Teoria del Senyal

Class Notes

Autor: JP Alumni

UPC

February 8, 2026

Legend and Visual Guide

This document uses colored boxes to highlight the nature of each concept or result.

Properties

Used for key **definitions**, **relations**, and **mathematical properties**. They contain rigorous results or fundamental equations.

Examples

Used for solved **examples** or **illustrative cases**. They help visualize the meaning of a concept and the application to the problem.

Special Phenomena

Used for important effects such as **aliasing**, **leakage**, or **quantization noise**. These sections highlight critical insights.

Whenever you see a colored box, its color indicates the type of information: property, example, or phenomenon.

Contents

Chapter 5

Fundamentals of estimation theory

5 Fundamentals of estimation theory

5.1 Introduction to estimation theory.

5.1.1 CONCEPT OF ESTIMATOR

Given a set of observations $x[n]$, we want to obtain the estimation of several parameters θ. Hence, we will use two different kind of estimations:

Classic Estimation Theory

Parameter model:
Parameters θ are **deterministic but unknown** (such as temperature, for example). They are fixed constants, not random variables.

Observation model:
The observation vector $x[n]$ is a **single realization** of a random vector with pdf $f_x(x;\theta)$, parameterized by θ.

Goal:
Estimate θ through a deterministic function

$$\hat{\theta}(x) = g(x)$$

Key idea:
Since θ affects the pdf of x, the value of x provides information about the true deterministic parameter.

Notes:
The estimator $\hat{\theta}$ is random (depends on x), but the true parameter is not. **Example:**
For $N = 1$

$$f_\theta(x[0]) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x[0]-\theta)^2}$$

We will then guess the value of θ from the observed x[0]

Bayesian Estimation Theory

Parameter model:
Parameters θ are treated as **random variables**. They have an a-priori distribution $f_\theta(\theta)$ (such as the example: i have to determine the average winter temperature, so I know it will be between 0º and 10º).

Observation model:
Given θ, the observations follow the conditional pdf

$$f_{x|\theta}(x \mid \theta)$$

Goal:
Estimate the realization of θ using both the data and the prior:

$$\hat{\theta}(x) = g(x)$$

Key idea:
Combine prior knowledge with the likelihood of the observations to form the posterior:

$$f_{\theta|x}(\theta|x) \propto f_{x|\theta}(x|\theta)\, f_\theta(\theta)$$

Notes:
Both θ and $\hat{\theta}$ are random variables. Estimation typically uses MAP or MMSE criteria.

Actually, for any distribution, there are multiple possible estimators for the same parameter, so we will need to guess which is the best one. Then, there is where the concepts of the following section play the main role.

5.1.2 Quality criteria

The estimators are functions of the observations x and, since the observations x[n] are random, the estimates are also random variables.
Hence, the quality of an estimator is measured in statistical terms, such as:

Bias

$$b_{\hat{\theta}} = E[\hat{\theta}(x) - \theta] = \mu_{\hat{\theta}} - \theta, \quad \mu_{\hat{\theta}} = E[\hat{\theta}(x)]$$

An estimator is **unbiased** if the bias is 0

Covariance/Variance

$$C_{\hat{\theta}} = E\left[(\hat{\theta}(x) - \mu_{\hat{\theta}})(\hat{\theta}(x) - \mu_{\hat{\theta}})^H\right], \quad \sigma^2_{\hat{\theta}} = \sum_{i=1}^{P} \sigma^2_{\hat{\theta}_i} = Tr(C_{\hat{\theta}})$$

An estimator is **consistent** if the variance tends to 0 when the number of samples tends to infinity

Mean Square Error (MSE)

$$M_{\hat{\theta}} = E\left[(\hat{\theta}(x) - \theta)(\hat{\theta}(x) - \theta)^H\right] = C_{\hat{\theta}} + b_{\hat{\theta}} b^H_{\hat{\theta}}$$

$$m_{\hat{\theta}_i} = E\left[|\hat{\theta}_i(x) - \theta_i|^2\right] = \sigma^2_{\hat{\theta}_i} + |b_{\hat{\theta}_i}|^2$$

$$m_{\hat{\theta}} = \sum_{i=1}^{P} m_{\hat{\theta}_i} = \mathrm{Tr}(C_{\hat{\theta}}) + \sum_{i=1}^{P} |b_{\hat{\theta}_i}|^2$$

The MSE includes **both variance and bias**: a low MSE requires small variance *and* small bias.

Example: Estimation of a Constant Signal in White Noise

Process model:

$$x[n] = A + w[n], \qquad n = 0, \ldots, N-1$$

$$E[w[n]] = 0, \qquad r_w[m] = \sigma_w^2 \delta[m], \qquad E[w[i]w[j]] = \sigma_w^2 \delta[i-j]$$

where:
- A: constant value to estimate.
- $w[n]$: white noise, zero-mean, variance σ_w^2

Estimator 1: Using a single sample

$$\hat{A}_1(x) = x[0]$$

Mean value:

$$E[\hat{A}_1(x)] = E[x[0]] = A$$

$$\Rightarrow \; \hat{A}_1 \text{ is } \textbf{unbiased}$$

Variance:

$$\sigma_{\hat{A}_1}^2 = E\left[(x[0]-A)^2\right] = E[w[0]^2] = \sigma_w^2$$

$$\Rightarrow \; \hat{A}_1 \text{ is } \textbf{not consistent}$$

Estimator 2: Sample mean

$$\hat{A}_2(x) = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

Mean value:

$$E[\hat{A}_2(x)] = \frac{1}{N}\sum_{n=0}^{N-1} E[x[n]] = \frac{1}{N}\sum A = A$$

$$\Rightarrow \; \hat{A}_2 \text{ is } \textbf{unbiased}$$

Variance:

$$\sigma_{\hat{A}_2}^2 = E\left[\left(\frac{1}{N}\sum w[n]\right)^2\right]$$

Since the noise is white:

$$E[w[i]w[j]] = \sigma_w^2 \delta[i-j]$$

$$\sigma_{\hat{A}_2}^2 = \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} E[w[i]w[j]] = \frac{1}{N^2}\sum_{i=0}^{N-1} \sigma_w^2 = \frac{\sigma_w^2}{N}$$

$$\Rightarrow \; \hat{A}_2 \text{ is } \textbf{consistent} \text{ and has lower variance than } \hat{A}_1$$

5.2 Classic estimation theory

5.2.1 Minimum variance unbiased estimation (MVUE) and efficient estimation (Cramer-Rao bound).

[1] However, in some cases the MSE estimator could not be useful, for example, if we take the previous process model with the estimator $\hat{A}(x) = \frac{a}{N}\sum_{n=0}^{N-1} x[n]$ we will see:

$$MSE(\hat{A}) = \frac{a^2\sigma^2}{N} + (a-1)^2 A^2$$

Hence, the minimum MSE is found by performing:

$$\left.\frac{\partial MSE(\hat{A})}{\partial a}\right|_{a=a_{optimal}} = 0 \implies a_{optimal} = \frac{A^2}{A^2 + \sigma^2/N}$$

which means the MSE function will depend upon the unknown A (this will lead, for large values of A, to have a "false" MSE value).
Then, we define the following estimator:

Minimum Variance Unbiased Estimator (MVUE)

An estimator is MVUE if it satisfies:

$$b_{\hat{\theta}} = 0 \qquad \text{and} \qquad \sigma^2_{\hat{\theta}} = \min \ \text{among all unbiased estimators.}$$

When the bias is zero, the MSE reduces to the variance:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta})$$

The MVUE **may or may not exist**. If it exists, it is the unbiased estimator that achieves the minimum possible variance.

We will have to use something that quantifies how "accurate" our estimation can be. This leads to the concept of *Sharpness*, defined as:

$$\boxed{\text{Sharpness} \ = \ -E\left[\frac{\partial^2}{\partial\theta^2}\ln f_\theta(x)\right]}$$

[1] an estimator $\hat{A}$ is found by $\partial \ln f_A(x)/\partial A|_{\hat{A}=A}$ and isolating $\hat{A}$

Example: Estimation accuracy

We consider again the model:

$$x[0] = A + w[0], \qquad w[0] \sim \mathcal{N}(0, \sigma_w^2)$$

Thus the likelihood is:

$$f_A(x[0]) = \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left[-\frac{(x[0]-A)^2}{2\sigma_w^2}\right]$$

Step 1: Compute the log-likelihood

$$\ln f_A(x[0]) = -\frac{1}{2}\ln(2\pi\sigma_w^2) - \frac{(x[0]-A)^2}{2\sigma_w^2}$$

Step 2: First derivative w.r.t. A

$$\frac{\partial}{\partial A}\ln f_A(x[0]) = \frac{x[0]-A}{\sigma_w^2}$$

Step 3: Second derivative

$$\frac{\partial^2}{\partial A^2}\ln f_A(x[0]) = -\frac{1}{\sigma_w^2}$$

Step 4: Sharpness

$$\boxed{-E\left[\frac{\partial^2}{\partial A^2}\ln f_A(x[0])\right] = \frac{1}{\sigma_w^2}}$$

Interpretation: A larger sharpness $(1/\sigma_w^2)$ means a "sharper" likelihood function around the true value of A, and therefore the estimation of A is more accurate.
Connection with variance: For this particular estimator,

$$\hat{A}_1(x) = x[0], \qquad \mathrm{Var}(\hat{A}_1) = \sigma_w^2$$

The inverse of the sharpness matches the variance of the estimator:

$$\mathrm{Var}(\hat{A}_1) = \frac{1}{\text{Sharpness}}$$

This idea generalizes into the **Cramér–Rao Lower Bound (CRLB)**, which states that the variance of **any** unbiased estimator is bounded below by the inverse of the sharpness.

5.2.2 Cramér–Rao Lower Bound (CRLB) for multiple parameters

For a vector of parameters

$$\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^T,$$

the CRLB is expressed using the *Fisher Information Matrix* (FIM).

Fisher Information Matrix

$$\boxed{I_{ij}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln f_{\boldsymbol{\theta}}(x)\right]}$$

Matrix CRLB For any unbiased estimator vector $\hat{\boldsymbol{\theta}}$:

$$\boxed{\mathrm{Cov}(\hat{\boldsymbol{\theta}}) \succeq \mathbf{I}^{-1}(\boldsymbol{\theta})}$$

meaning the covariance matrix must be at least as large (in the PSD sense) as the inverse FIM.

Example: estimating A and B in a linear model

Signal model:

$$x[n] = A + Bn + w[n], \qquad w[n] \sim \mathcal{N}(0, \sigma^2), \quad n = 0, \ldots, N-1.$$

The Fisher Information Matrix becomes:

$$\mathbf{I} = \frac{1}{\sigma^2}\begin{bmatrix} N & \sum n \\ \sum n & \sum n^2 \end{bmatrix}.$$

The CRLB gives:

$$\mathrm{Var}(\hat{A}) \geq [\mathbf{I}^{-1}]_{11}, \qquad \mathrm{Var}(\hat{B}) \geq [\mathbf{I}^{-1}]_{22}.$$

Interpretation: The parameters A and B are **coupled** through the off-diagonal terms. This means estimating one affects how well we can estimate the other.

5.2.3 Cramér–Rao Lower Bound (CRLB) for multiple parameters

When estimating a vector of parameters

$$\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^T,$$

the estimation accuracy is fundamentally limited by the **Fisher Information Matrix (FIM)**.

Fisher Information Matrix (FIM)

$$\boxed{\mathbf{J}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T} \ln f_{\boldsymbol{\theta}}(x)\right] = E\big[(\nabla_\theta \ln f_\theta(x))(\nabla_\theta \ln f_\theta(x))^T\big]}$$

It is a measure of how much information the data carries about the parameters.

Matrix CRLB For any unbiased estimator vector $\hat{\boldsymbol{\theta}}$,

$$\boxed{\mathrm{Cov}(\hat{\boldsymbol{\theta}}) \succeq \mathbf{J}^{-1}(\boldsymbol{\theta})}$$

which means that the covariance matrix must be at least as large (in the positive semidefinite sense) as the inverse FIM.

The more information (larger entries in $\mathbf{J}$), the **smaller the achievable variance**.

Example: explicit formula for 2 parameters For

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \qquad \mathbf{J} = \begin{bmatrix} J_{11} & J_{12} \\ J_{12} & J_{22} \end{bmatrix},$$

the inverse is:

$$\mathbf{J}^{-1} = \frac{1}{J_{11}J_{22} - J_{12}^2}\begin{bmatrix} J_{22} & -J_{12} \\ -J_{12} & J_{11} \end{bmatrix}.$$

Thus,

$$\boxed{\mathrm{Var}(\hat{\theta}_1) \;\geq\; \frac{J_{22}}{J_{11}J_{22} - J_{12}^2}} \qquad \boxed{\mathrm{Var}(\hat{\theta}_2) \;\geq\; \frac{J_{11}}{J_{11}J_{22} - J_{12}^2}}$$

Uncoupled case:

$$J_{12} = 0 \quad \Longrightarrow \quad \mathrm{Var}(\hat{\theta}_1) = \frac{1}{J_{11}} \text{ does not depend on whether } \theta_2 \text{ is known.}$$

The parameters are **uncoupled** and can be estimated independently.

Example: independent parameters

Suppose the likelihood factorizes:

$$f_{\theta_1,\theta_2}(x) = f_{\theta_1}(x_1)\, f_{\theta_2}(x_2).$$

Then the cross-terms vanish:

$$J_{12} = E\left[\frac{\partial}{\partial\theta_1} \ln f_{\theta_1}(x_1) \frac{\partial}{\partial\theta_2} \ln f_{\theta_2}(x_2)\right] = 0.$$

As a result:

$$\mathbf{J} = \begin{bmatrix} J_{11} & 0 \\ 0 & J_{22} \end{bmatrix}, \qquad \mathbf{J}^{-1} = \begin{bmatrix} 1/J_{11} & 0 \\ 0 & 1/J_{22} \end{bmatrix}.$$

Each parameter reaches its CRLB independently of the other. This happens, for example, when estimating:

$$x[n] = A + w[n], \qquad y[n] = B + v[n]$$

with independent noise sequences $w[n]$ and $v[n]$.

