_{m=-\infty}^{\infty} R_x[m]z^{-m}$$

If we sample a stationary random process uniformly we get a stationary random sequence.

Sampling theorem is valid in terms of PSD.

**Examples 2:**

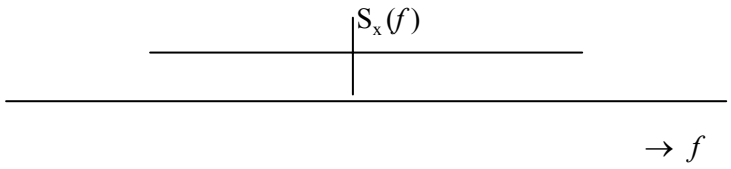
(1) $R_X(\tau) = e^{-a|\tau|} \quad a > 0$

$$S_X(w) = \frac{2a}{a^2 + w^2} \qquad -\infty < w < \infty$$

(2) $R_X(m) = a^{|m|} \quad |a| > 0$

$$S_X(w) = \frac{1-a^2}{1-2a\cos w + a^2} \qquad -\pi \le w \le \pi$$

## 2.6 White noise process




A white noise process $X(t)$ is defined by

$$S_X(f) = \frac{N}{2} \qquad -\infty < f < \infty$$

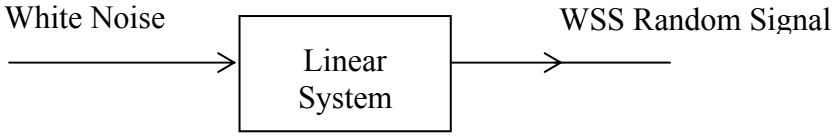
The corresponding autocorrelation function is given by

$$R_X(\tau) = \frac{N}{2}\delta(\tau) \quad \text{where } \delta(\tau) \text{ is the Dirac delta.}$$

The average power of white noise

$$P_{avg} = \int_{-\infty}^{\infty} \frac{N}{2} df \to \infty$$

- Samples of a white noise process are uncorrelated.
- White noise is an mathematical abstraction, it cannot be realized since it has infinite power
- If the system band-width(BW) is sufficiently narrower than the noise BW and noise PSD is flat , we can model it as a white noise process. Thermal and shot noise are well modelled as white Gaussian noise, since they have very flat psd over very wide band (GHzs
- For a zero-mean white noise process, the correlation of the process at any lag $\tau \neq 0$ is zero.
- White noise plays a key role in random signal modelling.
- Similar role as that of the impulse function in the modeling of deterministic signals.




## 2.7 White Noise Sequence

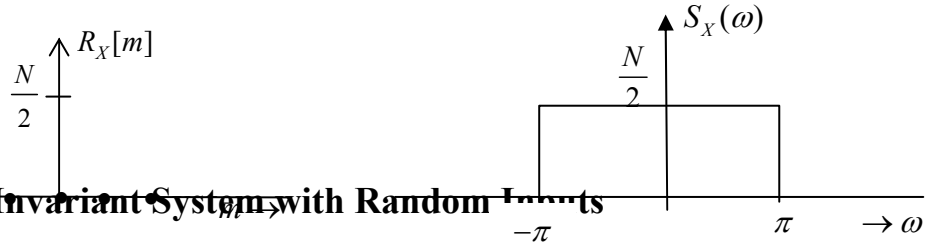
For a white noise sequence $x[n]$,

$$S_X(w) = \frac{N}{2} \qquad -\pi \le w \le \pi$$

Therefore

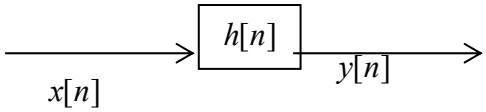
$$R_X(m) = \frac{N}{2}\delta(m)$$

where $\delta(m)$ is the unit impulse sequence.




## 2.8 Linear Shift Invariant System with Random Inputs

Consider a discrete-time linear system with impulse response $h[n]$.




$$y[n] = x[n] * h[n]$$
$$E\,y[n] = E\,x[n] * h[n]$$

For stationary input $x[n]$

$$\mu_Y = E\,y[n] = \mu_X * h[n] = \mu_X \sum_{k=0}^{l} h[n]$$

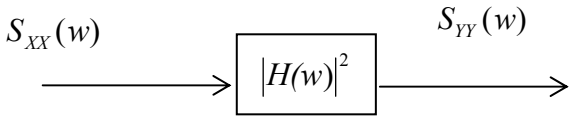
where $l$ is the length of the impulse response sequence

$$\begin{aligned} R_Y[m] &= E\, y[n]y[n-m] \\ &= E(x[n] * h[n]) * (x[n-m] * h[n-m]) \\ &= R_X[m] * h[m] * h[-m] \end{aligned}$$

$R_Y[m]$ is a function of lag $m$ only.

From above we get

$$S_Y(w) = |H(w)|^2 \; S_X(w)$$




**Example 3:**

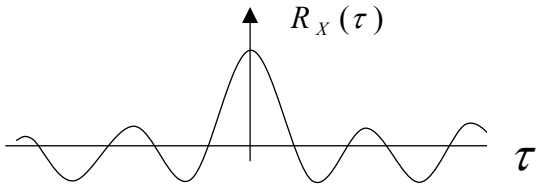
Suppose

$$\begin{aligned} H(w) &= 1 \quad -w_c \le w \le w_c \\ &= 0 \quad \text{otherwise} \end{aligned}$$

$$S_X(w) = \frac{N}{2} \qquad -\infty \le w \le \infty$$

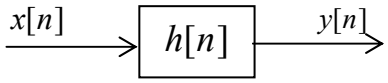
Then $$S_Y(w) = \frac{N}{2} \qquad -w_c \le w \le w_c$$

and $$R_Y(\tau) = \frac{N}{2}\,\text{sinc}(w_c\tau)$$




- Note that though the input is an uncorrelated process, the output is a correlated process.

Consider the case of the discrete-time system with a random sequence $x[n]$ as an input.



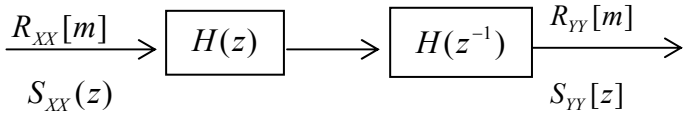

$$R_Y[m] = R_X[m] * h[m] * h[-m]$$

Taking the $z-$transform, we get

$$S_Y(z) = S_X(z)H(z)H(z^{-1})$$

Notice that if $H(z)$ is causal, then $H(z^{-1})$ is anti causal.

Similarly if $H(z)$ is minimum-phase then $H(z^{-1})$ is maximum-phase.



**Example 4:**

If $H(z)=\frac{1}{1-\alpha z^{-1}}$ and $x[n]$ is a unity-variance white-noise sequence, then

$$S_{YY}(z)=H(z)H(z^{-1})$$
$$=\left(\frac{1}{1-\alpha z^{-1}}\right)\left(\frac{1}{1-\alpha z}\right)\frac{1}{2\pi}$$

By partial fraction expansion and inverse $z-$ transform, we get

$$R_Y[m]=\frac{1}{1-\alpha^2}a^{|m|}$$

## 2.9 Spectral factorization theorem

A stationary random signal $X[n]$ that satisfies the Paley Wiener condition $\int_{-\pi}^{\pi}|\ln S_X(w)|\,dw<\infty$ can be considered as an output of a linear filter fed by a white noise sequence.

If $S_X(w)$ is an analytic function of $w$,

and $\int_{-\pi}^{\pi}|\ln S_X(w)|\,dw<\infty$, then $S_X(z)=\sigma_v^2 H_c(z)H_a(z)$

where

$H_c(z)$ is the causal minimum phase transfer function

$H_a(z)$ is the anti-causal maximum phase transfer function

and $\sigma_v^2$ a constant and interpreted as the variance of a white-noise sequence.

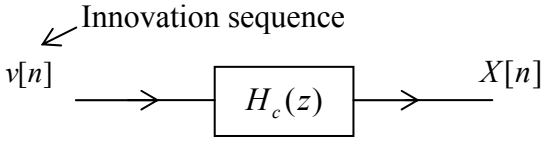



**Figure** Innovation Filter

Minimum phase filter => the corresponding inverse filter exists.

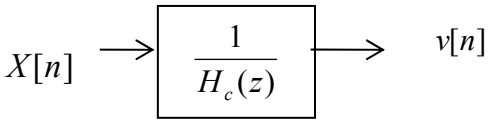



**Figure** whitening filter

Since $\ln S_{XX}(z)$ is analytic in an annular region $\rho<|z|<\frac{1}{\rho}$,

$$\ln S_{XX}(z)=\sum_{k=-\infty}^{\infty}c[k]z^{-k}$$

where $c[k]=\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln S_{XX}(w)e^{iwn}dw$ is the $kth$ order cepstral coefficient.

For a real signal $c[k]=c[-k]$

and $c[0]=\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln S_{XX}(w)dw$

$$S_{XX}(z)=e^{\sum_{k=-\infty}^{\infty}c[k]z^{-k}}$$

$$=e^{c[0]}e^{\sum_{k=1}^{\infty}c[k]z^{-k}}\;e^{\sum_{k=-\infty}^{-1}c[k]z^{-k}}$$

Let $H_C(z)=e^{\sum_{k=1}^{\infty}c[k]z^{-k}} \qquad |z|>\rho$

$$=1+h_c(1)z^{-1}+h_c(2)z^{-2}+.....$$

$(\because h_c[0]=Lim_{z\to\infty}H_C(z)=1$

$H_C(z)$ and $\ln H_C(z)$ are both analytic

=> $H_C(z)$ is a ***minimum phase filter***.

Similarly let

$$H_a(z)=e^{\sum_{k=-\infty}^{-1}c(k)z^{-k}}$$

$$=\;e^{\sum_{k=1}^{\infty}c(k)z^{k}}=H_C(z^{-1}) \qquad |z|<\frac{1}{\rho}$$

Therefore,

$S_{XX}(z)=\sigma_V^2H_C(z)H_C(z^{-1})$

where $\sigma_V^2=e^{c(0)}$

**<u>Salient points</u>**

- $S_{XX}(z)$ can be factorized into a minimum-phase and a maximum-phase factors i.e. $H_C(z)$ and $H_C(z^{-1})$.
- In general spectral factorization is difficult, however for a signal with rational power spectrum, spectral factorization can be easily done.
- Since is a minimum phase filter, $\frac{1}{H_C(z)}$ exists (=> stable), therefore we can have a filter $\frac{1}{H_C(z)}$ to filter the given signal to get the innovation sequence.
- $X[n]$ and $v[n]$ are related through an invertible transform; so they contain the same information.

## 2.10 Wold's Decomposition

Any WSS signal $X[n]$ can be decomposed as a sum of two mutually orthogonal processes

- a regular process $X_r[n]$ and a predictable process $X_p[n]$, $X[n] = X_r[n] + X_p[n]$
- $X_r[n]$ can be expressed as the output of linear filter using a white noise sequence as input.
- $X_p[n]$ is a predictable process, that is, the process can be predicted from its own past with zero prediction error.

# CHAPTER - 3: RANDOM SIGNAL MODELLING

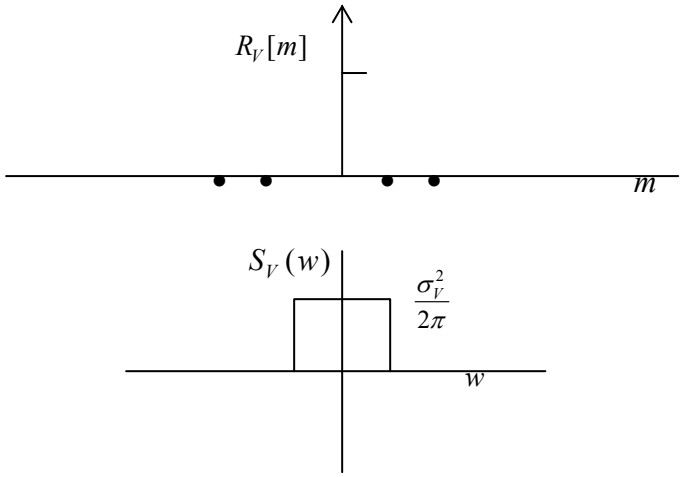
## 3.1 Introduction

The spectral factorization theorem enables us to model a regular random process as an output of a linear filter with white noise as input. Different models are developed using different forms of linear filters.

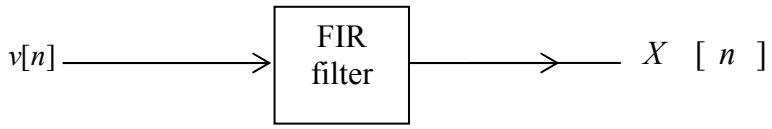
- These models are mathematically described by linear constant coefficient difference equations.
- In statistics, random-process modeling using difference equations is known as ***time series analysis***.

## 3.2 White Noise Sequence

The simplest model is the white noise $v[n]$. We shall assume that $v[n]$ is of 0-mean and variance $\sigma_V^2$.




## 3.3 Moving Average model $MA(q)$ model




The difference equation model is

$$X[n] = \sum_{i=o}^{q} b_i v[n-i]$$

$\mu_e = 0 \Rightarrow \mu_X = 0$

and $v[n]$ is an uncorrelated sequence means

$$\sigma_X{}^2 = \sum_{i=0}^{q} b_i{}^2 \sigma_V^2$$

The autocorrelations are given by

$$R_X[m] = E\,X[n]\,X[n-m]$$
$$= \sum_{i=0}^{q}\sum_{j=0}^{q} b_i b_j E v[n-i]v[n-m-j]$$
$$= \sum_{i=0}^{q}\sum_{j=0}^{q} b_i b_j R_V[m-i+j]$$

Noting that $R_V[m] = \sigma_V^2 \delta[m]$, we get

$R_V[m] = \sigma_V^2$ when
$m - i + j = 0$
$\Rightarrow\ i = m + j$

The maximum value for $m + j$ is $q$ so that

$$R_X[m] = \sum_{j=0}^{q-m} b_j b_{j+m} \sigma_V^2 \qquad 0 \le m \le q$$

and

$$R_X[-m] = R_X[m]$$

Writing the above two relations together $R_X[m] = \sum_{j=0}^{q-|m|} b_j b_{j+|m|} \sigma_v^2 \quad |m| \le q$
$= 0$ otherwise

Notice that, $R_X[m]$ is related by a nonlinear relationship with model parameters. Thus finding the model parameters is not simple.

The power spectral density is given by

$S_X(w) = \frac{\sigma_V^2}{2\pi}|B(w)|^2$ , where $B(w) = = b_o + b_1\, e^{-jw} + .....b_q e^{-jqw}$

FIR system will give some zeros. So if the spectrum has some valleys then MA will fit well.

### 3.3.1 Test for MA process

$R_X[m]$ becomes zero suddenly after some value of $m$.

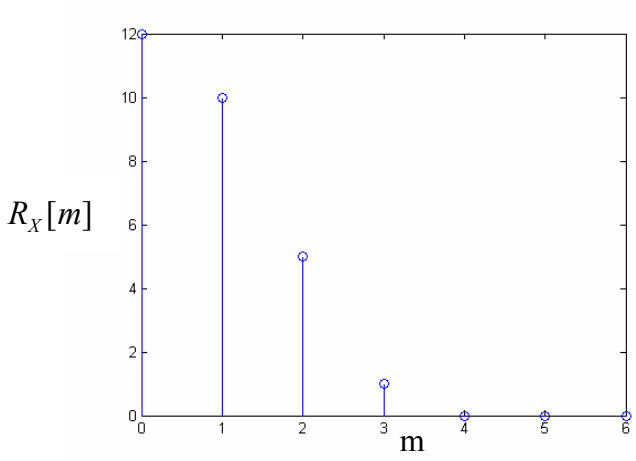



**Figure:** Autocorrelation function of a MA process

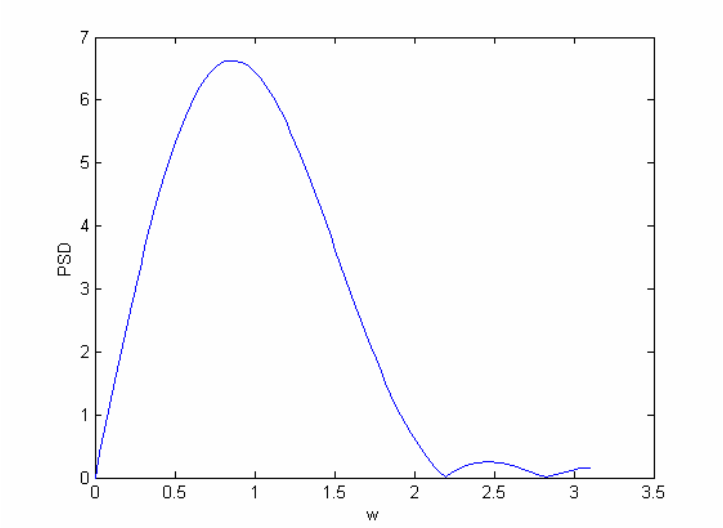
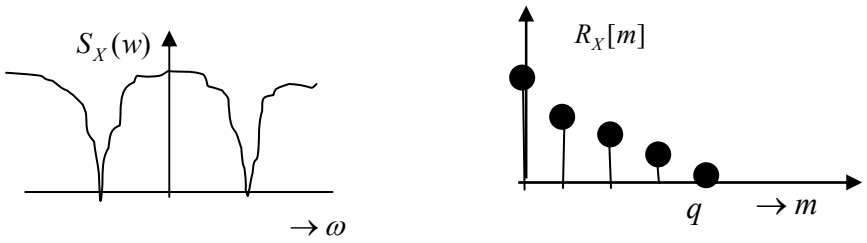



**Figure:** Power spectrum of a MA process




**Example 1: MA(1) process**

$X[n] = b_1 v[n-1] + b_0 v[n]$

Here the parameters $b_0$ and $b_1$ are tobe determined.

We have

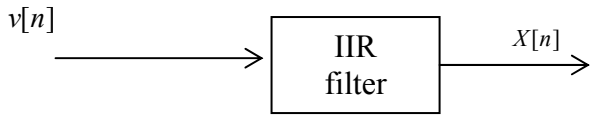
$\sigma_X^2 = b_1^2 + b_0^2$

$R_X[1] = b_1 b_0$

From above $b_0$ and $b_1$ can be calculated using the variance and autocorrelation at lag 1 of the signal.

## 3.4 Autoregressive Model

In time series analysis it is called AR(p) model.




The model is given by the difference equation

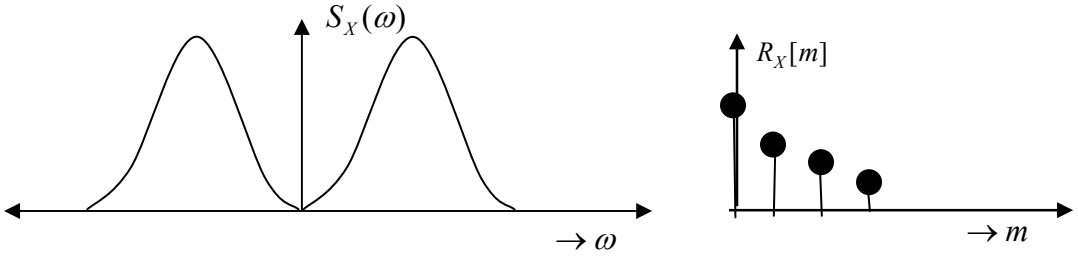
$$X[n] = \sum_{i=1}^{p} a_i X[n-i] + v[n]$$

The transfer function $A(w)$ is given by

$$A(w) = \frac{1}{1 - \sum_{i=1}^{n} a_i e^{-j\omega i}}$$

with $a_0 = 1$ (all poles model) and $S_X(w) = \frac{\sigma_e^{\,2}}{2\pi \mid A(\omega) \mid^2}$

If there are sharp peaks in the spectrum, the AR(p) model may be suitable.




The autocorrelation function $R_X[m]$ is given by

$$\begin{aligned} R_X[m] &= E\, X\,[n]\, X\,[n - m] \\ &= \sum_{i=1}^{p} a_i EX[n-i]X[n-m] + Ev[n]X[n-m] \\ &= \sum_{i=1}^{p} a_i R_X[m-i] + \sigma_V^2 \delta[m] \end{aligned}$$

$$\therefore R_X[m] = \sum_{i=1}^{p} a_i R_X[m-i] + \sigma_V^2 \delta[m] \qquad \forall\, m \in I$$

The above relation gives a set of linear equations which can be solved to find $a_i$s.

These sets of equations are known as Yule-Walker Equation.

**<u>Example 2</u>: AR(1) process**

$X[n] = a_1 X[n-1] + v[n]$

$R_X[m] = a_1 R_X[m-1] + \sigma_V^2 \delta[m]$

$\therefore R_X[0] = a_1 R_X[-1] + \sigma_V^2 \qquad (1)$

and $R_X[1] = a_1 R_X[0]$

so that $a_1 = \frac{R_X[1]}{R_X[0]}$

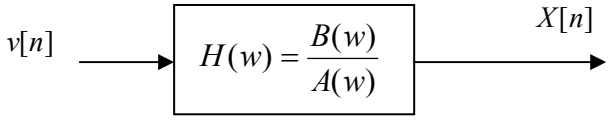
From (1) $\sigma_X^2 = R_X[0] = \frac{\sigma_V^2}{1 - a^2}$

After some arithmatic we get

$R_X[m] = \frac{a^{|m|}\sigma_V^2}{1 - a^2}$

## 3.5 ARMA(p,q) – Autoregressive Moving Average Model

Under the most practical situation, the process may be considered as an output of a filter that has both zeros and poles.




The model is given by

$$x[n] = \sum_{i=1}^{p} a_i X[n-i] + \sum_{i=0}^{q} b_i v[n-i] \qquad \text{(ARMA 1)}$$

and is called the $ARMA(p,q)$ model.

The transfer function of the filter is given by

$$H(w) = \frac{B(\omega)}{A(\omega)}$$

$$S_X(w) = \frac{|B(\omega)|^2 \sigma_V^2}{|A(\omega)|^2 2\pi}$$

How do get the model parameters?

For $m \geq \max(p,\ q+1)$, there will be no contributions from $b_i$ terms to $R_X[m]$.

$$R_X[m] = \sum_{i=1}^{p} a_i R_X[m-i] \qquad m \geq \max(p, q+1)$$

From a set of $p$ Yule Walker equations, $a_i$ parameters can be found out.

Then we can rewrite the equation

$$\widetilde{X}[n] = X[n] + \sum_{i=1}^{p} a_i X[n-i]$$

$$\therefore \widetilde{X}[n] = \sum_{i=0}^{q} b_i v[n-i]$$

From the above equation $b_i s$ can be found out.

The $ARMA(p,q)$ is an economical model. Only $AR(p)$ only $MA(q)$ model may require a large number of model parameters to represent the process adequately. This concept in model building is known as ***the parsimony of parameters***.

The difference equation of the $ARMA(p,q)$ model, given by eq. (ARMA 1) can be reduced to $p$ first-order difference equation give a state space representation of the random process as follows:

$$\mathbf{z[n]} = \mathbf{A}\mathbf{z}[\mathbf{n}-1] + Bu[n]$$
$$X[n] = \mathbf{C}\mathbf{z}[n]$$

where

$$\mathbf{z}[\mathbf{n}] = [x[n]\ X[n-1]....X[n-p]]'$$

$$\mathbf{A} = \begin{bmatrix} a_1\ a_2.....a_p \\ 0\ \ 1.....0 \\ ..... \\ 0\ \ 0.....1 \end{bmatrix},\ \mathbf{B} = [1\ \ 0...0]'\ \text{and}$$

$$\mathbf{C} = [b_0\ b_1...b_q]$$

Such representation is convenient for analysis.

## 3.6 General $ARMA(p,q)$ Model building Steps

- Identification of p and q.
- Estimation of model parameters.
- Check the modeling error.
- If it is white noise then stop.
  Else select new values for p and q
  and repeat the process.

## 3.7 Other model: To model nonstatinary random processes

- **ARIMA model**: Here after differencing the data can be fed to an ARMA model.
- **SARMA model**: Seasonal ARMA model etc. Here the signal contains a seasonal fluctuation term. The signal after differencing by step equal to the seasonal period becomes stationary and ARMA model can be fitted to the resulting data.

# SECTION – II

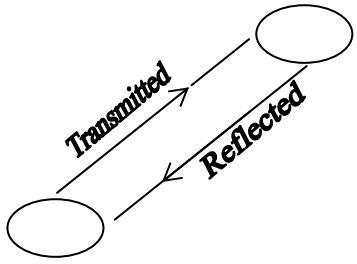
## ESTIMATION THEORY

# CHAPTER – 4: ESTIMATION THEORY

## 4.1 Introduction

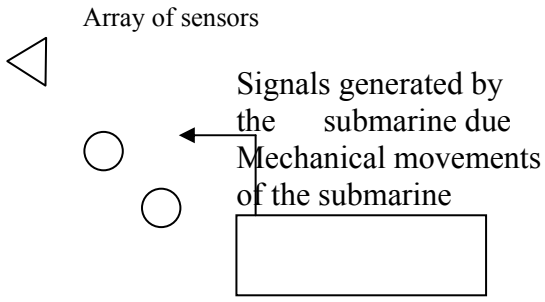
- For speech, we have LPC (linear predictive code) model, the LPC-parameters are to be estimated from observed data.
- We may have to estimate the correct value of a signal from the noisy observation.

In RADAR signal processing




- Estimate the target, target distance from the observed data

In sonar signal processing




- Estimate the location of the submarine.

Generally *estimation* includes parameter *estimation* and signal *estimation*.
We will discuss the problem of parameter estimation here.

We have a sequence of observable random variables $X_1, X_2, ...., X_N$, represented by the vector

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}$$

**X** is governed by a joint density junction which depends on some unobservable parameter θ given by

$f_{\mathbf{X}}(x_1, x_2, ..., x_N \mid \theta) = f_{\mathbf{X}}(\mathbf{x} \mid \theta)$

where $\theta$ may be deterministic or random. Our aim is to make an inference on $\theta$ from an observed sample of $X_1, X_2, ...., X_N$.

An estimator $\hat{\theta}$**(X)** is a rule by which we guess about the value of an unknown $\theta$ on the basis of **X**.

$\hat{\theta}$**(X)** is a random, being a function of random variables.

For a particular observation $x_1, x_2, ...., x_N$, we get what is known as an estimate (not estimator)

Let $X_1, X_2, ...., X_N$ be a sequence of independent and identically distributed (iid) random variables with mean $\mu_X$ and variance $\sigma_X^2$.

$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$ is an estimator for $\mu_X$.

$\hat{\sigma}_X^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_1 - \hat{\mu}_X)^2$ is an estimator for $\sigma_X^2$.

An estimator is a function of the random sequence $X_1, X_2, ...., X_N$ and if it does not involve any unknown parameters. Such a function is generally called a statistic.

## 4.2 Properties of the Estimator

A good estimator should satisfy some properties. These properties are described in terms of the mean and variance of the estimator.

## 4.3 Unbiased estimator

An estimator $\hat{\theta}$ of $\theta$ is said to be unbiased if and only if $E\hat{\theta} = \theta$.

The quantity $E\hat{\theta} - \theta$ is called the bias of the estimator.

Unbiased ness is necessary but not sufficient to make an estimator a good one.

Consider $\hat{\sigma}_1^{\,2} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{\mu}_X)^2$

and $\hat{\sigma}_2^{\,2} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \hat{\mu}_X)^2$

for an iid random sequence $X_1, X_2, ...., X_N$.

We can show that $\hat{\sigma}_2^2$ is an unbiased estimator.

$$E\sum_{i=1}^{N}(X_i - \hat{\mu}_X)^2 = E\sum(X_i - \mu_X + \mu_X - \hat{\mu}_X)^2$$

$$= E\sum\{(X_i - \mu_X)^2 + (\mu_X - \hat{\mu}_X)^2 + 2(X_i - \mu_X)(\mu_X - \hat{\mu}_X)\}$$

Now $E(X_i - \mu_X)^2 = \sigma^2$

and $E(\mu_X - \hat{\mu}_X)^2 = E\left(\mu_X - \frac{\sum X_i}{N}\right)^2$

$$= \frac{E}{N^2}(N\mu_X - \sum X_i)^2$$
$$= \frac{E}{N^2}(\sum(X_i - \mu_X))^2$$
$$= \frac{E}{N^2}\sum(X_i - \mu_X)^2 + \sum_i \sum_{j \neq i} E(X_i - \mu_X)(X_j - \mu_X)$$
$$= \frac{E}{N^2}\sum(X_i - \mu_X)^2 \quad \text{(because of independence)}$$
$$= \frac{\sigma_X^2}{N}$$

also $E(X_i - \mu_X)(\mu_X - \hat{\mu}_X) = -E(X_i - \mu_X)^2$

$$\therefore E\sum_{i=1}^{N}(X_i - \hat{\mu}_X)^2 = N\sigma^2 + \sigma^2 - 2\sigma^2 = (N-1)\sigma^2$$

So $E\hat{\sigma}_2^2 = \frac{1}{N-1}E\sum(X_i - \hat{\mu}_X)^2 = \sigma^2$

$\therefore \hat{\sigma}_2^2$ is an unbiased estimator of $\sigma.^2$

Similarly sample mean is an unbiased estimator.

$$\hat{\mu}_X = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$E\hat{\mu}_X = \frac{1}{N}\sum_{i=1}^{N} E\{X_i\} = \frac{N\mu_X}{N} = \mu_X$$

## 4.4 Variance of the estimator

The variance of the estimator $\hat{\theta}$ is given by

$$\text{var}(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2$$

For the unbiased case

$$\text{var}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

The variance of the estimator should be or low as possible.

