Estimation and Decision Theory

Wadih Sawaya and Giulia Cervia

IMT Nord Europe

September 23, 2025

Contents

- **Estimation Theory**
- **Detection: Hypothesis Testing**

1. Estimation Theory

Estimation theory- Introduction

- Given a sequence of real (or complex) values $\boldsymbol{x} = (x_1,\ x_2, \ldots,\ x_n)$ related to an unknown quantity θ or to an unknown vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ we seek to determine an estimation of θ or $\boldsymbol{\theta}$.
  - θ or $\boldsymbol{\theta}$ may be deterministic or random quantities. In this part we mainly focus on **deterministic** values or vectors, **unknown** from the observer .
    - We will generally adopt vector notation so that a scalar parameter is a one-component vector, unless we specify that it is indeed a scalar parameter.
  - Elements of vector $\boldsymbol{\theta}$ are the parameters of the estimation problem. The observer wishes to estimate them when observing $\mathbf{x}$.
  - Why do we talk about "estimating" the elements of $\boldsymbol{\theta}$ rather than computing their exact value?

Estimation theory- Introduction

- The answer is quite simple: the dependence of $\mathbf{x}$ on $\boldsymbol{\theta}$ is stochastic and $\mathbf{x}$ is an outcome of a random vector obeying this dependence.
- As a result, it is often very useful to consider a probabilistic model, and make use of the probability density function $p_{\boldsymbol{\theta}}(\mathbf{x})$ to propose an estimator for $\boldsymbol{\theta}$.

- However, in many situations this model is impossible to derive. The link between $\mathbf{x}$ and its parameters is rather made through another sequence, the source sequence $\mathbf{s}$.

    - This source sequence is a function of the parameters, and $\mathbf{x}$ is a noisy version of $\mathbf{s}$:

$$\mathbf{x} = \mathbf{s}(\boldsymbol{\theta}) + \mathbf{w}$$

Estimation theory- Introduction

- To illustrate these points, let's consider a few examples.
- In industrial engineering, exponential laws are widely used in the study of system reliability in order to control the risk of early failures, or the lifetime of a product.
  - Let $\mathbf{x}$ represents the lifetimes of a batch of n experimental items.
    - Assuming $p_\lambda(x) = \lambda \exp(-\lambda x)$, $x \geq 0$,ans $\lambda > 0$, the law of lifetime of an item, we seek to estimate λ after measuring $\mathbf{x}$ in order to control production and improve the environmental balance sheet.

Estimation theory- Introduction

- In communications systems and in automation engineering, signals are generally affected by Gaussian noise.
  - We assume that at a given instant $x_k = A + w_k$, where A is a fixed unknown amplitude and w_k a zero mean Gaussian noise with power P_N .
    - The probability density function of x_k is then:

$$p_{\boldsymbol{\theta}}(x_k) = \frac{1}{\sqrt{2\pi P_N}} \exp\left(-\frac{1}{2P_N}(x_k - A)^2\right).$$

    $\boldsymbol{\theta} = (A,\ P_N)$ and the receiver wishes to estimate them.

Estimation theory- Introduction

- Data engineers dealing with large amounts of data observe frequently their data after they has undergone distortions.
  - These distortions can either create interference between data or contrarily misses some of them making a parsimonious model to study. These distortions may often be represented linearly involving a known matrix $\mathbf{H}$ such that:

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}.$$

    The data engineer seeks to estimate $\boldsymbol{\theta}$ even without any information about the noise $\mathbf{w}$.

Estimation theory- Introduction

- Firstly, we will introduce the method of moment which is the most intuitive estimator based on arithmetic means.
- We will discuss the idea of “good” estimator and introduce the Cramér Rao Bound, a lower bound on performance.
- We present then the Maximum Likelihood estimator.
- We will end this part dedicated to estimation of a “deterministic and unknown” parameter with the least squares estimator.
- A discussion on Bayesian estimators, i.e., when the parameters are random variables, will conclude this section.

Parameter estimation

Context

- We observe a sequence $\mathbf{x} = (x_1,...,x_n)$:
  - In many situations, an observed **sample x** is considered: A sample is a vector of independent random variables X_i governed by the same probability density $p_{\boldsymbol{\theta}}(x)$. In this case we have:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \prod_{k=1}^{n} p_{\boldsymbol{\theta}}(x_k). \tag{1}$$

  - We recall that the parametric density function $p_{\boldsymbol{\theta}}(\mathbf{x})$ is known, but $\boldsymbol{\theta}$, the vector of K parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)$ is unknown.
    - the problem is to estimate $\boldsymbol{\theta}$ when observing $(x_1,...,x_n)$.
    - the estimated vector is $\hat{\boldsymbol{\theta}}$.

The method of moments (1)

- We will discuss first the method of moments for estimating $\boldsymbol{\theta}$ which, in our opinion, is the most intuitive one.
  - **Definition:** The $\ell-$th moment of a random variable X with density function $p_{\boldsymbol{\theta}}(x)$ is:

$$M_\ell = \mathbb{E}\left[X^\ell\right] = \int_{-\infty}^{\infty} x^\ell p_{\boldsymbol{\theta}}\left(x\right) dx. \tag{2}$$

  - Let μ and σ^2 be the mean and the variance of X respectively. We have:

$$\begin{aligned} M_1 &= \mu, \\ M_2 &= \sigma^2 + \mu^2. \end{aligned}$$

The method of moments (2)

- The method of moments is based on calculating the arithmetic mean of the observed $\mathbf{x}$ where each of its component is at power ℓ:

$$\hat{M}_\ell(n) = \frac{1}{n}\sum_{i=1}^{n} x_i^\ell. \tag{3}$$

  - Consider an outcome sequence $\mathbf{x} = (x_1,...,x_n)$ of a sample from a distribution with mean μ. The estimated 1st order moment is the sample mean:

$$\hat{M}_1(n) = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{4}$$

  - An estimator of the parameter μ is then simply $\hat{\mu}(n) = \hat{M}_1(n)$.

The method of moments (3)

- Numerical example: $\mu = 3$, $n = 10$ and $\mathbf{x} = (2.79, 2.42, 4, 6.04, 1.41, 3.15, 1.26, 4.56, 4.69, 1)$.

$$\hat{\mu}(10) = 3.132$$

  - Consider now $n = 1000$ observed data. An estimated mean for these data is $\hat{\mu}(1000) = 3.03$

The method of moments (3)

- This is not surprising indeed: from the strong law of large number we have:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s} \mu = \mathbb{E}[X] \tag{5}$$

  - Almost surely convergence: $\mathbb{P}\left(\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mu\right) = 1.$
    - In other words, the probability that a "sample mean" with infinity size will not be equal to μ, is zero.