5.2.4 Efficient estimator

Efficient estimator (scalar case)

An estimator is **efficient** if it is unbiased and achieves the CRLB (then this is the optimal estimator):

$$\mathrm{Var}(\hat{\theta}) = \mathrm{CRLB}(\theta)$$

Which we will characterize as:

$$a(x) = k \cdot b(x)$$

From the Cauchy–Schwarz equality condition ($E^2[a(x)b(x)] = E[a^2(x)]E[b^2(x)]$), efficiency requires that:

$$\hat{\theta}(x) - \theta = \mathrm{CRLB}(\theta)\, \frac{\partial \ln f_\theta(x)}{\partial\theta}.$$

Such estimators use the available data "in the most efficient way".
If they exist, efficient estimators are also **MVU estimators**.

Efficient estimator (vector case)

For a parameter vector $\theta \in \mathbb{R}^P$, efficiency requires:

$$\hat{\theta}(x) - \theta = J^{-1}(\theta)\, \nabla_\theta \ln f_\theta(x),$$

where $J(\theta)$ is the **Fisher information matrix**. If the equality holds, then:

$$\text{Cov}(\hat{\theta}) = J^{-1}(\theta).$$

Efficient estimators achieve the CRLB **for all parameters simultaneously**.

For a vector parameter $\theta \in \mathbb{R}^p$, the CRLB generalizes to a matrix form. Define the **Fisher Information Matrix (FIM)**:

$$\mathbf{J}(\theta) = E\Big[(\nabla_\theta \ln f_\theta(x))\, (\nabla_\theta \ln f_\theta(x))^T \Big].$$

The covariance matrix of any unbiased estimator $\hat{\theta}$ satisfies:

$$\boxed{\text{Cov}(\hat{\theta}) \ \succeq\ \mathbf{J}^{-1}(\theta)}$$

Meaning that $\text{Cov}(\hat{\theta}) - \mathbf{J}^{-1}(\theta)$ must be positive semidefinite.

Example 2: Estimating the phase φ of a sinusoid

Signal model:

$$x[n] = A\cos(\Omega n + \varphi) + w[n], \qquad w[n] \sim \mathcal{N}(0, \sigma^2)$$

where all parameters are known, except phase (φ).
Hence, we want to find the phase estimator:

$$f_\varphi(x) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - A\cos(\Omega n + \varphi))^2}$$

One would find applying the previous set of formula:

$$\sigma_\varphi^2 \geq CRLB = \frac{2\sigma^2}{NA^2}$$

The Fisher information:

$$\mathcal{I}(\varphi) = \frac{2A^2}{\sigma^2} \sum_{n=0}^{N-1} \sin^2(\Omega n + \varphi)$$

Thus the CRLB:

$$\boxed{\text{Var}(\hat{\varphi}) \ \geq\ \frac{\sigma^2}{2A^2 \sum_{n=0}^{N-1} \sin^2(\Omega n + \varphi)}}$$

In this case no estimator achieves the bound ⇒ no efficient estimator exists.
However, there's an MVUE estimator still existing and that is what we will have to find:

Example 3: Estimating the frequency F_0 of a sinusoid

Signal:

$$x[n] = A\cos(2\pi F_0 n + \varphi) + w[n], \qquad w[n] \sim \mathcal{N}(0, \sigma^2)$$

where the unknown here is the frequency (F_0).
The Fisher information for F_0:

$$\mathcal{I}(F_0) = \frac{A^2}{\sigma^2} \sum_{n=0}^{N-1} (2\pi n)^2 \sin^2(2\pi F_0 n + \varphi)$$

Hence the CRLB:

$$\boxed{\operatorname{Var}(\hat{F}_0) \;\geq\; \frac{\sigma^2}{A^2 \sum_{n=0}^{N-1} (2\pi n)^2 \sin^2(2\pi F_0 n + \varphi)}}$$

Frequency estimation is intrinsically harder: the denominator grows only quadratically with n.

Example 4: Joint estimation of amplitude, frequency and phase

Model:

$$x[n] = A\cos(2\pi F_0 n + \varphi) + w[n], \qquad \theta = \begin{bmatrix} A \\ F_0 \\ \varphi \end{bmatrix}$$

where here the unknowns are A, F_0 and φ.
The Fisher Information Matrix (FIM):

$$\mathbf{J}(\theta) = \frac{1}{\sigma^2} \begin{bmatrix} \frac{N}{2} & 0 & 0 \\ 0 & 2A^2\pi^2 \sum_0^{N-1} n^2 & A^2\pi \sum_0^{N-1} n \\ 0 & A^2\pi \sum_0^{N-1} n & \frac{NA^2}{2} \end{bmatrix}$$

('check class notes for mathematic reasoning').

Its inverse yields:

$$\operatorname{Var}(\hat{A}) \geq \frac{2\sigma^2}{N}, \qquad \operatorname{Var}(\hat{F}_0) \geq \frac{6\sigma^2}{(2\pi)^2 A^2 \, N(N^2-1)},$$

$$\operatorname{Var}(\hat{\varphi}) \geq \frac{(2N-1)\sigma^2}{A^2 N(N+1)}.$$

This generalizes the previous examples and shows interaction between parameters.

Example 5: Linear model $x = H\theta + w$

Observation model:

$$x = H\theta + w, \qquad w \sim \mathcal{N}(0, \sigma^2 I),$$

where H is a known $(N \times p)$ matrix with $N > p$ and $\mathrm{rank}(H) = p$, and θ is a $(p \times 1)$ parameter vector.

Likelihood:

$$\ln f(x;\theta) = -\frac{1}{2\sigma^2}\|x - H\theta\|^2 + c.$$

Score:

$$\nabla_\theta \ln f(x;\theta) = \frac{1}{\sigma^2} H^T (x - H\theta).$$

Fisher Information Matrix:

$$\mathbf{J}(\theta) = \frac{1}{\sigma^2} H^T H.$$

CRLB:

$$\mathrm{Cov}(\hat{\theta}) \;\geq\; \sigma^2 (H^T H)^{-1}.$$

Maximum Likelihood estimator:

$$\hat{\theta}_{\mathrm{ML}} = (H^T H)^{-1} H^T x.$$

This estimator attains the CRLB, so it is efficient and MVU.

5.2.5 CRLB for complex parameters

If parameters are complex-valued, the FIM extends to:

$$\mathbf{J}(\theta) = E\big[(\nabla_\theta \ln f)(\nabla_\theta \ln f)^H\big]$$

The CRLB becomes:

$$\boxed{\mathrm{Cov}(\hat{\theta}) \;\succeq\; \mathbf{J}^{-1}(\theta)}$$

Efficient estimators exist when:

$$\nabla_\theta \ln f = \mathbf{J}(\theta)\,(\hat{\theta} - \theta)$$

This ensures that the estimator reaches the CRLB simultaneously for all components.

5.2.6 Maximum likelihood (ML) estimation.

An alternative to the MVU estimator that is asymptotically efficient:

$$E[\hat{\theta}(x)] = 0 \qquad var[\hat{\theta}(x)] = CRLB \qquad N \to \infty$$

ML estimator

$$\hat{\theta}_{ML}(\mathbf{x}) = \arg[\max_\theta f_\theta(\mathbf{x})]$$

which consists essentially in finding the θ value that maximizes the pdf (or solving the optimization problem $\frac{\partial \ln f_\theta}{\partial \theta}|_{\theta=\hat{\theta}} = 0$).

Example 1: DC level in White Gaussian Noise

Let $A > 0$ be the unknown, where:

$$x[n] = A + w[n], \qquad w[n] \in N(0, A)$$

Then, since we already know:

$$f_A(x) = \frac{1}{(2\pi A)^{N/2}} \exp\left[-\frac{1}{2A}\sum_{n=0}^{N-1}(x[n]-A)^2\right]$$

Searching for an **efficient estimator**, we get, from the CRLB theorem, but we reach that an efficient estimator does not exist (since computing the log-derivative, we reach to a quadratic expression for A which disagrees with the linear expected behaviour). No matter this, we can still find the CRLB:

$$\frac{\partial \ln f_A(x)}{\partial A} = -\frac{N}{2A} + \frac{1}{A}\sum_{n=0}^{N-1}(x[n]-A) + \frac{1}{2A^2}\sum_{n=0}^{N-1}(x[n]-A)^2$$

$$\Longrightarrow \; \text{var}(\hat{A}(x)) \geq \frac{A^2}{N(A+\frac{1}{2})} = CRLB$$

Since an efficient estimator attaining the CRLB does not exist, we will look for the **ML-estimator**

$$\frac{\partial \ln f_A(x)}{\partial A} = 0$$

$$\Longrightarrow \; \hat{A}_{ML}(x) = -\frac{1}{2} \pm \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}}$$

By the law of large numbers (when $N \to \infty$), we get the expected results:

$$E[\hat{A}_{ML}(x)] \to A$$

$$\text{var}[\hat{A}_{ML}(x)] \to \frac{A^2}{N(A+\frac{1}{2})}$$

which implies the ML-estimator is asymptotically efficient.

Example 2: DC level in White Gaussian Noise

We consider the observation model

$$x[n] = A + w[n], \qquad n = 0, \dots, N-1,$$

where the parameter of interest is the DC level $A \in \mathbb{R}$, and

$$w[n] \sim \mathcal{N}(0, \sigma^2), \qquad \text{i.i.d.}$$

Therefore,

$$x[n] \sim \mathcal{N}(A, \sigma^2).$$

The joint pdf of the observation vector $x = (x[0], \dots, x[N-1])$ is

$$f_A(x) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \right).$$

Score function and Fisher Information

To study the existence of an **efficient estimator**, we compute the score function:

$$\frac{\partial}{\partial A} \ln f_A(x) = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A).$$

The Fisher Information is

$$J(A) = \mathbb{E}\left[\left(\frac{\partial}{\partial A} \ln f_A(x) \right)^2 \right] = \frac{1}{\sigma^2} N.$$

Thus the Cramér–Rao lower bound (CRLB) for *any unbiased estimator* $\hat{A}(x)$ is:

$$\boxed{\operatorname{var}(\hat{A}(x)) \ \geq \ \frac{\sigma^2}{N}}.$$

Observe that the score function is of the form

$$\frac{\partial}{\partial A} \ln f_A(x) = J(A)\,(\hat{A}(x) - A),$$

with

$$\hat{A}(x) = \frac{1}{N} \sum_{n=0}^{N-1} x[n],$$

which already reveals that an efficient estimator *does* exist.

Maximum Likelihood Estimator

The ML estimator is obtained from the likelihood equation

$$\frac{\partial}{\partial A} \ln f_A(x) = 0 \quad \Longrightarrow \quad \sum_{n=0}^{N-1} (x[n] - A) = 0.$$

Solving for A,

$$\boxed{\hat{A}_{\mathrm{ML}}(x) = \frac{1}{N} \sum_{n=0}^{N-1} x[n]}.$$

5.2.7 Properties of the MLE

Properties of MLE

1. Asymptotic normality. If the likelihood function satisfies the usual regularity conditions (derivable log–likelihood and nonzero Fisher information), then the maximum likelihood estimator of a scalar parameter θ is asymptotically Gaussian:

$$\hat{\theta}_{\mathrm{ML}}(x) \xrightarrow[N\to\infty]{} \mathcal{N}\big(\theta,\ J^{-1}(\theta)\big)\,,$$

where $J(\theta)$ is the Fisher information evaluated at the true value of θ. Thus, for large data records, the MLE behaves like an unbiased estimator whose variance attains the Cramér–Rao lower bound.

Vector case. For a parameter vector $\boldsymbol{\theta}$,

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(x) \xrightarrow[N\to\infty]{} \mathcal{N}\big(\boldsymbol{\theta},\ J^{-1}(\boldsymbol{\theta})\big)\,,$$

where $J(\boldsymbol{\theta})$ is the Fisher information matrix. Again, the asymptotic covariance equals the CRLB.

2. If an efficient estimator exists, the MLE will produce it. If an unbiased estimator achieves the Cramér–Rao lower bound, then this estimator must coincide with the MLE. Therefore, whenever an efficient estimator exists for a given statistical model, the MLE is that estimator.

Example (scalar parameter). Consider

$$x[n] = A + w[n], \qquad w[n] \sim \mathcal{N}(0, \sigma^2).$$

The log–likelihood is maximised by

$$\hat{A}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=0}^{N-1} x[n].$$

The Fisher information is $J(A) = N/\sigma^2$ and the CRLB is $\mathrm{var}(\hat{A}) \geq \sigma^2/N$. Since

$$\mathrm{var}(\hat{A}_{\mathrm{ML}}) = \frac{\sigma^2}{N},$$

the MLE attains the CRLB (it is efficient) and satisfies the asymptotic Gaussian property.