An unbiased estimator $\hat{\theta}$ is called a *minimum variance unbiased estimator* (MVUE) if

$E(\hat{\theta} - \theta)^2 \leq E(\hat{\theta}' - \theta)^2$

where $\hat{\theta}'$ is any other unbiased estimator.

## 4.5 Mean square error of the estimator

$MSE = E(\hat{\theta} - \theta)^2$

MSE should be as as small as possible. Out of all unbiased estimator, the MVUE has the minimum mean square error.

$MSE$ is related to the bias and variance as shown below.

$$\begin{aligned} MSE &= E(\hat{\theta} - \theta)^2 = E(\hat{\theta} - E\hat{\theta} + E\hat{\theta} - \theta)^2 \\ &= E(\hat{\theta} - E\hat{\theta})^2 + E(E\hat{\theta} - \theta)^2 + 2E(\hat{\theta} - E\hat{\theta})(E\hat{\theta} - \theta) \\ &= E(\hat{\theta} - E\hat{\theta})^2 + E(E\hat{\theta} - \theta)^2 + 2(E\hat{\theta} - E\hat{\theta})(E\hat{\theta} - \theta) \\ &= \text{var}(\hat{\theta}) + b^2(\hat{\theta}) + 0 \ (why?) \end{aligned}$$

So $\boxed{MSE = \text{var}(\hat{\theta}) + b^2(\hat{\theta})}$

## 4.6 Consistent Estimators

As we have more data, the quality of estimation should be better.

This idea is used in defining the consistent estimator.

An estimator $\hat{\theta}$ is called a consistent estimator of θ if $\hat{\theta}$ converges in probability to θ.

$\lim_{N\to\infty} P\left(\left|\hat{\theta}-\theta\right| \geq \varepsilon\right) = 0$ for any $\varepsilon > 0$

Less rigorous test is obtained by applying the Markov Inequality

MSE ←

$$P\left(\left|\hat{\theta}-\theta\right| \geq \varepsilon\right) \leq \frac{E(\hat{\theta}-\theta)^2}{\varepsilon^2}$$

If $\hat{\theta}$ is an unbiased estimator ( $b(\hat{\theta}) = 0$ ), then $MSE = \text{var}(\hat{\theta})$.

Therefore, if

$\underset{N\to\infty}{Lim}\, E(\hat{\theta}-\theta)^2 = 0$ , then $\hat{\theta}$ will be a consistent estimator.

Also note that

$$MSE = \text{var}(\hat{\theta}) + b^2(\hat{\theta})$$

Therefore, if the estimator is asymptotically unbiased (i.e. $b(\hat{\theta}) \to 0$ as $N \to \infty$) and $\text{var}(\hat{\theta}) \to 0$ as $N \to \infty$, then $MSE \to 0$.

$\therefore$ Therefore for an asymptotically unbiased estimator $\hat{\theta}$, if $\text{var}(\hat{\theta}) \to 0$, as $N \to \infty$, then $\hat{\theta}$ will be a consistent estimator.

**Example 1:** $X_1, X_2, ..., X_N$ is an iid random sequence with unknown $\mu_X$ and known variance $\sigma_X^2$.

Let $\hat{\mu}_X = \frac{1}{N}\sum_{i=1}^{N} X_i$ be an estimator for $\mu_X$. We have already shown that $\hat{\mu}_X$ is unbiased.

Also $\text{var}(\hat{\mu}_X) = \frac{\sigma_X^2}{N}$ Is it a consistent estimator?

Clearly $\lim_{N\to\infty} \text{var}(\hat{\mu}_X) = \lim_{N\to\infty} \frac{\sigma_X^2}{N} = 0$. Therefore $\hat{\mu}_X$ is a consistent estimator of $\mu_X$.

## 4.7 Sufficient Statistic

The observations $X_1, X_2 .... X_N$ contain information about the unknown parameter $\theta$. An estimator should carry the same information about $\theta$ as the observed data. This concept of sufficient statistic is based on this idea.

A measurable function $\hat{\theta}(X_1, X_2 .... X_N)$ is called a sufficient statistic of $\theta$ if it contains the same information about $\theta$ as contained in the random sequence $X_1, X_2 .... X_N$. In other word the joint conditional density $f_{X_1, X_2 .. X_N | \hat{\theta}(X_1, X_2 ... X_N)}(x_1, x_2, ... x_N)$ does not involve $\theta$.

There are a large number of sufficient statistics for a particular criterion. One has to select a sufficient statistic which has good estimation properties.

A way to check whether a statistic is sufficient or not is through the *Factorization theorem* which states:

$\hat{\theta}(X_1, X_2 .... X_N)$ is a sufficient statistic of $\theta$ if

$f_{X_1, X_2 .. X_N / \theta}(x_1, x_2, ... x_N) = g(\theta, \hat{\theta}) h(x_1, x_2, ... x_N)$

where $g(\theta, \hat{\theta})$ is a non-constant and nonnegative function of $\theta$ and $\hat{\theta}$ and $h(x_1, x_2, ... x_N)$ does not involve $\theta$ and is a nonnegative function of $x_1, x_2, ... x_N$.

**Example 2:** Suppose $X_1, X_2, ..., X_N$ is an iid Gaussian sequence with unknown mean $\mu_X$ and known variance 1.

Then $\hat{\mu}_X = \frac{1}{N}\sum_{i=1}^{N} X_i$ is a sufficient statistic of .

$$f_{X_1,X_2..X_N/\mu_X}(x_1,x_2,...x_N)=\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}\left(x_i-\mu_X\right)^2}$$

$$=\frac{1}{(\sqrt{2\pi})^N}\ e^{-\frac{1}{2}\sum_{i=1}^{N}\left(x_i-\mu_X\right)^2}$$

Because

$$=\frac{1}{(\sqrt{2\pi})^N}\ e^{-\frac{1}{2}\sum_{i=1}^{N}\left(x_i-\hat{\mu}+\hat{\mu}-\mu\right)^2}$$

$$=\frac{1}{(\sqrt{2\pi})^N}\ e^{-\frac{1}{2}\sum_{i=1}^{N}\left((x_i-\hat{\mu}_X)^2+(\hat{\mu}-\mu_X)^2+2(x_i-\hat{\mu})(\hat{\mu}-\mu_X)\right)}$$

$$=\frac{1}{(\sqrt{2\pi})^N}\ e^{-\frac{1}{2}\sum_{i=1}^{N}(x_i-\hat{\mu}_X)^2}e^{-\frac{1}{2}\sum_{i=1}^{N}\left((\hat{\mu}-\mu_X)^2\right)}e^0\ \ (why?)$$

The first exponential is a function of $x_1,x_2,...x_N$ and the second exponential is a function of $\mu_X$ and $\hat{\mu}_X$. Therefore $\hat{\mu}_X$ is a sufficient statistics of $\mu_X$.

## 4.8 Cramer Rao theorem

We described about the *Minimum Variance Unbiased Estimator (MVUE*) which is a very good estimator

$\hat{\theta}$ is an MVUE if

$$E(\hat{\theta})=\theta$$

and $Var(\hat{\theta})\le Var(\hat{\theta}')$

where $\hat{\theta}'$ is any other unbiased estimator of $\theta$.

Can we reduce the variance of an unbiased estimator indefinitely? The answer is given by the Cramer Rao theorem.

Suppose $\hat{\theta}$ is an unbiased estimator of random sequence. Let us denote the sequence by the vector

$$\mathbf{X}=\begin{bmatrix}X_1\\X_2\\\vdots\\X_N\end{bmatrix}$$

Let $f_{\mathbf{X}}(x_1,.....,x_N/\theta)$ be the joint density function which characterises $\mathbf{X}$. This function is also called likelihood function. $\theta$ may also be random. In that case likelihood function will represent conditional joint density function.

$L(\mathbf{x}/\theta)=\ln f_{\mathbf{X}}(x_1.....x_N/\theta)$ is called log likelihood function.

## 4.9 Statement of the Cramer Rao theorem

If $\hat{\theta}$ is an unbiased estimator of $\theta$, then

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

where $I(\theta) = E(\frac{\partial L}{\partial \theta})^2$ and $I(\theta)$ is a measure of average information in the random sequence and is called Fisher information statistic.

The equality of CR bound holds if $\frac{\partial L}{\partial \theta} = c(\hat{\theta} - \theta)$ where $c$ is a constant.

**Proof:** $\hat{\theta}$ is an unbiased estimator of $\theta$

$$\therefore E(\hat{\theta} - \theta) = 0 .$$

$$\Rightarrow \int_{-\infty}^{\infty} (\hat{\theta} - \theta) f_{\mathbf{X}}(\mathbf{x}/\theta) d\mathbf{x} = 0.$$

Differentiate with respect to $\theta$, we get

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} \{(\hat{\theta} - \theta) f_{\mathbf{X}}(\mathbf{x}/\theta)\} d\mathbf{x} = 0.$$

(Since line of integration are not function of θ.)

$$= \int_{-\infty}^{\infty} (\hat{\theta} - \theta) \frac{\partial}{\partial \theta} f_{\mathbf{X}}(\mathbf{x}/\theta)\} dy - \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}/\theta) d\mathbf{x} = 0.$$

$$\therefore \int_{-\infty}^{\infty} (\hat{\theta} - \theta) \frac{\partial}{\partial \theta} f_{\mathbf{X}}(\mathbf{x}/\theta)\} dy = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}/\theta) d\mathbf{x} = 1. \qquad (1)$$

Note that $\frac{\partial}{\partial \theta} f_{\mathbf{X}}(\mathbf{x}/\theta) = \frac{\partial}{\partial \theta} \{\ln f_{\mathbf{X}}(\mathbf{x}/\theta)\} f_{\mathbf{X}}(\mathbf{x}/\theta)$

$$= (\frac{\partial L}{\partial \theta}) f_{\mathbf{X}}(\mathbf{x}/\theta)$$

Therefore, from (1)

$$\int_{-\infty}^{\infty} (\hat{\theta} - \theta) \{\frac{\partial}{\partial \theta} L(\mathbf{x}/\theta)\} f_{\mathbf{X}}(\mathbf{x}/\theta)\} d\mathbf{x} = 1.$$

So that

$$\left\{ \int_{-\infty}^{\infty} (\hat{\theta} - \theta) \sqrt{f_{\mathrm{x}}(\mathbf{x}/\theta)} \frac{\partial}{\partial \theta} L(\mathbf{x}/\theta) \sqrt{f_{\mathrm{x}}(\mathbf{x}/\theta)} \mathrm{dx}\, d\mathbf{x} \right\}^2 = 1 . \qquad (2)$$

since $f_{\mathbf{X}}(\mathbf{x}/\theta)$ is $\geq 0$.

Recall the Cauchy Schawarz Ineaquality

$$|< \mathbf{a}, \mathbf{b} >|^2 \leq \|\mathbf{a}\|^2 \ \|\mathbf{b}\|^2$$

where the equality holds when $\mathbf{a} = c\mathbf{b}$ ( where c is any scalar ).

Applying this inequality to the L.H.S. of equation (2) we get

$$\left( \int_{-\infty}^{\infty} (\hat{\theta} - \theta) \sqrt{f_{\mathrm{x}}(\mathbf{x}/\theta)} \frac{\partial}{\partial \theta} L(\mathbf{x}/\theta) \sqrt{f_{\mathrm{x}}(\mathbf{x}/\theta)} \mathrm{dx}\, d\mathbf{x} \right)^2$$

$$\leq \int_{-\infty}^{\infty} (\hat{\theta}\text{-}\theta)^2 f_{\mathrm{x}}(\mathbf{x}/\theta)\, \mathrm{d}\mathbf{x} \quad \int_{-\infty}^{\infty} \left( \left( \frac{\partial}{\partial \theta} L(\mathbf{x}/\theta) \right)^2 f_{\mathrm{x}}(\mathbf{x}/\theta)\, \mathrm{d}\mathbf{x} \right.$$

$$= \mathrm{var}(\hat{\theta})\ \mathrm{I}(\theta)$$

$$\therefore L.H.S \leq \mathrm{var}(\hat{\theta})\, \mathrm{I}(\theta)$$

But R.H.S. = 1

$$\mathrm{var}(\hat{\theta})\, \mathrm{I}(\theta) \geq 1.$$

$$\therefore \mathrm{var}(\hat{\theta}) \geq \frac{1}{I(\theta)},$$

which is the Cramer Rao Inequality.

The equality will hold when

$$\frac{\partial}{\partial \theta} \{ L(\mathbf{x}/\theta) \sqrt{f_{\mathbf{X}}(\mathbf{x}/\theta)} \} = c(\hat{\theta} - \theta) \sqrt{f_{\mathbf{X}}(\mathbf{x}/\theta)}\ ,$$

so that

$$\boxed{\frac{\partial L(\mathbf{x}/\theta)}{\partial \theta} = c\ (\hat{\theta}\text{-}\theta)}$$

Also from $\int_{-\infty}^{\infty} f_{\mathrm{x}}(\mathbf{x}/\theta) d\mathbf{x} = 1,$ we get

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f_{\mathrm{x}}(\mathbf{x}/\theta) d\mathbf{x} = 0$$

$$\therefore \int_{-\infty}^{\infty} \frac{\partial L}{\partial \theta} f_{\mathrm{x}}(\mathbf{x}/\theta) d\mathbf{x} = 0$$

Taking the partial derivative with respect to $\theta$ again, we get

$$\int_{-\infty}^{\infty} \left\{ \frac{\partial^2 L}{\partial \theta^2} f_{\mathrm{x}}(\mathbf{x}/\theta) + \frac{\partial L}{\partial \theta} \frac{\partial}{\partial \theta} f_{\mathrm{x}}(\mathbf{x}/\theta) \right\} d\mathbf{x} = 0$$

$$\therefore \int_{-\infty}^{\infty} \left\{ \frac{\partial^2 L}{\partial \theta^2} f_{\mathrm{x}}(\mathbf{x}/\theta) + \left( \frac{\partial L}{\partial \theta} \right)^2 f_{\mathrm{x}}(\mathbf{x}/\theta) \right\} d\mathbf{x} = 0$$

$$\mathrm{E}\left( \frac{\partial \mathrm{L}}{\partial \theta} \right)^2 = -\,\mathrm{E} \frac{\partial^2 \mathrm{L}}{\partial \theta^2}$$

If $\hat{\theta}$ satisfies CR -bound with equality, then $\hat{\theta}$ is called an *efficient estimator.*

**Remark:**

(1) If the information $I(\theta)$ is more, the variance of the estimator $\hat{\theta}$ will be less.

(2) Suppose $X_1.....X_N$ are iid. Then

$$I_1(\theta) = E\left(\frac{\partial}{\partial\theta}\ln(f_{X_1/\theta}(x))\right)^2$$

$$\therefore I_N(\theta) = E\left(\frac{\partial}{\partial\theta}\ln(f_{X_1,X_2..X_N/\theta}(x_1x_2..x_N))\right)^2$$

$$= NI_1(\theta)$$

**Example 3:**

Let $X_1.....X_N$ are iid Gaussian random sequence with known variance $\sigma^2$ and unknown mean $\mu$.

Suppose $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$ which is unbiased.

Find CR bound and hence show that $\hat{\mu}$ is an efficient estimator.

Likelihood function

$f_{\mathbf{X}}(x_1, x_2,.....x_N/\theta)$ will be product of individual densities (since iid)

$$\therefore f_{\mathbf{X}}(x_1, x_2,.....x_N/\theta) = \frac{1}{(\sqrt{(2\pi)^N\ \sigma^{\mathrm{N}}})}\, e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2}$$

so that $L(\mathbf{X}/\mu) = -\ln(\sqrt{2\pi})^N\ \sigma^{\mathrm{N}}\ -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$

Now $\frac{\partial L}{\partial\mu} = 0 - \frac{1}{2\sigma^2}(-2)\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{X_i}-\mu)$

$$\therefore \frac{\partial^2 L}{\partial\mu^2} = -\frac{\mathrm{N}}{\sigma^2}$$

So that $\mathrm{E}\frac{\partial^2 L}{\partial\mu^2} = -\frac{\mathrm{N}}{\sigma^2}$

$$\therefore \text{CR Bound} = \frac{1}{I(\theta)} = \frac{1}{-\mathrm{E}\frac{\partial^2 L}{\partial\mu^2}} = \frac{1}{\frac{\mathrm{N}}{\sigma^2}} = \frac{\sigma^2}{\mathrm{N}}$$

$$\frac{\partial L}{\partial\theta} = \frac{1}{2\sigma^2}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{X_i}-\mu) = \frac{\mathrm{N}}{\sigma^2}\left(\sum_i \frac{X_i}{N} - \mu\right)$$

$$= \frac{N}{\sigma^2}(\hat{\mu} - \mu)$$

Hence $-\frac{\partial L}{\partial \theta} = c\,(\hat{\theta} - \theta)$

and $\hat{\mu}$ is an efficient estimator.

## 4.10 Criteria for Estimation

The estimation of a parameter is based on several well-known criteria. Each of the criteria tries to optimize some functions of the observed samples with respect to the unknown parameter to be estimated. Some of the most popular estimation criteria are:

- Maximum Likelihood
- Minimum Mean Square Error.
- Baye's Method.
- Maximum Entropy Method.

## 4.11 Maximum Likelihood Estimator (MLE)

Given a random sequence $X_1.....X_N$ and the joint density function $f_{X_1.....X_N/\theta}(x_1, x_2...x_N)$ which depends on an unknown nonrandom parameter $\theta$.

$f_{\mathbf{X}}(x_1, x_2, .....x_N / \theta)$ is called the likelihood function (for continuous function …., for discrete it will be joint probability mass function).

$L(\mathbf{x}/\theta) = \ln f_{\mathbf{X}}(x_1, x_2, ....., x_N / \theta)$ is called log likelihood function.

The maximum likelihood estimator $\hat{\theta}_{MLE}$ is such an estimator that

$$f_{\mathbf{X}}(x_1, x_2.....x_N / \hat{\theta}_{MLE}) \geq f_{\mathbf{X}}(x_1, x_2, ....., x_N / \theta), \forall \theta$$

If the likelihood function is differentiable w.r.t. $\theta$, then $\hat{\theta}_{MLE}$ is given by

$$\frac{\partial}{\partial \theta} f_{\mathbf{X}}(x_1, \ldots x_N / \theta)\Big|_{\hat{\theta}_{MLE}} = 0$$

or $$\frac{\partial L(\mathbf{x} \mid \theta)}{\partial \theta}\Big|_{\hat{\theta}_{MLE}} = 0$$

Thus the MLE is given by the solution of the likelihood equation given above.

If we have a number of unknown parameters given by $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_N \end{bmatrix}$

Then MLE is given by a set of conditions.

$$\frac{\partial \mathrm{L}}{\partial \theta_1}\Big]_{\theta_1=\hat{\theta}_{1\mathrm{MLE}}} = \frac{\partial \mathrm{L}}{\partial \theta_2}\Big]_{\theta_2=\hat{\theta}_{2\mathrm{MLE}}} = .... = \frac{\partial \mathrm{L}}{\partial \theta_{\mathrm{M}}}\Big]_{\theta_{\mathrm{M}}=\hat{\theta}_{\mathrm{MMLE}}} = 0$$

**Example 4:**

Let $X_1.....X_N$ are independent identically distributed sequence of $N(\mu,\sigma^2)$ distributed random variables. Find MLE for μ, $\sigma^2$.

$$f_{\mathbf{X}}(x_1,x_2,.....,x_N/\mu,\sigma^2) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

$$L(X/\mu,\sigma^2) = \ln f_{\mathbf{X}}(X_1,.....,X_N/\mu,\sigma^2)$$

$$= -\,\mathrm{N}\ln - N\ln\sqrt{2\pi} - \mathrm{N}\ln\sigma - \frac{1}{2}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\frac{\mathrm{x_i}-\mu}{\sigma}\right)^2$$

$$\frac{\partial L}{\partial \mu} = 0 \Rightarrow \sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\frac{\mathrm{x_i}-\mu}{\sigma}\right)^2 = \sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{x_i}-\hat{\mu}_{\mathrm{MLE}}) = 0$$

$$\frac{\partial \mathrm{L}}{\partial \sigma} = 0 = -\frac{\mathrm{N}}{\hat{\sigma}_{\mathrm{MLE}}} + \frac{\sum(\mathrm{x_i}-\hat{\mu}_{\mathrm{MLE}})^2}{\hat{\sigma}_{\mathrm{MLE}}} = 0$$

Solving we get

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N}x_i \quad \textit{and}$$

$$\hat{\sigma}_{MLE}{}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu}_{MLE})^2$$

**Example 5:**

Let $X_1.....X_N$ are independent identically distributed sequence with

$$f_{\mathbf{X}/\theta}(x) = \frac{1}{2}e^{-|x-\theta|} \quad -\infty < x < \infty$$

Show that the median of $X_1.....X_N$ is the MLE for $\theta$.

$$f_{X_1,X_2....X_N/\theta}(x_1,x_2....x_N) = \frac{1}{2^N}e^{-\sum_{i=1}^{N}|x_i-\theta|}$$

$$L(X/\theta) = \ln f_{\mathbf{X}/\theta}(x_1,.....,x_N) = -N\ln 2 - \sum_{i=1}^{N}|x_i-\theta|$$

$\sum_{i=1}^{N}|x_i-\theta|$ is minimized by $\textit{median}\ (X_1.....X_N)$

**Some properties of MLE (without proof)**

- MLE may be biased or unbiased, asymptotically unbiased.
- MLE is consistent estimator.
- If an efficient estimator exists, it is the MLE estimator.

An efficient estimator $\hat{\theta}$ exists =>

$$\frac{\partial}{\partial\theta} L(\mathbf{x}/\theta) = c(\hat{\theta} - \theta)$$

at $\theta = \hat{\theta}$,

$$\frac{\partial L(\mathbf{x}/\theta)}{\partial\theta}\Big|_{\hat{\theta}} = c(\hat{\theta} - \hat{\theta}) = 0$$

$\Rightarrow \hat{\theta}$ is the MLE estimator.

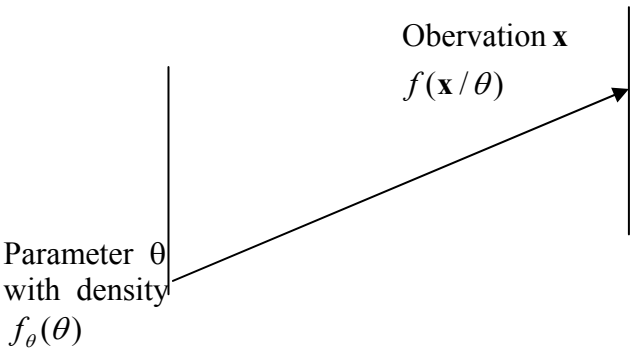
- Invariance Properties of MLE:

If $\hat{\theta}_{MLE}$ is the MLE of $\theta$, then $h(\hat{\theta}_{MLE})$ is the MLE of $h(\theta)$, where $h(\theta)$ is an invertible function of $\theta$.

## 4.12 Bayescan Estimators

We may have some prior information about θ in a sense that some values of θ are more likely (*a priori* information). We can represent this prior information in the form of a prior density function.

In the folowing we omit the suffix in density functions just for notational simplicity.




The likelihood function will now be the conditional density $f(\mathbf{x}/\theta)$.

$$f_{X,\Theta}(\mathbf{x},\theta) = f_\Theta(\theta) f_{\mathbf{X}/\Theta}(x)$$

Also we have the Bayes rule

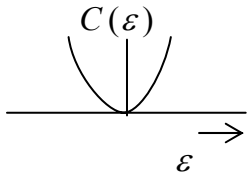
$$f_{\Theta/\mathbf{X}}(\theta) = \frac{f_\Theta(\theta) f_{\mathbf{X}/\Theta}(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}$$

where $f_{\Theta/\mathbf{X}}(\theta)$ is the *a posteriori* density function

The parameter $\theta$ is a random variable and the estimator $\hat{\theta}(\mathbf{x})$ is another random variable.

Estimation error $\varepsilon = \hat{\theta} - \theta$.

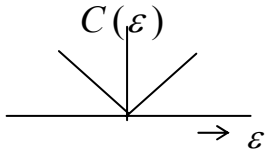
We associate a cost function $C(\hat{\theta}, \theta)$ with every estimator $\hat{\theta}$. It represents the positive penalty with each wrong estimation.




Thus $C(\hat{\theta}, \theta)$ is a non negative function.

The three most popular cost functions are:

**Quadratic cost function** $(\hat{\theta} - \theta)^2$



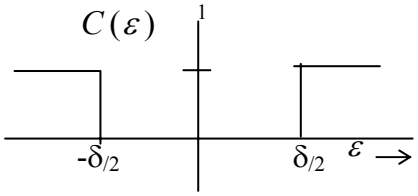

**Absolute cost function** $|\hat{\theta} - \theta|$

**Hit or miss cost function** (also called uniform cost function)

minimising means minimising on an average)

## 4.13 Bayesean Risk function or average cost

$$\overline{C} = EC(\theta, \hat{\theta}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} C(\theta, \hat{\theta})\, f_{\mathbf{X},\Theta}(\mathbf{x}, \theta) d\mathbf{x}\, d\theta$$




The estimator seeks to minimize the Bayescan Risk.

**<u>Case I</u>. Quadratic Cost Function**

$$C = (\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$$

Estimation problem is

Minimize $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\theta - \hat{\theta})^2 f_{\mathbf{X},\Theta}(\mathbf{x}, \theta) d\mathbf{x}\, d\theta$

with respect to $\hat{\theta}$.

This is equivalent to minimizing

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\theta - \hat{\theta})^2 f(\theta \mid \mathbf{x}) f(\mathbf{x}) d\theta\, d\mathbf{x}$$

$$= \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} (\theta - \hat{\theta})^2 f(\theta \mid \mathbf{x}) d\theta \right) f(\mathbf{x})\, d\mathbf{x}$$

Since $f(\mathbf{x})$ is always +ve, the above integral will be minimum if the inner integral is minimum. This results in the problem:

Minimize $\int_{-\infty}^{\infty} (\theta - \hat{\theta})^2 f(\theta \mid \mathbf{x}) d\theta)$

with respect to $\hat{\theta}$.

$$\Rightarrow \frac{\partial}{\partial \hat{\theta}} \int_{-\infty}^{\infty} (\hat{\theta} - \theta)^2 \, f_{\Theta/\mathbf{X}}(\theta) \, d\theta = 0$$

$$\Rightarrow -2 \int_{-\infty}^{\infty} (\hat{\theta} - \theta) \, f_{\Theta/\mathbf{X}}(\theta) \, d\theta = 0$$

$$\Rightarrow \hat{\theta} \int_{-\infty}^{\infty} f_{\Theta/\mathbf{X}}(\theta) \, d\theta = \int_{-\infty}^{\infty} \theta \, f_{\Theta/\mathbf{X}}(\theta) \, d\theta$$

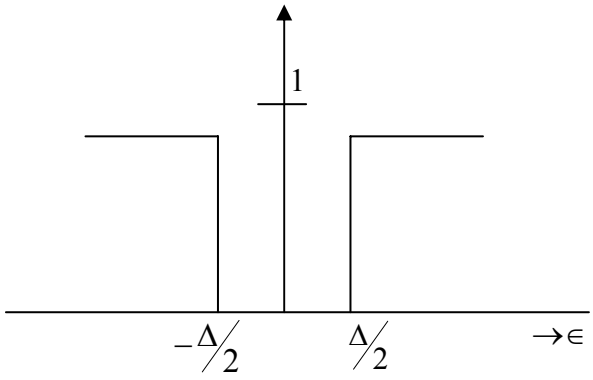
$$\Rightarrow \hat{\theta} = \int_{-\infty}^{\infty} \theta \, f_{\Theta/\mathbf{X}}(\theta) d\theta$$

$\therefore \hat{\theta}$ is the conditional mean or mean of the a posteriori density. Since we are minimizing quadratic cost it is also called *minimum mean square error estimator* (MMSE).