The method of moments (4)

- Recall the definition of a moment of order ℓ :

$$M_\ell = \mathbb{E}\left[X^\ell\right] = \int_{-\infty}^{\infty} x^\ell p_{\boldsymbol{\theta}}(x)\, dx. \tag{6}$$

  - M_ℓ is in fact a function of $\boldsymbol{\theta}$ so that we can write: $M_\ell = g_\ell(\boldsymbol{\theta})$.
  - For a model with K parameters we compute K moments $M_\ell = g_\ell(\boldsymbol{\theta})$:

$$\mathbf{M} = \begin{bmatrix} M_1 \\ \vdots \\ M_K \end{bmatrix} = \begin{bmatrix} \mathbb{E}[X] \\ \vdots \\ \mathbb{E}\left[X^K\right] \end{bmatrix} = \begin{bmatrix} g_1(\boldsymbol{\theta}) \\ \vdots \\ g_K(\boldsymbol{\theta}) \end{bmatrix} = \mathbf{g}(\boldsymbol{\theta}). \tag{7}$$

  - Find $\mathbf{g}^{-1}$ such that $\boldsymbol{\theta} = \mathbf{g}^{-1}(\mathbf{M})$.

The method of moments (5)

- Replace all M_ℓ by their $\ell-$th order sample mean $\hat{M}_\ell$ to get an estimate of $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\theta}} = \mathbf{g}^{-1}\left(\hat{\mathbf{M}}\right) = \mathbf{g}^{-1}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_i^K \end{pmatrix}. \tag{8}$$

  - Important remark: the choice of the first K moments is not always suitable for solving the latter system of equations.
    - We have to find a set of K moments allowing (8) to be solved.
    - However, error of the moment estimator increases with order. In consequence, we choose the set of K moments by ascending order.
    - The notion of estimator's error will be discussed later.

The method of moments (6)

- Example 1: let $p_\theta(x) = \frac{\theta}{x^{\theta+1}}$, $\theta > 1$, $x \geq 1$ (the Pareto law with parameter θ).
  - Here $K = 1$ with $M_1 = \frac{\theta}{\theta-1}$ then $\theta = \frac{\mu}{\mu-1}$.
    - $M_1 = g(\theta)$ and $\theta = \frac{M_1}{M_1-1} = g^{-1}(M_1)$
  - From an observed sample $(x_1, ..., x_n)$ we compute the sample mean $\hat{M}_1(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$.
  - The estimated parameter θ is then:

$$\hat{\theta}(n) = \frac{\hat{M}_1(n)}{\hat{M}_1(n) - 1}.$$

The method of moments (7)

- **Exercice 1**: Let us consider an observed sequence $(2.8,\ 2.2,\ 2.7,\ 2.3)$ from a Gaussian sample with known variance $\sigma^2 = 1$ but unknown mean $\mu > 0$.
  1. Compute M_1 and M_2.
  2. Let $M_1 = g_1(\mu)$ and $M_2 = g_2(\mu)$ such that $\mathbf{M} = \mathbf{g}(\mu)$. Give two expressions for $\mu = g_i^{-1}(M_i)$, $1 \leq i \leq 2$.
  3. Write two estimators of μ based on these two functions.
  4. Which one should we use ?

1. $M_1 = \mu$; $M_2 = 1 + \mu^2$
2. $\mu = g_1^{-1}(M_1) = M_1$, $\mu = g_2^{-1}(M_2) = \sqrt{M_2 - \sigma^2}$.
3. $\hat{\mu_1}(n) = \hat{M}_1 = \frac{1}{n}\sum_{k=1}^{n} x_k$, $\hat{\mu_2}(n) = \sqrt{\frac{1}{n}\sum_{k=1}^{n} x_k^2 - \sigma^2}$.
4. The estimation error of the moment estimator increases with order. In consequence we choose $\hat{\mu_1}(n) = \frac{1}{4}(2.8 + 2.2 + 2.7 + 2.3) = 2.5$.

The method of moments (8)

- **Exercice 2**: In communication systems noise is added to the signal. This noise has a Gaussian probability density function. We want to estimate the DC component of the noise and its variance. The distribution of the noise is $p_{\boldsymbol{\theta}}(x) = \frac{1}{\sqrt{2\pi\theta_2}} \exp\left(-\frac{1}{2\theta_2}(x-\theta_1)^2\right)$, with unknown parameters θ_1 and θ_2.
  1. Express $M_1$and M_2 with respect to θ_1 and θ_2.
  2. Let $M_1 = g_1(\boldsymbol{\theta})$ and $M_2 = g_2(\boldsymbol{\theta})$ such that $\mathbf{M} = \mathbf{g}(\boldsymbol{\theta})$. Express $\boldsymbol{\theta} = \mathbf{g}^{-1}(\mathbf{M})$.
  3. From an observed sample $(x_1, ..., x_n)$ write the sample mean of order 1 and 2.
  4. What is the estimate of $\boldsymbol{\theta}$?

     1. $M_1 = \theta_1$.and $M_2 = \theta_2 + \theta_1^2$. We write $\boldsymbol{\theta} = (\theta_1, \theta_2)$.
     2. $\theta_1 = M_1$ and $\theta_2 = M_2 - M_1^2$. $\boldsymbol{\theta} = \left(M_1, M_2 - M_1^2\right)$.
     3. $\hat{M}_1(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\hat{M}_2(n) = \frac{1}{n}\sum_{i=1}^{n} x_i^2$.
     4. $\hat{\theta}_1(n) = \hat{M}_1(n)$ and $\hat{\theta}_2(n) = \hat{M}_2(n) - \hat{M}_1^2(n)$.

The method of moments (9)

- **Exercice 3:** Let us consider an observed sequence of white and Gaussian noise $\mathbf{x}$ with unknown variance .
  1. Recall the definition of a white noise ?
  2. What is the consequence of the whiteness of the Gaussian noise on the observed sequence $\mathbf{x}$?
  3. White noise has zero mean. In many applications, like in a communication system we wish to estimate the power of the noise. Propose an estimator for this noise power.

1. White sequence has uncorrelated samples.
2. Whiteness of a Gaussian sequence implies independent samples. The sequence $\mathbf{x}$ is then an outcome of a Gaussian sample $(X_1, ..., X_n)$.
3. $M_1 = 0$ and $M_2 = \sigma^2$. The power is of a real stochastic process is $E\left[X^2\right] = M_2$.

We then need to estimate M_2: $\hat{M}_2 = \frac{1}{n} \sum_{k=1}^{n} x_i^2$.

Parameter estimation

The method of moments (10)

- **Exercise 4:** A system verifies the authenticity of documents by inspecting the distribution of a gray level of codes printed on them. A fake document can cause greater dispersion of the code's gray level around an average value. Several codes are read. Some codes undergo this spreading of gray levels, while others do not. The observed sequence then contains two modes with different proportions. We want to estimate this proportion.