Example 3: MLE of the Sinusoidal Phase

We observe a sinusoid corrupted with white Gaussian noise:

$$x[n] = A\cos(\Omega_0 n + \phi) + w[n], \qquad n = 0, \dots, N-1,$$

where the amplitude A and the frequency Ω_0 are **known**, and

$$w[n] \sim \mathcal{N}(0, \sigma^2).$$

Likelihood function. Since the noise is Gaussian and independent,

$$f_\phi(x) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} [x[n] - A\cos(\Omega_0 n + \phi)]^2 \right).$$

Maximizing $f_\phi(x)$ is equivalent to *minimizing the exponent*:

$$L(\phi) = \sum_{n=0}^{N-1} [x[n] - A\cos(\Omega_0 n + \phi)]^2 .$$

Derivative of the log-likelihood. We compute

$$\frac{\partial}{\partial\phi} \ln f_\phi(x) = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A\cos(\Omega_0 n + \phi))\, A\sin(\Omega_0 n + \phi)$$

Setting the derivative to zero gives the ML estimator:

$$\sum_{n=0}^{N-1} x[n]\sin(\Omega_0 n + \phi) = \sum_{n=0}^{N-1} A\cos(\Omega_0 n + \phi)\sin(\Omega_0 n + \phi)$$

Using the identity

$$\sin(\alpha + \phi) = \sin\alpha\cos\phi + \cos\alpha\sin\phi,$$

and the approximation (valid if Ω_0 is not near 0 or π)

$$\frac{1}{N} \sum_{n=0}^{N-1} \sin(2\Omega_0 n + 2\phi) \approx 0,$$

we arrive at the simplified condition:

$$\sum_{n=0}^{N-1} x[n]\sin(\Omega_0 n) = \hat{\phi} \sum_{n=0}^{N-1} x[n]\cos(\Omega_0 n).$$

ML Estimator. Thus,

$$\boxed{\hat{\phi}_{ML}(x) = -\arctan\left(\frac{\sum_{n=0}^{N-1} x[n]\sin(\Omega_0 n)}{\sum_{n=0}^{N-1} x[n]\cos(\Omega_0 n)} \right)}$$

Fisher Information and CRLB.

$$J(\phi) = \frac{NA^2}{2\sigma^2} \qquad \Rightarrow \qquad \operatorname{var}(\hat{\phi}) \geq \frac{1}{J(\phi)} = \frac{2\sigma^2}{NA^2} = \frac{1}{N\eta},$$

5.3 Bayesian estimation theory

5.3.1 Parameter characterization: a-priori distribution

Imagine that we want to estimate a DC component in a white Gaussian noise, we will proceed assuming our data satisfies:

$$\vec{x} = \vec{1}\theta + \vec{w}$$

Then, the ML estimator reads:

$$\hat{\theta}_{ML}(\vec{x}) = \arg\max_{\theta} f_{\theta}(x) = \frac{1}{N}\sum_{n=0}^{N-1} x[n] = \frac{1}{N}\vec{1}^T\vec{x}$$

However, let's assume now that we know, a priori, that $-A < \theta < A$. This will affect our estimator in such a way that:

$$\hat{\theta}(x) = \begin{cases} -A, & \hat{\theta}_{ML} < -A \\ \hat{\theta}_{ML}, & -A < \hat{\theta}_{ML} < A \\ +A, & \hat{\theta}_{ML} > +A \end{cases}$$

but makes the estimator biased.
Actually there is a more accurate way of determining such estimator for the same $a-priori$ information. For this, we use the Bayes Theorem:

Bayes Theorem

This is what we will call "a-posterior pdf", and gives us an idea of how informative is the "a-priori" information (as this function flattens, the prior information will be lower)

$$f_{\theta|x}(\theta|x) = \frac{f_{x,\theta}(x,\theta)}{f_x(x)} = \frac{f_{x|\theta}(x|\theta)f_{\theta}(\theta)}{\int f_{x|\theta}(x|\theta)f_{\theta}(\theta)d\theta}$$

5.3.2 Minimum Mean Square Error (MMSE) and Maximum A Posteriori (MAP) estimators.

All the following estimators will be based on the a-posteriori pdf found at the Bayesian Theorem:

Minimum Mean Square Error (MMSE) estimator

$$\hat{\theta}_{MMSE}(x) = \int \theta f_{\theta|x}(\theta|x)d\theta = E_{\theta|x}[\theta|x]$$

Maximum A Posteriori (MAP) estimator

$$\hat{\theta}_{MAP}(x) = \arg\max_{\theta} f_{\theta|x}(\theta|x) = \arg\max_{\theta} f_{x|\theta}(x|\theta)f_{\theta}(\theta) = \arg\max_{\theta} f_{\theta}(x)f_{\theta}(\theta)$$

Chapter 6

Spectral estimation

6 Spectral estimation

6.1 Introduction to spectral estimation

Actually, in a experiment we only have N samples of a single realization of the process. Then, in order to determine the PSD ($S_X(e^{j\Omega}) = \sum_{m=-\infty}^{+\infty} r_X[m]e^{-j\Omega m}$) that would imply making a sum of infinite terms, which is impossible in real data analysis. Hence, depending on the data we have, we could find two approaches:

- **Non-parametric:** where we do not have any a-priori model.
- **Parametric:** where we have the actual model of the process and estimate its parameters.

6.2 Non-parametric spectral estimation.

6.2.1 Periodogram: biased estimate of the auto-correlation

Let's window the sequence as we did many chapters before:

$$x_\nu[n] = x[n]\nu[n], \qquad \textit{usually} \quad \nu[n] = p_N[n]$$

Then the **bias estimation of the auto-correlation**:

$$\hat{r}_x[m] = \frac{1}{N}\, x_\nu[m] * x_\nu^*[-m] = \frac{1}{N}\sum_{k=-\infty}^{\infty} x_\nu[k]\, x_\nu^*(-(m-k))$$

$$= \frac{1}{N}\sum_{k=-\infty}^{\infty} x[k]\, x^*[k-m]\, \nu[k]\, \nu[k-m]$$

$$= \frac{1}{N}\sum_{n=-\infty}^{\infty} x[n+m]\, x^*[n]\, \nu[n+m]\, \nu[n]$$

with **mean value**:

$$E[\hat{r}_x[m]] = E\left[\frac{1}{N}\sum_{n=-\infty}^{\infty} x[n+m]\, x^*[n]\, \nu[n+m]\, \nu[n]\right] = \frac{1}{N}\sum_{n=-\infty}^{\infty} E[x[n+m]\, x^*[n]]\, \nu[n+m]\, \nu[n]$$

$$= \frac{1}{N}\sum_{n=-\infty}^{\infty} r_x[m]\, \nu[n+m]\, \nu[n] = r_x[m]\left(\frac{1}{N}\sum_{n=-\infty}^{\infty} \nu[n+m]\, \nu[n]\right) = r_x[m]\left(\frac{1}{N}\, \nu[m] * \nu[-m]\right)$$

where the convolution term is nothing but applying a **triangular windowing** to the actual auto-correlation:

$$w[n] \doteq \frac{1}{N}\, \nu[m] * \nu[-m]$$

Then, after this, we define the **Periodogram** as an estimation of the PSD, and the Fourier Transform of $\hat{r}_X[m]$:

$$\boxed{\hat{S}_P(e^{j\Omega}) = \mathcal{F}\{\hat{r}_X[m]\} = \frac{1}{N}\left|\sum_{n=0}^{N-1} x[n]e^{-j\Omega n}\right|^2}$$

with mean value:

$$\mathbb{E}[\hat{S}_P(e^{j\Omega})] = S_X(e^{j\Omega}) \circledast_{2\pi} W(e^{j\Omega})$$

Important results

Making the simulation, one would notice in the **bias**:
For a low number of samples N, the lobes are very wide and there is a loss of frequency resolution (leakage). The bias decreases as N increases

The periodogram is **asymptotically unbiased**:

$$\begin{cases} \lim_{N\to\infty} W(e^{j\Omega}) = 2\pi\delta(\Omega), & |\Omega| < \pi \\ \lim_{N\to\infty} \mathbb{E}[\hat{S}_P(e^{j\Omega})] = S_X(e^{j\Omega}) \end{cases}$$

In a similar way, the **variance** of the Periodogram is approximated by:

$$var[\hat{S}_P(e^{j\Omega})] = (S_X(e^{j\Omega}))^2$$

which **does not decrease** when increasing N, then **the periodogram is unconsistent**.

The **covariance** of the periodogram:

$$\operatorname{cov}\Big[\hat{S}_p(e^{j\Omega_1}),\, \hat{S}_p(e^{j\Omega_2})\Big] \simeq 0, \qquad |\Omega_2 - \Omega_1| \gg \frac{2\pi}{N}$$

$$E\Big[\hat{S}_p(e^{j\Omega_1})\, \hat{S}_p(e^{j\Omega_2})\Big] \simeq E\Big[\hat{S}_p(e^{j\Omega_1})\Big]\ E\Big[\hat{S}_p(e^{j\Omega_2})\Big].$$

Interpretation. For frequencies separated by more than $2\pi/N$, the periodogram values become essentially uncorrelated. Hence, their joint expectation factorizes. This reflects the limited resolution of the periodogram: estimates at well-separated frequencies behave like independent noisy measurements.

With all this results, we can eventully estimate **the power** based on the periodogram, since:

$$\hat{P}_X = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{S}_p(e^{j\Omega})d\Omega = \hat{r}_X[0] = \frac{1}{N}\sum_{n=0}^{N-1}|x[n]|^2$$

with mean (since it is unbiased):

$$\mathbb{E}[\hat{P}_X] = P_X$$

and variance:

$$var[\hat{P}_X] \xrightarrow{N\to\infty} 0$$

6.2.2 Estimator of the PSD based on the unbiased estimation of the auto-correlation

We now define the **unbiased estimate of the auto-correlation:**

$$\breve{r}_x[m] = \begin{cases} \dfrac{1}{N-m}\displaystyle\sum_{n=0}^{N-m-1} x[n+m]\, x^*[n], & 0 \le m \le N-1, \\ \dfrac{1}{N-|m|}\displaystyle\sum_{n=0}^{N-|m|-1} x[n]\, x^*[n+|m|] = \dfrac{N}{N-|m|}\hat{r}_X[m], & -(N-1) \le m < 0, \\ 0, & \text{otherwise.} \end{cases}$$

with **mean**:

$$\mathbb{E}[\bar{r}_X[m]] = r_X[m]w[m]$$

being w[m] the equivalent rectangular windowing applied to the auto-correlation. Then, applying the Fourier transform, we obtain the spectrum:

$$\breve{S}_X(e^{j\Omega}) = \mathcal{F}\{\breve{r}_X[m]\}$$

and the mean of this power:

$$\mathbb{E}\{\breve{S}_X(e^{j\Omega})\} = S_X(e^{j\Omega}) \textcircled{2\pi} W(e^{j\Omega}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_X(e^{j(\Omega-\theta)})W(e^{j\theta})d\theta$$

since this value is **not necessarily positive**, the estimator is **not used**.

6.2.3 Modified periodogram

Now consider applying a windows (not necessary rectangular) to the set of N sampling data $x_\nu[n] = x[n]\nu[n]$:

- The **modified autocorrelation:**

$$\hat{r}_{x_\nu}[m] = \frac{1}{N}\, x_\nu[m] * x_\nu^*[-m] = \frac{1}{N}\sum_{n=-\infty}^{\infty} x[n+m]\, x^*[n]\, \nu[n+m]\, \nu[n]$$

- Its mean

$$E[\hat{r}_{x_\nu}[m]] = r_x[m] * \left(\frac{1}{N}\,\nu[m] * \nu^*[-m]\right) = r_x[m] * w[m]$$

  So then, the resulting periodogram:

- **Modified Periodogram**

$$\hat{S}_{MP}(e^{j\Omega}) = \mathcal{F}[\hat{r}_{x,v}[m]] = \frac{1}{N}\,\mathcal{F}[x_v[m] * x_v^*[-m]] = \frac{1}{N}\left|X_v(e^{j\Omega})\right|^2 = \frac{1}{N}\left|\sum_{n=0}^{N-1} x[n]\, v[n]\, e^{-j\Omega n}\right|^2$$

Properties of the Modified Periodogram

- **Mean value:**

$$E\left[\hat{S}_{MP}(e^{j\Omega})\right] = S_X(e^{j\Omega}) \textcircled{2\pi} \frac{1}{N}\,|V(e^{j\Omega})|^2 = S_X(e^{j\Omega}) \textcircled{2\pi}\, W(e^{j\Omega}) \neq S_X(e^{j\Omega})$$

- **Asymptotic unbiasedness:** The modified periodogram becomes unbiased as $N \to \infty$ if the window satisfies

$$\sum_{n=0}^{N-1} |v[n]|^2 = N.$$

- **Inconsistency:** The variance does *not* decrease with N:

$$\operatorname{var}\left[\hat{S}_{MP}(e^{j\Omega})\right] \simeq \left(S_X(e^{j\Omega})\right)^2.$$

Then, the estimation of the power gives us the following result:

$$\hat{P}_X = \hat{r}_{X,\nu}[0] = \frac{1}{N}\sum_{n=0}^{N-1} |\nu[n]|^2 |x[n]|^2$$

with mean

$$\mathbb{E}[\hat{P}_X] = P_X \frac{1}{N} \sum_{n=0}^{N-1} |\nu[n]|^2$$

which is unbiased if the window is normalised: $\sum_{n=0}^{N-1} |\nu[n]|^2 = N$:

Window $v[n]$	**Secondary-to-main lobe ratio (dB)**	**Main-lobe width at** -3 **dB**
Rectangular	-13	$\Delta\Omega = 2\pi \frac{0.89}{N}$
Hanning	-32	$\Delta\Omega = 2\pi \frac{1.44}{N}$
Hamming	-43	$\Delta\Omega = 2\pi \frac{1.30}{N}$
Bartlett	-27	$\Delta\Omega = 2\pi \frac{1.28}{N}$

Table 1: Comparison of spectral window properties: side-lobe levels and main-lobe widths.

6.2.4 Smoothing the periodogram through windowing (Blackman-Tukey method).

Taking the periodogram:

$$\hat{S}_P(e^{j\Omega}) = \mathcal{F}\{\hat{r}_X[m]\}$$

which we have seen in 6.2.1 it is **asymptotically unbiased** and **inconsistent**.
Then, we will use an important method such as the Blackman-Tukey's method:

Data samples	**Biased ACF estimate**	**Windowing**	**Fourier Transform**
$x[0], x[1], \ldots, x[N-1]$	$\hat{r}_x[m],\ \|m\| \leq N-1$	$\hat{r}_x[m]\, w_a[m],\ \|m\| \leq L-1$	$\hat{S}_{BT}(e^{j\Omega})$

Table 2: Blackman-Tukey (BT) spectral estimator pipeline.

Then, the periodogram provided by the Blackman-Tukey's method:

$$\hat{S}_{BT}\left(e^{j\Omega}\right) = \mathcal{F}[\hat{r}_x[m]\, w_a[m]] = \hat{S}_p\left(e^{j\Omega}\right) \textcircled{\scriptsize 2π} W_a\left(e^{j\Omega}\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{S}_p\left(e^{j(\Omega-\theta)}\right) W_a\left(e^{j\theta}\right) d\theta.$$

but, some considerations have to be taken into account, since:

- The window has to fulfill $w_a[0] = 1$ for the estimation to be **unbiased**, since:

  $$\mathbb{E}[\hat{P}_X] = \mathbb{E}[\hat{r}_X[0]]w_a[0] = P_X \cdot w_a[0]$$

- The window has to be symmetric, real and with real non-negative Fourier transform, since the power has to be positive (by definition).