**Salient Points**

- Information about distribution of $\theta$ available.
- *a priori* density function $f_{\Theta}(\theta)$ is available. This denotes how observed data depend on $\theta$
- We have to determine a posteriori density $f_{\Theta/\mathbf{X}}(\theta)$. This is determined form the Bayes rule.

**Case II Hit or Miss Cost Function**




Risk $\overline{C} = EC(\theta, \hat{\theta}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} C(\theta, \hat{\theta}) \, f_{\mathbf{X},\Theta}(\mathbf{x}, \theta) d\mathbf{x} \, d\theta$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} c(\theta, \hat{\theta}) \; f_{\Theta/\mathbf{X}}(\theta) f_{\mathbf{X}}(\mathbf{x}) d\theta \, d\mathbf{x}$$

$$= \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} c(\theta, \hat{\theta}) \, f_{\Theta/\mathbf{X}}(\theta) d\theta \right) f_{\mathbf{X}}(\mathbf{x}) \, d\mathbf{x}$$

We have to minimize

$$\int_{-\infty}^{\infty} C(\theta, \hat{\theta}) \, f_{\Theta/\mathbf{X}}(\theta) d\theta \qquad \text{with respect to} \quad \hat{\theta}.$$

This is equivalent to minimizing

$$= 1 - \int_{\hat{\theta} - \frac{\Delta}{2}}^{\hat{\theta} + \frac{\Delta}{2}} f_{\Theta/\mathbf{X}}(\theta) \, d\theta$$

This minimization is equivalent to maximization of

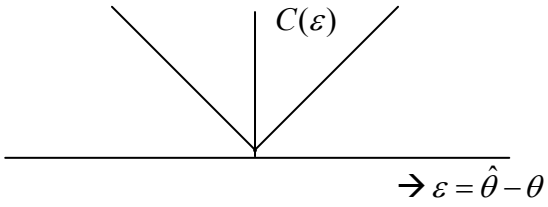
$$\int_{\hat{\theta}-\frac{\Delta}{2}}^{\hat{\theta}+\frac{\Delta}{2}} f_{\Theta/\mathbf{X}}(\theta)\,d\theta \cong \Delta f_{\Theta/\mathbf{X}}(\hat{\theta}) \qquad \text{when } \Delta \text{ is very small}$$

This will be maximum if $f_{\Theta/\mathbf{X}}(\theta)$ is maximum. That means select that value of $\hat{\theta}$ that maximizes the a posteriori density. So this is known as maximum *a posteriori* estimation (MAP) principle.

This estimator is denoted by $\hat{\theta}_{MAP}$ .

**Case III**

$$C(\hat{\theta},\theta) = \left|\hat{\theta}-\theta\right|$$




$\overline{C}$ = Average cost=E$\left|\hat{\theta}-\theta\right|$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left|\hat{\theta}-\theta\right| f_{\theta,\mathbf{X}}(\theta,\mathbf{x})\,d\theta\,d\mathbf{x}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left|\hat{\theta}-\theta\right| f_{\theta}(\theta) f_{\mathbf{X}/\theta}(\mathbf{x})\,d\theta\,d\mathbf{x}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left|\hat{\theta}-\theta\right| f_{\mathbf{X}/\theta}(\mathbf{x})\,d\mathbf{x} f_{\theta}(\theta)\,d\theta$$

For the minimum

$$\frac{\partial}{\partial\hat{\theta}} \int_{-\infty}^{\infty} C(\hat{\theta},\theta) f_{\theta/X}(\theta \mid x)\, d\theta \;=\; 0$$

$$\frac{\partial}{\partial\hat{\theta}} \left\{ \int_{-\infty}^{\hat{\theta}} (\hat{\theta}-\theta) f_{\theta/X}(\theta \mid x)\, d\theta \;+\; \int_{\hat{\theta}}^{\infty} (\theta-\hat{\theta}) f_{\theta/X}(\theta \mid x)\, d\theta = 0 \right.$$

Leibniz rule for differentiation of integration

$$\frac{\partial}{\partial u} \int_{\emptyset_1(u)}^{\emptyset_2(u)} h\,(u,v)\,dv \;=\; \int_{\emptyset_1(u)}^{\emptyset_2(u)} \frac{\partial h(u,v)}{\partial u}\; dv$$

$$+\frac{\mathrm{d}\emptyset_2(u)}{\mathrm{du}}\; h(u,\emptyset_2(u)) - \frac{\mathrm{d}\emptyset_1(u)}{\mathrm{du}}\; h(u,\emptyset_1(u))$$

Applying Leibniz rule we get

$$\int_{-\infty}^{\hat{\theta}} f_{\theta/X}(\theta \mid x)\, d\theta \;-\; \int_{\hat{\theta}}^{\infty} f_{\theta/X}(\theta \mid x)\, d\theta = 0$$

At the $\hat{\theta}_{MAE}$

$$\int_{-\infty}^{\hat{\theta}_{MAE}} f_{\theta/X}(\theta \mid x)\, d\theta \;-\; \int_{\hat{\theta}_{MAE}}^{\infty} f_{\theta/X}(\theta \mid x)\, d\theta = 0$$

So $\hat{\theta}_{MAE}$ is the median of the *a posteriori* density

**Example 6:**

Let $X_1, X_2 ...., X_N$ be an *iid* Gaussian sequence with unity Variance and unknown mean $\theta$. Further $\theta$ is known to be a 0-mean Gaussian with Unity Variance. Find the MAP estimator for $\theta$.

**Solution:** We are given

$$f_\Theta(\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\theta^2}$$

$$f_{\Theta/\mathbf{X}}(\theta) = \frac{1}{(\sqrt{2\pi})^N} e^{-\sum_{i=1}^{N}\frac{(x_i-\theta)^2}{2}}$$

Therefore $f_{\Theta/\mathbf{X}}(\theta) = \dfrac{f_\Theta(\theta) f_{\mathbf{X}/\Theta}(\mathbf{x})}{f_X(\mathbf{x})}$

We have to find $\theta$, such that $f_{\Theta/\mathbf{X}}(\theta)$ is maximum.

Now $f_{\Theta/\mathbf{X}}(\theta)$ is maximum when $f_\Theta(\theta) f_{\mathbf{X}/\Theta}(\mathbf{x})$ is maximum.

$\Rightarrow \ln f_\Theta(\theta) f_{\mathbf{X}/\Theta}(\mathbf{x})$ is maximum

$$\Rightarrow -\frac{1}{2}\theta^2 - \sum_{i=1}^{N}\frac{(x_i-\theta)^2}{2} \text{ is maximum}$$

$$\Rightarrow \theta - \sum_{i=1}^{N}(x_i - \theta)\Big]_{\theta=\hat{\theta}_{MAP}} = 0$$

$$\Rightarrow \hat{\theta}_{MAP} = \frac{1}{N+1}\sum_{i=1} x_i$$

**Example 7**:

Consider single observation $X$ that depends on a random parameter $\theta$. Suppose $\theta$ has a prior distribution

$f_\theta(\theta) = \lambda e^{-\lambda\theta}$ for $\theta \geq 0, \quad \lambda > 0$

and

$f_{X/\Theta}(x) = \theta e^{-\theta x} \quad |x| > 0$

find the MAP estimation for $\theta$.

$$f_{\Theta/X}(\theta) = \frac{f_\Theta(\theta) f_{X/\Theta}(x)}{f_X(x)}$$

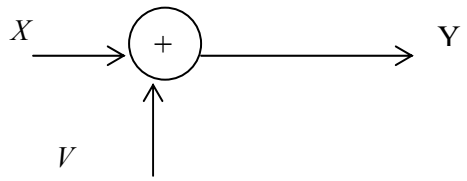
$$\ln(f(\theta \mid x)) = \ln(f_\Theta(\theta)) + \ln(f_{X/\Theta}(x)) - \ln f_X(x)$$

Therefore MAP estimator is given by.

$$\frac{\partial}{\partial\theta} \ln f_{X/\Theta}(x)\Big]_{\hat{\theta}_{MAP}} = 0$$

$$\Rightarrow \hat{\theta}_{MAP} = \frac{1}{\lambda + X}$$

**Example 8**: Binary Communication problem




$X$ is a binary random variable with

$$X = \begin{cases} 1 \text{ with probability } 1/2 \\ -1 \text{ with probability } 1/2 \end{cases}$$

$V$ is the Gaussian noise with mean 0 and variance $\sigma^2$.

To find the MSE for $X$ from the observed data $Y$.

Then

$$f_x(x) = \frac{1}{2}[\delta(x-1) + \delta(x+1)]$$

$$f_{Y/X}(y/x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-x)^2/2\sigma^2}$$

$$f_{X/Y}(x/y) = \frac{f_X(x) f_{Y/X}(y/x)}{\int_{-\infty}^{\infty} f_X(x) f_{Y/X}(y/x) fx}$$

$$= \frac{e^{-(y-x)^2/2\sigma^2}[\delta(x-1) + \delta(x+1)]}{e^{-(y-1)^2/2\sigma^2} + e^{-(y+1)^2/2\sigma^2}}$$

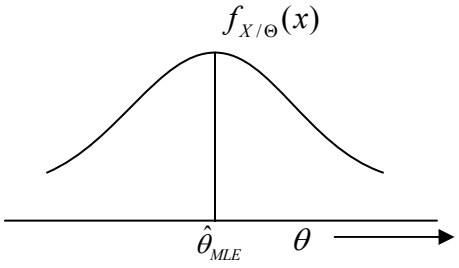
$$\hat{X}_{MMSE} = E(X/Y) = \int_{-\infty}^{\infty} x \frac{e^{-(y-x)^2/2\sigma^2}[\delta(x-1) + \delta(x+1)]}{e^{-(y-1)^2/2\sigma^2} + e^{-(y+1)^2/2\sigma^2}} dx$$

Hence

$$= \frac{e^{-(y-1)^2/2\sigma^2} - e^{-(y+1)^2/2\sigma^2}}{e^{-(y-1)^2/2\sigma^2} + e^{-(y+1)^2/2\sigma^2}}$$

$$= \tanh(y/\sigma^2)$$

## To summarise




**MLE:**

Simplest

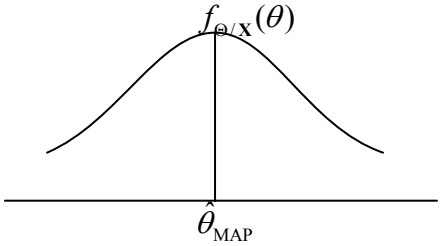
**MMSE:**

$\hat{\theta}_{MMSE} = E(\Theta / \mathbf{X})$ MMSE

- Find a posteriori density.
- Find the average value by integration
- Lots of calculation hence it is computationally exhaustive.

**MAP:**

$\hat{\theta}_{\text{MAP}}$ = Mode of the *a posteriori* density $f_{\Theta/\mathbf{X}}(\theta)$.




## 4.14 Relation between $\hat{\theta}_{\text{MAP}}$ and $\hat{\theta}_{\text{MLE}}$

From

$$f_{\Theta/\mathbf{X}}(\theta) = \frac{f_{\Theta}(\theta) f_{\mathbf{X}/\Theta}(\mathbf{x})}{f_X(\mathbf{x})}$$

$$\ln(f_{\Theta/\mathbf{X}}(\theta)) = \ln(f_{\Theta}(\theta)) + \ln(f_{\mathbf{X}/\Theta}(\mathbf{x})) - \ln(f_X(\mathbf{x}))$$

$\hat{\theta}_{\text{MAP}}$ is given by

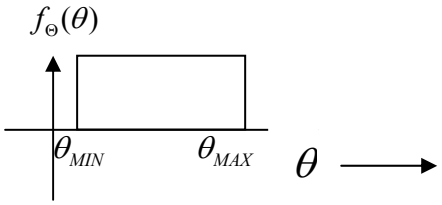
$$\frac{\partial}{\partial\theta}\ln f_{\Theta}(\theta) + \underbrace{\frac{\partial}{\partial\theta}\ln(f_{\mathbf{X}/\Theta}(\mathbf{x}))}_{} = 0$$

a priori density  likelihood function.

Supose $\theta$ is uniformly distributed between $\theta_{MIN}$ and $\theta_{MAX}$.

Then

$$\frac{\partial}{\partial\theta}\ln f_{\Theta}(\theta) = 0$$




If $\theta_{MIN} \le \hat{\theta}_{MLE} \le \theta_{MAX}$
then $\hat{\theta}_{MAP} = \hat{\theta}_{MLE}$

If $\hat{\theta}_{MAP} \le \theta_{MIN}$
then $\hat{\theta}_{MAP} = \theta_{MIN}$

If $\hat{\theta}_{MLE} \ge \theta_{MAX}$
then $\hat{\theta}_{MAP} = \theta_{MAX}$

# SECTION – III

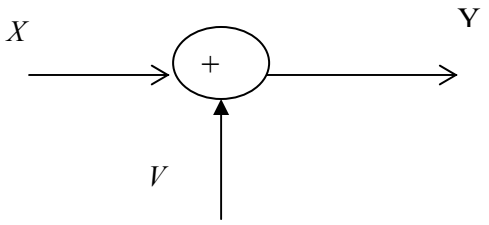
## OPTIMAL FILTERING

# CHAPTER – 5: WIENER FILTER

## 5.1 Estimation of signal in presence of white Gaussian noise (WGN)

Consider the signal model is

$$Y[n] = X[n] + V[n]$$




where $Y[n]$ is the observed signal, $X[n]$ is 0-mean Gaussian with variance 1 and $V[n]$ is a white Gaussian sequence mean 0 and variance 1. The problem is to find the best guess for $X[n]$ given the observation $Y[i], i = 1,2,...,n$

Maximum likelihood estimation for $X[n]$ determines that value of $X[n]$ for which the sequence $Y[i], i = 1,2,...,n$ is most likely. Let us represent the random sequence $Y[i], i = 1,2,...,n$ by the random vector $\mathbf{Y[n]} = [Y[n], Y[n-1],.., Y[1]]'$ and the value sequence $y[1], y[2],..., y[n]$ by $\mathbf{y[n]} = [y[n], y[n-1],.., y[1]]'$.

The likelihood function $f_{\mathbf{Y}[n]/X[n]}(\mathbf{y}[n]/x[n])$ will be Gaussian with mean $x[n]$

$$f_{\mathbf{Y}[n]/X[n]}(\mathbf{y}[n]/x[n]) = \frac{1}{\left(\sqrt{2\pi}\right)^n} e^{-\sum_{i=1}^{n}\frac{(y[i]-x[n])^2}{2}}$$

Maximum likelihood will be given by

$$\frac{\partial}{\partial x[n]}(f_{Y[1],Y[2],...,Y[n]/X[n]}(y[1], y[2],..., y[n])/x[n])\Big|_{\hat{x}_{MLE}[n]} = 0$$

$$\Rightarrow \hat{\chi}_{MLE}[n] = \frac{1}{n}\sum_{i=1}^{n} y[i]$$

Similarly, to find $\hat{\chi}_{MAP}[n]$ and $\hat{\chi}_{MMSE}[n]$ we have to find *a posteriori density*

$$f_{X[n]/\mathbf{Y}[n]}(x[n]/\mathbf{y}[\mathbf{n}]) = \frac{f_{X[n]}(x[n]) f_{\mathbf{Y}[n]/X[n]}(\mathbf{y}[n]/x[n])}{f_{\mathbf{Y}[n]}(\mathbf{y}[n])}$$

$$= \frac{1}{f_{\mathbf{Y}[n]}(\mathbf{y}[n])} e^{-\frac{1}{2}x^2[n]-\sum_{i=1}^{n}\frac{(y[i]-x[n])^2}{2}}$$

Taking logarithm

$$\log_e f_{X[n]/\mathbf{Y}[n]}(x[n]) = -\frac{1}{2}x^2[n] - \sum_{i=1}^{n}\frac{(y[i]-x[n])^2}{2} - \log_e f_{\mathbf{Y}[n]}(\mathbf{y}[n])$$

$\log_e f_{X[n]/\mathbf{Y}[n]}(x[n])$ is maximum at $\hat{x}_{MAP}[n]$. Therefore, taking partial derivative of $\log_e f_{X[n]/\mathbf{Y}[n]}(x[n])$ with respect to $x[n]$ and equating it to 0, we get

$$x[n] - \sum_{i=1}^{n}(y[i]-x[n])\Big|_{\hat{x}_{MAP}[n]} = 0$$

$$\hat{x}_{\text{MAP}}[n] = \frac{\sum_{i=1}^{n} y[i]}{n+1}$$

Similarly the minimum mean-square error estimator is given by

$$\hat{x}_{\text{MMSE}}[n] = E(X[n]/\mathbf{y}[n]) = \frac{\sum_{i=1}^{n} y[i]}{n+1}$$

- For MMSE we have to know the joint probability structure of the channel and the source and hence the *a posteriori pdf.*
- Finding pdf is computationally very exhaustive and nonlinear.
- Normally we may be having the estimated values first-order and second-order statistics of the data

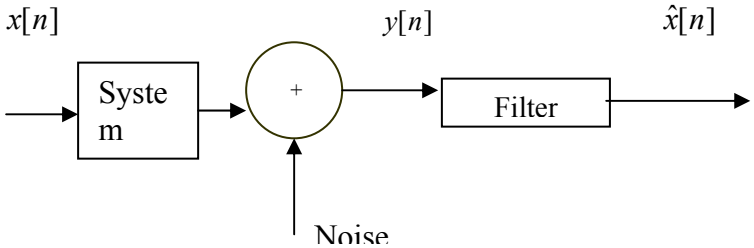
We look for a simpler estimator.

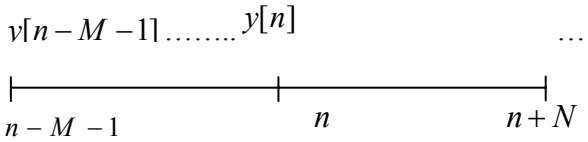
The answer is ***Optimal filtering or Wiener filtering***

We have seen that we can estimate an unknown signal (*desired signal*) $x[n]$ from an observed signal $y[n]$ on the basis of the known joint distributions of $y[n]$ and $x[n]$. We could have used the criteria like MMSE or MAP that we have applied for parameter es timations. But such estimations are generally non-linear, require the computation of *a posteriori* probabilities and involves computational complexities.

The approach taken by Wiener is to specify a form for the estimator that depends on a number of parameters. The minimization of errors then results in determination of an optimal set of estimator parameters. A mathematically sample and computationally easier estimator is obtained by assuming a linear structure for the estimator.

## 5.2 Linear Minimum Mean Square Error Estimator




The linear minimum mean square error criterion is illustrated in the above figure. The problem can be slated as follows:

Given observations of data $y[n-M+1], y[n-M+2]..y[n], ...y[n+N],$ determine a set of parameters $h[M-1], h[M-2]..., h[o], ...h[-N]$ such that

$$\hat{x}[n] = \sum_{i=-N}^{M-1} h[i]y[n-i]$$

and the mean square error $E\left(x[n]-\hat{x}[n]\right)^2$ is a minimum with respect to $h[-N], h[-N+1]...., h[o], h[i], ..., h[M-1].$

This minimization problem results in an elegant solution if we assume joint stationarity of the signals $x[n]$ and $y[n]$. The estimator parameters can be obtained from the second order statistics of the processes $x[n]$ and $y[n]$.

The problem of deterring the estimator parameters by the LMMSE criterion is also called the Wiener filtering problem. Three subclasses of the problem are identified

1. The optimal smoothing problem $N > 0$
2. The optimal filtering problem $N = 0$
3. The optimal prediction problem $N < 0$

In the smoothing problem, an estimate of the signal inside the duration of observation of the signal is made. The filtering problem estimates the curent value of the signal on the basis of the present and past observations. The prediction problem addresses the issues of optimal prediction of the future value of the signal on the basis of present and past observations.

## 5.3 Wiener-Hopf Equations

The mean-square error of estimation is given by

$$\begin{aligned} Ee^2[n] &= E(x[n]-\hat{x}[n])^2 \\ &= E(x[n]-\sum_{i=-N}^{M-1} h[i]y[n-i])^2 \end{aligned}$$

We have to minimize $Ee^2[n]$ with respect to each $h[i]$ to get the optimal estimation.

Corresponding minimization is given by

$$\frac{\partial E\left\{e^2[n]\right\}}{\partial h[j]} = 0, \text{ for } j = -N..0..M-1$$

( $E$ being a linear operator, $E$ and $\frac{\partial}{\partial h[j]}$ can be interchanged)

$$Ee[n]y[n-j] = 0, \quad j = -N...0,1,...M-1 \qquad (1)$$

or

$$E\left(\overbrace{x[n] - \sum_{i=-N_a}^{M-1} h[i]y[n-i]}^{e[n]}\right)y[n-j] = 0, \quad j = -N...0,1,...M-1 \qquad (2)$$

$$R_{XY}(j) = \sum_{i=-N_a}^{M-1} h[i]R_{YY}[j-i], \quad j = -N...0,1,...M-1 \qquad (3)$$

This set of $N+M+1$ equations in (3) are called *Wiener Hopf equations or Normal equations*.

- The result in (1) is the orthogonality principle which implies that the error is orthogonal to observed data.
- $\hat{x}[n]$ is the projection of $x[n]$ onto the subspace spanned by observations $y[n-M], y[n-M+1]..y[n],...y[n+N]$.
- The estimation uses second order-statistics i.e. autocorrelation and cross-correlation functions.

- If $x[n]$ and $y[n]$ are jointly Gaussian then MMSE and LMMSE are equivalent. Otherwise we get a sub-optimum result.
- Also observe that

$$x[n] = \hat{x}[n] + e[n]$$

where $\hat{x}[n]$ and $e[n]$ are the parts of $x[n]$ respectively correlated and uncorrelated with $y[n]$. Thus LMMSE separates out that part of $x[n]$ which is correlated with $y[n]$. Hence the Wiener filter can be also interpreted as the correlation canceller. (See Orfanidis).

## 5.4 FIR Wiener Filter

$$\hat{x}[n] = \sum_{i=0}^{M-1} h[i]y[n-i]$$

The model parameters are given by the orthogonality principlee

$$E\left(\overbrace{x[n] - \sum_{i=0}^{M-1} h[i]y[n-i]}^{e[n]}\right) y[n-j] = 0, \quad j = 0,1,...M-1$$

$$\sum_{i=0}^{M-1} h[i]R_{YY}[j-i] = R_{XY}(j), \quad j = 0,1,...M-1$$

In matrix form, we have

$$\mathbf{R_{YY}h} = \mathbf{r_{XY}}$$

where

$$\mathbf{R_{YY}} = \begin{bmatrix} R_{YY}[0] & R_{YY}[-1] & .... & R_{YY}[1-M] \\ R_{YY}[1] & R_{YY}[0] & .... & R_{YY}[2-M] \\ ... & & & \\ R_{YY}[M-1] & R_{YY}[N-2] & .... & R_{YY}[0] \end{bmatrix}$$

and

$$\mathbf{r_{XY}} = \begin{bmatrix} R_{XY}[0] \\ R_{XY}[1] \\ ... \\ R_{XY}[M-1] \end{bmatrix}$$

and

$$\mathbf{h} = \begin{bmatrix} h[0] \\ h[1] \\ ... \\ h[M-1] \end{bmatrix}$$

Therefore,

$$\mathbf{h} = \mathbf{R}_{\mathbf{YY}}^{-1}\mathbf{r}_{\mathbf{XY}}$$

## 5.5 Minimum Mean Square Error - FIR Wiener Filter

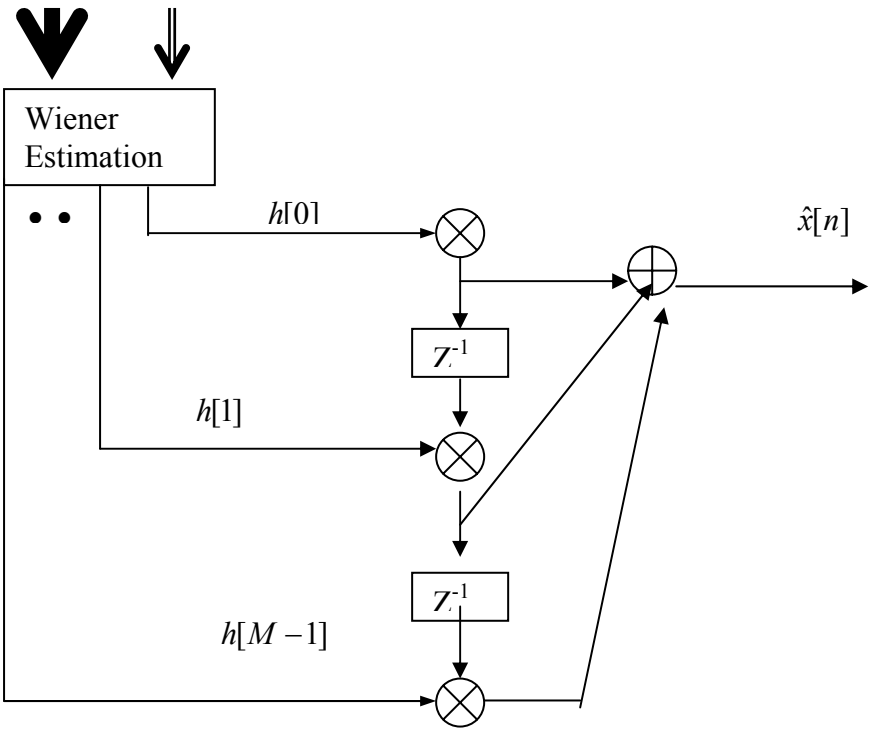
$$E(e^2[n] = Ee[n]\left( x[n] \; - \; \sum_{i=0}^{M-1} h[i]y[n-i] \right)$$

$$= Ee[n] \; x[n] \qquad \because \text{error isorthogonal to data}$$

$$= E\left( x[n] \; - \; \sum_{i=0}^{M-1} h[i]y[n-i] \right) x[n]$$

$$= R_{XX}[0] - \sum_{i=0}^{M-1} h[i]R_{XY}[i]$$

$\mathbf{R}_{XX}$ $\mathbf{r}_{\mathbf{XY}}$



**Example1: Noise Filtering**

Consider the case of a carrier signal in presence of white Gaussian noise

$$x[n] = A\,cos[w_0 n + \phi], \quad w_0 = \frac{\pi}{4}$$

$$y[n] = x[n] + v[n]$$

here $\phi$ is uniformly distributed in $(1, 2\pi)$.