  Consider the Gaussian mixture hereafter where μ, σ_1, σ_2 are known parameters but the proportion a is unknown. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be an outcome from the sample $(X_1, \ldots, X_n)$ where $X_i \sim p_\theta$:

$$p_\theta(x) = a\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma_1^2}\right) + (1-a)\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma_2^2}\right).$$

  - Find an estimator for a based on the observation of $\mathbf{x}$.

Parameter estimation

- **Solution for exercise 4:**
  $M_1 = a\mu + (1-a)\mu = \mu$ is a constant and doesn't depend on a.
  The moment of order 1 is not appropriate for estimating a.

  $$M_2 = a\left(\sigma_1^2 + \mu^2\right) + (1-a)\left(\sigma_2^2 + \mu^2\right) = a\left(\sigma_1^2 - \sigma_2^2\right) + \sigma_2^2 + \mu^2.$$

  We can conclude that $a = g^{-1}(M_2) = \frac{M_2 - \sigma_2^2 - \mu^2}{\sigma_1^2 - \sigma_2^2}$.

  An estimator of a is then $\hat{a} = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \sigma_2^2}{\sigma_1^2 - \sigma_2^2}$.

The method of moments (11)

- **Exercise 5:** In a radar system, several echoes from a target can be observed at different times. This received signal is corrupted by White Gaussian and additive noise. When there is no target, only noise is present, whereas in a presence of a target, the signal has an unknown amplitude μ added to the noise. The received signal therefore has two modes..

  Consider the Gaussian mixture hereafter where σ is a known parameter but a and μ are unknown. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be an outcome from the sample $(X_1, \ldots, X_n)$ wher $X_i \sim p_{\boldsymbol{\theta}}$

$$p_{\boldsymbol{\theta}}(x) = a\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right) + (1-a)\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

  1. Find an estimator for a and μ based on the observation of $\mathbf{x}$.

Parameter estimation

- **Solution for exercise 5:**

$M_1 = (1-a)\,\mu$

$M_2 = a\sigma^2 + (1-a)\,(\sigma^2+\mu^2) = \sigma^2 + (1-a)\,\mu^2$

$\mathbf{M} = (M_1,\, M_2) = ((1-a)\,\mu,\, \sigma^2 + (1-a)\,\mu^2)$

$a = 1 - \frac{M_1^2}{M_2 - \sigma^2}$

$\mu = \frac{M_2 - \sigma^2}{M_1}$

Replacing M_1 andM_2 by their sample mean we have:

$\hat{a} = 1 - \frac{\left(\frac{1}{n}\sum x_i\right)^2}{\frac{1}{n}\sum x_i^2 - \sigma^2}$ and $\hat{\lambda} = \frac{\frac{1}{n}\sum x_i^2 - \sigma^2}{\frac{1}{n}\sum x_i}$.

The method of moments (12)

- **Exercice** 6: Devices from an automotive industry are tested under stress conditions and operate at their optimal performance to estimate the levels of their potential defects and estimate the parameter of a lifetime prediction model.
  Several models are available. Let us take for this exercise the simplest one, that is the exponential law.
  For $t \geq 0$ and $\lambda > 0$ this law is $p_\lambda(t) = \lambda \exp(-\lambda t)$.
  1. Based on $\mathbf{t} = (t_1, ..., t_n)$ an outcome of a sample, give an estimate for λ.

     $M_1 = \frac{1}{\lambda}$, $\lambda = \frac{1}{M_1}$.
     $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i}$.

The method of moments (13)

- **Exercise 7:** Let X be a uniform random variable in the interval $[0,\ \theta]$.

  1. Find an estimator for based on the observation of an outcome of a sample $(X_1, ..., X_n)$.

     $M_1 = \frac{\theta}{2}$,then $\theta = 2M_1$ and $\hat{\theta} = \frac{2}{n}\sum_{i=1}^{n} x_i$

  - Remark: Thanks to weak law of large numbers applied on iid sequences, the method of moments achieves good performance when n, the length of the observed sequence $\mathbf{x}$, is large enough. It is then crucial to observe **samples** for this method of estimation.

Parameter estimation

Formal characterization of good estimators (1)

- An estimator is a function that associates each point in the n-dimensional observation space a point in the K-dimensional parameter space:

$$\hat{\boldsymbol{\theta}}(n) = T\left(x_1, ..., x_n\right). \tag{9}$$

  - $T\left(.\right)$ is called a statistic: it is a function of the sequence of random variables $(X_1, ..., X_n)$.
    - $T\left(\cdot\right)$ is a $K-$ dimensional random vector $\hat{\Theta}(n)$:

$$\hat{\Theta}(n) = T\left(X_1, ..., X_n\right) \tag{10}$$

    - $\hat{\Theta}$ is an estimator of $\boldsymbol{\theta}$ obtained with the statistic $T(.)$ (we simplify notations by dropping n).

Formal characterization of good estimators (2)

- Definition 1 (unbiased estimator): An unbiased estimator has:

$$\mathbb{E}\left[\hat{\Theta}\right] = \boldsymbol{\theta}. \tag{11}$$

  - Recall the sample mean for estimating the mean value μ of a sample $\mathbf{x}$ and consider a second estimator $\hat{\theta}_2 = x_1$.

$$\begin{aligned} \hat{\theta}_1 &= \frac{1}{n}\sum_{i=1}^{n} x_i, \\ \hat{\theta}_2 &= x_1. \end{aligned}$$

    - $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are both unbiased estimators, i.e. $E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu.$ and $E[X_1] = \mu.$

Formal characterization of good estimators (3)

- But even if $\frac{1}{n}\sum_{i=1}^{n} x_i$ =0.89 and $x_1 = 0.95$ for an exact value of $\mu = 1$, we are intuitively inclined to trust $\hat{\theta}_1$rather than $\hat{\theta}_2$.
  - And we are right because, as we will see, $var(\hat{\Theta}_1) < var(\hat{\Theta}_2)$.
    - $var(\hat{\Theta})$ is the square error of the estimator with respect to the (average) estimated value. If the estimator is unbiased this error will be the error with respect to the exact value of the parameter.
    - $\hat{\Theta}_1$ deviations around μ is less than deviations of $\hat{\Theta}_2$ around μ : $var(\hat{\Theta}_1) = \frac{\sigma^2}{n}$ and $var(\hat{\Theta}_2) = \sigma^2$,where σ^2 is the variance of X.
    - Being unbiased is not sufficient to be a “good” estimator.