- If the length of the window ($[-L+1, L-1]$) is much lower than N, then:

  $$\mathbb{E}[\hat{r}_X[m]w_a[m]] \approx r_X[m]w_a[m]$$

  and

  $$\mathbb{E}[\hat{S}_{BT}(e^{j\Omega})] = \cdots = S_X(e^{j\Omega}) \textcircled{\scriptsize 2π} W_a(e^{j\Omega})$$

  This shows that the method introduces spectral smoothing, whose effect is entirely governed by the choice of $w_a[m]$.

- The estimator is asymptotically unbiased if: $N \to \infty$, $L \to \infty$ and $w_a[0] = 1$
- The variance of the Blackman-Tukey's estimator is approximated by the following expression:

  $$var[\hat{S}_{BT}(e^{j\Omega})] \approx \frac{E_{W_a}}{N} S_X^2(e^{j\Omega})$$

  being E_{W_a} the energy of the window:

  $$E_{W_a} = \sum_{m=-L+1}^{L-1} |w_a[m]|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |W_a(e^{j\Omega})|^2 d\Omega$$

Example of window: Triangular window $w_a[n]$ (Blackman–Tukey method)

Definition of the window. The triangular (Bartlett) window of length $2L-1$ is defined as

$$w_a[m] = \begin{cases} 1 - \dfrac{|m|}{L}, & |m| \le L-1, \\ 0, & \text{otherwise.} \end{cases}$$

It smoothly tapers the biased autocorrelation estimate $\hat{r}_x[m]$ to zero, reducing spectral leakage.

Energy of the window. The normalized energy is required to evaluate the bias and variance of the BT estimator:

$$E_{w_a} = \sum_{m=-(L-1)}^{L-1} w_a^2[m].$$

Since $w_a[m]$ is symmetric, we compute

$$E_{w_a} = 1 + 2\sum_{m=1}^{L-1}\left(1 - \frac{m}{L}\right)^2 = 1 + 2\sum_{m=1}^{L-1}\left(1 - \frac{2m}{L} + \frac{m^2}{L^2}\right).$$

Evaluating each term:

$$\sum_{m=1}^{L-1} 1 = L-1, \qquad \sum_{m=1}^{L-1} m = \frac{(L-1)L}{2}, \qquad \sum_{m=1}^{L-1} m^2 = \frac{(L-1)L(2L-1)}{6}.$$

Substituting:

$$E_{w_a} = 1 + 2\left[(L-1) - \frac{(L-1)L}{L} + \frac{(L-1)L(2L-1)}{6L^2}\right].$$

Simplifying:

$$E_{w_a} = \frac{2L-1}{3}.$$

Variance of the Blackman–Tukey estimator. For a general window $w[m]$, the variance is approximately

$$\operatorname{var}\left[\hat{S}_{BT}(e^{j\Omega})\right] \simeq \frac{1}{N}\left(\sum_{m=-\infty}^{\infty} w^2[m]\right) S_X^2(e^{j\Omega}).$$

Applying the triangular window energy result:

$$\operatorname{var}\left[\hat{S}_{BT}(e^{j\Omega})\right] \simeq \frac{E_{w_a}}{N}\, S_X^2(e^{j\Omega}) = \frac{2L-1}{3N}\, S_X^2(e^{j\Omega}).$$

Interpretation. A triangular window reduces spectral leakage more effectively than a rectangular window, but increases the main-lobe width. The variance decreases proportionally to $(2L-1)/3N$, showing that longer windows improve frequency resolution but increase estimator variance.

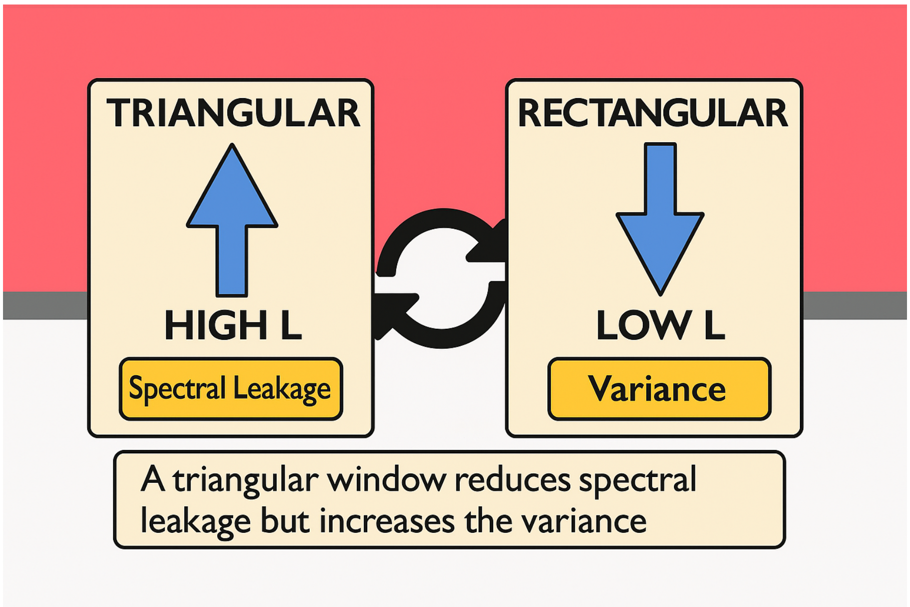

TRIANGULAR
HIGH L
Spectral Leakage
RECTANGULAR
LOW L
Variance
A triangular window reduces spectral leakage but increases the variance

Figure 1: As L increases, the bias is decreased. When L decreases, then the variance decreases, too $L\uparrow \implies \mathit{bias}\downarrow \qquad L\downarrow \implies \mathit{var}\downarrow$.

6.2.5 Bartlett-Welch spectral estimation techniques: average of periodograms

In this section we will deal with a new kind of spectral estimator called: **Bartlett-Welch's estimator**.

- It reduces the variance by averaging modified periodograms calculated using several segments of data samples.

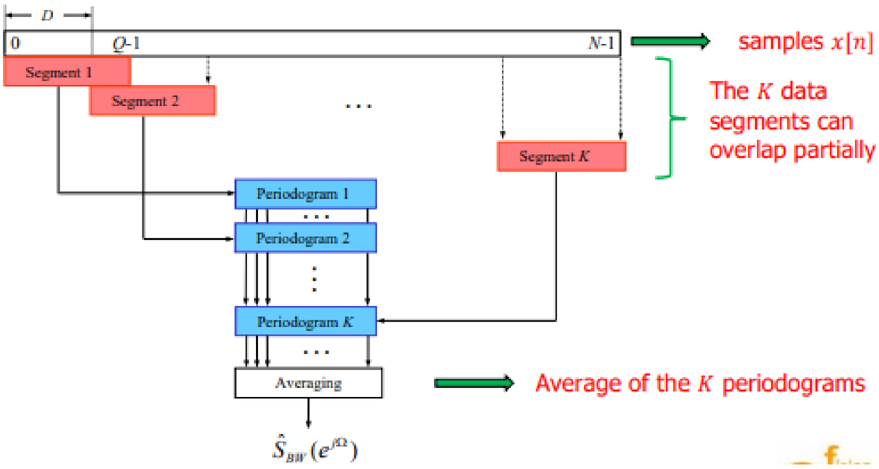


Figure 2: Bartlett-Welch's block diagram

Then, the expression of the estimator is the following:

- **Number of intervals:** K
- **Length of each interval:** Q samples
- **Distance between consecutive intervals:** D samples

$$\text{If } Q = D, \text{ there is no overlapping;} \qquad \text{if } Q \neq D, \text{ overlap} = Q - D.$$

- **Relation between parameters:**

  $$N = Q + (K-1)D \;\approx\; K \cdot D,$$

  where N is the total length of the data vector.

Bartlett–Welch estimator (averaging of modified periodograms).

$$\hat{S}_{BW}(e^{j\Omega}) = \frac{1}{K}\sum_{k=0}^{K-1}\left[\frac{1}{Q}\left|\sum_{n=0}^{Q-1} x[n+kD]\, v[n]\, e^{-j\Omega n}\right|^2\right]$$

where $v[n]$ is the window applied to each subsegment of length Q.

Note: we perform the average in order to make the variance zero so as to make it consistent.

Bartlett estimator (no overlapping, rectangular window).

$$\hat{S}_{B}(e^{j\Omega}) = \frac{1}{K}\sum_{k=0}^{K-1}\left[\frac{1}{Q}\left|\sum_{n=0}^{Q-1} x[n+kQ]\, e^{-j\Omega n}\right|^2\right].$$

Note.

- Bartlett is a particular case of Welch with:

  $$v[n] = 1, \qquad D = Q \quad \Rightarrow \quad \text{no overlap.}$$

- Welch reduces variance with respect to the classical periodogram by averaging K modified periodograms, at the cost of frequency resolution.

Then, the **mean** of this estimators:

$$E\Big[\hat{S}_{BW}(e^{j\Omega})\Big] = E\left[\frac{1}{K}\sum_{k=0}^{K-1}\left|\frac{1}{Q}\sum_{n=0}^{Q-1} x[n+kD]\, v[n]\, e^{-j\Omega n}\right|^2\right]$$

$$= \frac{1}{K}\sum_{k=0}^{K-1} E\left[\left|\frac{1}{Q}\sum_{n=0}^{Q-1} x[n+kD]\, v[n]\, e^{-j\Omega n}\right|^2\right]$$

$$= E\left[\left|\frac{1}{Q}\sum_{n=0}^{Q-1} x[n]\, v[n]\, e^{-j\Omega n}\right|^2\right] = S_x(e^{j\Omega}) \textcircled{\scriptsize 2\pi} \frac{1}{Q}\,\left|V(e^{j\Omega})\right|^2$$

if we want to **improve the resolution** (i.e. decrease the bias), we have to **increase the length Q** of each segment.
For the case of the **variance**; if the overlapping $(Q-D)$ between consecutive segments is not very high (or we use data windows; giving the overlapped segments a lower weight):

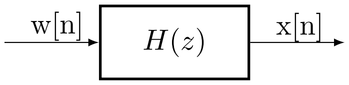
$$var[\hat{S}_{BW}(e^{j\Omega})] \approx \frac{1}{K} var[\hat{S}_P(e^{j\Omega})] \approx \frac{S_X^2(e^{j\Omega})}{K}$$

Note: if we increase the overlapping we can increase the number of segments K without decreasing the length of each segment Q but if the overlapping is very high, then consecutive segments will be correlated and the averaging will not reduce the variance.

6.3 Parametric spectral estimation.

6.3.1 Linear models of processes: AR, MA, ARMA.

Let's consider a **WSS stochastic process** $x[n]$ which can be modeled as the output of a LTI with input $w[n]$ (zero-mean white noise):



such that:

$$x[n] = \sum_{k=-\infty}^{+\infty} h[k]w[n-k]$$

and fulfills:

$$\begin{cases} \mathbb{E}[w[n]] = 0 \\ r_w = \mathbb{E}[w[n+m]w^*[n]] = \sigma^2\delta[m] \\ S_w(e^{j\Omega}) = \mathcal{F}\{r_w[m]\} = \sigma^2 \end{cases}$$

for the output data, then:

$$S_X(e^{j\Omega}) = S_w(e^{j\Omega})|H(e^{j\Omega})|^2 = \sigma^2|H(e^{j\Omega})|^2$$

Depending on the form of the $H(z)$ function, we have three different fractional parametric models:

Model	**System function** $H(z)$	**Power spectral density** $S_x(e^{j\Omega})$
AR(p) (all–pole)	$H(z) = \dfrac{1}{1+\sum_{k=1}^{p} a_k z^{-k}}$	$S_x(e^{j\Omega}) = \dfrac{\sigma^2}{\left\|1+\sum_{k=1}^{p} a_k e^{-j\Omega k}\right\|^2}$
MA(q) (all–zero)	$H(z) = 1+\sum_{k=1}^{q} b_k z^{-k}$	$S_x(e^{j\Omega}) = \sigma^2 \left\|1+\sum_{k=1}^{q} b_k e^{-j\Omega k}\right\|^2$
ARMA(p,q) (pole–zero)	$H(z) = \dfrac{1+\sum_{k=1}^{q} b_k z^{-k}}{1+\sum_{k=1}^{p} a_k z^{-k}}$	$S_x(e^{j\Omega}) = \sigma^2 \dfrac{\left\|1+\sum_{k=1}^{q} b_k e^{-j\Omega k}\right\|^2}{\left\|1+\sum_{k=1}^{p} a_k e^{-j\Omega k}\right\|^2}$

Table 3: System functions and power spectral densities for AR, MA, and ARMA models.

Objective and Methodology

Given N samples of a single realization of the process $x[n]$, we follow these steps:

1. **Estimation of the autocorrelation:**

$$\hat{r}_x[m]$$

2. **Estimation of the parameters of the fractional model:**

$$r_x[m] \to \hat{\sigma}^2\,, \left\{\hat{a}_k\right\}_{k=1}^{p} \left\{\hat{b}_k\right\}_{k=1}^{q}$$

3. **Estimation of the PSD:**

$$\hat{S}_{AR}(e^{j\Omega}) = \frac{\hat{\sigma}^2}{\left|1 + \sum_{k=1}^{p} \hat{a}_k e^{-j\Omega k}\right|^2}, \qquad \hat{S}_{MA}(e^{j\Omega}) = \hat{\sigma}^2 \left|1 + \sum_{k=1}^{q} \hat{b}_k e^{-j\Omega k}\right|^2,$$

$$\hat{S}_{ARMA}(e^{j\Omega}) = \hat{\sigma}^2 \, \frac{\left|1 + \sum_{k=1}^{q} \hat{b}_k e^{-j\Omega k}\right|^2}{\left|1 + \sum_{k=1}^{p} \hat{a}_k e^{-j\Omega k}\right|^2}.$$

Main advantage: If the model is correct, the estimation is significantly improved (lower bias and variance), because only a small number of parameters must be estimated.