$v[n]$ is white Gaussian noise sequence of variance 1 and is independent of $x[n]$. Find the parameters for the FIR Wiener filter with M=3.

$$R_{XX}[m] = \frac{A^2}{2}\cos\ w_0 m$$

$$\begin{aligned} R_{YY}[m] &= E\ y[n]y[n-m] \\ &= E\big(x[n]+v[m]\big)\big(x[n-m]+v[n-m]\big) \\ &= R_{XX}[m] + R_{VV}[m] + 0 + 0 \\ &= \frac{A^2}{2}cos(w_0 m) + \delta[m] \end{aligned}$$

$$\begin{aligned} R_{XY}[m] &= E\ x[n]y[n-m] \\ &= E\ x[n]\big(x[n-m]+v[n-m]\big) \\ &= R_{XX}[m] \end{aligned}$$

Hence the Wiener Hopf equations are

$$\begin{bmatrix} R_{YY}[0] & R_{YY}[1] & Ryy[2] \\ R_{YY}[1] & R_{YY}[0] & R_{YY}[1] \\ R_{YY}[2] & R_{YY}1] & R_{YY}[0] \end{bmatrix} \begin{bmatrix} h[0] \\ h[1] \\ h[2] \end{bmatrix} = \begin{bmatrix} R_{XX}[0] \\ R_{XX}[1] \\ R_{XX}[2] \end{bmatrix}$$

$$\begin{bmatrix} \frac{A^2}{2}+1 & \frac{A^2}{2}\cos\frac{\pi}{4} & \frac{A^2}{2}\cos\frac{\pi}{2} \\ \frac{A^2}{2}\cos\frac{\pi}{4} & \frac{A^2}{2}+1 & \frac{A^2}{2}\cos\frac{\pi}{4} \\ \frac{A^2}{2}\cos\frac{\pi}{2} & \frac{A^2}{2}\cos\frac{\pi}{4} & \frac{A^2}{2}+1 \end{bmatrix} \begin{bmatrix} h[0] \\ h[1] \\ h[2] \end{bmatrix} = \begin{bmatrix} \frac{A^2}{2} \\ \frac{A^2}{2}\cos\frac{\pi}{4} \\ \frac{A^2}{2}\cos\frac{\pi}{2} \end{bmatrix}$$

suppose A = 5v then

$$\begin{bmatrix} 13.5 & \frac{12.5}{\sqrt{2}} & 0 \\ \frac{12.5}{\sqrt{2}} & 13.5 & \frac{12.5}{\sqrt{2}} \\ 0 & \frac{12.5}{\sqrt{2}} & 13.5 \end{bmatrix} \begin{bmatrix} h[0] \\ h[1] \\ h[2] \end{bmatrix} = \begin{bmatrix} 12.5 \\ \frac{12.5}{\sqrt{2}} \\ 0 \end{bmatrix}$$

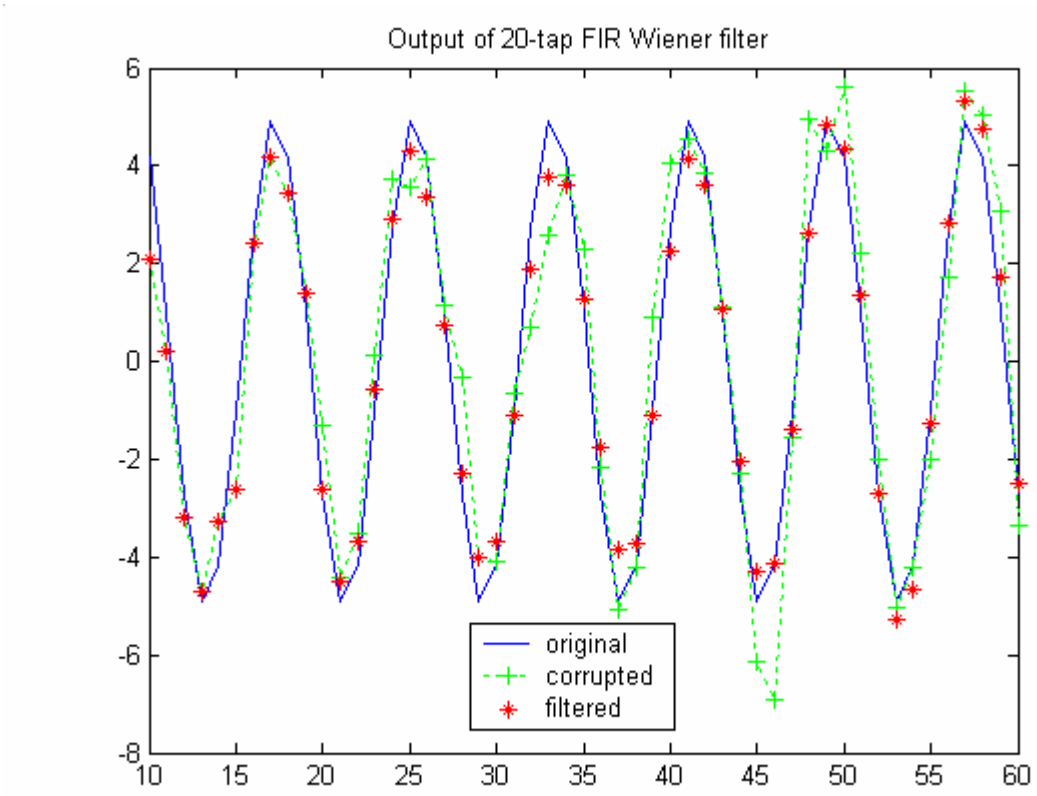
$$\begin{bmatrix} h[0] \\ h[1] \\ h[2] \end{bmatrix} = \begin{bmatrix} 13.5 & \frac{12.5}{\sqrt{2}} & 0 \\ \frac{12.5}{\sqrt{2}} & 13.5 & \frac{12.5}{\sqrt{2}} \\ 0 & \frac{12.5}{\sqrt{2}} & 13.5 \end{bmatrix}^{-1} \begin{bmatrix} 12.5 \\ \frac{12.5}{\sqrt{2}} \\ 0 \end{bmatrix}$$

$h[0] = 0.707$

$h[1 = 0.34$

$h[2] = -0.226$

Plot the filter performance for the above values of $h[0]$, $h[1]$ and $h[2]$. The following figure shows the performance of the 20-tap FIR wiener filter for noise filtering.

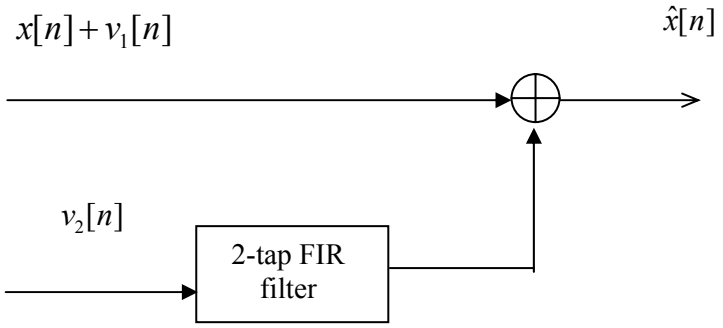


**Example 2** : **Active Noise Control**

Suppose we have the observation signal $y[n]$ is given by

$$y[n] = 0.5\cos(w_0 n + \phi) + v_1[n]$$

where $\phi$ is uniformly distrubuted in $(0, 2\pi)$ and $v_1[n] = 0.6v[n-1] + v[n]$ is an MA(1) noise. We want to control $v_1[n]$ with the help of another correlated noise $v_2[n]$ given by $v_2[n] = 0.8v[n-1] + v[n]$




The Wiener Hopf Equations are given by

$\mathbf{R}_{V_2V_2}\mathbf{h} = \mathbf{r}_{V_1V_2}$

where $\mathbf{h} = [h[0] \;\; h[1]]'$

and

$\mathbf{R}_{V_2V_2} = \begin{bmatrix} 1.64 & 0.8 \\ 0.8 & 1.64 \end{bmatrix}$ and

$\mathbf{r}_{V_1V_2} = \begin{bmatrix} 1.48 \\ 0.6 \end{bmatrix}$

$\therefore \begin{bmatrix} h[0] \\ h[1] \end{bmatrix} = \begin{bmatrix} 0.9500 \\ -0.0976 \end{bmatrix}$

**Example 3:**

(Continuous time prediction) Suppose we want to predict the continuous-time process $X(t)$ at time $(t+\tau)$ by

$\hat{X}(t+\tau) = aX(t)$

Then by orthogonality principle

$E(X(t+\tau) - aX(t))X(t) = 0$

$\Rightarrow a = \dfrac{R_{XX}(\tau)}{R_{XX}(0)}$

As a particular case consider the first-order Markov process given by

$$\frac{d}{dt}X(t) = AX(t) + v(t)$$

In this case,

$$R_{XX}(\tau) = R_{XX}(0)e^{-A\tau}$$

$$\therefore a = \frac{R_{XX}(\tau)}{R_{XX}(0)} = e^{-A\tau}$$

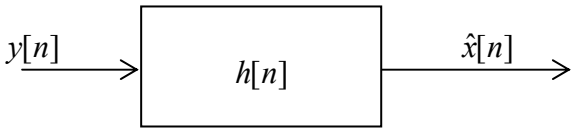
Observe that for such a process

$$E(X(t+\tau) - aX(t))X(t-\tau_1) = 0$$
$$= R_{XX}(\tau+\tau_1) - aR_{XX}(\tau_1)$$
$$= R_{XX}(0)e^{-A(\tau+\tau_1)} - e^{-A(\tau)}R_{XX}(0)e^{-A(\tau_1)}$$
$$= 0$$

Therefore, the linear prediction of such a process based on any past value is same as the linear prediction based on current value.

## 5.6 IIR Wiener Filter (Causal)

Consider the IIR filter to estimate the signal $x[n]$ shown in the figure below.




The estimator $\hat{x}[n]$ is given by

$$\hat{x}(n) = \sum_{i=0}^{\infty} h(i)y(n-i)$$

The mean-square error of estimation is given by

$$Ee^2[n] = E(x[n] - \hat{x}[n])^2$$
$$= E(x[n] - \sum_{i=0}^{\infty} h[i]y[n-i])^2$$

We have to minimize $Ee^2[n]$ with respect to each $h[i]$ to get the optimal estimation.

Applying the orthogonality principle, we get the WH equation.

$$E(x[n] - \sum_{i=0}^{\infty} h(i)y[n-i])y[n-j] = 0,\ j = 0,\ 1,\ .....$$

From which we get

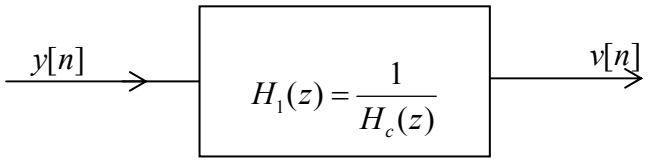
$$\sum_{i=0}^{\infty} h[i]\ R_{YY}[j-i]\ = R_{XY}[j],\ j = 0,\ 1,\ .....$$

- We have to find $h[i], \quad i = 0,1,...\infty$ by solving the above infinite set of equations.
- This problem is better solved in the z-transform domain, though we cannot directly apply the convolution theorem of z-transform.

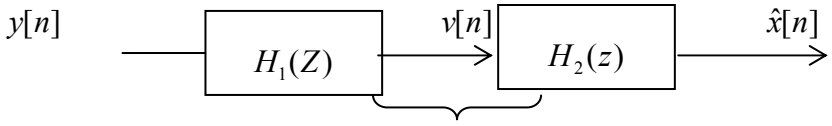
Here comes Wiener's contribution.

The analysis is based on the spectral Factorization theorem:

$$S_{YY}(z) = \sigma_v^2 H_c(z) H_c(z^{-1})$$




Whitening filter




Wiener filter

Now $h_2[n]$ is the coefficient of the Wiener filter to estimate $x[n]$I from the innovation sequence $v[n]$. Applying the orthogonality principle results in the Wiener Hopf equation

$$\hat{x}(n) = \sum_{i=0}^{\infty} h_2(i) v(n-i)$$

$$E\left\{ x[n] - \sum_{i=0}^{\infty} h_2[i] v[n-i] \right\} v[n-j]) = 0$$

$$\therefore \sum_{i=0}^{\infty} h_2[i] R_{VV}[j-i] = R_{XV}[j], \; j = 0,1,...$$

$$R_{VV}[m] = \sigma_V^2 \delta[m]$$

$$\therefore \sum_{i=0}^{\infty} h_2(i) \sigma_V^2 \delta[j-i] = R_{XV}(j), \quad j = 0,1,...$$

So that

$$h_2[j] = \frac{R_{XV}[j]}{\sigma_V^{\,2}} \quad \text{j} \geq 0$$

$$H_2(z) = \frac{[S_{XV}(z)]+}{\sigma_V^{\,2}}$$

where $[S_{XV}(z)]+$ is the positive part (i.e., containing non-positive powers of $z$ ) in power series expansion of $S_{XV}(z)$.

$$v[n] = \sum_{i=0}^{\infty} h_1[i]y[n-i]$$

$$R_{XV}[j] = Ex[n]v[n-j]$$

$$= \sum_{i=0}^{\infty} h_i[i]E\ x[n]\ y[n - j - i]$$

$$= \sum_{i=0}^{\infty} h_1[i]R_{XY}[\mathrm{j}+\mathrm{i}]$$

$$S_{XV}(z) = H_1(z^{-1})S_{XY}(z) = \frac{1}{H_c(z^{-1})}S_{XY}(z)$$

$$\therefore H_2(z) = \frac{1}{\sigma_V^2}\left[\frac{S_{XY}(z)}{H_c(z^{-1})}\right]_+$$

Therefore,

$$H(z) = H_1(z)H_2(z) = \frac{1}{\sigma_V^2 H_c(z)}\left[\frac{S_{XY}(z)}{H_c(z^{-1})}\right]_+$$

We have to

- find the power spectrum of data and the cross power spectrum of the of the desired signal and data from the available model or estimate them from the data
- factorize the power spectrum of the data using the spectral factorization theorem

## 5.7 Mean Square Estimation Error – IIR Filter (Causal)

$$E(e^2[n] = Ee[n]\left( x[n] \ - \ \sum_{i=0}^{\infty} h[i]y[n-i]\right)$$

$$= Ee[n]\ x[n] \qquad \because \text{error isorthogonal to data}$$

$$= E\left( x[n] \ - \ \sum_{i=0}^{\infty} h[i]y[n-i]\right)x[n]$$

$$= R_{XX}[0] - \sum_{i=0}^{\infty} h[i]R_{XY}[i]$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_X(w)dw - \frac{1}{2\pi}\int_{-\pi}^{\pi} H(w)S_{XY}^*(w)dw$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} (S_X(w) - H(w)S_{XY}^*(w))dw$$

$$= \frac{1}{2\pi}\oint_C (S_X(z) - H(z)S_{XY}(z^{-1}))z^{-1}dz$$

**Example 4:**

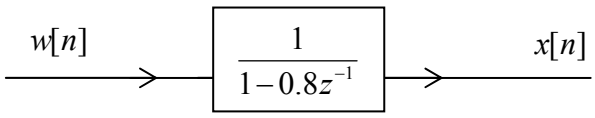
$y[n] = x[n] + v_1[n]$ observation model with

$x[n] = 0.8\,x[n-1] + w[n]$

where $v_1[n]$ is and additive zero-mean Gaussian white noise with variance 1 and $w[n]$ is zero-mean white noise with variance 0.68. Signal and noise are uncorrelated.

Find the optimal Causal Wiener filter to estimate $x[n]$.

**Solution:**




$$S_{XX}(z) = \frac{0.68}{\left(1-0.8z^{-1}\right)\left(1-0.8z\right)}$$

$$\begin{aligned} R_{YY}[m] &= E\; y[n]y[n-m] \\ &= E\left(x[n]+v[m]\right)\left(x[n-m]+v[n-m]\right) \\ &= R_{XX}[m] + R_{VV}[m] + 0 + 0 \end{aligned}$$

$$S_{YY}(z) = S_{XX}(z) + 1$$

Factorize

$$S_{YY}(z) = \frac{0.68}{\left(1-0.8z^{-1}\right)\left(1-0.8z\right)} + 1$$

$$= \frac{2(1-0.4z^{-1})(1-0.4z)}{\left(1-0.8z^{-1}\right)\left(1-0.8z\right)}$$

$$\therefore H_c(z) = \frac{(1-0.4z^{-1})}{\left(1-0.8z^{-1}\right)}$$

*and*

$$\sigma_V^2 = 2$$

Also

$$\begin{aligned} R_{XY}[m] &= E\; \mathrm{x}[n]y[n-m] \\ &= E\left(x[n]\right)\left(x[n-m]+v[n-m]\right) \\ &= R_{XX}[m] \end{aligned}$$

$$S_{XY}(z) = S_{XX}(z)$$

$$= \frac{0.68}{\left(1-0.8z^{-1}\right)\left(1-0.8z\right)}$$

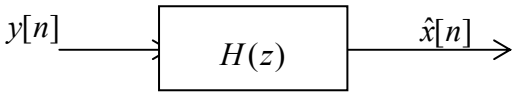
$$\therefore H(z) = \frac{1}{\sigma_V^2 H_c(z)} \left[ \frac{S_{XY}(z)}{H_c(z^{-1})} \right]_+$$

$$= \frac{1}{2} \frac{(1-0.8z^{-1})}{\left(1-0.4z^{-1}\right)} \left[ \frac{0.68}{(1-0.8z^{-1})(1-0.8z)} \right]_+$$

$$= \frac{0.944}{\left(1-0.4z^{-1}\right)}$$

$$h[n] = 0.944(0.4)^n \quad n \geq 0$$

## 5.8 IIR Wiener filter (Noncausal)




The estimator $\hat{x}[n]$ is given by

$$\hat{x}[n] = \sum_{i=-\alpha}^{\alpha} h[i] y[n-i]$$

For LMMSE, the error is orthogonal to data.

$$E\left( x\,[n] \; - \sum_{i=-\alpha}^{\alpha} h\,[i] \; y\,[n-i] \right) y[n-j] = 0 \quad \forall \mathrm{j} \in \mathrm{I}$$

$$\sum_{i=-\infty}^{\infty} h[i] \; R_{YY}[j-i] \; = R_{XY}[j], \; j = -\infty, ...0, \; 1, \; ...\infty$$

- This form Wiener Hopf Equation is simple to analyse.
- Easily solved in frequency domain. So taking Z transform we get
- Not realizable in real time

$$H(z) S_{YY}(z) = S_{XY}(z)$$

so that

$$H(z) = \frac{S_{XY}(z)}{S_{YY}(z)}$$

or

$$H(w) = \frac{S_{XY}(w)}{S_{YY}(w)}$$

## 5.9 Mean Square Estimation Error – IIR Filter (Noncausal)

The mean square error of estimation is given by

$$E(e^2[n] = Ee[n]\left( x[n] \; - \; \sum_{i=-\infty}^{\infty} h[i]y[n-i] \right)$$

$$= Ee[n] \; x[n] \qquad \because \text{error isorthogonal to data}$$

$$= E\left( x[n] \; - \; \sum_{i=-\infty}^{\infty} h[i]y[n-i] \right) x[n]$$

$$= R_{XX}[0] - \sum_{i=-\infty}^{\infty} h[i]R_{XY}[i]$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_X(w)dw - \frac{1}{2\pi}\int_{-\pi}^{\pi} H(w)S_{XY}^*(w)dw$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} (S_X(w) - H(w)S_{XY}^*(w))dw$$

$$= \frac{1}{2\pi}\oint_C (S_X(z) - H(z)S_{XY}(z^{-1}))z^{-1}dz$$

**<u>Example 5:</u> Noise filtering by noncausal IIR Wiener Filter**

Consider the case of a carrier signal in presence of white Gaussian noise

$$y[n] = x[n] + v[n]$$

where $v[n]$ is and additive zero-mean Gaussian white noise with variance $\sigma_V^2$. Signal and noise are uncorrelated

$$S_{YY}(w) = S_{XX}(w) + S_{VV}(w)$$

and

$$S_{XY}(w) = S_{XX}(w)$$

$$\therefore H(w) = \frac{S_{XX}(w)}{S_{XX}(w) + S_{VV}(w)}$$

$$= \frac{\frac{S_{XY}(w)}{S_{VV}(w)}}{\frac{S_{XX}(w}{S_{VV}(w)}) + 1}$$

Suppose SNR is very high

$$H(w) \cong 1$$

(i.e. the signal will be passed un-attenuated).

When SNR is low

$$H(w) = \frac{S_{XX}(w)}{S_{VV}(w)}$$

(i.e. If noise is high the corresponding signal component will be attenuated in proportion of the estimated SNR.

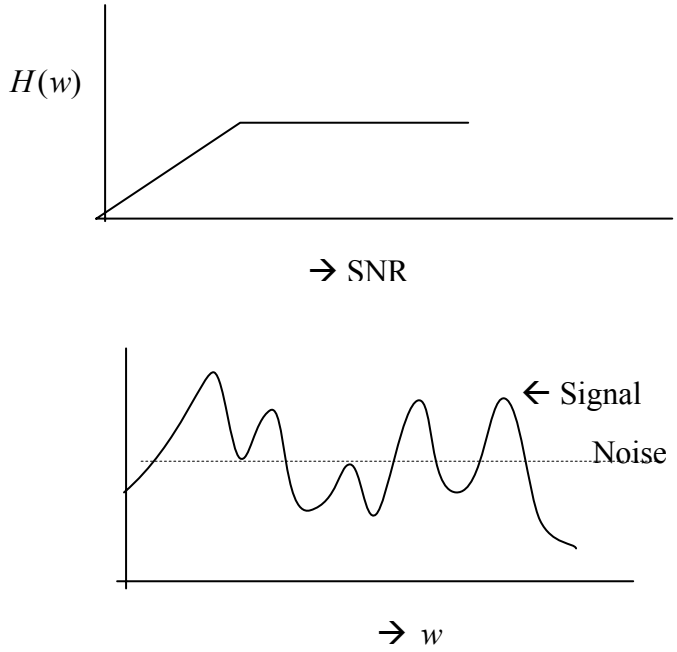



**Figure** - (a) a high-SNR signal is passed unattended by the IIR Wiener filter
(b) Variation of SNR with frequency

**Example 6: Image filtering by IIR Wiener filter**

$S_{YY}(w)$ = power spectrum of the corrupted image
$S_{VV}(w)$ = power spectrum of the noise, estimated from the noise model
or from the constant intensity ( like back-ground) of the image

$$H(w) = \frac{S_{XX}(w)}{S_{XX}(w) + S_{VV}(w)}$$
$$= \frac{S_{YY}(w) - S_{VV}(w)}{S_{YY}(w)}$$

**Example 7:**

Consider the signal in presence of white noise given by

$x[n] = 0.8\, x\,[n-1] + w[n]$

where $v[n]$ is and additive zero-mean Gaussian white noise with variance 1 and $w[n]$ is zero-mean white noise with variance 0.68. Signal and noise are uncorrelated.

Find the optimal noncausal Wiener filter to estimate $x[n]$.

$$H(z) = \frac{S_{XY}(z)}{S_{YY}(z)} = \frac{\dfrac{0.68}{(1-0.8z^{-1})(1-0.8z)}}{\dfrac{2(1-0.4z^{-1})(1-0.4z)}{(1-0.6z^{-1})(1-0.6z)}}$$

$$= \frac{0.34}{(1-0.4z^{-1})(1-0.4z)}$$ One pole outside the unit circle

$$= \frac{0.4048}{1-0.4z^{-1}} + \frac{0.4048}{1-0.4z}$$

$$\therefore h[n] = 0.4048(0.4)^{n}u(n) + 0.4048(0.4)^{-n}u(-n-1)$$

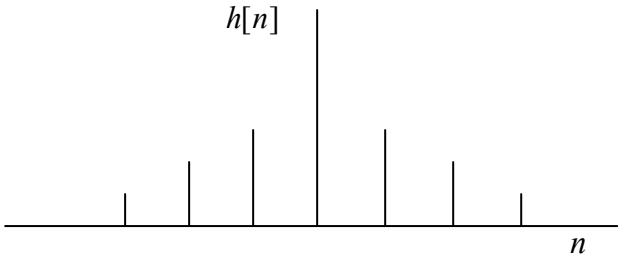



**Figure -** Filter Impulse Response

# CHAPTER – 6: LINEAR PREDICTION OF SIGNAL

## 6.1 Introduction

Given a sequence of observation

$y[n-1], y[n-2], \ldots . y[n-M]$, what is the best prediction for $y[n]$?

(one-step ahead prediction)

The minimum mean square error prediction $\hat{y}[n]$ for $y[n]$ is given by $\hat{y}[n] = E\{y[n] | y[n-1], y[n-2], \ldots, y[n-M]\}$ which is a nonlinear predictor.

A linear prediction is given by

$$\hat{y}[n] = \sum_{i=1}^{M} h[i] y[n-i]$$

where $h[i], i = 1 \ldots . M$ are the prediction parameters.

- Linear prediction has very wide range of applications.
- For an exact AR (M) process, linear prediction model of order M and the corresponding AR model have the same parameters. For other signals LP model gives an approximation.

## 6.2 Areas of application

- Speech modeling
- Low-bit rate speech coding
- Speech recognition
- ECG modeling
- DPCM coding
- Internet traffic prediction

LPC (10) is the popular linear prediction model used for speech coding. For a frame of speech samples, the prediction parameters are estimated and coded. In CELP (Code book Excited Linear Prediction) the prediction error $e[n] = y[n] - \hat{y}[n]$ is vector quantized and transmitted.

$\hat{y}[n] = \sum_{i=1}^{M} h[i] y[n-i]$ is a FIR Wiener filter shown in the following figure. It is called the *linear prediction filter***.**

Therefore

$$e[n] = y[n] - \hat{y}[n]$$
$$= y[n] - \sum_{i=1}^{M} h[i]\, y[n-i]$$

is the prediction error and the corresponding filter is called prediction error filter.

Linear Minimum Mean Square error estimates for the prediction parameters are given by the orthogonality relation

$$E\, e[n]\, y[n-j] = 0 \quad for\ j = 1, 2, \ldots, M$$

$$\therefore E(y[n] - \sum_{i=1}^{M} h[i] y[n-i]) y[n-j] = 0 \qquad j = 1, 2, \ldots, M$$

$$\Rightarrow R_{YY}[j] - \sum_{i=1}^{M} h[i]\, R_{YY}[j-i] = 0$$

$$\Rightarrow R_{YY}[j] = \sum_{i=1}^{M} h[i]\, R_{YY}[j-i] \qquad j = 1, 2, \ldots, M$$

which is the Wiener Hopf equation for the linear prediction problem and same as the Yule Walker equation for AR (M) Process.

In Matrix notation

$$\begin{bmatrix} R_{YY}[0] & R_{YY}[1] & \ldots & R_{YY}[M-1] \\ R_{YY}[1] & R_{YY}[o] & \ldots & R_{YY}[M-2] \\ . & & & \\ . & & & \\ . & & & \\ R_{YY}[M-1] & R_{YY}[M-2] & \ldots & R_{YY}[0] \end{bmatrix} \begin{bmatrix} h[1] \\ h[2] \\ . \\ . \\ . \\ h[M] \end{bmatrix} = \begin{bmatrix} R_{YY}[1] \\ R_{YY}[2] \\ . \\ . \\ . \\ R_{YY}[M] \end{bmatrix}$$

$$\mathbf{R_{YY}} \mathbf{h} = \mathbf{r_{YY}}$$

$$\therefore \mathbf{h} = (\mathbf{R_{YY}})^{-1} \mathbf{r_{YY}}$$

## 6.3 Mean Square Prediction Error (MSPE)

$$E(e^2[n]) = E(y[n] - \sum_{i=1}^{M} h[i] y[n-i]) e[n]$$
$$= Ey[n] e[n]$$
$$= Ey[n](y[n] - \sum_{i=1}^{M} h[i] y[n-i])$$
$$= R_{YY}[0] - \sum_{i=1}^{M} h[i] R_{YY}[i])$$

## 6.4 Forward Prediction Problem

The above linear prediction problem is the forward prediction problem. For notational simplicity let us rewrite the prediction equation as

$$\hat{y}\,[n] \;=\; \sum_{i=1}^{M} h_M[i]\, y[n-i]$$

where the prediction parameters are being denoted by $h_M[i], i = 1...M$.

## 6.5 Backward Prediction Problem

Given $y[n], y[n-1], \ldots. y[n-M+1],$ we want to estimate $y[n-M]$.

The linear prediction is given by

$$\hat{y}[n-M] \;=\; \sum_{i=1}^{M} b_M\,[i]\; y[n+1-i]]$$

Applying orthogonality principle.

$$E(y[n-M]-\sum_{i=1}^{M} b_M\,[i] y[n+1-i])\; y\,[n+1-j]\;=0 \quad j=1,2...,M.$$

This will give

$$R_{YY}\;[M+1-j] = \;\sum_{i=1}^{M} b_M[i]\, R_{YY}[j-i] \quad j = 1,\ 2,\ ...,\ M$$

Corresponding matrix form

$$\begin{bmatrix} R_{YY}[0] & R_{YY}[1] & .... & R_{YY}\;[M-1] \\ R_{YY}[1] & R_{YY}[o] & ... & R_{YY}[M-2] \\ . & & & \\ . & & & \\ . & & & \\ R_{YY}[M-1] & R_{YY}[M-2] & ..... & R_{YY}[0] \end{bmatrix} \begin{bmatrix} b_M[1] \\ b_M[2] \\ . \\ . \\ . \\ b_M[M] \end{bmatrix} = \begin{bmatrix} R_{YY}\;[M] \\ R_{YY}[M-1] \\ . \\ . \\ . \\ R_{YY}\;[1] \end{bmatrix} \qquad (1)$$

## 6.6 Forward Prediction

Rewriting the Mth-order forward prediction problem, we have

$$\begin{bmatrix} R_{YY}[0] & R_{YY}[1] & .... & R_{YY}\;[M-1] \\ R_{YY}[1] & R_{YY}[o] & ... & R_{YY}[M-2] \\ . & & & \\ . & & & \\ . & & & \\ R_{YY}[M-1] & R_{YY}[M-2] & ..... & R_{YY}[0] \end{bmatrix} \begin{bmatrix} h_M[1] \\ h_M[2] \\ . \\ . \\ . \\ h_M[M] \end{bmatrix} = \begin{bmatrix} R_{YY}\;[1] \\ R_{YY}[2] \\ . \\ . \\ . \\ R_{YY}\;[M] \end{bmatrix} \qquad (2)$$

From (1) and (2) we conclude

$b_M[i] = h_M[M+1-i], i = 1,2...,M$

Thus forward prediction parameters in reverse order will give the backward prediction parameters.