Formal characterization of good estimators (4)

- As another example let us consider the estimation of μ and σ^2, the mean and the variance of an outcome $\mathbf{x}$ of a Gaussian sample. We know already that the method of moments estimator is:

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i, \\ \hat{\theta}_2 = \widehat{(\sigma^2)} &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \hat{\mu}^2. \end{aligned}$$

Formal characterization of good estimators (5)

- The sample mean estimator is unbiased but $\hat{\Theta}_2$ as expressed above isn't:

$$\begin{aligned}
E\left[\hat{\theta}_2 - \sigma^2\right] &= E\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2\right] - E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] - \sigma^2, \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2\right] - \frac{1}{n}E\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2\right] - \frac{1}{n}E\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}^{n} x_i x_j\right] - \sigma^2, \\
&= \left(\sigma^2 + \mu^2\right) - \frac{1}{n}\left(\sigma^2 + \mu^2\right) - \frac{1}{n}\left((n-1)\mu^2\right) - \sigma^2, \\
&= -\frac{\sigma^2}{n}.
\end{aligned}$$

  - We note that the bias $b(\sigma^2) = E\left[\hat{\theta}_2 - \sigma^2\right] = -\frac{\sigma^2}{n}$ goes to zero as n goes towards infinity. The figure below illustrates this asymptotic unbiasedness.

Formal characterization of good estimators (6)

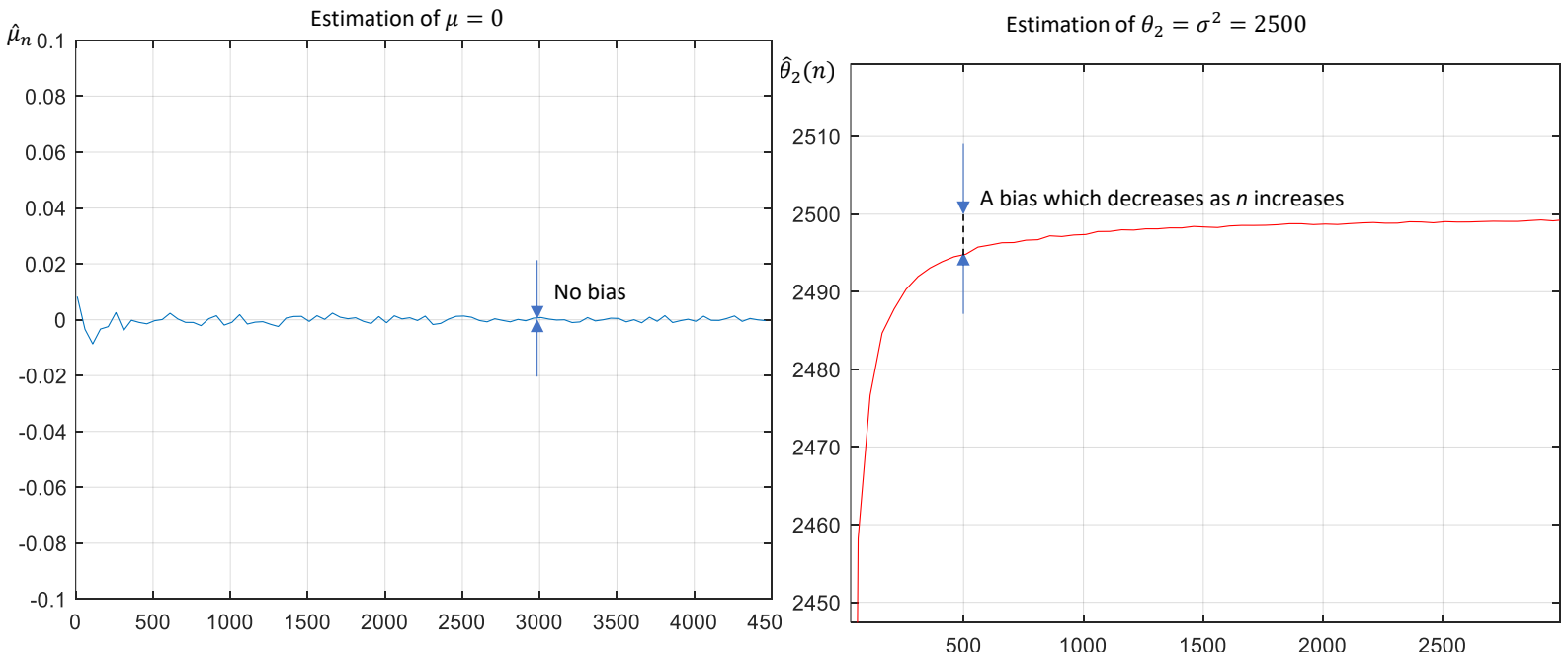


Figure: Method of moments for estimating the mean and the variance: $\mu = \theta_1$ and $\sigma^2 = \theta_2$.

Parameter estimation

Formal characterization of good estimators (7)

- Their exists an estimator for $\theta_2 = \sigma^2$ without a bias when averaging over $n-1$ instead of n :

$$\check{\theta}_2(n) = \frac{1}{n-1}\left[\left(\sum_{i=1}^{n} x_i^2\right) - \left(n\hat{M}_1^2(n)\right)\right].$$

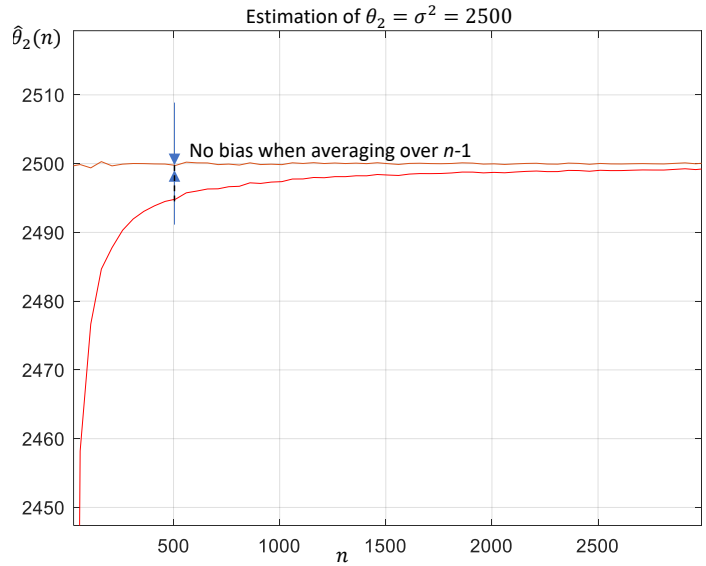


Figure: Estimating the variance by averaging over $n-1$

Formal characterization of good estimators (8)

- Theorem 1: this theorem expresses a lower bound on the estimation error. For a scalar parameter, any unbiased estimator has:

$$var(\hat{\Theta}) = E\left[\hat{\Theta}^2\right] - \left(E\left[\hat{\Theta}\right]\right)^2 \geq CRB \tag{12}$$

  where CRB = Cramér-Rao Bound is a positive scalar and will be expressed later.