6.3.2 Yule-Walker equations

Auto-Regressive (AR) Model

1. Model Definition

An Auto-Regressive model of order p is defined as:

$$x[n] = w[n] - \sum_{k=1}^{p} a_k \, x[n-k],$$

where $w[n]$ is white noise with variance σ^2.
The transfer function is:

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}}.$$

2. Causality of the Filter

Using the impulse response $h[n]$:

$$x[n] = \sum_{s=0}^{\infty} h[s]\, w[n-s], \qquad H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} = 1 - a_1 z^{-1} - a_2 z^{-2} - \cdots$$

Initial values:

$$h[0] = 1, \qquad h[1] = -a_1, \qquad \dots$$

3. Autocorrelation Derivation

Starting from:

$$w[n] = x[n] + \sum_{k=1}^{p} a_k x[n-k],$$

multiply by $x[n-m]$ and take expectations:

$$r_w[m] = r_x[m] + \sum_{k=1}^{p} a_k \, r_x[m-k].$$

Since

$$r_w[m] = \sigma^2 \delta[m],$$

we obtain the AR autocorrelation recursion:

$$\sigma^2 \delta[m] = r_x[m] + \sum_{k=1}^{p} a_k r_x[m-k].$$

Explicitly:

$$\begin{cases} \sigma^2 = r_x[0] + \sum_{k=1}^{p} a_k r_x[k], & m = 0, \\ 0 = r_x[m] + \sum_{k=1}^{p} a_k r_x[m-k], & m > 0. \end{cases}$$

4. Yule–Walker Equations

In matrix form:

$$\begin{bmatrix} r_x[0] & r_x[1] & \cdots & r_x[p-1] \\ r_x[1] & r_x[0] & \cdots & r_x[p-2] \\ \vdots & \vdots & \ddots & \vdots \\ r_x[p-1] & r_x[p-2] & \cdots & r_x[0] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} r_x[1] \\ r_x[2] \\ \vdots \\ r_x[p] \end{bmatrix}.$$

With:

$$\sigma^2 = r_x[0] + \sum_{k=1}^{p} a_k r_x[k].$$

If $r_x[m]$ is unknown, we use the biased estimate $\hat{r}_x[m]$, instead.

5. PSD Expression

The theoretical Power Spectral Density of an AR(p) model is:

$$S_x(e^{j\Omega}) = \frac{\sigma^2}{\left|1 + \sum_{k=1}^{p} a_k e^{-jk\Omega}\right|^2}.$$

Equivallently, we can use the estimations instead of the parameters. Such that we will obtain:

$$\begin{cases} \mathbb{E}[\hat{S}_{AR,p}(e^{j\Omega})] \approx S_X(e^{j\Omega}) \\ var[\hat{S}_{AR,p}(e^{j\Omega})] \approx \frac{2p}{N} S_X^2(e^{j\Omega}) \end{cases}$$

6. Practical Example

We consider an AR(2) process and 512 samples:

$$\hat{S}_p(e^{j\Omega}) \quad \text{(periodogram, noisy and high variance)}$$

Using Yule–Walker estimation with order $p = 4$:

$$\hat{S}_{\mathrm{AR}(4)}(e^{j\Omega})$$

Results:

- The AR model smooths the estimate significantly.
- Bias–variance tradeoff improves dramatically.
- The PSD estimate becomes close to the true $S_x(e^{j\Omega})$.

Main Advantage

If the model is correct, the estimation is much better (lower variance, clearer peaks) because only a few parameters need to be estimated.

ARMA(p, q) Model – Technical Datasheet

1. Model definition

An ARMA(p, q) process $x[n]$ is defined as:

$$x[n] = w[n] + \sum_{k=1}^{q} b_k\, w[n-k] \; - \; \sum_{k=1}^{p} a_k\, x[n-k],$$

where $w[n]$ is white noise with variance σ^2.
Its transfer function is:

$$H(z) = \frac{1 + \sum_{k=1}^{q} b_k z^{-k}}{1 + \sum_{k=1}^{p} a_k z^{-k}}.$$

2. Causality and impulse response

The filter is causal and BIBO-stable if all poles lie strictly inside the unit circle.
The impulse response satisfies:

$$x[n] = \sum_{s=0}^{\infty} h[s]\, w[n-s], \qquad h[0] = 1, \quad h[1] = b_1 - a_1, \dots$$

Thus the ARMA model can be interpreted as:
- an **MA part** (b_k) shaping the noise,
- an **AR part** (a_k) feeding back past samples.

3. Autocorrelation expression

Starting from the ARMA equation:

$$w[n] + \sum_{k=1}^{q} b_k w[n-k] = x[n] + \sum_{k=1}^{p} a_k x[n-k],$$

multiply both sides by $x[n-m]$ and take expectations:

$$E[w[n]\,x[n-m]] + \sum_{k=1}^{q} b_k\, E[w[n-k]\,x[n-m]] = r_x[m] + \sum_{k=1}^{p} a_k\, r_x[m-k].$$

Using whiteness:

$$E[w[n]\,x[n-m]] = \sigma^2 h[m], \qquad E[w[n-k]\,x[n-m]] = \sigma^2 h[m-k],$$

So the general ARMA autocorrelation identity becomes:

$$\sigma^2 h[m] + \sigma^2 \sum_{k=1}^{q} b_k\, h[m-k] = r_x[m] + \sum_{k=1}^{p} a_k r_x[m-k].$$

4. Delayed Yule–Walker equations (estimation of AR part)

For lags $m > q$, all MA terms vanish:

$$0 = r_x[m] + \sum_{k=1}^{p} a_k\, r_x[m-k], \qquad m > q.$$

This yields the **delayed Yule–Walker system**:

$$\begin{bmatrix} r_x[q] & r_x[q-1] & \dots & r_x[q-p+1] \\ r_x[q+1] & r_x[q] & \dots & r_x[q-p+2] \\ \vdots & & & \vdots \\ r_x[q+p-1] & r_x[q+p-2] & \dots & r_x[q] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} r_x[q+1] \\ r_x[q+2] \\ \vdots \\ r_x[q+p] \end{bmatrix}.$$

Once the AR coefficients $\{a_k\}$ are found, the variance is:

$$\sigma^2 = r_x[0] + \sum_{k=1}^{p} a_k r_x[k].$$

5. Estimation of MA coefficients

Estimating the MA parameters $\{b_k\}_{k=1}^{q}$ is significantly more involved because the autocorrelation equations are nonlinear in b_k.

6. Example

Consider an ARMA process with real poles inside the unit circle and finite-order MA part. Once the AR and MA parameters have been computed, the PSD is:

$$S_x(e^{j\Omega}) = \sigma^2 \frac{\left|1 + \sum_{k=1}^{q} b_k e^{-jk\Omega}\right|^2}{\left|1 + \sum_{k=1}^{p} a_k e^{-jk\Omega}\right|^2}.$$

This PSD is typically much smoother than nonparametric estimators such as the periodogram, and it converges faster because only a small number of parameters needs to be estimated.

6.4 Summary

Statistical Meaning of Key Spectral-Estimation Terms

Resolution Ability to distinguish close frequency components.

$$\text{Higher resolution} \iff \text{narrower main lobe} \iff \textbf{lower bias but higher variance}.$$

Spectral leakage Energy from one frequency spills into others due to window sidelobes.

$$\text{More leakage} \iff \textbf{higher bias}.$$

Smoothing Averaging across frequencies or across segments (Welch/BT).

$$\text{More smoothing} \iff \textbf{lower variance but higher bias}.$$

Bias Difference between the expected estimator and the true PSD.

$$\text{Larger window sidelobes} \iff \textbf{higher bias (more leakage)}.$$

Variance Random fluctuations of the estimator around its mean.

$$\text{Averaging over } K \text{ segments} \implies \text{var} \propto \frac{1}{K}.$$

Window main-lobe width Controls frequency resolution.

$$\text{Wider main lobe} \iff \textbf{lower resolution, more smoothing}.$$

Window side-lobe level Controls leakage.

$$\text{Higher sidelobes} \iff \textbf{more leakage (higher bias)}.$$

Consistency Variance goes to zero as $N \to \infty$. The periodogram is *not* consistent because

$$\text{var}[\hat{S}_P] \not\to 0.$$

Periodogram

Definition:

$$\hat{S}_P(e^{j\Omega}) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x[n]e^{-j\Omega n}\right|^2.$$

Mean value:

$$E[\hat{S}_P] = S_X \circledast W.$$

Asymptotic unbiasedness:

$$\lim_{N\to\infty} E[\hat{S}_P] = S_X.$$

Variance:

$$\mathrm{var}[\hat{S}_P] \approx S_X^2 \quad \text{(inconsistent)}.$$

Power:

$$\hat{P}_X = \frac{1}{N}\sum |x[n]|^2, \quad E[\hat{P}_X] = P_X.$$

Notes: High variance, spectral leakage, resolution fixed by N.

Modified Periodogram

Definition:

$$\hat{S}_{MP}(e^{j\Omega}) = \frac{1}{N}\left|\sum x[n]v[n]e^{-j\Omega n}\right|^2.$$

Mean:

$$E[\hat{S}_{MP}] = S_X \circledast \frac{1}{N}|V|^2.$$

Variance:

$$\mathrm{var}[\hat{S}_{MP}] \simeq S_X^2 \quad \text{(inconsistent)}.$$

Window effect: Lower leakage but lower resolution.
Unbiased power iff:

$$\sum |v[n]|^2 = N.$$

Notes: Trades leakage vs. resolution. Variance unchanged.

Bartlett & Welch

Bartlett:

$$\hat{S}_B = \frac{1}{K}\sum_k \frac{1}{Q}\left|\sum_{n=0}^{Q-1} x[n+kQ]e^{-j\Omega n}\right|^2.$$

Welch:

$$\hat{S}_{BW} = \frac{1}{K}\sum_k \frac{1}{Q}\left|\sum x[n+kD]v[n]e^{-j\Omega n}\right|^2.$$

Mean:

$$E[\hat{S}_{BW}] = S_X \circledast \frac{1}{Q}|V|^2.$$

Variance:

$$\mathrm{var}[\hat{S}_{BW}] \approx \frac{S_X^2}{K}.$$

Notes:

- Consistent (variance ↓ with K).
- Welch improves bias vs Bartlett.
- Resolution depends on Q, not on N.

Blackman–Tukey

Definition:

$$\hat{S}_{BT} = \mathcal{F}\{\hat{r}_x[m]w_a[m]\} = \hat{S}_P \circledast W_a.$$

Mean:

$$E[\hat{S}_{BT}] = S_X \circledast W_a.$$

Variance:

$$\mathrm{var}[\hat{S}_{BT}] \approx \frac{E_{W_a}}{N} S_X^2.$$

Unbiased if:

$$w_a[0] = 1, \quad N \to \infty, \quad L \to \infty.$$

Tradeoff:

$$L \uparrow \Rightarrow \text{bias} \downarrow, \quad \text{var} \uparrow.$$

Notes: Consistent, controllable smoothing, bias set by lag window.

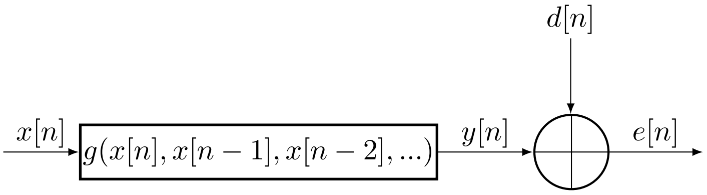
Chapter 7

Optimal Wiener filtering

7 Optimal Wiener filtering

7.1 Minimum Mean Square Error (MMSE) Bayesian estimator.

We suggest now a new way of designing an estimation function based on minimizing the MMSE. In particular, the diagram of the process is the following:



where $e[n] = d[n] - y[n]$ the estimation error. Then, the key question is related to:

How do we have to design the estimator $g(\cdot)$?

Design based on the Minimum Mean Square Error (MMSE) criterion Let the estimation error be defined as

$$e[n] = d[n] - g\big(x[n], x[n-1], x[n-2], \dots\big),$$

where:

- $d[n]$ is the desired signal,
- $x[n]$ is the observed signal,
- $g(\cdot)$ is the estimator to be designed.

The MMSE estimator is obtained by minimizing the mean square error:

$$g^{\text{MMSE}} = \arg\min_{g} \ \mathbb{E}\big[\,|e[n]|^2\,\big] = \arg\min_{g} \ \mathbb{E}\Big[\big|d[n] - g\big(x[n], x[n-1], x[n-2], \dots\big)\big|^2\Big]\,.$$

The solution of this optimization problem is the **Bayesian posterior mean estimator**:

$$g^{\text{MMSE}}\big(x[n], x[n-1], x[n-2], \dots\big) = \mathbb{E}\big[d[n] \,\big|\, x[n], x[n-1], x[n-2], \dots\big]\,.$$

Gaussian case. If $d[n]$ and $x[n]$ are *jointly Gaussian random processes*, the conditional expectation is a **linear function**. Therefore, the MMSE estimator is linear:

$$g^{\text{MMSE}}\big(x[n], x[n-1], \dots\big) = h_0^* x[n] + h_1^* x[n-1] + h_2^* x[n-2] + \dots = \mathbf{h}^H \mathbf{x}[n].$$

Hence, for Gaussian processes:

- The MMSE estimator is linear.
- The optimal system is a linear filter.

Design based on a linear FIR filter with Q coefficients In practice, we often *force* the estimator to be linear (for non-gaussian signals), even when optimality is not guaranteed.

We restrict the estimator to be a finite impulse response (FIR) filter with Q coefficients.

The output of the filter is

$$y[n] = \sum_{i=0}^{Q-1} h^*[i]\, x[n-i] = \mathbf{h}^H \mathbf{x}[n],$$

where

$$\mathbf{h} = \begin{bmatrix} h[0] \\ h[1] \\ \vdots \\ h[Q-1] \end{bmatrix}, \qquad \mathbf{x}[n] = \begin{bmatrix} x[n] \\ x[n-1] \\ \vdots \\ x[n-Q+1] \end{bmatrix}.$$

Then, the error signal is defined as

$$e[n] = d[n] - y[n].$$

So the main objective is to find the optimal filter **h** that minimizes the mean square error:

$$\mathbf{h}_{\text{MMSE}} = \arg\min_{\mathbf{h}} \; \mathbb{E}\left[\, |d[n] - \mathbf{h}^H \mathbf{x}[n]|^2 \,\right].$$

Remarks.