M S prediction error

$$\varepsilon_M = E\left( y[n-M] \; - \sum_{i=1}^{M} b_M\,[i]\; y\,[n+1-i] \right) y\,[n-M]$$

$$= R_{YY}[0] - \sum_{i=1}^{M} b_M\,[i]\; R_{YY}\,[M+1-i]$$

$$= R_{YY}[0] - \sum_{i=1}^{M} h_M\,[M+1-i]\; R_{YY}\,[M+1-i]$$

which is same as the forward prediction error.

Thus

| Backward prediction error = Forward Prediction error. |
|---|

**Example 1:**

Find the second order predictor for $y[n]$ given $y[n] = x[n] + v[n]$, where $v[n]$ is a 0-mean white noise with variance 1 and uncorrelated with $x[n]$ and $x[n] = 0.8x[n-1] + w[n]$, $w[n]$ is a 0-mean random variable with variance 0.68

The linear predictor is given by

$$\hat{y}\,[n] \;=\; h_2\,[1]\; y[n-1] + h_2\,[2]\; y[n-2]$$

We have to find $h_2\,[1]$ and $h_2\,[2]$.

Corresponding Yule Walker equations are

$$\begin{bmatrix} R_{YY}[0] & R_{YY}[1] \\ R_{YY}[1] & R_{YY}[0] \end{bmatrix} \begin{bmatrix} h_2[1] \\ h_2[2] \end{bmatrix} = \begin{bmatrix} R_{YY}\;\mathbf{[1]} \\ R_{YY}\;\mathbf{[2]} \end{bmatrix}$$

To find out $R_{YY}[0]$ , $R_{YY}[1]$ and $R_{YY}[2]$

$y[n] = x[n] + v[n],$

$R_{YY}[m] = R_{XX}[m] + \delta[m]$

$x[n] = 0.8x[n-1] + w[n]$

$$\therefore R_{XX}[m] = \frac{0.68}{1-(0.8)^2}(0.8)^{|m|} = 1.89 \times (0.8)^{|m|}$$

$R_{YY}[0] = 2.89,\; R_{YY}[1] = 1.51$ and $R_{YY}[2] = 1.21$

Solving $h_2[1] = 0.4178$ and $h_2[2] =$ 0.2004.

## 6.7 Levinson Durbin Algorithm

Levinson Durbin algorithm is the most popular technique for determining the LPC parameters from a given autocorrelation sequence.

Consider the Yule Walker equation for $mth$ order linear predictor.

$$\begin{bmatrix} R_{YY}[0] & R_{YY}[1] & .... & R_{YY}[m-1] \\ R_{YY}[1] & R_{YY}[o] & .... & R_{yy}[m\text{-}2] \\ . & & & \\ . & & & \\ . & & & \\ R_{YY}[m\text{-}1] & ..... & & R_{YY}[0] \end{bmatrix} \begin{bmatrix} h_{\mathrm{m}}[1] \\ h_{\mathrm{m}}[2] \\ . \\ . \\ . \\ h_{\mathrm{m}}[m] \end{bmatrix} = \begin{bmatrix} R_{YY}\ [1] \\ . \\ . \\ . \\ . \\ R_{YY}\ [\mathrm{m}] \end{bmatrix} \qquad (1)$$

Writing in the reverse order

$$\begin{bmatrix} R_{YY}[0] & R_{YY}[1] & .... & R_{YY}[m-1] \\ R_{YY}[1] & R_{YY}[o] & .... & R_{yy}[m\text{-}2] \\ . & & & \\ . & & & \\ . & & & \\ R_{YY}[m\text{-}1] & ..... & & R_{YY}[0] \end{bmatrix} \begin{bmatrix} h_{\mathrm{m}}[m] \\ h_{\mathrm{m}}[m-1] \\ . \\ . \\ . \\ h_{\mathrm{m}}[1] \end{bmatrix} = \begin{bmatrix} R_{YY}\ [\mathrm{m}] \\ . \\ . \\ . \\ . \\ R_{YY}\ [1] \end{bmatrix} \qquad (2)$$

Then (m+1) the order predictor is given by

$$\left[\begin{array}{cccc|c} R_{YY}[0] & R_{YY}[1] & .... & R_{YY}[m-1] & R_{YY}[m] \\ R_{YY}[1] & R_{YY}[0] & .... & R_{YY}[m-2] & R_{YY}[m-1] \\ . & & & & \\ . & & & & \\ . & & & & \\ R_{YY}[m-1] & R_{YY}[m-2] & .... & R_{YY}[0] & R_{YY}[1] \\ \hline R_{YY}[m] & R_{YY}[m-1] & .... & R_{YY}[1] & R_{YY}[0] \end{array}\right] \left[\begin{array}{c} h_{\mathrm{m+1}}[1] \\ h_{\mathrm{m+1}}[2] \\ . \\ . \\ . \\ h_{m+1}[m] \\ \hline h_{m+1}[m+1] \end{array}\right] = \left[\begin{array}{c} R_{YY}[1] \\ R_{YY}[2] \\ . \\ . \\ . \\ R_{\mathrm{YY}}[\mathrm{m}] \\ \hline R_{\mathrm{YY}}[\mathrm{m}+1] \end{array}\right]$$

Let us partition equation (2) as shown. Then

$$\begin{bmatrix} R_{YY}[0] & R_{YY}[1] & .... & R_{YY}[m-1] \\ R_{YY}[1] & R_{YY}[0] & .... & R_{YY}[m-2] \\ . & & & \\ . & & & \\ R_{YY}[m-1] & R_{YY}[m-2] & .... & R_{YY}[0] \end{bmatrix} \begin{bmatrix} h_{\mathrm{m+1}}[1] \\ h_{\mathrm{m+1}}[2] \\ . \\ . \\ . \\ h_{m+1}[m] \end{bmatrix} + \mathrm{h}_{\mathrm{m+1}}[\mathrm{m}+1] \begin{bmatrix} R_{YY}[\mathrm{m}] \\ R_{YY}[\mathrm{m}\text{ - }1] \\ . \\ . \\ . \\ R_{YY}[1] \end{bmatrix} = \begin{bmatrix} R_{YY}[1] \\ R_{YY}[2] \\ . \\ . \\ . \\ R_{YY}[m] \end{bmatrix} \qquad (3)$$

and

$$\sum_{i=1}^{m} h_{m+1}[i]\, R_{YY}[m+1-i] + h_{m+1}[m+1] R_{YY}[0] = R_{YY}[m+1] \qquad (4)$$

From equation (3) premultiplying by $\mathbf{R}_{\mathbf{YY}}^{-1}$, we get

$$\begin{bmatrix} h_{m+1}[1] \\ h_{m+1}[2] \\ . \\ . \\ . \\ h_{m+1}[m] \end{bmatrix} + \mathrm{h}_{m+1}[\mathrm{m}+1]\mathbf{R}_{\mathbf{YY}}^{-1} \begin{bmatrix} R_{YY}[\mathrm{m}] \\ R_{YY}[\mathrm{m}-1] \\ . \\ . \\ . \\ R_{YY}[1] \end{bmatrix} = \mathbf{R}_{\mathbf{YY}}^{-1} \begin{bmatrix} R_{YY}[1] \\ R_{YY}[2] \\ . \\ . \\ . \\ R_{YY}[m] \end{bmatrix}$$

$$\begin{bmatrix} h_{m+1}[1] \\ h_{m+1}[2] \\ . \\ . \\ . \\ h_{m+1}[m] \end{bmatrix} + h_{m+1}[\mathrm{m}+1] \begin{bmatrix} h_{\mathrm{m}}[m] \\ h_{\mathrm{m}}[m-1] \\ . \\ . \\ . \\ h_{\mathrm{m}}[1] \end{bmatrix}^{=} = \begin{bmatrix} h_m[1]) \\ h_{\mathrm{m}}[2] \\ . \\ . \\ . \\ h_{\mathrm{m}}[m] \end{bmatrix}$$

The equations can be rewritten as

$$h_{m+1}[i] = h_m[i] + k_{m+1}h_m[m+1-i] \quad i = 1,2,...m \tag{5}$$

where $k_m = -h_m[m]$ is called the reflection coefficient or the PARCOR (partial correlation) coefficient.

From equation (4) we get

$$\sum_{i=1}^{m} h_{m+1}[i]\, R_{YY}[m+1-i] + h_{m+1}[m]R_{YY}[0] = R_{YY}[m+1]$$

using equation (5)

$$\sum_{i=1}^{m} \{h_m[i] + k_{m+1}h_m[m+1-i]\}R_{YY}[m+1-i] - k_{m+1}R_{YY}[0] = R_{YY}[m+1]$$

$$\sum_{i=1}^{m} h_m[i]R_{YY}[m+1-i] - k_{m+1}R_{YY}[0] + k_{m+1}\sum_{i=1}^{m} h_m[m+1-i]\, R_{YY}[m+1-i] = R_{YY}[m+1]$$

$$k_{m+1}\{R_{YY}[0] - \sum_{i=1}^{m} h_m[m+1-i]\, R_{YY}[m+1-i]\} = -R_{YY}[m+1] + \sum_{i=1}^{m} h_m[i]R_{YY}[m+1-i]$$

$$k_{m+1} = \frac{-R_{YY}[m+1] + \sum_{i=1}^{m} h_m[i]R_{YY}[m+1-i]}{\varepsilon[m]}$$

$$= \frac{\sum_{i=0}^{m} h_m[i]R_{YY}[m+1-i]}{\varepsilon[m]}$$

where

$\varepsilon[m] = R_{YY}[0] - \sum_{i=1}^{m} h_m[m+1-i]\, R_{YY}[m+1-i]$ is the mean - square prediction error.

Here we have used the assumption that $h_m[0] = -1$

$$\therefore \varepsilon[m+1] = R_{YY}[0] - \sum_{i=1}^{m+1} h_{m+1}[m+2-i]\, R_{YY}[m+2-i]$$

Using the recursion for $h_{m+1}[i]$

We get

$$\varepsilon[m+1] = \varepsilon[m]\left(1 - k_{m+1}{}^2\right)$$

will give MSE recursively. Since MSE is non negative

$$k_m^2 \le 1$$
$$\therefore |k_m| \le 1$$

- If $|k_m| < 1,$ the LPC error filter will be minimum-phase, and hence the corresponding syntheses filter will be stable.
- Efficient realization can be achieved in terms of $k_m$.
- $k_m$ represents the direct correlation of the data $y[n-m]$ on $y[n]$ when the correlation due to the intermediate data $y[n-m+1], y[n-m+2], ..., y[n-1]$ is removed. It is defined by

$$k_m = \frac{E e_m^f[n] e_m^f[n]}{R_{yy}(0)}$$

where

$$e_m^f[n] = \text{forward prediction error} = y[n] - \sum_{i=1}^{n} h_m[i] y[n-i]$$

and

$$e_m^b[n] = \text{backward prediction error} = y[n-m] - \sum_{i=1}^{n} h_m[m+1-i] y[n+1-i]$$

## 6.8 Steps of the Levinson- Durbin algorithm

Given $R_{YY}[m], m = 0, 1, 2, \ldots.$

Initialization

Take $h_m[0] = -1$ for all $m$

For $m = 0,$

$$\varepsilon[0] = R_{YY}[0]$$

For $m = 1,2,3...$

$$k_m = \frac{\sum_{i=0}^{m-1} h_{m-1}[i] R_{YY}[m-i]}{\varepsilon[m-1]}$$

$h_m[i] = h_{m-1}[i] + k_m h_{m-1}[m-i]$, i = 1,2..., m - 1

$h_m[m] = -k_m$

$\varepsilon_m = \varepsilon_{m-1}(1-k_m^2)$

Go on computing up to given final value of $m$.

**Some salient points**

- The reflection parameters and the mean-square error completely determine the LPC coefficients. Alternately, given the reflection coefficients and the final mean-square prediction error, we can determine the LPC coefficients.
- The algorithm is order recursive. By solving for *m-th* order linear prediction problem we get all previous order solutions
- If the estimated autocorrelation sequence satisfy the properties of an autocorrelation functions, the algorithm will yield stable coefficients

## 6.9 Lattice filer realization of Linear prediction error filters

$e_m^f[n]$ = prediction error due to *mth* order forward prediction.

$e_m^b[n]$ = prediction error due to *mth* order backward prediction.

Then,

$$e_m^f[n] = y[n] - \sum_{i=1}^{m} h_m[i]\ y[n-i] \qquad (1)$$

$$e_m^b[n] = y[n-m] - \sum_{i=1}^{m} h_m[m+1-i]\ y[n+1-i]$$

From (1), we get

$$e_m^f[n] = y[n] - h_m[m]y[n-m] - \sum_{i=1}^{m-1} h_m[i]y[n-i]$$

$$= y[n] + k_m y[n-m] - \sum_{i=1}^{m-1} \left(h_{m-1}[i] + k_m h_{m-1}[m-i]\right)y[n-i]$$

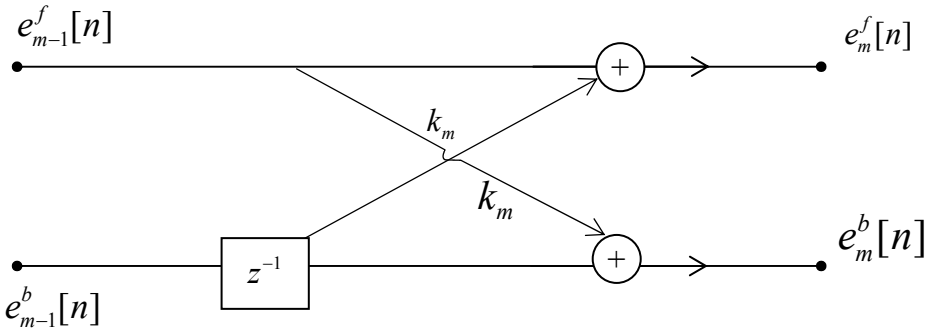
$$= y[n] - \sum_{i=1}^{m-1} h_{m-1}[i]y[n-i] + k_m \left( y[n-m] - \sum_{i=1}^{m-1} h_{m-1}[m-i]y[n-i] \right)$$

$$= e_{m-1}^f[n] + k_m e_{m-1}[n-1]$$

$$\therefore e_m^f[n] = e_{m-1}^f[n] + k_m e_{m-1}^b[n-1]$$

Similarly we can show that

$$e_m^b[n] = e_{m-1}^b[n-1] + k_m e_{m-1}^f[n]$$




**How to initialize the lattice?**

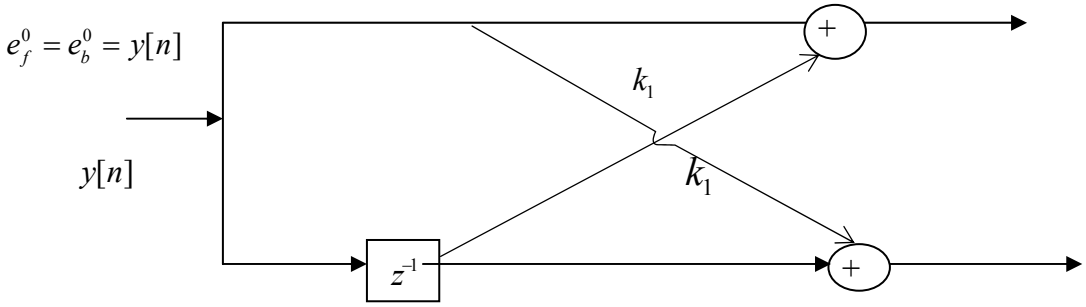
We have

$e_0^f = y[n] - 0$

*and*

$e_0^b = y[n] - 0$

Hence




## 6.10 Advantage of Lattice Structure

- Modular structure can be extended by first cascading another section. New stages can be added without modifying the earlier stages.
- Same elements are used in each stage. So efficient for VLSI implementation.
- Numerically efficient as $|k_m| < 1$.
- Each stage is decoupled with earlier stages.

=> It follows from the fact that for W.S.S. signal, $e_m[n]$ sequences as a function of m are uncorrelated.

$e_i^b[n]$ and $e_m^b[n]$ $\quad 0 \leq i < m$

are uncorrelated (Gram Schmidt orthogonalisation may be obtained through Lattice filter).

$$e_k^b[n] = \text{y[n-k]} - \sum_{i=1}^{k} h_k[k+1-i]\ y\ [n+1-i]$$

$$E\ \ e_m^b[n]e_k^b[n] = E\ \ e_m^b[n](\text{ y[n-k]} - \sum_{i=1}^{k} h_k[k+1-i]\ y\ [n+1-i])$$

$$= 0 \qquad \text{for } 0 \le k < m$$

Thus the lattice filter can be used to whiten a sequence.

With this result, it can be shown that

$$k_m = -\frac{E\left(e_{m-1}^f[n]e_{m-1}^b[n-1]\right)}{E\left(e_{m-1}^b[n-1]\right)^2} \qquad \text{(i)}$$

and

$$k_m = -\frac{E\left(e_{m-1}^f[n]e_{m-1}^b[n-1]\right)}{E\left(e^f{}_{m-1}[n]\right)^2} \qquad \text{(ii)}$$

**Proof:**

Mean Square Prediction Error

$$= E\left(e_m^f[n]\right)^2$$
$$= E\left(e_{m-1}^f[n] + k_m e_{m-1}^b[n-1]\right)^2$$

This is to be minimized w.r.t. $k_m$

$$\Rightarrow 2\left(e_{m-1}^f[n] + k_m e_{m-1}^b[n-1]\right)e_{m-1}^b[n-i] = 0$$
$$\Rightarrow (i)$$

Minimizing $E\left(e_m^b[n]\right)^2 \Rightarrow (ii)$

**Example 2:**

Consider the random signal model $y[n] = x[n] + v[n]$, where $v[n]$ is a 0-mean white noise with variance 1 and uncorrelated with $x[n]$ and $x[n] = 0.8x[n-1] + w[n]$, $w[n]$ is a 0-mean random variable with variance 0.68

a) Find the second –order linear predictor for $y[n]$
b) Obtain the lattice structure for the prediction error filter
c) Use the above structure to design a second-order FIR Wiener filter to estimate $x[n]$ from $y[n]$.

# CHAPTER – 7: ADAPTIVE FILTERS

## 7.1 Introduction

In practical situations, the system is operating in an uncertain environment where the input condition is not clear and/or the unexpected noise exists. Under such circumstances, the system should have the flexible ability to modify the system parameters and makes the adjustments based on the input signal and the other relevant signal to obtain optimal performance.

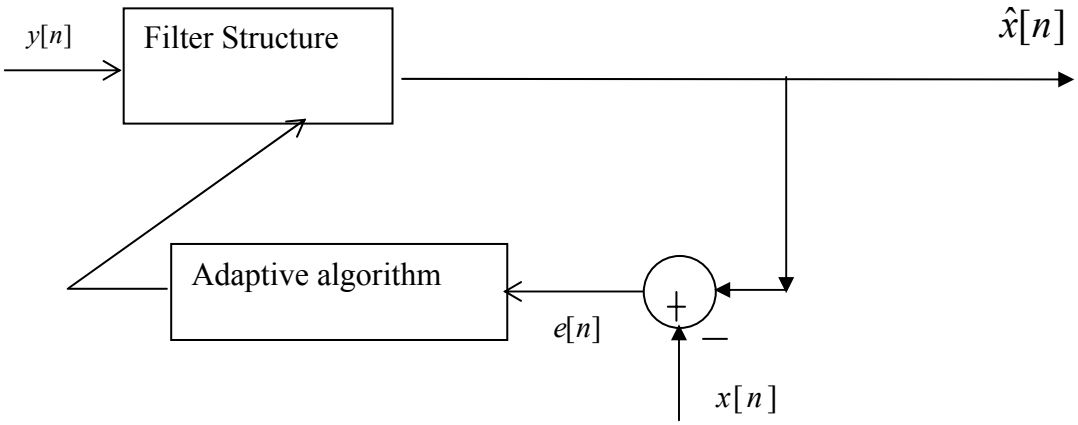
A system that searches for improved performance guided by a computational algorithm for adjustment of the parameters or weights is called an ***adaptive system***. The adaptive system is time-varying.

Wiener filter is a linear time-invariant filter.

In practical situation, the signal is non-stationary. Under such circumstances, optimal filter should be time varying. The filter should have the ability to modify its parameters based on the input signal and the other relevant signal to obtain optimal performance.

How to do this?

- Assume stationarity within certain data length. Buffering of data is required and may work in some applications.
- The time-duration over which stationarity is a valid assumption, may be short so that accurate estimation of the model parameters is difficult.
- One solution is *adaptive filtering*. Here the filter coefficients are updated as a function of the filtering error. The basic filter structure is as shown in Fig. 1.




The filter structure is FIR of known tap-length, because the adaptation algorithm updates each filter coefficient individually.

## 7.2 Method of Steepest Descent

Consider the FIR Wiener filter of length M. We want to compute the filter coefficients iteratively.

Let us denote the time-varying filter parameters by

$$h_i[n], \mathrm{i} = 0,1,\ldots \mathrm{M}\text{-}1$$

and define the filter parameter vector by

$$\boldsymbol{h}[n] = \begin{bmatrix} h_0[n] \\ h_1[n] \\ \vdots \\ h_{M-1}[n] \end{bmatrix}$$

We want to find the filter coefficients so as to minimize the mean-square error $Ee^2[n]$

where

$$\begin{aligned} e[n] &= x[n] - \hat{x}[n] \\ &= \mathrm{x[n]} - \sum_{i=0}^{M-1} h_i[n]y[n-i] \\ &= \mathrm{x[n]} - \mathbf{h}'[n]\mathbf{y}[n] \\ &= \mathrm{x[n]} - \mathbf{y}'[n]\mathbf{h}[n] \end{aligned}$$

$$\text{where } \mathbf{y}[n] = \begin{bmatrix} y[n] \\ y[n-1] \\ \vdots \\ y[n-M+1] \end{bmatrix}$$

Therefore

$$\begin{aligned} Ee^2[n] &= E(x[n] - \mathbf{h}'[n]\mathbf{y}[n])^2 \\ &= R_{XX}[0] - 2\mathbf{h}'[n]\mathbf{r}_{xy} + \mathbf{h}'[n]\mathbf{R}_{\mathbf{YY}}\mathbf{h}[n] \end{aligned}$$

$$\text{where } \mathbf{r}_{XY} = \begin{bmatrix} R_{XY}[0] \\ R_{XY}[1] \\ \vdots \\ R_{XY}[M-1] \end{bmatrix}$$

and

$$\mathbf{R}_{\mathbf{YY}} = \begin{bmatrix} R_{YY}[0] & R_{YY}[-1] & \ldots. & R_{YY}[1-M] \\ R_{YY}[1] & R_{YY}[0] & \ldots. & R_{YY}[2-M] \\ \ldots & & & \\ R_{YY}[M-1] & R_{YY}[M-2] & \ldots. & R_{YY}[0] \end{bmatrix}$$

- The cost function represented by $Ee^2[n]$ is a quadratic in $\mathbf{h}[\mathbf{n}]$
- A unique global minimum exists
- The minimum is obtained by setting the gradient of $Ee^2[n]$ to zero.

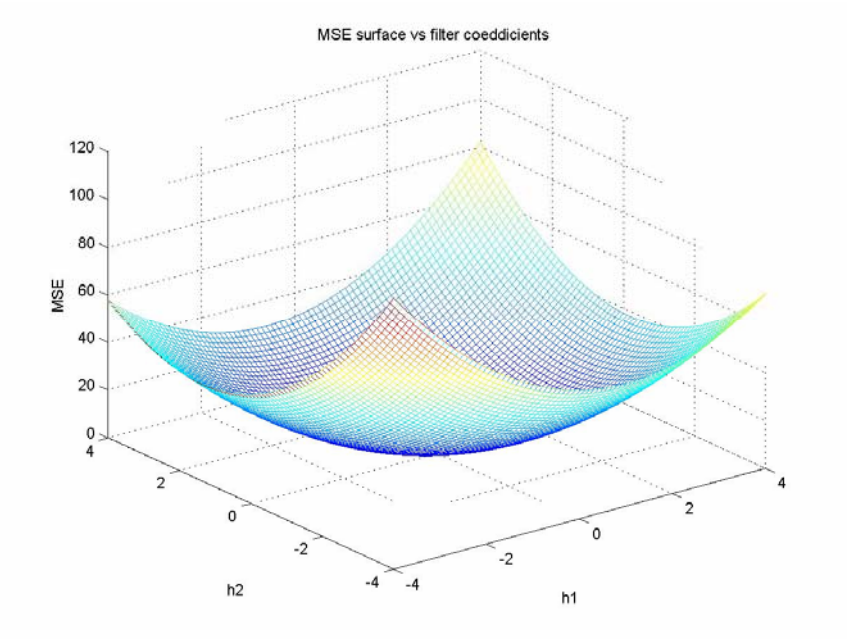



**Figure** - Cost Function $Ee^2[n]$

The optimal set of filter parameters are given by

$$\mathbf{h}_{\mathbf{opt}} = \mathbf{R}_{\mathbf{XY}}^{-1}\mathbf{r}_{\mathbf{XY}}$$

which is the FIR Wiener filter.

Many of the adaptive filter algorithms are obtained by simple modifications of the algorithms for deterministic optimization. Most of the popular adaptation algorithms are based on gradient-based optimization techniques, particularly the steepest descent technique.

The optimal Wiener filter can be obtained iteratively by the method of steepest descent. The optimum is found by updating the filter parameters by the rule

$$\mathbf{h}[n+1] = \mathbf{h}[n] + \frac{\mu}{2}(-\nabla Ee^2[n])$$

where

$$\nabla Ee^2[n] = \begin{bmatrix} \frac{\partial Ee^2[n]}{\partial h_0} \\ \cdots\cdots\cdots \\ \cdots\cdots\cdots \\ \cdots\cdots\cdots \\ \frac{\partial Ee^2[n]}{\partial h_{M-1}} \end{bmatrix}$$

$$= -2\mathbf{r}_{XY} + 2\mathbf{R}_{\mathbf{YY}}\mathbf{h}[n]$$

and $\mu$ is the step-size parameter.

So the steepest descent rule will now give

$$\mathbf{h}[n+1] = \mathbf{h}[n] + \mu(\mathbf{r}_{XY} - \mathbf{R}_{YY}\mathbf{h}[n])$$

## 7.3 Convergence of the steepest descent method

We have

$$\begin{aligned}\mathbf{h}[n+1] &= \mathbf{h}[n] + \mu(\mathbf{r_{XY}} - \mathbf{R_{YY}}\mathbf{h}[n]) \\ &= \mathbf{h}[n] - \mu\mathbf{R_{YY}}\mathbf{h}[n] + \mu\mathbf{r}_{XY} \\ &= (\mathbf{I} - \mu\mathbf{R_{YY}})\mathbf{h}[n] + \mu\mathbf{r}_{XY}\end{aligned}$$

where $\mathbf{I}$ is the $MxM$ identity matrix.

This is a coupled set of linear difference equations.

Can we break it into simpler equations?

$\mathbf{R}_{yy}$ can be digitalized (KL transform) by the following relation

$$\mathbf{R_{YY}} = \mathbf{Q}\Lambda\mathbf{Q}'$$

where $\mathbf{Q}$ is the orthogonal matrix of the eigenvectors of $\mathbf{R}_{YY}$.

$\Lambda$ is a diagonal matrix with the corresponding eigen values as the diagonal elements.

Also $\mathbf{I} = \mathbf{Q}\mathbf{Q}' = \mathbf{Q}'\mathbf{Q}$

Therefore

$$\mathbf{h}[n+1] = (\mathbf{Q}\mathbf{Q}' - \mu\mathbf{Q}\Lambda\mathbf{Q}')\mathbf{h}[n] + \boldsymbol{\mu}\cdot\mathbf{r}_{XY}$$

Multiply by $\mathbf{Q}'$

$$\mathbf{Q}'\mathbf{h}[n+1] = (\mathbf{I} - \mu\Lambda)\mathbf{Q}'\mathbf{h}[n] + \mu\mathbf{Q}'\mathbf{r}_{XY}$$

**Define a new variable**

$\bar{\mathbf{h}}[n] = \mathbf{Q}'\mathbf{h}[n]$ and $\bar{\mathbf{r}}_{XY} = \mathbf{Q}'\mathbf{r}_{XY}$

Then

$$\bar{\mathbf{h}}[n+1] = (\mathbf{I} - \mu\Lambda)\bar{\mathbf{h}}[n] + \mu\bar{\mathbf{r}}_{XY}$$

$$= \begin{bmatrix} 1-\mu\lambda_1 & 0 \cdots\cdots & \cdots & .0 \\ 0 & & & \\ \vdots & & & \\ \vdots & & & \\ 0 & ..... & ... & 1-\mu\lambda_M \end{bmatrix} \bar{\mathbf{h}}[n] + \mu\bar{\mathbf{r}}_{\mathbf{xy}}$$

This is a decoupled set of linear difference equations

$$\bar{h}_i[n+1] = (1-\mu\lambda_i).\bar{h}_i[n] + \mu.\bar{r}_{xy}[i] \qquad i = 0,1,...M-1$$

and can be easily solved for stability. The stability condition is given by

$$|1-\mu\lambda_i| < 1$$

$$\Rightarrow -1 < 1-\mu\lambda_i < 1$$

$$\Rightarrow 0 < \mu < 2/\lambda_i, i = 1,.....M$$

Note that all the eigen values of $\mathbf{R_{YY}}$ are positive.