  - CRB is a positive scalar for one parameter and a $K \times K$ matrix in case of K parameters.
    - for the latter case the ordered relation «$\geq$» must be interpreted as «$var(\hat{\Theta}) - CRB$ is positive semi-definite», and $var\left(\hat{\Theta}\right)$ is the covariance matrix of the estimator.

Formal characterization of good estimators (9)

- An unbiased scalar parameter estimator with $var(\hat{\Theta}) = CRB$ is an **efficient estimator**.
  - Consider two different unbiased estimators $\hat{\Theta}_1 = T_1(X_1, ..., X_n)$ and $\hat{\Theta}_2 = T_2(X_1, ..., X_n)$. The smaller the variance the better the estimator. $\hat{\Theta}_1$ is then preferred if :

$$var(\hat{\Theta}_1) < var(\hat{\Theta}_2).$$

Parameter estimation

Formal characterization of good estimators (10)

- Definition 2 (consistent or convergent estimator): An estimator is said to be convergent if $\hat{\Theta}(n) \underset{n\to\infty}{\overset{P}{\to}} \theta$.
  - convergence in probability: "the probability that the estimated value differs from the true value with an arbitrary amount $\epsilon > 0$ goes to zero as n increases" or $\lim_{n\to\infty} P\left(\left|\hat{\Theta}(n) - \theta\right| \geq \epsilon\right) = 0 \; \forall \epsilon > 0.$
    - Property for a convergent unbiased estimator: $\lim_{n\to\infty} var(\hat{\Theta}(n)) = 0$. The estimation error goes to zero as n increases.
    - Example : the first order moment estimator obtained by sample mean is convergent. For this estimator we have the following:

$$\begin{aligned} \mathbb{E}\left[\hat{\Theta}(n)\right] &= \mu, \Rightarrow \text{unbiased} \\ var(\hat{\Theta}(n)) &= \frac{\sigma^2}{n} = CRB \Rightarrow \text{efficient} \\ \lim_{n\to\infty} var(\hat{\Theta}(n)) &= 0 \Rightarrow \text{consistent.} \end{aligned}$$

Formal characterization of good estimators (11)

- Let us express now the Cramér-Rao bound (CRB).
  - Estimation is a result of a statistic, i.e. a function applied on observed data. These data in turn are closely related, via the distribution law, to the parameter we wish to estimate.
  - The more information the data contains about the parameter, the more accurate the estimator will be. This information can be extracted from the distribution law (the probability density function) p_θ.
  - Ronald Fisher (1890 – 1962) brought the idea of quantifying the sensitivity of the distribution law to parameter variability, and introduce his Fisher Information.

Parameter estimation

Formal characterization of good estimators (12)

- The distribution law $p_\theta(x)$ is viewed as a function of θ with x fixed.
- When a change in θ causes a noticeable change in $p_\theta(x)$, outcomes forming the observed sequence will naturally be affected by this change. The information content in the observed sample about this parameter is high.
- Conversely, if a change in θ has very little effect on $p_\theta(x)$, then the observed sequences contain very little information about the parameter.

Fisher Information (1)

- Define a score $S(X;\theta) = \frac{\partial \ln p_\theta(X)}{\partial \theta}$ to study the variations of the distribution law with respect to θ. Notice that for each value of θ the score is a random variable. Fisher Information measure the variance of this random variable.
  - The mean of this score is $\mathbb{E}\left[\left(\frac{\partial \ln p_\theta(X)}{\partial \theta}\right)\right] = 0$.
    - Its variance is the Fisher Information $I(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln p_\theta(X)}{\partial \theta}\right)^2\right]$.
    - We traditionally note $I(\theta)$ but it is not necessarily dependent on θ. $I(\theta)$ should be understood as " the Information about θ" .
    - One can show that $I(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln p_\theta(X)}{\partial \theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \ln p_\theta(X)}{\partial \theta^2}\right]$. Which is easier to calculate.

Fisher Information (2)

- In case of a scalar parameter θ, Fisher Information is a scalar but for $K > 1$ it is a $K \times K$ matrix.
  - for a scalar parameter θ, Fisher Information for a sample of n random variables is:

$$I_n(\theta) \quad = \quad nI(\theta) \tag{13}$$

  - The Cramér-Rao Bound is:

$$CRB = \frac{1}{I_n(\theta)} \tag{14}$$

  - Case for K parameters: The score is a random vector
    - Its mean is the K column zeros vector,
    - Fisher's information matrix is its co-variance matrix.

Maximum Likelihood Estimator - MLE (1)

- The method of moment is easy and intuitive.
  - but this advantage may lack optimality .
  - The most popular estimator for iid outcomes of random variables is ML estimator.
    - In case of $K = 1$, «if an efficient estimate exists, it is the ML estimator... [ ], (the properties of ML) provide some motivation for using it even when an efficient estimate does not exist».
    - Harry L. Van Trees in «Detection, Estimation, and Modulation Theory», J. Wiley and Sons Inc.
  - ML estimator uses p_θ to perform estimation.

Maximum Likelihood Estimator - MLE (2)

- Recall that for an outcome $(x_1, ..., x_n)$ of a sample we have for a scalar parameter $\theta \in \Omega$:

$$p_\theta(x_1, ..., x_n) = \prod_{i=1}^{n} p_\theta(x_i) \tag{15}$$

  - where $p_\theta(x_i)$ is the marginal distribution .
  - Definition 3: The log-likelihood function is a function of θ based on the observed sample $(x_1, ..., x_n)$:

$$\mathcal{L}(\theta\,|x_i) \quad = \quad \ln\left(p_\theta(x_i)\right) \tag{16}$$

  - Writing concisely $(x_1, ..., x_n) = \mathbf{x}$ we have:

$$\mathcal{L}(\theta\,|\mathbf{x}) \quad = \quad \sum_{i=1}^{n} \mathcal{L}(\theta\,|x_i).$$

Maximum Likelihood Estimator - MLE (3)

- Maximum Likelihood estimator is the solution of maximizing $\mathcal{L}(\theta \,|\mathbf{x})$ in the range $\theta \in \Omega$:

$$\hat{\theta}_{ML} = \arg \left\{ \max_{\theta \in \Omega} \left(\mathcal{L}(\theta \,|\mathbf{x}) \right) \right\} \tag{17}$$

  where Ω is the parameter space.

  - ML estimator is then the most likely parameter in Ω regarding the observed data.

Maximum Likelihood Estimator - MLE (4)

- **Exercise 8:**
  1. Compute the ML estimator of the mean of a Gaussian distribution $\theta \in \mathbb{R}$ when observing an outcome $\mathbf{x}$ of a sample $(X_1, \ldots, X_n)$.
  2. Is it unbiased ?
  3. Is it an efficient estimator ?