- The linear MMSE estimator is **optimal** when the signals are Gaussian.
- For non-Gaussian signals, it may be **suboptimal**, but it is still widely used due to its simplicity.
- The solution leads to the **Wiener–Hopf equations**, which provide a closed-form expression for $\mathbf{h}_{\text{MMSE}}$.

7.2 Wiener filter and Wiener–Hopf equations

Recalling the definition of the error signal

$$e[n] = d[n] - \mathbf{h}^H \mathbf{x}[n],$$

the mean square error (MSE) is given by

$$\xi = \mathbb{E}\left[|e[n]|^2\right] = \mathbb{E}\left[\big(d[n] - \mathbf{h}^H \mathbf{x}[n]\big)\big(d^*[n] - \mathbf{x}^H[n]\mathbf{h}\big)\right].$$

Expanding the product and using linearity of expectation,

$$\xi = \mathbb{E}\left[|d[n]|^2\right] - \mathbb{E}\left[d[n]\mathbf{x}^H[n]\right]\mathbf{h} - \mathbf{h}^H \mathbb{E}[\mathbf{x}[n] d^*[n]] + \mathbf{h}^H \mathbb{E}\left[\mathbf{x}[n]\mathbf{x}^H[n]\right]\mathbf{h}.$$

Defining the following statistical quantities:

$$P_d = \mathbb{E}\left[|d[n]|^2\right], \qquad \mathbf{R}_x = \mathbb{E}\left[\mathbf{x}[n]\mathbf{x}^H[n]\right] = \begin{bmatrix} r_x[0] & r_x[1] & \cdots & r_x[Q-1] \\ r_x^*[1] & r_x[0] & \cdots & r_x[Q-2] \\ \vdots & \vdots & \ddots & \vdots \\ r_x^*[Q-1] & r_x^*[Q-2] & \cdots & r_x[0] \end{bmatrix},$$

$$\mathbf{r}_{xd} = \mathbb{E}[\mathbf{x}[n] d^*[n]] = \begin{bmatrix} r_{xd}[0] \\ r_{xd}[-1] \\ \vdots \\ r_{xd}[-Q+1] \end{bmatrix},$$

the MSE can be written compactly as

$$\boxed{\xi = P_d + \mathbf{h}^H \mathbf{R}_x \mathbf{h} - \mathbf{h}^H \mathbf{r}_{xd} - \mathbf{r}_{xd}^H \mathbf{h}}.$$

7.2.1 Minimization of the MSE

To find the optimal filter coefficients, we minimize ξ with respect to $\mathbf{h}$. Taking the gradient with respect to $\mathbf{h}^*$ and setting it to zero,

$$\nabla_{\mathbf{h}^*} \xi = \mathbf{R}_x \mathbf{h} - \mathbf{r}_{xd} = \mathbf{0}.$$

This yields the **Wiener–Hopf equations**:

$$\boxed{\mathbf{R}_x \mathbf{h}_{\text{opt}} = \mathbf{r}_{xd}}.$$

Provided that $\mathbf{R}_x$ is invertible, the optimal Wiener filter is

$$\boxed{\mathbf{h}_{\text{opt}} = \mathbf{R}_x^{-1} \mathbf{r}_{xd}.}$$

7.2.2 Minimum achievable MSE

Substituting $\mathbf{h}_{\text{opt}}$ into the MSE expression, the minimum error power is

$$\xi_{\min} = P_d - \mathbf{r}_{xd}^H \mathbf{R}_x^{-1} \mathbf{r}_{xd}.$$

This represents the *minimum achievable MSE* using a linear FIR estimator.

7.2.3 Orthogonality principle

An equivalent characterization of the Wiener filter is given by the **orthogonality principle**. At optimality, the estimation error is orthogonal to the data:

$$\boxed{\mathbb{E}[e[n]\mathbf{x}[n]] = \mathbf{0}.}$$

This condition is fully equivalent to the Wiener–Hopf equations and provides a geometric interpretation: the error vector lies in the subspace orthogonal to the signal space spanned by $\mathbf{x}[n]$.

7.2.4 MSE for a non-optimal filter

For an arbitrary filter $\mathbf{h}$, the MSE can be written as

$$\xi = \xi_{\min} + (\mathbf{h} - \mathbf{h}_{\text{opt}})^H \mathbf{R}_x (\mathbf{h} - \mathbf{h}_{\text{opt}}).$$

This expression shows that:

- $\xi_{\min}$ is the **minimum achievable MSE**.
- Any deviation from $\mathbf{h}_{\text{opt}}$ increases the MSE by a positive quadratic form.

Key observations.

- The Wiener filter provides the **globally optimal linear estimator**.
- The MSE increase due to suboptimal coefficients depends on both the mismatch $(\mathbf{h} - \mathbf{h}_{\text{opt}})$ and the signal statistics $\mathbf{R}_x$.
- If $\mathbf{r}_{xd} = \mathbf{0}$ (orthogonality between $x[n]$ and $d[n]$), then $\mathbf{h}_{\text{opt}} = \mathbf{0}$ and $\xi_{\min} = P_d$.

7.2.5 Wiener filter using a finite number of samples to estimate the MSE.

The main problem we can front to consists in the computation of the second order moments R_x and r_{xd}. In some situations, this information is not available. Hence, in some situations we only have a finite number of samples of the data $x[n]$ and the reference $d[n]$, and the objective is to minimize an estimation of the MSE:

$$MSE \equiv \frac{1}{N}\sum_{n=0}^{N-1}|e[n]|^2 = \frac{1}{N}\sum_{n=0}^{N-1}|d[n]-\mathbf{h}^H\mathbf{x}[n]|^2$$

For the case we have:

- N samples of the reference $d[n]$.
- $N+Q-1$ samples of the input $x[n]$.
- Q samples of the filter response $h[n]$

then for each value of n:

$$e[n] = d[n] - y[n] = d[n] - \mathbf{h}^H \cdot \mathbf{x}[n]$$

and if we arrange the N samples in one vector:

$$\begin{bmatrix} e[0] & e[1] & \cdots & e[N-1] \end{bmatrix} = \begin{bmatrix} d[0] & d[1] & \cdots & d[N-1] \end{bmatrix} - \begin{bmatrix} y[0] & y[1] & \cdots & y[N-1] \end{bmatrix}$$

$$= \begin{bmatrix} d[0] & d[1] & \cdots & d[N-1] \end{bmatrix} - \begin{bmatrix} h^*[0] & h^*[1] & \cdots & h^*[Q-1] \end{bmatrix} \begin{bmatrix} x[0] & x[1] & \cdots & x[N-1] \\ x[-1] & x[0] & \cdots & x[N-2] \\ \vdots & \vdots & \ddots & \vdots \\ x[-Q+1] & x[-Q+2] & \cdots & x[N-Q] \end{bmatrix}$$

Then, the mean square error can be estimated as

$$\widehat{\mathrm{MSE}} = \frac{1}{N}\sum_{n=0}^{N-1}|e[n]|^2 = \frac{1}{N}\mathbf{e}^H\mathbf{e} = \frac{1}{N}(\mathbf{d}-\mathbf{Xh})^H(\mathbf{d}-\mathbf{Xh})$$
$$= \frac{1}{N}\Big(\mathbf{d}^H\mathbf{d} - \mathbf{h}^H\mathbf{X}^H\mathbf{d} - \mathbf{d}^H\mathbf{Xh} + \mathbf{h}^H\mathbf{X}^H\mathbf{Xh}\Big).$$

Taking the gradient with respect to $\mathbf{h}^*$ and setting it to zero,

$$\nabla_{\mathbf{h}^*}\widehat{\mathrm{MSE}} = \frac{1}{N}\left(\mathbf{X}^H\mathbf{Xh} - \mathbf{X}^H\mathbf{d}\right) = \mathbf{0}.$$

This yields the **sample Wiener–Hopf equations**:

$$\boxed{\widehat{\mathbf{R}}_x\,\mathbf{h}_{\text{opt}} = \widehat{\mathbf{r}}_{xd}}$$

with

$$\widehat{\mathbf{R}}_x = \frac{1}{N}\mathbf{X}^H\mathbf{X} = \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{x}[n]\mathbf{x}^H[n], \qquad \widehat{\mathbf{r}}_{xd} = \frac{1}{N}\mathbf{X}^H\mathbf{d} = \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{x}[n]d^*[n].$$

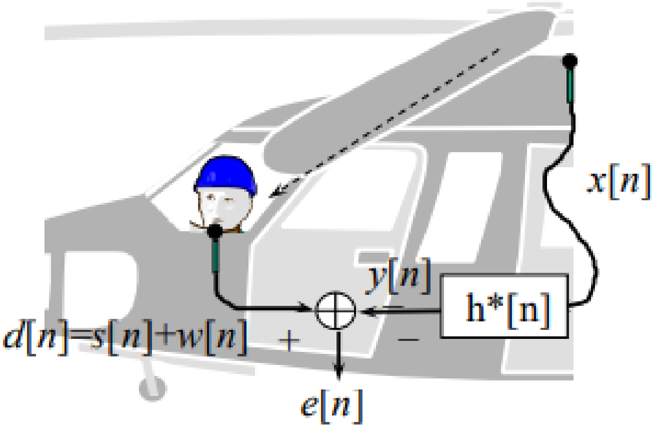
Provided that $\widehat{\mathbf{R}}_x$ is invertible, the optimal filter is

$$\boxed{\mathbf{h}_{\text{opt}} = \widehat{\mathbf{R}}_x^{-1}\,\widehat{\mathbf{r}}_{xd}.}$$

Remark. The matrices $\widehat{\mathbf{R}}_x$ and $\widehat{\mathbf{r}}_{xd}$ are *sample estimates* of the true auto-correlation matrix and cross-correlation vector. As $N \to \infty$, they converge to their ensemble counterparts.

7.3 Aplications

7.3.1 Application 1: Noise / Interference Cancellation



One of the most illustrative applications of the Wiener filter is **noise (or interference) cancellation**. The goal is to recover a desired signal that is corrupted by an additive interference, using an auxiliary measurement that is correlated with the interference.

Signal model We observe a signal

$$d[n] = s[n] + w[n],$$

where:

- $s[n]$ is the *desired signal* (e.g. the pilot's voice),
- $w[n]$ is an *interference or noise* signal (e.g. engine noise).

In addition, we have access to a *reference signal* $x[n]$ that:

- is **correlated** with the interference $w[n]$,
- is **uncorrelated** with the desired signal $s[n]$.

All signals are assumed to have zero mean.

Filtering structure We process the reference signal $x[n]$ through a linear FIR filter $H(z)$:

$$y[n] = \sum_{i=0}^{Q-1} h^*[i]\, x[n-i],$$

and subtract the filter output from the observed signal:

$$e[n] = d[n] - y[n].$$

The signal $e[n]$ is the estimate of the desired signal $s[n]$. The filter coefficients $\mathbf{h}$ are chosen to minimize the mean square error

$$J = \mathbb{E}\big[\,|e[n]|^2\,\big] = \mathbb{E}[|d[n] - y[n]|^2].$$

MMSE criterion The optimization problem is

$$\mathbf{h}_{\text{opt}} = \arg\min_{\mathbf{h}} \ \mathbb{E}\big[\, |d[n] - \mathbf{h}^H \mathbf{x}[n]|^2 \,\big] .$$

Minimizing this cost function leads to the **Wiener–Hopf equations**:

$$\mathbf{R}_x \mathbf{h}_{\text{opt}} = \mathbf{r}_{xd},$$

where:

$$\mathbf{R}_x = \mathbb{E}\big[\mathbf{x}[n]\mathbf{x}^H[n]\big] , \qquad \mathbf{r}_{xd} = \mathbb{E}[\mathbf{x}[n]d^*[n]] .$$

If $\mathbf{R}_x$ is invertible, the optimal filter is

$$\mathbf{h}_{\text{opt}} = \mathbf{R}_x^{-1}\mathbf{r}_{xd}.$$

Role of the uncorrelated-signal assumption Using the signal model $d[n] = s[n] + w[n]$, the cross-correlation becomes

$$\mathbf{r}_{xd} = \mathbb{E}[\mathbf{x}[n]s^*[n]] + \mathbb{E}[\mathbf{x}[n]w^*[n]] .$$

Since $x[n]$ and $s[n]$ are assumed uncorrelated,

$$\mathbb{E}[\mathbf{x}[n]s^*[n]] = \mathbf{0},$$

and therefore

$$\mathbf{r}_{xd} = \mathbf{r}_{xw}.$$

This shows that the filter is *entirely driven by the correlation between the reference signal and the interference.* As the filter adapts, $y[n]$ becomes an increasingly accurate estimate of $w[n]$, so that

$$e[n] \approx s[n].$$

Consistency of the solution As the interference is progressively canceled, the error signal $e[n]$ becomes closer to $s[n]$. Since $s[n]$ and $x[n]$ are uncorrelated, the optimality condition

$$\mathbb{E}[e[n]\mathbf{x}[n]] = \mathbf{0}$$

is satisfied. Thus, the same Wiener solution remains valid throughout the cancellation process.

Special case: uncorrelated reference If the reference signal $x[n]$ is uncorrelated with the interference $w[n]$, then

$$\mathbf{r}_{xw} = \mathbf{0},$$

and the Wiener–Hopf equations reduce to

$$\mathbf{R}_x \mathbf{h}_{\text{opt}} = \mathbf{0} \quad \Rightarrow \quad \mathbf{h}_{\text{opt}} = \mathbf{0}.$$

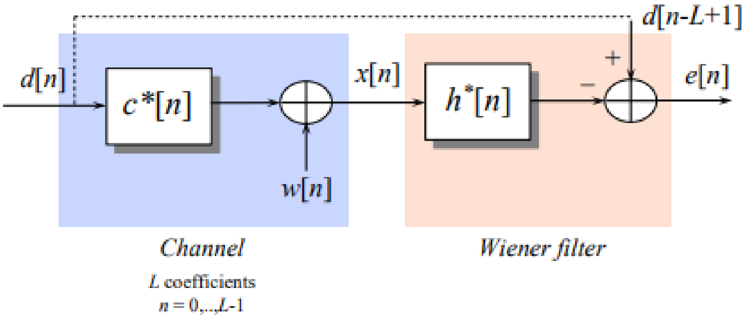
In this case, the filter does nothing: no noise cancellation is possible, but importantly, *no additional noise is introduced into the output.* This highlights the robustness of the Wiener filter.