Let $\lambda_{\max}$ be the maximum eigen value. Then,

$$\lambda_{\max} < \lambda_1 + \lambda_2 + .....\lambda_M$$
$$= \text{Trace}(\mathbf{R_{YY}})$$

$$\therefore 0 < \mu < \frac{2}{Trace(\mathbf{R_{yy}})}$$
$$= \frac{2}{\text{M.R}_{\text{YY}}[0]}$$

The steepest decent algorithm converges to the corresponding Wiener filter

$$\lim_{n\to\infty} \mathbf{h}[n] = \mathbf{R_{YY}^{-1}}\ \mathbf{r_{XY}}$$

if the step size $\mu$ is within the range of specified by the above relation.]

## 7.4 Rate of Convergence

The rate of convergence of the Steepest Descent Algorithm will depend on the factor $(1-\mu\lambda_i)$ in

$$\overline{h}_i[n+1] = (1-\mu\lambda_i).\overline{h}_i[n] + \mu.\overline{R}_{xy}[i] \qquad i = 0,1,...M-1$$

Thus the rate of convergence depends on the statistics of data and is related to the eigen value spread for the autocorrelation matrix. This rate is expressed using the condition number of $\mathbf{R_{YY}}$, defined as $k = \frac{\lambda_{\max}}{\lambda_{\min}}$ where $\lambda_{\max}$ and $\lambda_{\min}$ are respectively the maximum and the minimum eigen values of $\mathbf{R_{YY}}$. The fastest convergence of this system occurs when $k$ = 1, corresponding to white noise.

## 7.5 LMS algorithm (Least – Mean –Square) algorithm

Consider the steepest descent relation

$$\mathbf{h}[n+1] = \mathbf{h}[n] - \frac{\mu}{2}\nabla\mathbf{E}e^2[n]$$

Where

$$\nabla \mathbf{E}e^2[n] = \begin{bmatrix} \frac{\partial Ee^2[n]}{\partial h_0} \\ ..... \\ ..... \\ ..... \\ \frac{\partial Ee^2[n]}{\partial h_{M-1}} \end{bmatrix}$$

In the LMS algorithm $Ee^2[n]$ is approximated by $e^2[n]$ to achieve a computationally simple algorithm.

$$\nabla \mathbf{E}e^2[n] \cong 2.e[n]. \begin{bmatrix} \frac{\partial e[n]}{\partial h_0} \\ ..... \\ ..... \\ ..... \\ \frac{\partial e[n]}{\partial h_{M-1}} \end{bmatrix}$$

Now consider

$$e[n] = x[n] - \sum_{i=0}^{M-1} h_i[n] y[n-i]$$

$$\frac{\partial e[n].}{\partial h_j} = -y[n-j], j = 0,1,.....M-1$$

$$\therefore \begin{bmatrix} \frac{\partial e[n]}{\partial h_0} \\ ..... \\ ..... \\ \frac{\partial e[n]}{\partial h_{M-1}} \end{bmatrix} = - \begin{bmatrix} y[n] \\ y[n-1] \\ ..... \\ ..... \\ ..... \\ y[n-M+1] \end{bmatrix} = -\mathbf{y}[n]$$

$$\therefore \nabla \mathbf{E}\mathbf{e}^2[n] = -2e[n]\mathbf{y}[n]$$

The steepest descent update now becomes

$$\mathbf{h[n+1]} = \mathbf{h[n]} + \mu e[n]\mathbf{y[n]}$$

This modification is due to Widrow and Hopf and the corresponding adaptive filter is known as the ***LMS filter.***

Hence the LMS algorithm is as follows

Given the input signal $y[n]$, reference signal $x[n]$ and step size $\mu$

1. Initialization $h_i[0] = 0, i = 0,1,2.....M-1$

2 For $n > 0$

Filter output

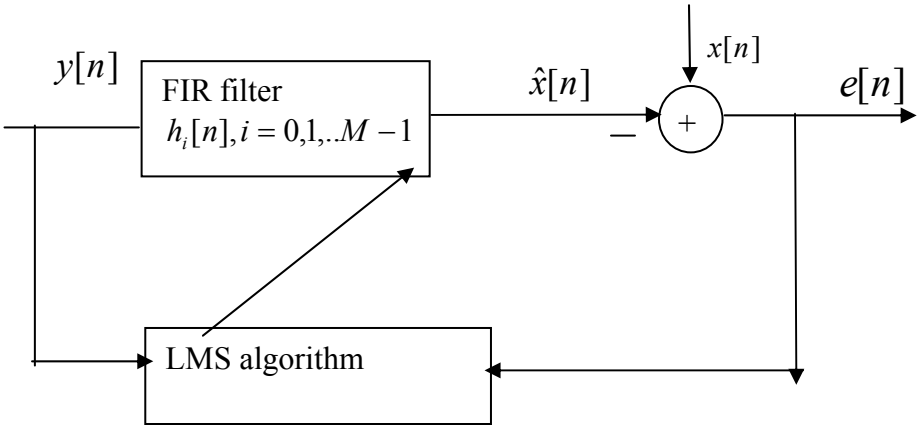
$$\hat{x}[n] = \mathbf{h}'[n]\mathbf{y}[n]$$

Estimation of the error

$$e[n] = x[n] - \hat{x}[n]$$

3. Tap weight adaptation

$$\mathbf{h[n+1]} = \mathbf{h[n]} + \mu \mathbf{e[n]y[n]}$$




## 7.6 Convergence of the LMS algorithm

As there is a feedback loop in the adaptive algorithm, convergence is generally not assured. The convergence of the algorithm depends on the step size parameter $\mu$.

- The LMS algorithm is convergent in the mean if the step size parameter $\mu$ satisfies the condition.

$$0 < \mu < \frac{2}{\lambda_{\max}}$$

**Proof:**

$$\mathbf{h[n+1]} = \mathbf{h[n]} + \mu e[n]\mathbf{y[n]}$$

$$\begin{aligned} \therefore E\mathbf{h[n+1]} &= E\mathbf{h[n]} + \mu E\mathbf{y[n]}e[n] \\ &= E\mathbf{h[n]} + \mu E\mathbf{y[n]}(x[n] - \mathbf{y'[n]}h[\mathbf{n}]) \\ &= E\mathbf{h[n]} + \mu \mathbf{r}_{\mathbf{XY}} - \mu E\mathbf{y[n]y'[n]h[n]} \end{aligned}$$

Assuming the coefficient to be independent of data (Independence Assumption), we get

$$\begin{aligned} E\mathbf{h[n+1]} &= E\mathbf{h[n]} + \mu \mathbf{r}_{\mathbf{XY}} - \mu E\mathbf{y[n]y'[n]}E\mathbf{h[n]} \\ &= E\mathbf{h[n]} + \mu \mathbf{r}_{\mathbf{XY}} - \mu \mathbf{R}_{\mathbf{XY}} E\mathbf{h[n]} \end{aligned}$$

Hence the mean value of the filter coefficients satisfies the steepest descent iterative relation so that the same stability condition applies to the mean of the filter coefficients.

- In the practical situation, knowledge of $\lambda_{\max}$ is not available and Trace $\mathbf{R}_{\mathbf{yy}}$ can be taken as the conservative estimate of $\lambda_{\max}$ so that for convergence
- $0 < \mu < \dfrac{2}{Trace(\mathbf{R}_{\mathbf{YY}})}$
- Also note that trace, $Trace(\mathbf{R}_{\mathbf{YY}}) = M\mathbf{R}_{\mathbf{YY}}[0]$ = Tape input power of the LMS filter.

Generally, a too small value of $\mu$ results in slower convergence where as big values of $\mu$ will result in larger fluctuations from the mean. Choosing a proper value of $\mu$ is very important for the performance of the LMS algorithm.

In addition, the rate of convergence depends on the statistics of data and is related to the eigenvalue spread for the autocorrelation matrix. This is defined using the condition number of $\mathbf{R_{YY}}$, defined as $k = \frac{\lambda_{\max}}{\lambda_{\min}}$ where $\lambda_{\min}$ is the minimum eigenvalue of $\mathbf{R_{YY}}$. The fastest convergence of this system occurs when $k$ = 1, corresponding to white noise. This states that the fastest way to train a LMS adaptive system is to use white noise as the training input. As the noise becomes more and more colored, the speed of the training will decrease.

The average of each filter tap –weight converges to the corresponding optimal filter tap-weight. But this does not ensure that the coefficients converge to the optimal values.

## 7.7 Excess mean square error

Consider the LMS difference equation:

$$\mathbf{h[n+1]} = \mathbf{h[n]} + \mu e[n]\mathbf{y[n]}$$

We have seen that the mean of LMS coefficient converges to the steepest descent solution. But this does not guarantee that the mean square error of the LMS estimator will converge to the mean square error corresponding to the wiener solution. There is a fluctuation of the LMS coefficient from the wiener filter coefficient.

Let $\mathbf{h_{opt}}$ = optimal wiener filter impulse response.

The instantaneous deviation of the LMS coefficient from $\mathbf{h_{opt}}$ is

$$\Delta\mathbf{h}[n] = \mathbf{h}[n] - \mathbf{h}_{opt}$$

$$\begin{aligned}\varepsilon[n] = Ee^2[n] &= E\{x[n] - \mathbf{h}'_{opt}\mathbf{y}[n] - \Delta\mathbf{h}'[n]\mathbf{y}[n]\}^2 \\ &= E\{x[n] - \mathbf{h}'_{\mathbf{opt}}y[n]\}^2 + E\Delta\mathbf{h}'[n]\mathbf{y}[n]\mathbf{y}'[n]\Delta\mathbf{h}[n] - 2E(e_{opt}[n]\Delta\mathbf{h}'[n]\mathbf{y}[n]) \\ &= \varepsilon_{\min} + E\Delta\mathbf{h}'[n]\mathbf{y}[n]\mathbf{y}'[n]\Delta\mathbf{h}[n] - 2E(e_{opt}[n]\Delta\mathbf{h}'[n]\mathbf{y}[n]) \\ &= \varepsilon_{\min} + E\Delta\mathbf{h}'[n]\mathbf{y}[n]\mathbf{y}'[n]\Delta\mathbf{h}[n]\end{aligned}$$

assuming the independence of deviation with respect to data and at $E\Delta\mathbf{h}[n] = 0$.

Therefore,

$$\varepsilon_{excess} = E\Delta\mathbf{h}'[n]\mathbf{y}[n]\mathbf{y}'[n]\Delta\mathbf{h}[n]$$

An exact analysis of the excess mean-square error is quite complicated and its approximate value is given by

$$\varepsilon_{excess} = \varepsilon_{\min} \frac{\sum_{i=1}^{M} \frac{\mu\lambda_i}{2-\mu\lambda_i}}{1-\sum_{i=1}^{M} \frac{\mu\lambda_i}{2-\mu\lambda_i}}$$

The LMS algorithm is said to converge in the mean-square sense provided the step-length parameter satisfies the relations

$$\mu \sum_{i=1}^{M} \frac{\mu\lambda_i}{2-\mu\lambda_i} < 1$$

and $\quad 0 < \mu < \frac{2}{\lambda_{\max}}$

If $\sum_{i=1}^{M} \frac{\mu\lambda_i}{2-\mu\lambda_i} << 1$

then $\varepsilon_{excess} = \varepsilon_{\min} \sum_{i=1}^{M} \frac{\mu\lambda_i}{2-\mu\lambda_i}$

Further, if

$$\mu << 1$$

$$\varepsilon_{excess} = \varepsilon_{\min} \mu \frac{\frac{1}{2} Trace(\mathbf{R_{YY}})}{1-0}$$

$$\simeq \varepsilon_{\min} \mu \frac{1}{2} Trace(\mathbf{R_{YY}})$$

The factor $\frac{\varepsilon_{excess}}{\varepsilon_{\min}} = \sum_{i=1}^{M} \frac{\mu\lambda_i}{2-\mu\lambda_i}$ is called the *misadjustment factor* for the LMS filter.

## 7.8 Drawback of the LMS Algorithm

Convergence is slow when the eigenvalue spread of the autocorrelation matrix is large.

Misadjustment factor given by

$$\frac{\varepsilon_{excess}}{\varepsilon_{\min}} \approx \frac{1}{2} \mu Trace(\mathbf{R_{YY}})$$

is large unless $\mu$ is much smaller. Thus the selection of the step-size parameter is crucial in the case of the LMS algorithm.

When the input signal is nonstationary the eigenvalues also change with time and selection of $\mu$ becomes more difficult.

**Example 1:**

The input to a communication channel is a test sequence $x[n] = 0.8x[n-1] + w[n]$ where $w[n]$ is a 0 mean unity variance white noise. The channel transfer function is given by $H(z) = z^{-1} - 0.5z^{-2}$ and the channel is affected white Gaussian noise of variance 1.

(a) Find the FIR Wiener filter of length 2 for channel equalization

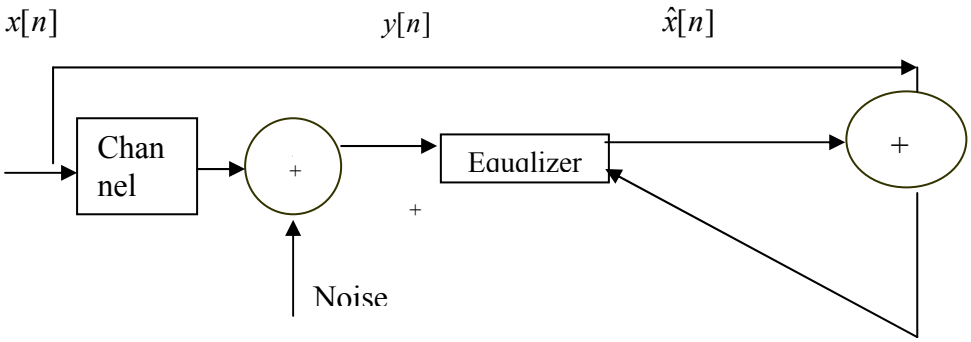
(b) Write down the LMS filter update equations

(c) Find the bounds of LMS step length parameters

(d) Find the excess mean square error.

**Solution:**

From the given model




$$R_{XX}[m] = \frac{\sigma_W^{\,2}}{1-0.8^2}(0.8)^{|m|}$$

$$R_{XX}[0] = 2.78$$

$$R_{XX}[1] = 2.22$$

$$R_{XX}[2] = 1.78$$

$$R_{XX}[3] = 1.42$$

$$y[n] = x[n-1] - 0.5x[n-2] + v(n)$$

$$\therefore\ R_{YY}[m] = 1.25R_{XX}[m] - 0.5R_{XX}[m+1] - 0.5R_{XX}[m-1] + \delta[m]$$

$$R_{YY}[0] = 2.255$$

$$R_{YY}[1] = 0.5325$$

also $R_{XY}[m] = Ex[n]\big(x[n-1-m] - 0.5\ x[n-2-m] + v[n-m]\big)$

$$= R_{XX}[m+1] - 0.5\ R_{XX}[m+2]$$

$\therefore\ R_{XY}[0] = 1.33$

and $R_{XY}[1] = 1.07$

Therefore, the Wiener solution is given by

$$\begin{bmatrix} 2.255 & 0.533 \\ 0.533 & 2.255 \end{bmatrix} \begin{bmatrix} h_0 \\ h_1 \end{bmatrix} = \begin{bmatrix} 1.33 \\ 1.07 \end{bmatrix}$$

$$MSE = R_{XX}[0] - h_0\ R_{XY}[1] - h_1 R_{XY}[2]$$

$h_0 = 0.51$

$h_1 = 0.35$

$\lambda_1, \lambda_2 = 2.79, 1.72$

$\mu \angle \dfrac{2}{2.79}$

$= 0.72$

Excess mean square error

$$= \xi mn \frac{\sum_{I-1}^{2} \frac{2\lambda i}{2-\mu\lambda i}}{1-\sum_{I-1}^{2} \frac{\mu\lambda i}{2-2\lambda i}}$$

## 7.9 Leaky LMS Algorithm

Minimizes $e^2[n] + \alpha \|\mathbf{h}[n]\|^2$

where $\|\mathbf{h}[n]\|$ is the modulus of the LMS weight vector and $\alpha$ is a positive quantity.

The corresponding algorithm is given by

$$\mathbf{h}[n+1] = (1 - \mu\alpha)\mathbf{h}[n] + \mu\alpha\mathbf{e}[n]\mathbf{y}[n]$$

where $\mu\alpha$ is chosen to be less than 1. In such a situation the pole will be inside the unit circle, instability problem will not be there and the algorithm will converge.

## 7.10 Normalized LMS Algorithm

For convergence of the LMS algorithm

$$0 < \mu < \frac{2}{\lambda_{MAX}}$$

and the conservative bound is given by

$$0 < \mu < \frac{2}{Trace(\mathbf{R_{YY}})}$$

$$= \frac{2}{M\mathbf{R_{YY}}[0]}$$

$$= \frac{2}{ME(Y^2[n])}$$

We can estimate the bound by estimating $E(Y^2[n])$ by

$\frac{1}{M}\sum_{n=0}^{M} Y^2[n]$ so that we get the bound

$$0 < \mu < \frac{2}{\sum_{i=0}^{M-1} y^2[n-i]} = \frac{2}{\|\mathbf{y}[n]\|^2}$$

Then we can take

$$\mu = \beta \frac{2}{\|\mathbf{y}[n]\|^2}$$

and the LMS updating becomes

$$\mathbf{h}[n+1] = \mathbf{h}[n] + \beta \frac{1}{\|\mathbf{y}[n]\|^2} e[n]\mathbf{y}[n]$$

where $0 < \beta < 2$

## 7.11 Discussion - LMS

The NLMS algorithm has more computational complexity compared to the LMS algorithm

Under certain assumptions, it can be shown that the NLMS algorithm converges for $0 < \beta < 2$

$\|\mathbf{y}[n]\|^2$ can be efficiently estimated using the recursive relation

$$\|\mathbf{y}[n]\|^2 = \|\mathbf{y}[n-1]\|^2 + y^2[n] - y^2[n-M-1]$$

Notice that the NLMS algorithm does not change the direction of updation in steepest descent algorithm

If $\mathbf{y}[n]$ is close to zero, the denominator term ( $\|\mathbf{y}[n]\|^2$ ) in NLMS equation becomes very small and

$$\mathbf{h}[n+1] = \mathbf{h}[n] + \beta \frac{1}{\|\mathbf{y}[n]\|^2} e[n]\mathbf{y}[n] \qquad \text{may diverge}$$

To overcome this drawback a small positive number $\varepsilon$ is added to the denominator term the NLMS equation. Thus

$$\mathbf{h}[n+1] = \mathbf{h}[n] + \beta \frac{1}{\varepsilon + \|\mathbf{y}[n]\|^2} e[n]\mathbf{y}[n]$$

For computational efficiency, other modifications are suggested to the LMS algorithm. Some of the modified algorithms are *blocked-LMS* algorithm, *signed LMS* algorithm etc. LMS algorithm can be obtained for IIR filter to adaptively update the parameters of the filter

$$y[n] = \sum_{i=1}^{M-1} a_i[n]y[n-i] + \sum_{i=0}^{N-1} b_i[n]x[n-i]$$

How ever, IIR LMS algorithm has poor performance compared to FIR LMS filter.

## 7.12 Recursive Least Squares (RLS) Adaptive Filter

- LMS convergence slow
- Step size parameter is to be properly chosen
- Excess mean-square error is high
- LMS minimizes the instantaneous square error $e^2[n]$
- Where $e[n] = x[n] - \mathbf{h}'[n]\,\mathbf{y}[n] = x[n] - \mathbf{y}'[n]\,\mathbf{h}[n]$

The RLS algorithm considers all the available data for determining the filter parameters. The filter should be optimum with respect to all the available data in certain sense.

Minimizes the cost function

$$\varepsilon[n] = \sum_{k=0}^{n} \lambda^{\,n-k} e^2[k]$$

with respect to the filter parameter vector $\mathbf{h}[n] = \begin{bmatrix} h_0[n] \\ h_1[n] \\ h_{M-1}[n] \end{bmatrix}$

where λ is the weighing factor known as the forgetting factor

- Recent data is given more weightage
- For stationary case λ = 1 can be taken
- λ ≅ 0.99 is effective in tracking local nonstationarity

The minimization problem is

Minimize $\varepsilon[n] = \sum_{k=0}^{n} \lambda^{n-k} \left(x[k] - \mathbf{y}'[k]\ \mathbf{h}[n]\right)^2$ with respect to $\mathbf{h}[n]$

The minimum is given by

$$\frac{\partial \varepsilon(n)}{\partial \mathbf{h}(n)} = \mathbf{0}$$

$$\Rightarrow 2\sum_{k=0}^{n} \lambda^{n-k} \left(x[k]\,\mathbf{y}[k] - \mathbf{y}[k]\,\mathbf{y}'[k]\,\mathbf{h}[n]\right) = 0$$

$$\Rightarrow \mathbf{h}[n] = \left(\sum_{k=0}^{n} \lambda^{n-k} \mathbf{y}[k]\mathbf{y}'[k]\right)^{-1} \sum_{k=0}^{n} \lambda^{n-k} x[k]\mathbf{y}[k]$$

Let us define $\hat{\mathbf{R}}_{YY}[n] = \sum_{k=0}^{n} \lambda^{n-k} \mathbf{y}[k]\mathbf{y}'[k]$

which is an estimator for the autocorrelation matrix $\mathbf{R}_{\mathbf{YY}}$.

Similarly $\hat{\mathbf{r}}_{\mathbf{XY}}[n] = \sum_{k=0}^{n} \lambda^{n-k} x[k]\mathbf{y}[k] =$ estimator for the autocorrelation vector $\mathbf{r}_{XY}[n]$

Hence $\mathbf{h}[n] = \left(\hat{\mathbf{R}}_{\mathbf{XY}}[n]\right)^{-1} \hat{\mathbf{r}}_{\mathbf{XY}}[n]$

Matrix inversion is involved which makes the direct solution difficult. We look forward for a recursive solution.

## 7.13 Recursive representation of $\hat{\mathbf{R}}_{\mathbf{YY}}[n]$

$\hat{\mathbf{R}}_{\mathbf{YY}}[n]$ can be rewritten as follows

$$\begin{aligned}\hat{\mathbf{R}}_{\mathbf{YY}}[n] &= \sum_{k=0}^{n-1} \lambda^{n-k} \mathbf{y}[k]\,\mathbf{y}'[k] + \mathbf{y}[n]\,\mathbf{y}'[n] \\ &= \lambda \sum_{k=0}^{n-1} \lambda^{n-1-k} \mathbf{y}[k]\,\mathbf{y}'[k] + \mathbf{y}[n]\,\mathbf{y}'[n] \\ &= \lambda \hat{\mathbf{R}}_{\mathbf{YY}}[n-1] + \mathbf{y}[n]\,\mathbf{y}'[n]\end{aligned}$$

This shows that the autocorrelation matrix can be recursively computed from its previous values and the present data vector.

Similarly $\hat{\mathbf{r}}_{\mathbf{XY}}[n] = \lambda \hat{\mathbf{r}}_{\mathbf{XY}}[n-1] + x[n]\mathbf{y}[n]$

$$\begin{aligned}\mathbf{h}[n] &= \left[\hat{\mathbf{R}}_{\mathbf{YY}}[n]\right]^{-1} \hat{\mathbf{r}}_{\mathbf{XY}}[n] \\ &= \left(\lambda \hat{\mathbf{R}}_{\mathbf{YY}}[n-1] + \mathbf{y}[n]\mathbf{y}[n]'\right)^{-1} \hat{\mathbf{r}}_{\mathbf{XY}}[n]\end{aligned}$$

For the matrix inversion above the matrix inversion lemma will be useful.

## 7.14 Matrix Inversion Lemma

If **A, B, C** and **D** are matrices of proper orders, **A** and C nonsingular

$$\mathbf{(A + BCD)}^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\,(\mathbf{DA}^{-1}\mathbf{B} + \mathbf{C}^{-1})^{-1}\,\mathbf{DA}^{-1}$$

Taking $\mathbf{A} = \lambda \hat{\mathbf{R}}_{\mathbf{YY}}[n-1],\ \mathbf{B} = \mathbf{y}[n],\ \mathbf{C} = 1$ and $\mathbf{D} = \mathbf{y}'[n]$

we will have

$$\left(\hat{\mathbf{R}}_{\mathbf{yy}}[n]^{-1}\right) = \frac{1}{\lambda}\hat{\mathbf{R}}_{\mathbf{YY}}^{-1}[n-1] - \frac{1}{\lambda}\hat{\mathbf{R}}_{\mathbf{YY}}^{-1}[n-1]\mathbf{y}[n]\left(\mathbf{y}'[n]\frac{1}{\lambda}\left[\hat{\mathbf{R}}_{\mathbf{YY}}^{-1}[n-1]\right]\mathbf{y}[n] + 1\right)^{-1}\mathbf{y}'[n]\hat{\mathbf{R}}_{\mathbf{YY}}^{-1}[n-1]$$

$$= \frac{1}{\lambda}\left(\hat{\mathbf{R}}_{\mathbf{YY}}^{-1}[n-1] - \frac{\hat{\mathbf{R}}_{\mathbf{YY}}^{-1}[n-1]\mathbf{y}[n]\mathbf{y}[n']\hat{\mathbf{R}}_{\mathbf{YY}}^{-1}[n-1]}{\lambda + \mathbf{y}'[n]\hat{\mathbf{R}}_{\mathbf{YY}}^{-1}[n-1]\mathbf{y}[n]}\right)$$

Rename $\mathbf{P}[n] = \hat{\mathbf{R}}_{\mathbf{YY}}^{-1}[n]$. Then

$$\mathbf{P}[n] = \frac{1}{\lambda}\left(\mathbf{P}[n-1] - \mathbf{k}[n]\mathbf{y}'[n]\mathbf{P}[n-1]\right)$$

where $\mathbf{k}[n]$ is called the 'gain vector' and given by

$$\mathbf{k}[n] = \frac{\mathbf{P}[n-1]\mathbf{y}[n]}{\lambda + \mathbf{y}'[n]\mathbf{P}[n-1]\mathbf{y}[n]}$$

$\mathbf{k}[n]$ important to interpret adaptation is also related to the current data vector $\mathbf{y}[n]$ by

$$\mathbf{k}[n] = \mathbf{P}[n]\mathbf{y}[n]$$

To establish the above relation consider

$$\mathbf{P}[n] = \frac{1}{\lambda}\left(\mathbf{P}[n-1] - \mathbf{k}[n]\mathbf{y}'[n]\mathbf{P}[n-1]\right)$$

Multiplying by $\lambda$ and post-multiplying by $\mathbf{y}[n]$ and simplifying we get

$$\begin{aligned}\lambda\mathbf{P}[n]\mathbf{y}[n] &= \left(\mathbf{P}[n-1] - \mathbf{k}[n]\mathbf{y}'[n]\mathbf{P}[n-1]\right)\mathbf{y}[n] \\ &= \mathbf{P}[n-1]\mathbf{y}[n] - \mathbf{k}[n]\mathbf{y}'[n]\mathbf{P}[n-1]\mathbf{y}[n] \\ &= \lambda\mathbf{k}[n]\end{aligned}$$

Therefore

$$\begin{aligned}\mathbf{h}[n] &= \left(\hat{\mathbf{R}}_{\mathbf{YY}}[n]\right)^{-1}\ \hat{\mathbf{r}}_{\mathbf{XY}}[n] \\ &= \mathbf{P}[n]\left(\lambda\hat{\mathbf{r}}_{\mathbf{XY}}[n-1] + x[n]\mathbf{y}[n]\right) \\ &= \lambda\mathbf{P}[n]\hat{\mathbf{r}}_{\mathbf{XY}}[n-1] + x[n]\mathbf{P}[n]\mathbf{y}[n] \\ &= \lambda\tfrac{1}{\lambda}\left[\mathbf{P}[n-1] - \mathbf{k}[n]\mathbf{y}'[n]\mathbf{P}[n-1]\right]\hat{\mathbf{r}}_{\mathbf{XY}}[n-1] + x[n]\mathbf{P}[n]\mathbf{y}[n] \\ &= \mathbf{h}[n-1] - \mathbf{k}[n]\,\mathbf{y}'[n]\,\mathbf{h}[n-1] + x[n]\,\mathbf{k}[n] \\ &= \mathbf{h}[n-1] + \mathbf{k}[n]\left(x[n] - \mathbf{y}'[n]\mathbf{h}[n-1]\right)\end{aligned}$$

## 7.15 RLS algorithm Steps

**Initialization:**

At n = 0

$\mathbf{P}[0] = \delta\ \mathbf{I}_{\mathbf{MxM}}$, $\delta$ a postive number

$\mathbf{y}[0] = \mathbf{0}$, $\mathbf{h}[0] = \mathbf{0}$ ,

Choose $\lambda$

**Operation:**

For 1 to n = Final do

1. Get $x[n], \mathbf{y}[n]$
2. Get $e[n] = x[n] - \mathbf{h}'[n-1]\mathbf{y}[n]$
3. Calculate gain vector $\mathbf{k}[n] = \dfrac{\mathbf{P}[n-1]\mathbf{y}[n]}{\lambda + \mathbf{y}'[n]\mathbf{P}[n-1]\mathbf{y}[n]}$

4. Update the filter parameters

$$\mathbf{h}[n] = \mathbf{h}[n-1] + \mathbf{k}[n]e[n]$$

5. Update the $\mathbf{P}$ matrix

$$\mathbf{P}[n] = \frac{1}{\lambda}\left(\mathbf{P}[n-1] - \mathbf{k}[n]\mathbf{y}'[n]\mathbf{P}[n-1]\right)$$

end do

## 7.16 Discussion – RLS

### 7.16.1 Relation with Wiener filter

We have the optimality condition analysis for the RLS filters

$$\hat{\mathbf{R}}_{\mathbf{YY}}[n]\mathbf{h}[n] = \hat{\mathbf{r}}_{\mathbf{xy}}[n]$$

where $\hat{\mathbf{R}}_{\mathbf{YY}}[n] = \sum_{k=0}^{n} \lambda^{n-k}\mathbf{y}[k]\mathbf{y}'[k]$

Dividing by $n+1$

$$\frac{\hat{\mathbf{R}}_{\mathbf{YY}}[n]}{n+1} = \frac{\sum_{k=0}^{n} \lambda^{n-k}\mathbf{y}[k]\mathbf{y}''[k]}{n+1}$$

if we consider the elements of $\frac{\hat{\mathbf{R}}_{\mathbf{YY}}[n]}{n+1}$, we see that each is an estimator for the auto-correlation of specific lag.

$$\lim_{n\to\infty} \frac{\hat{\mathbf{R}}_{\mathbf{YY}}[n]}{n+1} = \mathbf{R}_{\mathbf{YY}}[n]$$

in the mean square sense. In other words, weighted sample autocorrelation is a consistent estimator of the auto-correlation function of a WSS process.