  - We have $\hat{\theta}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$.
  - It is unbiased.
  - Its variance is $var(\hat{\Theta}_{MV}) = \frac{\sigma^2}{n}$.
  - The Fisher Information about the mean μ is $I_n(\mu) = \frac{n}{\sigma^2}$.
    - $CRB = \frac{\sigma^2}{n}$ and $var(\hat{\Theta}_{MV}) = CRB$
    - The sample mean and ML estimators of the mean of a Gaussian distribution are then efficient.
  - Note about the Fisher Information value $\frac{n}{\sigma^2}$:
    - The greater n is, the more information about the mean is available
    - The smaller σ^2, the greater the information about the mean.

Maximum Likelihood Estimator - MLE (5)

- **Exercise 9** : The Poisson distribution is usually used to model the law of manufacturing defects of one product. A manufacturing entity collects all its defective products that have failed within a month of their manufacture which happens the 1st day of the month.
  - Based on these samples, we seek to estimate the parameter of the Poisson distribution to infer the "probability that k items are defective at manufacture". It should be a rare event, the Poisson distribution fits well this law.
  - Let $k_i \in \mathbb{N}$ be the number of defective products collected within the $i-$th month and $\mathbf{k} = (k_1, ..., k_n)$ data collected over n months.
  - The Poisson distribution is defined as follows ($\lambda > 0$) :

$$P_\lambda(X = k) = \frac{\lambda^k \exp(-\lambda)}{k!}.$$

1. Compute the ML estimator of $\lambda > 0$.
2. Is it unbiased ?
3. Is it an efficient estimator ?

Maximum Likelihood Estimator - MLE (6)

- **Solution to exercise 9:**
  - The k_i are obtained from a sample (iid data): $\hat{\lambda}_{n,MV} = \frac{1}{n}\sum_{i=1}^{n} k_i$, since $k_i \in \mathbb{N}$ then $\frac{1}{n}\sum_{i=1}^{n} k_i \geq 0$.
  - $\mathbb{E}\left[\hat{\Lambda}_{n,MV}\right] = \lambda$, it is unbiased .
  - $var\left(\hat{\Lambda}_{n,MV}\right) = \frac{\lambda}{n}$,
  - The Fisher Information is: $I_n(\lambda) = \frac{n}{\lambda}$, and the $CRB = \frac{1}{I_n(\lambda)} = \frac{\lambda}{n}$
  - $var\left(\hat{\Lambda}_{n,MV}\right) = CRB$, the estimator is then efficient.
    - Note that this estimator is also convergent.

Maximum Likelihood Estimator - MLE (7)

- **Exercise 10:** Let $\mathbf{k}$, $k_i \in \mathbb{N} \leq N$, be an observed sequence of a sample $(K_1, ..., K_n)$, from a Binomial distribution with parameters (N, p).
  1. Compute the ML estimator of $p \in [0, 1]$,N is known.
  2. Is it unbiased ?
  3. Is it an efficient estimator ?
     - $\hat{p}_{n,MV} = \frac{1}{nN} \sum_{i=1}^{n} k_i \leq \frac{nN}{nN} = 1.$
     - $\mathbb{E}\left[\hat{P}_{n,MV}\right] = p$, it is unbiased .
     - $var\left(\hat{P}_{n,MV}\right) = \frac{1}{nN} p(1-p),$
     - The Fisher Information is: $I_n(p) = \frac{nN}{p(1-p)}$, and the $CRB = \frac{1}{I_n(p)} = \frac{p(1-p)}{nN}$
     - $var\left(\hat{P}_{n,MV}\right) = CRB$, the estimator is then efficient.
  - Note that this estimator is also convergent.

Parameter estimation

- **Exercise 11:**
  1. Compute the ML estimator of the parameter λ of the exponential distribution modeling the lifetime law of an item, based on the observation of an outcome $\mathbf{t}$ of a sample $(T_1, \quad , T_n)$. We have for $\lambda > 0$:
  $$p_\lambda(t) = \lambda \exp(-\lambda t)\, I_{[0,\,\infty[}(t). \tag{18}$$
  2. Is this estimator unbiased ?

Parameter estimation

- **Solution to exercise 11:**

  For $t_i \geq 0$, $\mathcal{L}(\lambda\,|\mathbf{t}) = Log\left(\prod_{i=1}^{n} \lambda \exp\left(-\lambda t_i\right)\right) = Log\lambda^n - \sum_{i=1}^{n} \lambda t_i$,

  $= nLog\lambda - \lambda\sum_{i=1}^{n} \lambda t_i$

  $\frac{\partial \mathcal{L}(\lambda|\mathbf{t})}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i$

  $\left.\frac{\partial \mathcal{L}(\lambda|\mathbf{t})}{\partial \lambda}\right|_{\hat{\lambda}_{ML}} = 0$, then $\hat{\lambda}_{ML} = \frac{n}{\sum_{i=1}^{n} t_i}$

  This estimator is biased. From Jensen Inequality for convex functions

  $nE\left[\frac{1}{\sum_{i=1}^{n} T_i}\right] \geq \frac{n}{E\left[\sum_{i=1}^{n} T_i\right]}$,

  Since for exponential distribution $E\left[T\right] = \frac{1}{\lambda}\Big]$, $E\left[\hat{\Lambda}_{ML}\right] \geq \frac{n}{n\frac{1}{\lambda}}$ and $E\left[\hat{\Lambda}_{ML}\right] \geq \lambda$.

Parameter estimation

- **Exercise 12:** Through this exercise, we want to propose a simple approach to ML when the parameter belongs to a given parameter space. The solution we will propose will be pragmatic, and we will therefore not address the problem of optimization with constraints, which is the general way to find the ML estimator in this context.

1. Compute the ML estimator of the mean of a Gaussian distribution $\theta \leq \theta_0$ when observing an outcome $\mathbf{x}$ of a sample $(X_1, \ldots, X_n)$.