Interpretation This example shows how the Wiener filter exploits correlation structure:

- useful information is extracted from the reference signal,
- uncorrelated components are automatically ignored,
- optimal noise cancellation is achieved in the MMSE sense.

It provides a clear and physically intuitive illustration of optimal linear filtering.

7.3.2 Application 2: Channel Equalization



We now apply the MMSE/Wiener filtering framework to the problem of **channel equalization** in the time domain.

The objective is to recover a transmitted symbol sequence $d[n]$ from a received signal that has been distorted by a linear channel and corrupted by additive noise.

Signal model We assume a linear time-invariant channel with impulse response $c[n]$ of length L, and additive noise $w[n]$. The received signal is

$$x[n] = \sum_{k=0}^{L-1} c[k]\, d[n-k] + w[n].$$

In vector form, this can be written as

$$\mathbf{x}[n] = \mathbf{C}^H \mathbf{d}[n] + \mathbf{w}[n],$$

where $\mathbf{C}$ is a Toeplitz convolution matrix constructed from the channel coefficients $c[n]$.

Equalizer structure We design a linear FIR equalizer with coefficients $\mathbf{h}$ that produces the estimate

$$\hat{d}[n-L+1] = \mathbf{h}^H \mathbf{x}[n].$$

The delay $L-1$ is introduced to account for the channel memory and ensure a causal implementation.

The estimation error is defined as

$$e[n] = d[n-L+1] - \mathbf{h}^H \mathbf{x}[n].$$

MMSE formulation The optimal equalizer minimizes the mean square error

$$\mathbf{h}_{\text{opt}} = \arg\min_{\mathbf{h}} \mathbb{E}\left[\left| d[n-L+1] - \mathbf{h}^H \mathbf{x}[n] \right|^2 \right].$$

From the Wiener theory, the solution is given by

$$\boxed{\mathbf{h}_{\text{opt}} = \mathbf{R}_x^{-1} \mathbf{r}_{xd}}$$

where

$$\mathbf{R}_x = \mathbb{E}\left[\mathbf{x}[n]\mathbf{x}^H[n]\right], \qquad \mathbf{r}_{xd} = \mathbb{E}[\mathbf{x}[n] d^*[n-L+1]].$$

Evaluation of the correlation terms If the transmitted symbols and the noise are uncorrelated, we have

$$\mathbf{R}_x = \mathbb{E}\big[\mathbf{C}^H \mathbf{d}[n]\mathbf{d}^H[n]\mathbf{C}\big] + \mathbb{E}\big[\mathbf{w}[n]\mathbf{w}^H[n]\big] = \mathbf{C}^H \mathbf{R}_d \mathbf{C} + \mathbf{R}_w.$$

If the transmitted symbols are white with variance σ_d^2,

$$\mathbf{R}_d = \sigma_d^2 \mathbf{I}.$$

The cross-correlation vector becomes

$$\mathbf{r}_{xd} = \mathbb{E}[\mathbf{x}[n]d^*[n-L+1]] = \sigma_d^2 \begin{bmatrix} c^*[L-1] \\ c^*[L-2] \\ \vdots \\ c^*[0] \end{bmatrix}.$$

Final expression of the Wiener equalizer Substituting the previous expressions into the Wiener solution, the optimal equalizer can be written as

$$\boxed{\mathbf{h}_{\text{opt}} = \left(\mathbf{C}^H \mathbf{C} + \frac{1}{\sigma_d^2}\mathbf{R}_w\right)^{-1} \mathbf{c}}$$

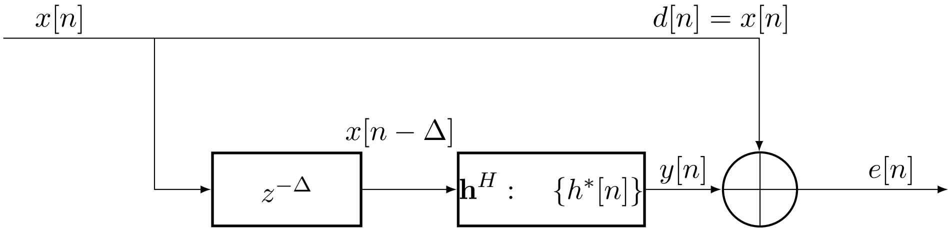
where $\mathbf{c} = [c^*[L-1],\ c^*[L-2],\ \ldots,\ c^*[0]]^T$.

Interpretation.

- The equalizer inverts the channel while accounting for noise.
- In the absence of noise, the solution approaches a zero-forcing equalizer.
- In the presence of noise, the MMSE solution balances channel inversion and noise amplification.

7.4 Linear prediction

We now suggest a new kind of prediction that estimate the samples of a process using past samples. Then the new modeling of the filter estimation:



with an output:

$$y[n] = \hat{x}[n] = \sum_{i=0}^{Q-1} h_i^* x[n-\Delta-i] = \mathbf{h}^H \mathbf{x}[n-\Delta]$$

Then, as we have already done in the previous chapters, the solution of the MSE problem is obtained by solving the Wiener–Hopf equations. For the linear predictor, the relevant correlation quantities are

$$\mathbf{R}_x = \mathbb{E}\big[\mathbf{x}[n-\Delta]\mathbf{x}^H[n-\Delta]\big], \qquad \mathbf{r}_{xd} = \mathbb{E}[\mathbf{x}[n-\Delta]\, x^*[n]] .$$

The optimal predictor coefficients are therefore given by

$$\boxed{\mathbf{h}_{\text{opt}} = \mathbf{R}_x^{-1}\mathbf{r}_{xd}.}$$

Prediction error The prediction error is defined as

$$e[n] = x[n] - \hat{x}[n].$$

The corresponding minimum prediction error power is then

$$\xi_{\min} = \mathbb{E}\big[\, |e[n]|^2 \,\big] = \mathbb{E}\big[\, |x[n] - \mathbf{h}_{\text{opt}}^H \mathbf{x}[n-\Delta]|^2 \,\big] .$$

It can be shown that this reduces to

$$\xi_{\min} = r_x[0] - \mathbf{h}_{\text{opt}}^H \mathbf{r}_{xd},$$

where $r_x[k]$ denotes the autocorrelation function of the process $x[n]$.

7.5 Filter generating the prediction error

The predictor can be interpreted as a Wiener filter whose output is subtracted from the current sample. The transfer function of the filter generating the prediction error is

$$H_{\text{pred,err}}(z) = 1 - z^{-\Delta} \sum_{i=0}^{Q-1} h_i^* z^{-i}.$$

This representation highlights that the prediction error is obtained by filtering the process $x[n]$ with a suitable filter that removes its predictable component.

7.5.1 Linear prediction of an AR process

Let us now assume that $x[n]$ follows an autoregressive (AR) process of order p,

$$x[n] = -\sum_{k=1}^{p} a_k x[n-k] + w[n],$$

where $w[n]$ is a white noise process with variance σ^2.

In this case, the Yule–Walker equations relate the AR coefficients to the autocorrelation sequence:

$$\begin{bmatrix} r_x[0] & r_x[1] & \cdots & r_x[p-1] \\ r_x[1] & r_x[0] & \cdots & r_x[p-2] \\ \vdots & \vdots & \ddots & \vdots \\ r_x[p-1] & r_x[p-2] & \cdots & r_x[0] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} r_x[1] \\ r_x[2] \\ \vdots \\ r_x[p] \end{bmatrix}.$$

For a prediction delay $\Delta = 1$ and a predictor length $Q \geq p$, the optimal linear predictor satisfies

$$h_{\text{opt},i} = \begin{cases} -a_{i+1}, & 0 \leq i \leq p-1, \\ 0, & p \leq i \leq Q-1. \end{cases}$$

White prediction error In this situation, the prediction error coincides with the excitation noise:

$$e[n] = w[n].$$

Therefore, the prediction error is a white noise process and the minimum achievable error power is

$$\xi_{\min} = \mathbb{E}\big[|w[n]|^2\big] = \sigma^2.$$

This result shows that linear prediction completely removes the predictable structure of an AR process, leaving only the innovation term.

7.6 Adaptive implementation of the Wiener filter.

7.6.1 Steepest Descent (SD)

In many practical situations, the statistical properties of the signals involved are not perfectly known or may vary with time. In such cases, a fixed (non-adaptive) filter is no longer optimal. This motivates the use of *adaptive systems.*

An adaptive system automatically adjusts its parameters in order to optimize a given performance criterion, typically based on the observation of the input and output signals. Conceptually, an adaptive system consists of two parts:

- A **processing system**, usually a linear filter with impulse response $\mathbf{h}$.
- A **learning system**, which updates $\mathbf{h}$ based on a cost function measuring performance.

A common example is the automatic control of the gain of an amplifier, where the system adapts its parameters to achieve a desired output power level.

The MSE function We consider a linear FIR filter with coefficient vector $\mathbf{h}$. The filter output is

$$y[n] = \mathbf{h}^H \mathbf{x}[n],$$

and the estimation error is

$$e[n] = d[n] - y[n].$$

The mean square error (MSE) is defined as

$$\xi(\mathbf{h}) = \mathbb{E}\left[|e[n]|^2\right].$$

Using the results derived previously, the MSE can be written as

$$\xi(\mathbf{h}) = P_d + \mathbf{h}^H \mathbf{R}_x \mathbf{h} - \mathbf{h}^H \mathbf{r}_{xd} - \mathbf{r}_{xd}^H \mathbf{h},$$

where $\mathbf{R}_x$ is the autocorrelation matrix of the input and $\mathbf{r}_{xd}$ is the cross-correlation vector between the input and the desired signal.

This expression can be rearranged as

$$\xi(\mathbf{h}) = \xi_{\min} + (\mathbf{h} - \mathbf{h}_{\text{opt}})^H \mathbf{R}_x (\mathbf{h} - \mathbf{h}_{\text{opt}}),$$

where

$$\mathbf{h}_{\text{opt}} = \mathbf{R}_x^{-1} \mathbf{r}_{xd}, \qquad \xi_{\min} = P_d - \mathbf{r}_{xd}^H \mathbf{R}_x^{-1} \mathbf{r}_{xd}.$$

This shows that the MSE is a **positive semidefinite quadratic function** of the filter coefficients. Its minimum is unique and is attained at $\mathbf{h} = \mathbf{h}_{\text{opt}}$.

The quadratic structure of $\xi(\mathbf{h})$ implies that:

- Any one-dimensional cut of the MSE is a parabola.
- The level curves of the MSE are ellipsoids in the coefficient space.
- The center of the ellipsoids corresponds to the optimal Wiener solution $\mathbf{h}_{\text{opt}}$.

The gradient of the MSE points in the direction of maximum increase of the error and is given by

$$\nabla_{\mathbf{h}^*} \xi(\mathbf{h}) = \mathbf{R}_x \mathbf{h} - \mathbf{r}_{xd}.$$

7.6.2 Gradient method: steepest descent

To adaptively minimize the MSE, we can update the filter coefficients by moving in the opposite direction of the gradient. This leads to the **steepest descent algorithm**:

$$\mathbf{h}^{(k+1)} = \mathbf{h}^{(k)} - \mu \nabla_{\mathbf{h}^*} \xi(\mathbf{h}^{(k)}) = \mathbf{h}^{(k)} - \mu\big(\mathbf{R}_x \mathbf{h}^{(k)} - \mathbf{r}_{xd}\big),$$

where $\mu > 0$ is the step size controlling the speed of convergence.

To analyze convergence, it is convenient to define the error vector

$$\tilde{\mathbf{h}}^{(k)} = \mathbf{h}^{(k)} - \mathbf{h}_{\text{opt}}.$$

In this new variable, the MSE becomes

$$\xi(\tilde{\mathbf{h}}^{(k)}) = \xi_{\min} + \tilde{\mathbf{h}}^{(k)H} \mathbf{R}_x \tilde{\mathbf{h}}^{(k)}.$$

Role of eigenvalues

Let the eigen-decomposition of the correlation matrix be

$$\mathbf{R}_x = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^H,$$

where $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_Q)$.
In this basis, the level curves of the MSE are ellipsoids whose principal axes are aligned with the eigenvectors of $\mathbf{R}_x$. The length of each axis is inversely proportional to $\sqrt{\lambda_i}$. This has an important consequence:

- If the eigenvalues are similar, the ellipsoids are nearly circular and convergence is fast.
- If there is a large eigenvalue spread, the ellipsoids are elongated and convergence becomes slow and oscillatory.

Error surface examples for a FIR filter with $Q = 2$

We consider the mean square error (MSE) function for a FIR filter with two coefficients, $\mathbf{h} = [h_0 \; h_1]^T$. The MSE can be written as

$$\xi(h_0, h_1) = r_x[0] - 2\,\Re\{\mathbf{h}^H \mathbf{r}_{xd}\} + \mathbf{h}^H \mathbf{R}_x \mathbf{h}.$$

Case 1: low eigenvalue dispersion

$$\mathbf{R}_x = \begin{bmatrix} 1.1 & 0.1 \\ 0.1 & 1.1 \end{bmatrix}, \qquad \mathbf{r}_{xd} = \begin{bmatrix} 0.5272 \\ -0.4458 \end{bmatrix}, \qquad r_x[0] = 0.9486$$

The optimal Wiener filter is

$$\mathbf{h}_{\text{opt}} = \mathbf{R}_x^{-1} \mathbf{r}_{xd} = \begin{bmatrix} 0.5204 \\ -0.4526 \end{bmatrix}.$$

The eigenvalues of $\mathbf{R}_x$ are

$$\lambda_1 = 1.2, \qquad \lambda_2 = 1.0,$$

which implies a well-conditioned error surface with nearly circular level curves and fast convergence.