Similarly

$$\lim_{n\to\infty} \frac{\hat{\mathbf{r}}_{\mathbf{XY}}[n]}{n+1} = \mathbf{r}_{\mathbf{XY}}[n]$$

Hence as $n \to \infty$, optimality condition can be written as

$$\mathbf{R}_{\mathbf{YY}}[n]\mathbf{h}[n] = \mathbf{r}_{\mathbf{XY}}[n].$$

## 7.16.2. Dependence condition on the initial values

Consider the recursive relation

$$\hat{\mathbf{R}}_{YY}[n] = \lambda \hat{\mathbf{R}}_{YY}[n-1] + \mathbf{y}[n]\mathbf{y}'[n]$$

Corresponding to

$$\hat{\mathbf{R}}_{YY}^{-1}[0] = \delta \mathbf{I}$$

we have $$\hat{\mathbf{R}}_{\mathbf{YY}}[0] = \frac{\mathbf{I}}{\delta}$$

With this initial condition the matrix difference equation has the solution

$$\begin{aligned}\tilde{\mathbf{R}}[n] &= \lambda^{n+1}\hat{\mathbf{R}}_{YY}[-1] + \sum_{k=0}^{n}\lambda^{n-k}\mathbf{y}[k]\mathbf{y}'[k] \\ &= \lambda^{n+1}\hat{\mathbf{R}}_{YY}[-1] + \hat{\mathbf{R}}_{YY}[n] \\ &= \lambda^{n+1}\frac{\mathbf{I}}{\delta} + \hat{\mathbf{R}}_{YY}[n]\end{aligned}$$

Hence the optimality condition is modified as

$$(\lambda^{n+1}\frac{\mathbf{I}}{\delta} + \hat{\mathbf{R}}_{YY}[n])\tilde{\mathbf{h}}[n] = \hat{\mathbf{r}}_{XY}[n]$$

where $\tilde{\mathbf{h}}[n]$ is the modified solution due to assumed initial value of the P-matrix.

$$\frac{\lambda^{n+1}\hat{\mathbf{R}}_{YY}^{-1}[n]\tilde{\mathbf{h}}[n]}{\delta} + \tilde{\mathbf{h}}[n] = \mathbf{h}[n]$$

If we take $\lambda$ as less than 1,then the bias term in the left-hand side of the above equation will be eventually die down and we will get

$$\tilde{\mathbf{h}}[n] = \mathbf{h}[n]$$

## 7.16.3. Convergence in stationary condition

- If the data is stationary, the algorithm will converge in mean at best in M iterations, where M is the number of taps in the adaptive FIR filter.
- The filter coefficients converge in the mean to the corresponding Wiener filter coefficients.
- Unlike the LMS filter which converges in the mean at infinite iterations, the RLS filter converges in a finite time. Convergence is less sensitive to eigen value spread. This is a remarkable feature of the RLS algorithm.
- The RLS filter can also be shown to converge to the Wiener filter in the mean-square sense so that there is zero excess mean-square error.

### 7.16.4. Tracking non-staionarity

If $\lambda$ is small $\lambda^{n-i} \cong 0$ for $i << n$

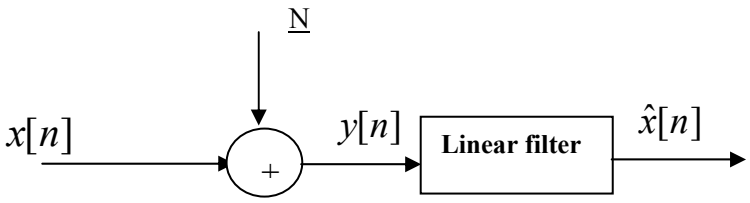
$\Rightarrow$ the filter is based on most recent values. This is also qualitatively explains that the filter can track non stationary in data.

### 7.16.5. Computational Complexity

Several matrix multiplications result in $\cong 7M^2$ arithmetic operations, which are quite large, if the filter tap-length is high. So we have to go for the best implementation of the RLS algorithm.

# CHAPTER – 8: KALMAN FILTER

## 8.1 Introduction




To estimate a signal $x[n]$ in the presence of noise,

- FIR Wiener Filter is optimum when the data length and the filter length are equal.
- IIR Wiener Filter is based on the assumption that infinite length of data sequence is available.

Neither of the above filters represents the physical situation. We need a filter that adds a tap with each addition of data.

The basic mechanism in Kalman filter is to estimate the signal recursively by the following relation

$$\hat{x}[n] = A_n \hat{x}[n-1] + K_n y[n]$$

The whole of Kalman filter is also based on the innovation representation of the signal. We used this model to develop causal IIR Wiener filter.

## 8.2 Signal Model

The simplest Kalman filter uses the first-order AR signal model

$x[n] = ax[n-1] + w[n]$

where $w[n]$ is a white noise sequence.

The observed data is given by

$y[n] = x[n] + v[n]$

where $v[n]$ is another white noise sequence independent of the signal.

The general stationary signal is modeled by a difference equation representing the ARMA (p,q) model. Such a signal can be modeled by the state-space model and is given by

$$\mathbf{x}[n] = \mathbf{A}\mathbf{x}[n-1] + \mathbf{B}\mathbf{w}[n] \qquad (1)$$

And the observations can be represented as a linear combination of the 'states' and the observation noise.

$$y[n] = \mathbf{c}'\mathbf{x}[n] + v[n] \qquad (2)$$

Equations (1) and (2) have direct relation with the state space model in the control system where you have to estimate the 'unobservable' states of the system through an observer that performs well against noise.

**Example 1:**

Consider the $AR(M)$ model

$$x[\mathrm{n}] = a_1 x[\mathrm{n}-1] + a_2 x[\mathrm{n}-2] + .... + a_{\mathrm{M}} x[\mathrm{n}-M] + w[n]$$

Then the state variable model for $x[n]$ is given by

$$\mathbf{x}[\mathrm{n}] = \mathbf{A}\mathbf{x}[\mathrm{n}-1] + \mathbf{B}w[n]$$

where

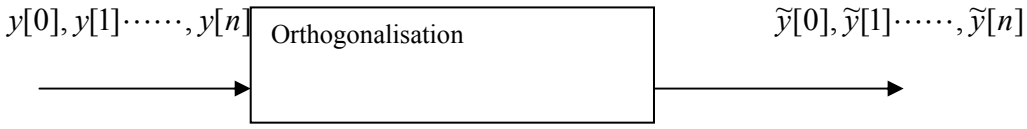
$$\mathbf{x}[\mathbf{n}] = \begin{bmatrix} x_1[n] \\ x_2[n] \\ \vdots \\ x_M[n] \end{bmatrix}, \quad x_1[n] = x[n],\ x_2[n] = x[n-1] .... \text{ and } x_M[n] = x[n-M+1],$$

$$\mathbf{A} = \begin{bmatrix} a_1 & a_2 .. & .. a_M \\ 1 & 0 .. & .. \ 0 \\ 0 & 1 .. & .. \ 0 \\ 0 & 0 .. & .. \ 1 \end{bmatrix}$$

$$\text{and } \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ .. \\ 0 \end{bmatrix}$$

Our analysis will include only the simple (scalar) Kalman filter

The Kalman filter also uses the innovation representation of the stationary signal as does by the IIR Wiener filter. The innovation representation is shown in the following diagram.




Let $\hat{x}[n]$ be the LMMSE of $x[n]$ based on the data $y[0], y[1] \cdots\cdots, y[n]$.

In the above representation $\tilde{y}[n]$ is the innovation of $y[n]$ and contains the same information as the original sequence. Let $\hat{E}(y[n] \mid y[n-1].....,y[0])$ be the linear prediction of $y[n]$ based on $y[n-1].....,y[0].$

Then

$$\begin{aligned}
\tilde{y}[n] &= y[n] - \hat{E}(y[n] \mid y[n-1].....,y[0]) \\
&= y[n] - \hat{E}(x[n] + v[n] \mid y[n-1].....,y[0]) \\
&= y[n] - \hat{E}(ax[n-1] + w[n] + v[n] \mid y[n-1].....,y[0]) \\
&= y[n] - \hat{E}(ax[n-1] \mid y[n-1].....,y[0]) \\
&= y[n] - a\hat{x}[n-1] \\
&= y[n] - x[n] + x[n] - a\hat{x}[n-1] \\
&= y[n] - x[n] + ax[n-1] + w[n] - a\hat{x}[n-1] \\
&= y[n] - x[n] + a(x[n-1] - \hat{x}[n-1]) + w[n] \\
&= v[n] + ae[n-1] + w[n]
\end{aligned}$$

which is a linear combination of three mutually independent orthogonal sequences.
It can be easily shown that $\tilde{y}[n]$ is orthogonal to each of $\tilde{y}[n-1], \tilde{y}[n-2],.....$and $\tilde{y}[0].$

The LMMSE estimation of $x[n]$ based on $y[0], y[1]\cdots\cdots, y[n],$ is same as the estimation based on the innovation sequence , $\tilde{y}[0], y[1],....\tilde{y}[n-1], \tilde{y}[n]$. Therefore,

$$\hat{x}[n] = \sum_{i=0}^{n} k_i \tilde{y}[n]$$

where $k_i s$ can be obtained by the orthogonality relation.
Consider the relation

$$\begin{aligned}
x[n] &= \hat{x}[n] + e[n] \\
&= \textstyle\sum_{i=0}^{n} k_i \tilde{y}[i] + e[n]
\end{aligned}$$

Then

$$E(x[n] - \textstyle\sum_{i=0}^{n} k_i \tilde{y}[i])\tilde{y}[j] = 0 \qquad j = 0,1..n$$

so that

$$k_j = Ex[n]\tilde{y}[j] / \sigma_j^{\,2} \qquad j = 0,1..n$$

Similarly

$$x[n-1] = \hat{x}[n-1] + e[n-1]$$
$$= \sum_{i=0}^{n-1} k_i' \tilde{y}[i] + e[n-1]$$
$$k_j' = Ex[n-1]\tilde{y}[j]/\sigma_j^2 \qquad j = 0,1..n-1$$
$$= E(x[n] - w[n])\tilde{y}[j]/a\sigma_j^2 \qquad j = 0,1..n-1$$
$$= E(x[n])\tilde{y}[j]/a\sigma_j^2 \qquad j = 0,1..n-1$$
$$\therefore k_j' = k_j / a \qquad j = 0,1..n-1$$

Again,

$$\hat{x}[n] = \sum_{i=0}^{n} k_i \tilde{y}[i] = \sum_{i=0}^{n-1} k_i \tilde{y}[i] + k_n \tilde{y}[n]$$
$$= a\sum_{i=0}^{n-1} k_i' \tilde{y}[i] + k_n \tilde{y}[n]$$
$$= a\hat{x}[n-1] + k_n(y[n] - \hat{E}(y[n]/y[n-1].....,y[0]))$$
$$= a\hat{x}[n-1] + k_n(y[n] - a\hat{x}[n-1])$$
$$= (1-k_n)a\hat{x}[n-1]) + k_n y[n]$$

where $\hat{E}(y[n]/y[n-1].....,y[0]))$ is the linear prediction of $y[n]$ based on obseravtions $y[n-1].....,y[0])$ and which is same as $a\hat{x}[n-1]$.

$\therefore \hat{x}[n] = A_n\hat{x}[n-1] + k_n y[n]$
with $A_n = (1-k_n)a$

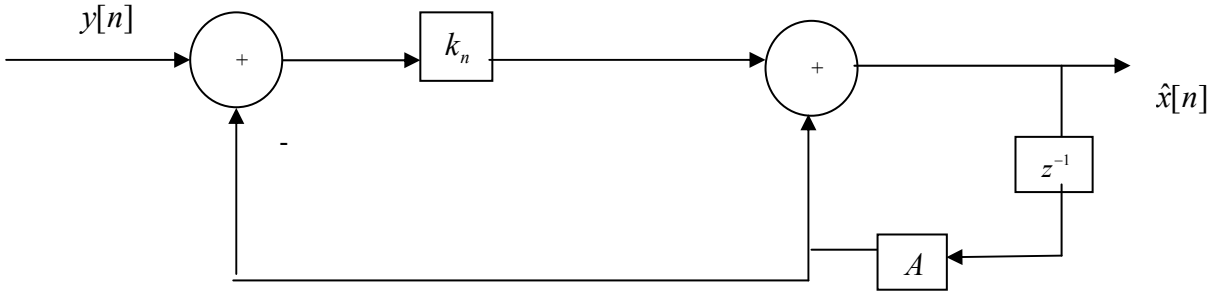
Thus the recursive estimator $\hat{x}[n]$ is given by

$$\boxed{\hat{x}[n] = A_n\hat{x}[n-1] + k_n y[n]}$$

Or

$$\boxed{\hat{x}[n] = a\hat{x}[n-1] + k_n(y[n] - a\hat{x}[n-1])}$$

The filter can be represented in the following diagram



## 8.3 Estimation of the filter-parameters

Consider the estimator

$$\hat{x}[n] = A_n \hat{x}[n-1] + k_n y[n]$$

The estimation error is given by

$$e[n] = x[n] - \hat{x}[n]$$

Therefore $e[n]$ must orthogonal to past and present observed data .

$$Ee[n]y[n-m] = 0, m \geq 0$$

We want to find $A_n$ and the $k_n$ using the above condition.

$e[n]$ is orthogonal to current and past data. First consider the condition that $e[n]$ is orthogonal to the current data.

$$\therefore Ee[n]y[n] = 0$$
$$\Rightarrow Ee[n](x[n] + v[n]) = 0$$
$$\Rightarrow Ee[n]x[n] + Ee[n]v[n] = 0$$
$$\Rightarrow Ee[n](\hat{x}[n] + e[n]) + Ee[n]v[n] = 0$$
$$\Rightarrow Ee^2[n] + Ee[n]v[n] = 0$$
$$\Rightarrow \varepsilon^2[n] + E(x[n] - A_n\hat{x}[n-1] - k_n y[n])v[n] = 0$$
$$\Rightarrow \varepsilon^2[n] - k_n\sigma_V^2 = 0$$
$$\Rightarrow \boxed{k_n = \frac{\varepsilon^2[n]}{\sigma_V^2}}$$

We have to estimate $\varepsilon^2[n]$ at every value of n.

**How to do it?**

Consider

$$\begin{aligned}\varepsilon^2[n] &= Ex[n]e[n] \\ &= Ex[n](x[n] - (1-k_n)a\hat{x}[n-1] - k_n y[n]) \\ &= \sigma_X^2 - (1-k_n)aEx[n]\hat{x}[n-1] - k_n Ex[n]y[n] \\ &= (1-k_n)\sigma_X^2 - (1-k_n)aE(ax[n-1] + w[n])\hat{x}[n-1] \\ &= (1-k_n)\sigma_X^2 - (1-k_n)a^2 Ex[n-1]\hat{x}[n-1]\end{aligned}$$

Again

$$\begin{aligned}\varepsilon^2[n-1] &= Ex[n-1]e[n-1] \\ &= Ex[n-1](x[n-1] - \hat{x}[n-1]) \\ &= \sigma_X^2 - Ex[n-1]\hat{x}[n-1]\end{aligned}$$

Therefore,

$$Ex[n-1]\hat{x}[n-1] = \sigma_X^2 - \varepsilon^2[n-1]$$

Hence $$\boxed{\varepsilon^2[n] = \frac{\sigma_W^2 + a^2\varepsilon^2[n-1]}{\sigma_W^2 + \sigma_V^2 + a^2\varepsilon^2[n-1]}\sigma_V^2}$$

where we have substituted $\sigma_W^2 = (1-a^2)\sigma_X^2$

We have still to find $\varepsilon[0].$ For this assume $x[-1] = \hat{x}[-1] = 0.$ Hence from the relation

$\varepsilon[n] = (1-k_n)\sigma_X{}^2 - (1-k_n)a^2 Ex[n-1]\hat{x}[n-1]$

we get

$\varepsilon^2[0] = (1-k_0)\sigma_X{}^2$

Substituting we get

$$k_0 = \frac{\varepsilon^2[0]}{\sigma_V^2}$$

in the expression for $\varepsilon^2[n]$

We get $$\varepsilon^2[0] = \frac{\sigma_X{}^2\sigma_V{}^2}{\sigma_X{}^2 + \sigma_V{}^2}$$

## 8.4 The Scalar Kalman filter algorithm

Given: Signal model parameters $a$ and $\sigma_{\mathrm{W}}^2$ and the observation noise variance $\sigma_{\mathrm{V}}^2$.

Initialisation $\hat{x}[-1] = 0$

Step 1. $n = 0.$ Calculate $$\varepsilon^2[0] = \frac{\sigma_X^2\sigma_V^2}{\sigma_X^2 + \sigma_V^2}$$

Step 2. Calculate $$k_n = \frac{\varepsilon^2[n]}{\sigma_V^2}$$

Step 3. Input $y[n].$ Estimate $\hat{x}[n]$ by

$$\hat{x}[n] = a\hat{x}[n-1] + k_n(y[n] - a\hat{x}[n-1]$$

Step 4. $n = n+1.$

Step 5. $$\varepsilon^2[n] = \frac{\sigma_W^2 + a^2\varepsilon^2[n-1]}{\sigma_W^2 + \sigma_V^2 + a^2\varepsilon^2[n-1]}\sigma_V^2$$

Step 6. Go to Step 2

**Example 2:**

Given

$$x[\mathrm{n}] = 0.6x[\mathrm{n}-1] + w[n] \qquad \mathrm{n} \geq 0$$

$$y[n] = x[n] + v[n] \qquad \mathrm{n} \geq 0$$

$$\sigma_W^2 = 0.25,\ \sigma_V^2 = 0.5$$

Find the expression for the Kalman filter equations at convergence and the corresponding mean square error.

Using $$\varepsilon^2[n] = \frac{\sigma_W^2 + a^2\varepsilon^2[n-1]}{\sigma_W^2 + \sigma_V^2 + a^2\varepsilon^2[n-1]}\sigma_V^2$$

We get $$\varepsilon^2 = \frac{0.25 + 0.6^2\varepsilon^2}{0.25 + 0.5 + 0.6\varepsilon^2}0.5$$

Solving and taking the positive root

$$\varepsilon^2 = \quad 0.320$$

$$k_n = \varepsilon^2 = 0.390$$

We have to initialize $\varepsilon^2[0]$.

Irrespective of this initialization, $k[n]$ and $\varepsilon[n]$ converge to final values.

## 8.5 Vector Kalman Filter

$$\mathbf{x}[\mathrm{n}] = \mathbf{A}\mathbf{x}[\mathrm{n}-1] + \mathbf{w}[n]$$

$$\mathbf{y}[\mathrm{n}] = \mathbf{c}\mathbf{x}[\mathrm{n}] + \mathbf{v}[n]$$

The vector Kalman filter can be derived in a similar fashion. We will not discuss this derivation.

# SECTION – IV

# SPECTRAL ESTIMATION

# CHAPTER – 9 : SPECTRAL ESTIMATION TECHNIQUES FOR STATIONARY SIGNALS

## 9.1 Introduction

The aim of spectral analysis is to determine the spectral content of a random process from a finite set of observed data.

- Spectral analysis is s very old problem: Started with the Fourier Series (1807) to solve the wave equation.
- Strum generalized it to arbitrary function (1837)
- Schuster devised periodogram (1897) to determine frequency content numerically.

Consider the definition of the power spectrum of a random sequence $\{x[n], -\infty < n < \infty\}$

$$S_{XX}(w) = \sum_{m=-\infty}^{\infty} R_{XX}[m] e^{-j2\pi wm}$$

where $R_{XX}[m]$ is the autocorrelation function.

- The power spectral density is the discrete Fourier Transform (DFT) of the autocorrelation sequence
- The definition involves infinite autocorrelation sequence.
- But we have to use only finite data. This is not only for our inability to handle infinite data, but also for the fact that the assumption of stationarity is valid only for a sort duration. For example, the speech signal is stationary for 20 to 80-ms.

Spectral analysis is a preliminary tool – it says that the particular frequency content may be present in the observed signal. Final decision is to be made by 'Hypothesis Testing'.

Spectral analysis may be broadly classified as *parametric* and non-*parametric*. In the parametric method, the random sequence is modeled by a time-series model, the *model parameters* are estimated from the given data and the spectrum is found by substituting parameters in the model spectrum

We will first discuss the *non-parametric techniques* for spectral estimation. These techniques are based on the Fourier transform of the sample autocorrelation function which is an estimator for the true autocorrelation function.

## 9.2 Sample Autocorrelation Functions

Two estimators of the autocorrelation function exist

$$\hat{R}_{XX}[m] = \frac{1}{N}\sum_{n=0}^{N-1-|m|} x[n]x[n+m] \qquad \text{(biased)}$$

$$\hat{R}'_{XX}[m] = \frac{1}{N-|m|}\sum_{n=0}^{N-1-|m|} x[n]x[n+m] \qquad \text{(unbiased)}$$

Note that

$$\begin{aligned} E\,\hat{R}_{\mathrm{XX}}[m] &= \frac{1}{N}\sum E\sum_{n=0}^{N-1-|m|} x[n]x[n+m] \\ &= \frac{N-|m|}{N}\,R_{XX}[m] \\ &= R_{XX}[m] - \frac{|m|}{N}R_{XX}[m] \\ &= \left(\frac{N-|m|}{N}\right)R_{XX}[m] \end{aligned}$$

Hence $\hat{R}_{XX}[m]$ is a biased estimator of $R_{XX}[m]$. Had we divided the terms under summation by $N-|m|$ instead of N, the corresponding estimator would have been unbiased. Therefore, $\hat{R}'_{XX}[m]$ is an unbiased estimator.

$\hat{R}_{\mathrm{XX}}[m]$ is an asymptotically unbiased estimator. As N→ ∞, the bias of $\hat{R}_{\mathrm{XX}}[m]$ will tend to 0. The variance of $\hat{R}_{\mathrm{XX}}[m]$ is very difficult to be determined, because it involves fourth-order moments. An approximate expression for the covariance of $\hat{R}_{\mathrm{XX}}[m]$ is given by Jenkins and Watt (1968) as

$$\begin{aligned} &Cov(\hat{R}_{XX}[m_1], \hat{R}_{XX}[m_2]) \\ &\cong \frac{1}{N}\sum_{n=-\infty}^{\infty}\left(R_{XX}[n]R_{XX}[n+m_2-m_1] + R_{XX}[n-m_1]R_{XX}[n+m_2]\right) \end{aligned}$$

This means that the estimated autocorrelation values are highly correlated.

The variance of $\hat{R}_{\mathrm{XX}}[m]$ is obtained from above as

$$var(\hat{R}_{XX}[m] \cong \frac{1}{N}\sum_{n=-\infty}^{\infty}\left(R_{XX}{}^{2}[n] + R_{XX}[n-m]R_{XX}[n+m]\right)$$

Note that the variance of $\hat{R}_{\mathrm{XX}}[m]$ is large for large lag $m$, especially as $m$ approaches N.

Also as $N \to \infty$ , $\text{var}\left(\hat{R}_{XX}[m]\right) \to 0$ provided $\sum_{n=-\infty}^{\infty}\left(R_{XX}{}^{2}[n]\right) < \infty$.

As $N \to \infty$, $\text{var}(\hat{R}_{XX}[m]) \to 0$. Though sample autocorrelation function is a consistent estimator, its Fourier transform is not and here lies the problem of spectral estimation.

Though unbiased and consistent estimators for $R_{XX}[m]$, $\hat{R}'_{XX}[m]$ is not used for spectral estimation because $\hat{R}'_{XX}[m]$ does not satisfy the non-negative definiteness criterion.

## 9.3 Periodogram (Schuster, 1898)

$$\hat{S}^{p}_{XX}(w) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x[n]e^{-jnw}\right|^2 , \quad -\pi \le w \le \pi$$

$\hat{S}^{p}_{XX}(w)$ gives the power output of band pass filters of impulse response

$$h_i[n] = \frac{1}{\sqrt{N}} e^{-w_i n} rect(\frac{n}{N})$$

$h_i[n]$ is a very poor band-pass filter.

Also

$$\hat{S}^{p}_{XX}(w) = \sum_{m=-(N-1)}^{N-1} \hat{R}_{XX}[m]e^{-jwm}$$

$$\text{where} \quad \hat{R}_{XX}[m] = \frac{1}{N}\sum_{n=0}^{N-1-|m|} x[n]x[n+m]$$

The periodogram is the Fourier transform of the sample autocorrelation function.

To establish the above relation

Consider

$$x_N[n] = x[n] \text{ for } n < N$$
$$= 0 \quad \text{otherwise}$$
$$\therefore \hat{R}_{XX}[m] = \frac{x_N[m] * x_N[-m]}{N}$$
$$\hat{S}^{p}_{XX}(w) = \frac{1}{N}\left(\sum_{n=0}^{N-1} x[n]e^{-jwn}\right)\left(\sum_{n=0}^{N-1} x[n]e^{jwn}\right)$$
$$= \frac{1}{N}\left|\sum_{n=0}^{N-1} x[n]e^{-jwn}\right|^2$$

$$So \quad \hat{S}_{XX}^{p}(w) = \sum_{m=-(N-1)}^{N-1} \hat{R}_{XX}[m]e^{-jwm}$$

$$E\ \hat{S}_{XX}^{p}(w) = \sum_{m=-(N-1)}^{N-1} E\left[\hat{R}_{XX}[m]\right]e^{-jwm}$$

$$= \sum_{m=-(N-1)}^{N-1}\left(1-\frac{|m|}{N}\right)R_{XX}[m]e^{-jwm}$$

as $N \to \infty$ the right hand side approaches true power spectral density $S_{XX}(f)$. Thus the periodogram is an asymptotically unbiased estimator for the power spectral density.

To prove consistency of the periodogram is a difficult problem.

We consider the simple case when a sequence of *Gaussian white noise* in the following example.