- **Solution to exercise 12:**
- The constrained optimization problem is to find $\hat{\theta}_{ML}$ solution of:

$$\max_{\theta \leq \theta} \mathcal{L}(\theta \,|\mathbf{x})$$

$\frac{\partial \mathcal{L}(\theta|\mathbf{x})}{\partial \theta} = 0$ has its solution without any constraint which is $\frac{1}{n}\sum_{i=1}^{n} x_i$ we compare this quantity to θ_0:

$$\begin{aligned} \text{if} \quad & \tfrac{1}{n}\sum_{i=1}^{n} x_i \leq \theta_0 \Rightarrow \hat{\theta}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i, \\ \text{if} \quad & \tfrac{1}{n}\sum_{i=1}^{n} x_i > \theta_0 \Rightarrow \hat{\theta}_{ML} = \theta_0. \end{aligned}$$

Comment on optimality - Minimum Variance Unbiased Estimator

- Recall again the sample mean estimator of a scalar parameter θ. Let $\hat{\Theta}_1$ be this estimator:

$$\hat{\Theta}_1 = \frac{1}{n} \sum_{i=1}^{n} X_i \, .$$

  - We showed that it is unbiased:

$$E\left[\hat{\Theta}_1\right] = \theta.$$

  - For many practical and widely used distributions (Gaussian, Poisson, Binomial, Exponential family, Generalized Gaussian...) it is efficient.

$$var\left(\hat{\Theta}_1\right) \quad = \quad CRB.$$

Comment on optimality - Minimum Variance Unbiased Estimator

- But one can perspicaciously notice that a scale of Θ_1 by a factor $0 \leq \alpha \leq 1$ will reduce its variance. However it generates a bias:

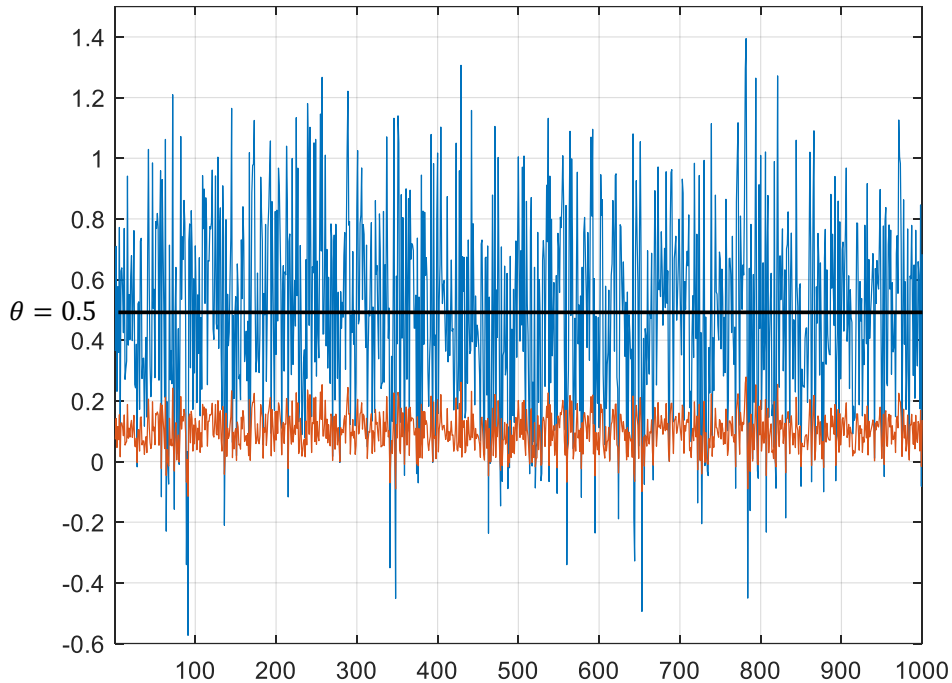
$$\begin{aligned} \hat{\Theta}_2 &= \frac{\alpha}{n} \sum_{i=1}^{n} X_i \,, \\ E\left[\hat{\Theta}_2\right] - \theta &= (\alpha - 1)\,\theta. \\ var\left(\hat{\Theta}_2\right) &= \alpha^2 var\left(\hat{\Theta}_1\right). \end{aligned}$$

  - There is no contradiction with the Cramér Rao lower bound theorem since the assumption of the latter is the unbiasedness of the estimator (note that there exist another lower bound formulation for biased estimators).

Parameter estimation

Comment on optimality - Minimum Variance Unbiased Estimator

- Which of these two estimators is better?



Blue : $\widehat{\Theta}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$

Red : : $\widehat{\Theta}_2 = \frac{\alpha}{n}\sum_{i=1}^{n} X_i$

Comment on optimality - Minimum Variance Unbiased Estimator

- It will be judicious to calculate the error with respect to the true value θ rather than the estimated value. This error encompasses the variance and the bias.
  - This can be done by considering the mean square error (*MSE*) which, for a scalar parameter is:

$$MSE = E\left[\left(\hat{\Theta} - \theta\right)^2\right]. \tag{19}$$

  - Using classical calculations, we obtain:

$$E\left[\left(\hat{\Theta} - \theta\right)^2\right] = var\left(\hat{\Theta}\right) + b^2\left(\theta\right). \tag{20}$$

where $b\left(\theta\right)$ is the bias, i.e. $b\left(\theta\right) = E\left[\hat{\Theta}\right] - \theta$.

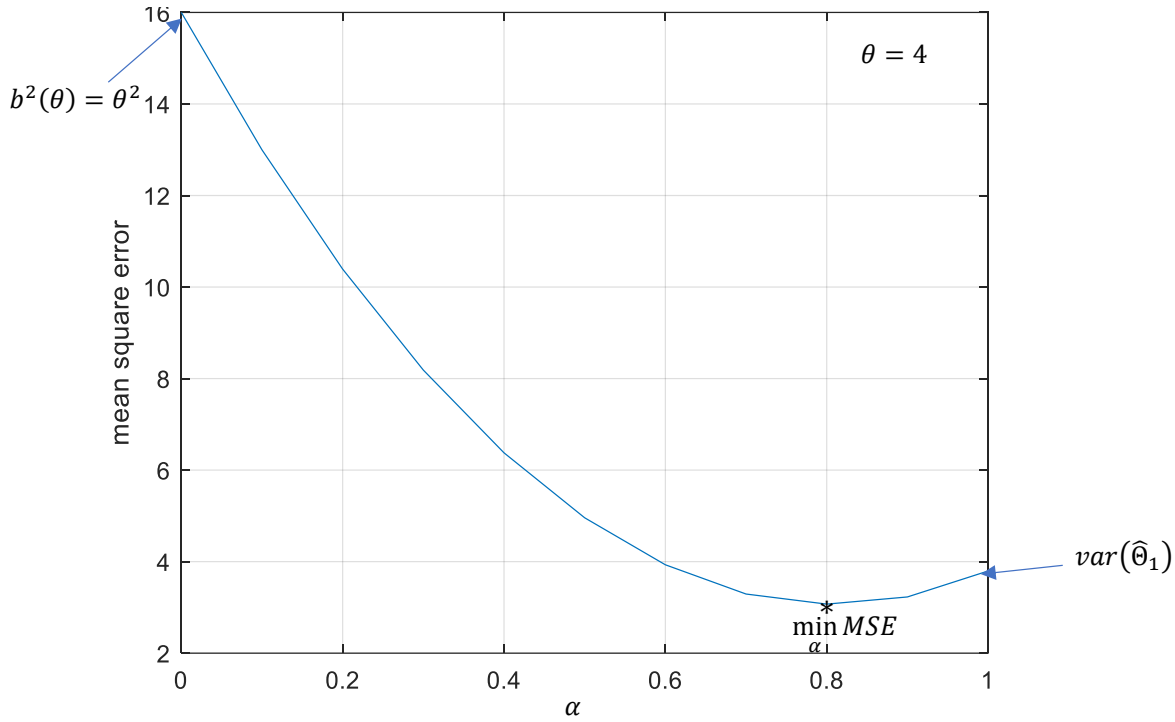
Comment on optimality - Minimum Variance Unbiased Estimator