Case 2: high eigenvalue dispersion

$$\mathbf{R}_x = \begin{bmatrix} 40 & 39 \\ 39 & 40 \end{bmatrix}, \qquad \mathbf{r}_{xd} = \begin{bmatrix} 0.5272 \\ -0.4458 \end{bmatrix}, \qquad r_x[0] = 0.9486$$

The optimal Wiener filter is

$$\mathbf{h}_{\text{opt}} = \begin{bmatrix} 0.487 \\ -0.486 \end{bmatrix}.$$

The eigenvalues are

$$\lambda_1 = 79, \qquad \lambda_2 = 1,$$

leading to a highly elongated error surface. In this case, gradient-based algorithms suffer from slow convergence and zig-zag trajectories.

Key takeaway. The eigenvalue spread of $\mathbf{R}_x$ determines the geometry of the MSE surface and directly impacts the convergence speed and stability of adaptive algorithms such as steepest descent or LMS.

7.6.3 Convergence and step-size in steepest descent

Convergence will be achieved whenever:

$$\lim_{k\to\infty} \mathbf{z}^{(k)} = 0 \iff \lim_{k\to\infty} \mathbf{h}^{(k)} = \mathbf{h}_{opt}$$

Then, since:

$$z_i^{(k+1)} = (1 - \mu\lambda_i) z_i^{(k)} = (1 - \mu\lambda_i)^2 z_i^{(k-1)} = \cdots = (1 - \mu\lambda_i)^{k+1} z_i^{(0)}$$

we get the condition:

$$\lim_{k\to\infty} z_i^{(k)} = 0 \iff |1 - \mu\lambda_i| < 1$$

so, the sufficient and necessary condition for convergence is:

$$\boxed{|1-\mu\lambda_i| < 1 \iff 0 < \mu < \frac{2}{\lambda_i} \implies 0 < \mu < \frac{2}{\lambda_{max}}}$$

since $\text{Tr}(\mathbf{R}) = \lambda_1 + \lambda_2 + \cdots + \lambda_{max}$, we define a new bounder:

$$0 < \mu < \frac{2}{Tr(\mathbf{R})} \leq \frac{2}{\lambda_{max}}$$

Speed of convergence, since $z_i^{(k+1)} = (1-\mu\lambda_i)^{k+1} z_i^{(0)}$ is then determined, for a given μ, by:

$$\max_{\forall\lambda_i} |1-\mu\lambda_i|$$

and the number of iterations necessary to converge, associated to the i-th component:

$$|1-\mu\lambda_i|^N = \varepsilon \implies N = \frac{\ln\varepsilon}{\ln|1-\mu\lambda_i|}$$

Then, since N depends intrinsically on μ, we will first select the value of μ that minimizes the convergence time:

$$\mu_{opt} = \arg\min_{\forall\mu}\{\max_{\forall i=1,\ldots,Q} |1-\mu\lambda_i|\}$$

$$= 1-\mu\lambda_{min} = -(1-\mu\lambda_{max}) \implies \boxed{\mu_{opt} = \frac{2}{\lambda_{min}+\lambda_{max}}}$$

with number of iterations:

$$N = \frac{\ln\varepsilon}{\ln\left(\frac{\lambda_{max}/\lambda_{min}-1}{\lambda_{max}/\lambda_{min}+1}\right)}$$

where we can notice here the dispersion ($\chi = \frac{\lambda_{max}}{\lambda_{min}}$) controls the number of iterations.

7.7 Adaptative implementation of the Wiener filter

7.7.1 Least Mean Square (LMS) and Normalized LMS (NLMS)

The Wiener filter can then be implemented using the Steepest Descent method (exact gradient) or the Least Mean Square (LMS) algorithm (stochastic estimate).

1. Exact Steepest Descent Method

When the statistical properties of the signals are known, the filter coefficients **h** are updated using the exact gradient of the Mean Square Error (MSE).

- **MSE Cost Function:** $\xi(\mathbf{h}^{(k)}) = E\left[|e[n]|^2\right] = P_d + \mathbf{h}^{(k)H}\mathbf{R}_x\mathbf{h}^{(k)} - \mathbf{h}^{(k)H}\mathbf{r}_{xd} - \mathbf{r}_{xd}^H\mathbf{h}^{(k)}$
- **Exact Gradient:** $\nabla_{\mathbf{h}}\xi(\mathbf{h}^{(k)}) = \mathbf{R}_x\mathbf{h}^{(k)} - \mathbf{r}_{xd} = E\left[\mathbf{x}[n]\mathbf{x}^H[n]\right]\mathbf{h}^{(k)} - E\left[\mathbf{x}[n]d^*[n]\right]$
- **Update Rule:** $\mathbf{h}^{(k+1)} = \mathbf{h}^{(k)} - \mu\nabla_{\mathbf{h}}\xi(\mathbf{h}^{(k)})$

2. Least Mean Square (LMS) Algorithm

The LMS algorithm provides an instantaneous estimate of the gradient when signal expectations are unavailable.

- **Instantaneous Gradient Estimate:** $\hat{\nabla}_{\mathbf{h}}\xi(\mathbf{h}^{(n)}) = \mathbf{x}[n]\mathbf{x}^H[n]\mathbf{h}^{(n)} - \mathbf{x}[n]d^*[n]$
- **Simplified Gradient (using error signal):** $\hat{\nabla}_{\mathbf{h}}\xi(\mathbf{h}^{(n)}) = \mathbf{x}[n]y^*[n] - \mathbf{x}[n]d^*[n] = -\mathbf{x}[n]e^*[n]$
- **LMS Update Equation:** $\mathbf{h}^{(n+1)} = \mathbf{h}^{(n)} + \mu\mathbf{x}[n]e^*[n]$

About the convergence of this method, because the LMS gradient is random, convergence is studied in the average sense.

- **Weight Error Equation:** Subtracting the optimal weights $\mathbf{h}_{opt}$ from both sides of the update rule:

$$\mathbf{h}^{(n+1)} - \mathbf{h}_{opt} = (\mathbf{I} - \mu\mathbf{x}[n]\mathbf{x}^H[n])(\mathbf{h}^{(n)} - \mathbf{h}_{opt}) + \mu\mathbf{x}[n]e^*_{opt}[n]$$

- **Independence Assumption:** We assume $\mathbf{x}[n]$ and $\mathbf{h}^{(n)}$ are approximately independent.
- **Orthogonality Principle:** The term $E[\mathbf{x}[n]e^*_{opt}[n]]$ is equal to 0 by the orthogonality principle of the optimum Wiener filter.
- **Average Convergence Rule:**

$$E[\mathbf{h}^{(n+1)} - \mathbf{h}_{opt}] = (\mathbf{I} - \mu\mathbf{R}_x)E[\mathbf{h}^{(n)} - \mathbf{h}_{opt}]$$

Convergence Conditions in the Mean Sense The stability and convergence of the LMS algorithm depend on the choice of the step size μ. The condition for convergence is identical to that of the steepest descent method:

$$0 < \mu < \frac{2}{\lambda_{\max}} \tag{1}$$

Where $\lambda_{\max}$ is the largest eigenvalue of the autocorrelation matrix $\mathbf{R}_x$.

A more conservative (practical) condition is often used to ensure stability without needing to calculate eigenvalues:

$$0 < \mu < \frac{2}{\text{trace}(\mathbf{R}_x)} = \frac{2}{Q \cdot r_x[0]} \leq \frac{2}{\lambda_{\max}} \tag{2}$$

In practice, a common choice is $\mu = \frac{2\alpha}{Q \cdot r_x[0]}$ where $0 < \alpha < 1$.

3. Normalized LMS (NLMS)

When the input signal is non-stationary, its dynamics change over time. To guarantee convergence, the step size must be updated dynamically.

- **Dynamic Step-size:** $\mu[n] = \frac{2\alpha}{Q \cdot \hat{r}_x[0;n]}$ with $0 < \alpha < 1$.
- **Instantaneous Power Estimation:** $Q \cdot \hat{r}_x[0;n] = \mathbf{x}^H[n]\mathbf{x}[n]$.
- **Time-averaged Estimation:** $\hat{r}_x[0;n] = \gamma\hat{r}_x[0;n-1] + (1-\gamma)|x[n]|^2$.

Avoiding Instabilities To avoid numerical saturation or instabilities when the estimated input power is very low, a small positive constant $P_{x,0}$ (minimum threshold) is added to the denominator:

$$\mu[n] = \frac{2\alpha}{P_{x,0} + Q \cdot \hat{r}_x[0;n]} \tag{3}$$

LMS

1. Gradient Implementation Methods

- **Exact Steepest Descent:** Uses the true statistical gradient.
  - **Cost Function:** $\xi(\mathbf{h}^{(k)}) = E\left[|e[n]|^2\right] = P_d + \mathbf{h}^{(k)H}\mathbf{R}_x\mathbf{h}^{(k)} - \mathbf{h}^{(k)H}\mathbf{r}_{xd} - \mathbf{r}_{xd}^H\mathbf{h}^{(k)}$.
  - **Exact Gradient:** $\nabla_{\mathbf{h}}\xi(\mathbf{h}^{(k)}) = \mathbf{R}_x\mathbf{h}^{(k)} - \mathbf{r}_{xd}$.
  - **Update Rule:** $\mathbf{h}^{(k+1)} = \mathbf{h}^{(k)} - \mu\nabla_{\mathbf{h}}\xi(\mathbf{h}^{(k)})$.
- **Least Mean Square (LMS):** Employs an instantaneous estimate of the gradient.
  - **Gradient Estimate:** $\hat{\nabla}_{\mathbf{h}}\xi(\mathbf{h}^{(n)}) = -\mathbf{x}[n]e^*[n]$.
  - **LMS Update Rule:** $\mathbf{h}^{(n+1)} = \mathbf{h}^{(n)} + \mu\mathbf{x}[n]e^*[n]$.

2. Convergence Study (Mean Sense)

Convergence is studied in statistical terms due to the randomness of the estimated gradient.

- **Independence Assumption:** We assume $\mathbf{x}[n]$ and $\mathbf{h}^{(n)}$ are approximately independent.
- **Mean Convergence Equation:** $E[\mathbf{h}^{(n+1)} - \mathbf{h}_{opt}] = (\mathbf{I} - \mu\mathbf{R}_x)E[\mathbf{h}^{(n)} - \mathbf{h}_{opt}]$.
- **Convergence Condition:** $0 < \mu < \frac{2}{\lambda_{\max}}$ or conservatively $0 < \mu < \frac{2}{\text{trace}(\mathbf{R}_x)}$.

3. Steady-State Regime

The covariance matrix of the coefficients $\mathbf{S}^{(n)} = E[(\mathbf{h}^{(n)} - \mathbf{h}_{opt})(\mathbf{h}^{(n)} - \mathbf{h}_{opt})^H]$:

- **Recursive Expression:** $\mathbf{S}^{(n+1)} = (\mathbf{I} - \mu\mathbf{R}_x)\mathbf{S}^{(n)}(\mathbf{I} - \mu\mathbf{R}_x) + \mu^2\xi_{\min}\mathbf{R}_x$.
- **Steady-State Approximation (where $\mathbf{S}^{(n)} = \mathbf{S}^{(n+1)}$):** If $\mu \ll \frac{1}{\lambda_{\max}}$, then

  $$\mathbf{S} = \frac{\mu}{2}\xi_{\min}\mathbf{I}$$

  .

4. Normalized LMS (NLMS)

Dynamic adaptation of the step-size to handle non-stationary input signals:

- **NLMS Step-size:** $\mu[n] = \frac{2\alpha}{P_{x,0} + \mathbf{x}^H[n]\mathbf{x}[n]}$ with $0 < \alpha < 1$.
- $P_{x,0}$ is a small threshold added to avoid numerical instabilities.

Misadjustment in LMS

1. The Steady-State Regime

In the LMS algorithm, filter coefficients are updated according to a random estimate of the gradient. The progress is tracked via the covariance matrix of the coefficients:

- **Definition:** $\mathbf{S}^{(n)} = E\left[(\mathbf{h}^{(n)} - \mathbf{h}_{opt})(\mathbf{h}^{(n)} - \mathbf{h}_{opt})^H\right]$.
- **Recursive Expression:** Assuming independence between $\mathbf{x}[n]$ and $\mathbf{h}_n$, the update is $\mathbf{S}^{(n+1)} = (\mathbf{I} - \mu\mathbf{R}_x)\mathbf{S}^{(n)}(\mathbf{I} - \mu\mathbf{R}_x) + \mu^2\xi_{\min}\mathbf{R}_x$.
- **Convergence:** When steady-state is achieved ($\mathbf{S}^{(n+1)} = \mathbf{S}^{(n)} = \mathbf{S}$), and assuming a small step size $\mu \ll \frac{1}{\lambda_{\max}}$, the matrix simplifies to $\mathbf{S} = \frac{\mu}{2}\xi_{\min}\mathbf{I}$.

2. Mean Value of the MSE

Because coefficients are non-deterministic, the Mean Square Error (MSE) is viewed as a mean value:

- **Instantaneous MSE:** $\xi(\mathbf{h}^{(n+1)}) = \xi_{\min} + (\mathbf{h}^{(n+1)} - \mathbf{h}_{opt})^H\mathbf{R}_x(\mathbf{h}^{(n+1)} - \mathbf{h}_{opt})$.
- **Mean MSE at Steady-State:** $E[\xi] = \xi_{\min} + \mathrm{trace}[\mathbf{R}_x\mathbf{S}] = \xi_{\min} + \frac{\mu}{2}\xi_{\min}Qr_x[0]$.

3. Misadjustment (M)

Misadjustment measures how far the final MSE is from the optimal Wiener solution ($\xi_{\min}$):

- **Formula:** $M = \frac{E[\xi]-\xi_{\min}}{\xi_{\min}} = \frac{\mu}{2}Qr_x[0] = \alpha$.
- **The Tradeoff:** The parameter α (where $\mu = \frac{2\alpha}{Qr_x[0]}$) directly controls the tradeoff between the accuracy of the final result (misadjustment) and the speed of convergence.