Let us examine the periodogram only at the DFT frequencies $w_k = \frac{2\pi k}{N}, k = 0,1,...N-1.$

**Example 1:**

The periodogram of a zero-mean white Gaussian sequence $x[n]$, $n = 0, \ldots, N$-1.

The power spectral density is given by

$$S_{XX}(w) = \frac{\sigma_x^2}{2\pi} \qquad -\pi < w \le \pi$$

The periodogram is given by

$$\hat{S}_{XX}^{p}(w) = \frac{1}{N}\left(\left|\sum_{n=0}^{N-1} x[n]e^{-jwn}\right|\right)^2$$

Let us examine the periodogram only at the DFT frequencies $w_k = \frac{2\pi k}{N}, k = 0,1,...N-1.$

$$\hat{S}_{XX}^{p}(k) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x[n]e^{-j\frac{2\pi}{N}kn}\right|^2$$

$$= \frac{1}{N}\left|\sum_{n=0}^{N-1} x[n]e^{-jw_K n}\right|^2$$

$$= \left(\sum_{n=0}^{N-1}\frac{1}{\sqrt{N}}x[n]\cos\ w_K n\right)^2 + \left(\sum_{n=0}^{N-1}\frac{x[n]\sin\ w_K n}{\sqrt{N}}\right)^2$$

$$= C_X^2(w_K) + S_X^2(w_K)$$

where $C_X^2(w_K)$ and $S_X^2(w_K)$ are the consine and sine parts of $\hat{S}_{XX}^{p}(w)$.

Let us consider

$$C_X^2(w_K) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n] \cos w_K n$$

which is a linear combination of a Gaussian process.

Clearly $\mathrm{E}\, C_X^2(w_K) = 0$

$$\mathrm{var}\, C_X^2(w_K) = \frac{1}{N} E\left( \sum_{n=0}^{N-1} x[n] \cos w_k n \right)^2$$

$$= \frac{1}{N} \sum_{n=0}^{N-1} E x^2[n] \cos^2 w_K n + E(\text{Cross terms})$$

Assume the sequence $x[n]$ to be independent.

Therefore,

$$\mathrm{var}\left(\mathrm{C_K}(\mathrm{w_K})\right) = \frac{\sigma_X^2}{N} \sum_{n=0}^{N-1} \cos^2 w_K n = \frac{\sigma_X^2}{N} \sum_{n=0}^{N-1} \left( \frac{1 + \cos 2w_K n}{2} \right)$$

$$= \frac{\sigma_X^2}{N} \left( \frac{N}{2} + \cos\ (N-1) w_K \frac{\sin N w_K}{2 \sin w_K} \right)$$

$$= \sigma_X^2 \left( \frac{1}{2} + \cos\ (N-1) w_K \frac{\sin N w_K}{2N \sin w_K} \right)$$

For $w_K = \frac{2\pi}{N} k$

$\frac{\sin N w_K}{2 \sin w_K} = 0 \qquad k \neq 0, k \neq \frac{N}{2}$ (assuming $N$ even).

$\therefore \mathrm{var}\left(\mathrm{C_K}(\mathrm{w_K})\right) = \frac{\sigma_X^2}{2}$ *for* $k \neq 0$

Again $\frac{\sin N w_K}{N \sin w_K} = 1$ *for* $k = 0.$

$\therefore \mathrm{var}\left(\mathrm{C_K}(\mathrm{w_K})\right) = \sigma_X^2$ for $k = 0, k = \frac{N}{2}$

Similarly considering the sine part

$$S_K(w_K) = \frac{1}{\sqrt{N}} \sum_{\mathrm{n=0}}^{\mathrm{N\text{-}1}} x[n] \sin w_K n$$

$E\ S_K(w_K) = 0$ $\quad \because x[n]$ is zero mean.

$\mathrm{var}\ S_K(w_K) = 0$ *for* $k = 0$ (for dc part has no sine term)

$= \frac{\sigma_X^2}{2}$ *for* $k \neq 0.$

*Therefore for* $k = 0,$ *Distribution.*

$C_K[k] \sim N\ (0, \quad \sigma_X^2)$

$S_K[k] = 0.$

*for* $k \neq 0$

$$C_K \sim N\left(0, \frac{\sigma_X^2}{2}\right)$$

$$S_K \sim N\left(0, \frac{\sigma_X^2}{2}\right).$$

We can also show that

$$Cov(S_X(w_K)C_X(w_K)) = 0.$$

So $C_X(w_K)$ and $S_X(w_K)$ are independent Gaussian R.V.s..

The periodogram

$\hat{S}_{XX}^{p}[k] = C_X^2[k] + S_X^2[k]$ has a chi-square distribution

## 9.4 Chi square distribution

Let $X_1, X_2 \ldots X_N$ be independent zero-mean Gaussian variables each with variance $\sigma_X^2$. Then $Y = X_1^2 + X_2^2 + \ldots + X_N^2$ has (chi square) $\chi_N^2$ distribution with mean

$$EY = E(X_1^2 + X_2^2 + \ldots + X_N^2) = N\sigma_X^2 \text{ and variance } 2N\sigma_X^4.$$

$\hat{S}_{XX}[k] = C_X[k] + S_X^2[k].$

It is a $\chi_2^2$ distribution.

$E\ \hat{S}_{XX}[k] = \frac{\sigma_X^2}{2} + \frac{\sigma_X^2}{2} = \sigma_X^2 = S_{XX}[k]$

$\Rightarrow\ \hat{S}_{XX}[k]$ is unbiased

var $(\hat{S}_{XX}[k]) = 2 \times 2\left(\frac{\sigma_X^2}{2}\right)^2 = S_{XX}^2[k]$ which is independent of $N$

$\hat{S}_{XX}[0] = C_0^2[0]$ is a $\chi_1^2$ distribution of degree of freedom 1

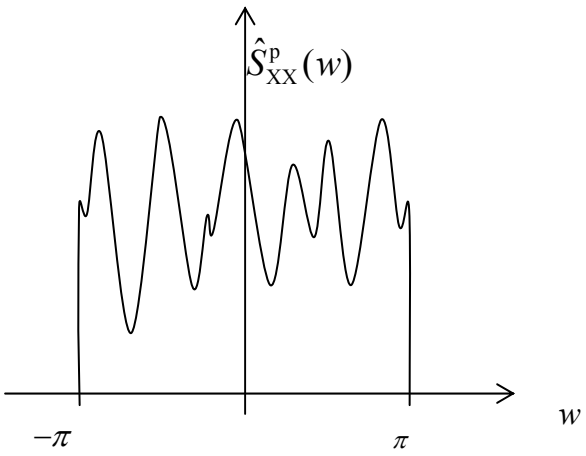
$E\ \hat{S}_{XX}[0] = \sigma_x^2$

$\Rightarrow \hat{S}_{XX}[0]$ is unbiased

and var $\left(\hat{S}_{XX}[0]\right) = \sigma_X^4 = S_{XX}^2[0].$

It can be shown that for the Gaussian independent white noise sequence at any frequency $w$,

var $(\hat{S}_{XX}^{p}(w)) = S_{XX}^2(w)$




**For general case**

$$\hat{S}^{p}_{XX}(w) = \sum_{m=-(N-1)}^{N-1} \hat{R}_{XX}[m]e^{-jwm}$$

*where* $\hat{R}_{XX}[m] = \frac{1}{N}\sum_{n=0}^{N-1-|m|} x[n]x[n+m]$, the biased estimator of autocorrelation fn.

$$= \sum_{m=-(N-1)}^{N-1}\left(1-\frac{|m|}{N}\right)\hat{R}'_{XX}[m]e^{-jwm}$$

$$= \sum_{\mathrm{m=-(N-1)}}^{\mathrm{N-1}} w_B[m]\hat{R}'_{XX}[m]e^{-jwm}$$

where $\hat{R}'_{XX}[m]$ is the unbiased estimator of auto correlation

and $w_B[m] = 1-\frac{|m|}{N}$ is the Bartlett Window.

( Fourier Transform of product of two functions)

So, $\hat{S}^{p}_{XX}(w) = W_B(w) * FT\left\{\hat{R}'_{XX}[m]\right\}$

$$\mathrm{E}\,\hat{S}^{p}_{XX}(w) = W_B(w) * FT\left\{E\,\hat{R}'_{XX}[m]\right\}$$

$$= W_B(w) * S_{XX}(w)$$

$$= \int_{-\pi}^{\pi} W_B(w-\xi)S_{XX}(\xi)d\xi$$

As $N \to \infty$, $\mathrm{E}\,\hat{S}^{p}_{XX}(w) \to S_{XX}(w)$

Now $\mathrm{var}\left(\hat{S}^{p}_{XX}(w)\right)$ cannot be found out exactly (no analytical tool). But an approximate expression for the covariance is given by

$$Cov\left(\hat{S}^{p}_{XX}(w_1),\hat{S}^{p}_{XX}(w_2)\right)$$

$$\simeq S_{XX}(w_1)S_{XX}(w_2)\left[\left(\frac{\sin\frac{N(w_1+w_2)}{2}}{N\sin(\frac{w_1+w_2}{2})}\right)^2 + \left(\frac{\sin\frac{N\,(w_1-w_2)}{2}}{N\sin(\frac{w_1-w_2}{2})}\right)^2\right]$$

$$\underline{\text{var}\left(\hat{S}_{XX}^{p}(w)\right) \simeq} \; S_{XX}^{2}(w)\left[1+\left(\frac{\sin Nw}{N\sin w}\right)^{2}\right]$$

$$\therefore \text{var}\left\{\hat{S}_{XX}^{p}(w)\right\} \simeq 2S_{XX}^{2}(w) \quad \text{for w} = 0, \pi$$

$$\cong S_{XX}^{2}(w) \quad \text{for } 0 < \text{w} < \pi$$

Consider $\text{w}_1 = 2\pi\frac{k_1}{N}$ and $f_2 = 2\pi\frac{k_2}{N}$ , $k_1$ , $k_2$ integers.

Then

$$Cov\left(\hat{S}_{XX}^{p}(w_1), \hat{S}_{XX}^{p}(w_2)\right) \cong 0$$

This means that there will be no correlation between two neighboring spectral estimates. Therefore periodogram is not a reliable estimator for the power spectrum for the following two reasons:

(1) The periodogram is not a consistent estimator in the sense that $\text{var}\left(\hat{S}_{XX}^{p}(w)\right)$ does not tend to zero as the data length approaches infinity.

(2) For two frequencies, the covariance of the periodograms decreases data length increases. Thus the peridogram is erratic and widely fluctuating.

Our aim will be to look for spectral estimator with variance as low as possible, and without increasing much the bias.

## 9.5 Modified Periodograms

**Data windowing**

Multiply *data* with a *more suitable window* rather *than* rectangular before finding the perodogram. The resulting periodogram will have less bias.

### 9.5.1 Averaged Periodogram: The Bartlett Method

The motivation for this method comes from the observation that

$$\lim_{N\to\infty,} \; \text{E}\,\hat{S}_{XX}^{p}(w) \to \; S_{XX}(w)$$

We have to modify $\hat{S}_{XX}^{p}(w)$ to get a consistent estimator for $S_{XX}(f)$.

Given the data $x[n], n = 0,1,. \; . \; . \; N\text{-}1$

Divide the data in $K$ non-overlapping segments each of length $L$.

Determine periodogram of each section.

$$\hat{S}_{XX}^{(k)}(w) = \frac{1}{L}\left|\sum_{n=0}^{L-1} x[n]e^{-jwn}\right|^2 \quad \text{for each section}$$

Then $\hat{S}_{XX}^{(av)}(w) = \frac{1}{K}\sum_{m=0}^{K-1}\hat{S}_{XX}^{(k)}(w).$

As shown earlier,

$$E\hat{S}_{XX}^{(k)}(w) = \int_{-\pi}^{\pi} \mathrm{W_B}(w-\xi)S_{XX}(\xi)d\xi$$

where $w_B[m] = \begin{cases} 1-\frac{|m|}{L}, & |m| \le M-1 \\ 0, & otherwise \end{cases}$

$$\mathrm{W_B}(w) = \left(\frac{\sin\frac{wL}{2}}{\sin\frac{w}{2}}\right)^2$$

$$\hat{S}_{XX}^{(k)}(w) = \sum_{m=-(L-1)}^{L-1}\hat{R}_{XX}[m]e^{-jwm}$$

$$E\hat{S}_{XX}^{av}(w) = \frac{1}{K}\sum_{k=0}^{K-1}E\hat{S}_{XX}^{(k)}(w) = E\left\{S_{XX}^{(k)}(w)\right\}$$

$$E\hat{S}_{XX}^{av}(w) = \int_{-\pi}^{\pi} W_B(w-\xi)S_{XX}(\xi)d\xi.$$

To find the mean of the averaged periodogram, the true spectrum is now convolved with the frequency $W_B(f)$ of the Barlett window. The effect of reducing the length of the data from $N$ points to $L = N/K$ results in a window whose spectral width has been increased by a factor K consequently the frequency resolution has been reduced by a factor K.

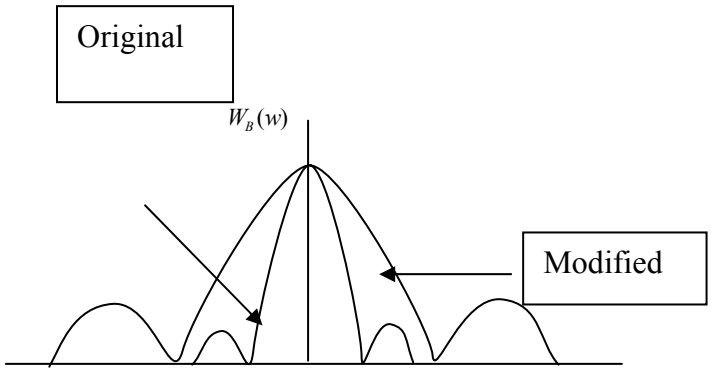



**Figure Effect** of reducing window size

## 9.5.2 Variance of the averaged periodogram

$\text{var}(\hat{S}_{XX}^{av}(w))$ is not simple to evaluate as 4th order moments are involved.

**Simplification :**

Assume the K data segments are independent.

$$\text{Then } \text{var}(\hat{S}_{XX}^{av}(w)) = \text{var}\left\{\sum_{k=0}^{K-1} \hat{S}_{XX}^{(k)}(w)\right\}$$

$$= \frac{1}{K^2}\sum_{k=0}^{K-1} \text{var}\, \hat{S}_{XX}^{(m)}(w)$$

$$\simeq \frac{1}{K^2} \times K\left(1 + \left(\frac{\sin wL}{L \sin w}\right)^2\right) S_{XX}^2(w)$$

$$\simeq \frac{1}{K} \text{ original variance of the periodogram.}$$

So variance will be reduced by a factor less than $K$ because in practical situations, data segments will not be independent.

For large $L$ and large $K$, $\text{var}(\hat{S}_{XX}^{av}(w))$ will tend to zero and $\hat{S}_{XX}^{av}(w)$ will be a consistent estimator.

**The Welch Method** (Averaging modified periodograms)

(1) Divide the data into overlapping segments ( overlapping by about 50 to 75%).

(2) Window the data so that the *modified periodogram* of each segment is

$$\hat{S}_{XX}^{(\text{mod})}(w) = \frac{1}{UL}\left|\sum_{n=0}^{L-1} x[n]w[n]e^{-jwn}\right|^2$$

$$= \frac{1}{L}\sum_{n=0}^{L-1} w^2[n]$$

The window $w[n]$ need not be an is a even function and is used to control spectral leakage.

$\sum_{n=0}^{L-1} x[n]w[n]e^{-jwn}$ is the DTFT of $x[n]\, w[n]$ where $w[n]$ = 1 for $n = 0$ , . . . $L$ - 1

= 0 otherwise

(3) Compute $\hat{S}_{XX}^{(Welch)}(w) = \frac{1}{K}\sum_{k=0}^{K-1} \hat{S}_{XX}^{(\text{mod})}(w)$.

## 9.6 Smoothing the periodogram : The Blackman and Tukey Method

- Widely used non parametric method
- Biased autocorrelation sequence is used $\hat{R}_{XX}[m]$.
- Higher order or large lags autocorrelation estimation involves less data ($N$ - $m$, $m$ large) and so more prone to error, we give less importance to higher-order autocorrelation.

$\hat{R}_{XX}[m]$ is multiplied by a window sequence $w[m]$ which under weighs the autocorrelation function at large lags. The window function $w[m]$ has the following properties.

$$0 < w[m] < 1$$
$$w[0] = 1$$
$$w[-m] = w[m]$$
$$w[m] = 0 \quad for \ |m| > M.$$

$w[0] = 1$ is a consequence of the fact that the smoothed periodogram should not modify a smooth spectrum and so $\int_{-\pi}^{\pi} W(w)dw = 1.$

The smoothed periodogram is given by

$$\hat{S}_{XX}^{BT}(w) = \sum_{m=-(M-1)}^{M-1} w[m]\hat{R}_{XX}[m]e^{-jwm}$$

**<u>Issues concerned</u> :**

1. **How to select** $w[m]$?

There are a large numbers of windows available. Use a window with small side-lobes. This will reduce bias and improve resolution.

2. **How to select M. – Normally**

$M \simeq \frac{N}{5}$ or $M \simeq 2\sqrt{N}$ (Mostly based on experience)

$N \tilde{} 1000$ if N is large 10,000

$\hat{S}_{XX}^{BT}(w)$ = convolution of $\hat{S}_{XX}^{p}(w)$ and $W(w)$, the F.T. of the window sequence.

= Smoothing the periodogram, thus decreasing the variance in the estimate at the expense of reducing the resolution.

$$E(\hat{S}_{XX}^{BT}(w)) = E\ \hat{S}_{XX}^{p}(w) * W(w)$$

where $E\ \hat{S}_{XX}^{p}(w) = \int_{-\pi}^{\pi} S_{XX}(\theta) W_B(w-\theta)\ d\theta$

or from time domain

$$E\ \hat{S}_{XX}^{BT}(w) = E \sum_{m=-(M-1)}^{M-1} W[m]\hat{R}_{XX}[m]e^{-jwm} \quad \text{arymtotically unbiased}$$

$$= \sum_{m=-(M-1)}^{M-1} W[m]\underbrace{E\ \hat{R}_{XX}[m]}e^{-jwm}$$

$\hat{S}_{XX}^{BT}(w)$ can be proved to be asymptotically unbiased.

and variance of $\hat{S}_{XX}^{BT}(w) \simeq \frac{S_{XX}^2(w)}{N} \sum_{K=-(M-1)}^{M-1} w^2[k]$

Some of the popular windows are rectangular window, Bartlett window, Hamming window, Hanning window, etc.

**Procedure**

1. Given the data $x[n], n = 0,1..,N-1$
2. Find periodogram. $\hat{S}_{XX}^{p}(2\pi k/N) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x[n]e^{-j2\pi k/N}\right|^2$,
3. By IDFT find the autocorrelation sequence.
4. Multiply by proper window and take FFT.

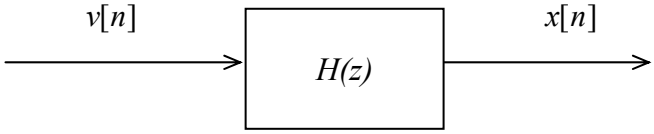
## 9.7 Parametric Method

Disadvantage classical spectral estimators like Blackman Tuckey Method using windowed autocorrelation function.

- Do not use a priori information about the spectral shape
- Do not make realistic assumptions about $x[n]$ for $n < 0$ and $n \geq N$.

Normally $\hat{R}_X[m], m = 0, \pm 1, ..... \pm (N-1)$ can be estimated from $N$ sample values. From these autocorrelations we can estimate a model for the signal – basically a time series model. Once the model is available it is equivalent to know the autocorrelation for all lags and hence will give better spectral estimation. (better resolution) A stationary signal can be represented by ARMA (p, q) model.

$$x[n] = \sum_{i=1}^{p} a_i x[n-i] + \sum_{i=0}^{q} b_i v[n-i]$$




$$H(z) = \frac{B(z)}{A(z)} = \frac{\sum_{i=0}^{q} b_i z^{-i}}{1 - \sum_{i=1}^{p} a_i z^{-i}}$$

*and*

$$S_{XX}(w) = |H(w)|^2 \sigma_V^2$$

- Model signal as AR/MA/ARMA process
- Estimate *model parameters* from the given data
- Find the power spectrum by substituting the values of the model parameters in expression of power spectrum of the model

## 9.8 AR spectral estimation

$AR(p)$ process is the most popular model based technique.

- Widely used for parametric spectral analysis because:
- Can approximate continuous power spectrum arbitrarily well
- (but might need very large $p$ to do so)
- Efficient algorithms available for model parameter estimation
- if the process Gaussian maximum entropy spectral estimate is $AR(p)$ spectral estimate
- Many physical models suggest AR processes (e.g. speech)
- Sinusoids can be expressed as AR-like models

The spectrum is given by

$$\hat{S}_{XX}(w) = \frac{\sigma_V^2}{\left|1 - \sum_{i=1}^{P} a_i e^{-jwi}\right|^2}$$

where $a_i$s are $AR(p)$ process parameters.

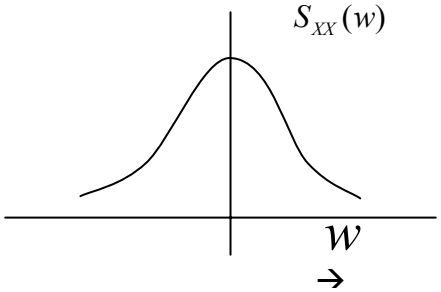



Figure an AR spectrum

## 9.9 The Autocorrelation method

The $a_i$s are obtained by solving the Yule Walker equations corresponding to estimated autocorrelation functions.

$$\begin{bmatrix} \hat{R}_{XX}[0] & \hat{R}_{XX}[1] ..... & \hat{R}_{XX}[p-1] \\ \hat{R}_{XX}[1] & \hat{R}_{XX}[0] ..... & \hat{R}_{XX}[p-2] \\ ..... & & ..... \\ \hat{R}_{XX}[p-1] & \hat{R}_{XX}[1] ..... & \hat{R}_{XX}[0] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ .. \\ a_P \end{bmatrix} = \begin{bmatrix} \hat{R}_X[1] \\ \hat{R}_X[2] \\ .. \\ \hat{R}_X[P+1] \end{bmatrix}$$

$$\sigma_V^2 = \hat{R}_X[0] - \sum_{i=1}^{P} a_i \hat{R}_{XX}[i]$$

## 9.10 The Covariance method

The problem of estimating the AR-parmeters may be considered in terms of minimizing the sum-square error

$$\varepsilon[n] = \sum_{n=p}^{N-1} e^2[n]$$

with respect to the AR parameters, where

$$e[n] = x[n] - \sum_{i=1}^{p} a_i x[n-i]$$

Theis formulation results in

$$\begin{bmatrix} \hat{R}_{XX}[1,1] & \hat{R}_{XX}[2,1] ..... & \hat{R}_{XX}[p,1] \\ \hat{R}_{XX}[1,2] & \hat{R}_{XX}[2,2] ..... & \hat{R}_{XX}[p,2] \\ ..... & & ..... \\ \hat{R}_{XX}[1,p] & \hat{R}_{XX}[2,p] ..... & \hat{R}_{XX}[p,p] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ .. \\ a_P \end{bmatrix} = \begin{bmatrix} \hat{R}_X[0,1] \\ \hat{R}_X[0,2] \\ .. \\ \hat{R}_X[0,P+1] \end{bmatrix}$$

$$\sigma_V^2 = \hat{R}_X[0,0] - \sum_{i=1}^{P} a_i \hat{R}_{XX}[0,i]$$

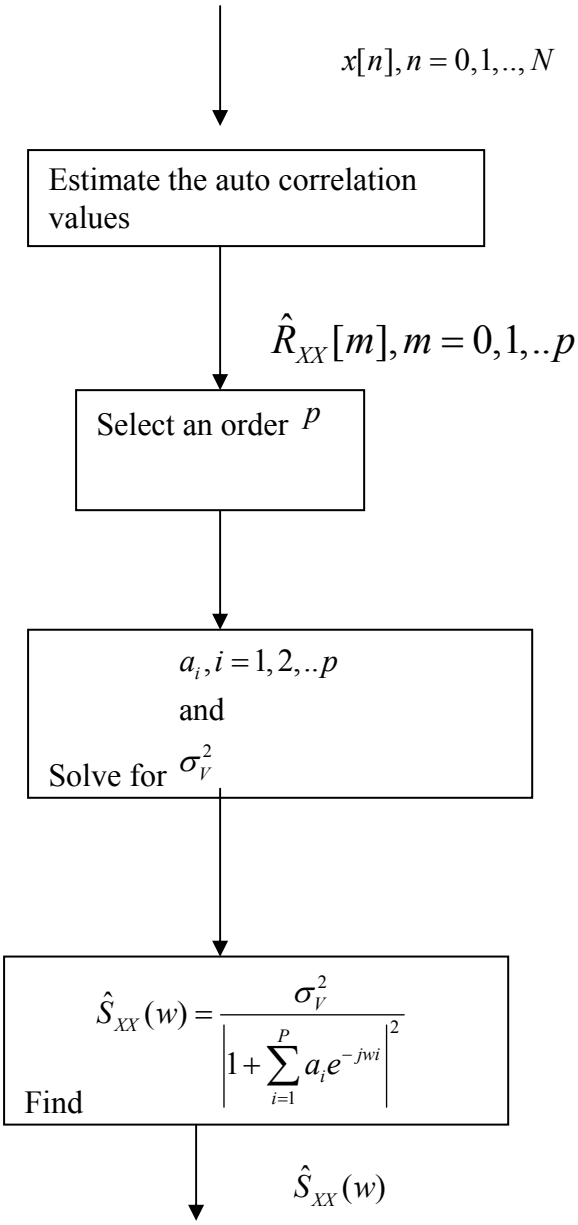
where

$$\hat{R}_{XX}[k,l] = \frac{1}{N-p} \sum_{n=p}^{N} x[n-k]x[n-l]$$

is an estimate for the autocorrelation function.

Note that the autocorrelation matrix in the *Covariance method* is not Toeplitz and cannot be solved by efficient algorithms like the Levinson Durbin recursion.

The flow chart for the AR spectral estimation is given below:




**Some questions :**

*Can an $AR(p)$ process model a band pass signal?*

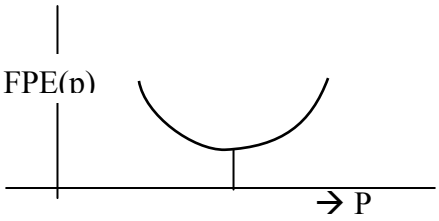
- If we use $AR(1)$ model, it will never be able to model a band-pass process. If one sinusoid is present then $AR(2)$ process will be able to discern it. If there is a strong frequency component $w_0$, an $AR(2)$ process with poles at $re^{\pm jw_0}$ with $r \to 1$ will be able to discriminate the frequency components.

*How do we select the order of the AR(p) process?*

- MSE will give some guidance regarding the selected order is proper or not.

  For spectral estimation, some criterion function with respect to the order parameter $p$ are to be minimized. For example,

- Minimize Forward Prediction Error.




$$FPE\ (p) = \hat{\sigma}_P^2 \frac{N+p+1}{N-p-1}$$

where N = No of data

$\hat{\sigma}_P^2$ = mean square prediction error (variance for non zero mean case)

- Akike Information Criteria

  -Minimize

$$AIC(p) = \ln(\hat{\sigma}_v^2) + \frac{2p}{N}$$

## 9.11 Frequency Estimation of Harmonic signals

A class of spectral estimators is based on eigen decomposition of the autocorrelation matrix of the data vector. Either the eigen values or the eigen vectors can be used. Notable of these algorithms is the MUSIC (Multiple Signal Classification) algorithm. We will discuss this algorithm also.

## 10. Text and Reference


1. **M. Hays, *Statistical Digital Signal Processing and Modelling,* John Willey and Sons, 1996.**

2. M.D. Srinath, P.K. Rajasekaran and R. Viswanathan, S*tatistical Signal Processing with Applications,* PHI, 1996.

3. Simon Haykin, *Adaptive Filter Theory,* Prentice Hall, 1996

4. D.G. Manolakis, V.K. Ingle and S.M. Kogon, *Statistical and Adaptive Signal Processing*, McGraw Hill, 2000

5. S. M. Kay, *Modern Spectral Estimation,* Prentice Hall, 1987

6. S. J. Orfanidis, *Optimum Signal Processing*, Second Edition, MacMillan Publishing, 1989.

*7.* H. Stark and J.W. Woods, *Probability and Random Processes with Applications to Signal Processing*, Prentice Hall 2002.

8. A. Papoulis and S.U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th Edition, McGraw-Hill, 2002