- It is quite often the case that techniques employed to reduce variance results in an increase in bias (the absolute value), and vice versa. This phenomenon is called the Bias Variance Tradeoff.
  - Remark: The mean square error is the error with respect to the true value θ whereas the variance of $\hat{\Theta}$ is the error with respect to the estimated value.
  - For $\hat{\Theta}_2$ the bias will equal $b(\theta) = (\alpha - 1)\,\theta$ and the mean square error is:

$$E\left[\left(\hat{\Theta} - \theta\right)^2\right] = \alpha^2 var\left(\hat{\Theta}_1\right) + (\alpha - 1)^2\,\theta. \tag{21}$$

Parameter estimation

Comment on optimality - Minimum Variance Unbiased Estimator



- The mean square error curve with respect to α for $\theta = 4$. Notice that the mean square error can be less than the variance of the sample mean estimator.

Comment on optimality - Minimum Variance Unbiased Estimator

- Unfortunately, minimum of MSE depends on the unknown parameter to estimate so that developing a minimum mean square error estimator is not possible in general.
  - A quick look on (20) shows this dependency as the bias depends on θ.
    - case when this is possible: the variance $var\left(\hat{\Theta}_1\right)$ is proportional to θ,as for the exponential distribution.

Comment on optimality - Minimum Variance Unbiased Estimator

- Otherwise it is useful to have some information about θ.
  - θ may be random with probability density $p(\theta)$.The Bayesian framework for estimation will help us to compute, among others, optimal estimators in the sense of MMSE (Minimum Mean Square Error)
    - θ takes values from a set of finite possible values. The framework of testing hypothesis will answer this issue.
  - Since the bias depends on the unknown parameter we constrain the bias to be zero:
    - We will then consider unbiased estimator with minimal variance (MVU estimator).

Comment on optimality - Minimum Variance Unbiased Estimator

- The MVU estimator may not exist. Even if it does exist, we may not be able to find it.

  - To achieve MVU estimator, several strategies may be used. For this course, we mention two of the three situations that are generally proposed for finding an MVU estimator:
    - compute the *CRB* and check if some estimator satisfies it. It will be pertinent to check first with the ML estimator.
    - restrict the class of estimators to linear ones and find the best linear unbiased minimum variance estimator.

Least Squares Estimator (1)

- In the previous chapters we have proposed parameter estimation with the knowledge about the probability distribution of the observed sample.
  - We also focused on determining the ML estimator, which is often unbiased and efficient..
  - We now investigate a class of estimators where information about the sampling distribution is missing.
  - We suppose that the observed sequence $(x_1, x_2, ..., x_n)$ is a perturbed version of an original signal $(s_1, s_2, ...s_n)$.
  - To remedy the lack of information about the sampling distribution we study the practical case where we can define an algebraic relationship between the set of parameters and the original signal $(s_1, s_2, ...s_n)$.

Least Squares Estimator (2)

- We restrict ourselves to cases where the original signal has a linear relation with the unknown parameters that we wish to estimate. In addition the original signal is deterministic.
  - Vector notations: $(s_1, s_2, ...s_n)$ is a vector $\mathbf{s} = [s_1\ s_2 ... s_n]^T$, $(x_1, x_2, ..., x_n)$ is $\boldsymbol{x} = [x_1\ x_2 ... x_n]^T$, $(\theta, \theta_2, ...\theta_K)$ is $\boldsymbol{\theta} = [\theta_1\ \theta_2 ... \theta_K]^T$ of parameters and $(w_1, w_2, ..., w_n)$ is $\boldsymbol{w} = [w_1\ w_2 ... w_n]^T$ the noise values.
  - The linear dependency between $\boldsymbol{\theta}$ and $\mathbf{s}$ has a general form given by:

$$\boldsymbol{s} = \mathbf{H}\boldsymbol{\theta}, \tag{22}$$

$$\mathbf{x} = \mathbf{s} + \mathbf{w}, \tag{23}$$

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \tag{24}$$

  - $\mathbf{H}$ is a known $n \times K$ matrix.

Least Squares Estimator (3)

- Least Squares Estimator (LSE) is based on minimizing the error function:

$$\begin{aligned} \mathcal{J}(\boldsymbol{\theta}) &= \sum_{k=1}^{n} \left(x_k - s_k\right)^2, \\ &= \sum_{k=1}^{n} \left(x_k - \boldsymbol{h}_k\boldsymbol{\theta}\right)^2. \end{aligned} \tag{25}$$

  where $\mathbf{h}_k$ is the i th row of $\mathbf{H}$.

Least Squares Estimator (4)

- We will present the simplest case where $s_k = \theta$ for all k such that $x_k = \theta + w_k$. The general expression gives:

$$\mathbf{x} = \theta\mathbf{h} + \mathbf{w}. \tag{26}$$

  - where $\mathbf{H} = \mathbf{h} = [1,\ 1,\ ...,\ 1]^t$.
  - In matrix notation one can express $\mathcal{J}(\boldsymbol{\theta})$ as:

$$\begin{aligned} \mathcal{J}(\theta) &= (\mathbf{x} - \theta\mathbf{h})^T (\mathbf{x} - \theta\mathbf{h}). \\ &= \mathbf{x}^T\mathbf{x} - \theta\mathbf{x}^T\mathbf{h} - \theta\mathbf{h}^T\mathbf{x} + \theta^2\mathbf{h}^T\mathbf{h}, \\ &= \mathbf{x}^T\mathbf{x} - 2\theta\mathbf{h}^T\mathbf{x} + \theta^2\mathbf{h}^T\mathbf{h}. \end{aligned}$$

Least Squares Estimator (5)

- To minimize $\mathcal{J}(\theta)$ we have to solve $\frac{\partial \mathcal{J}(\theta)}{\theta} = 0$

$$\begin{aligned} \frac{\partial \mathcal{J}(\theta)}{\theta} &= -2\mathbf{h}^T\mathbf{x} + 2\theta\mathbf{h}^T\mathbf{h}, \\ \mathbf{h}^T\mathbf{h}\hat{\theta}_{LS} &= \mathbf{h}^T\mathbf{x}, \\ \hat{\theta}_{LS} &= \frac{\mathbf{h}^T\mathbf{x}}{\mathbf{h}^T\mathbf{h}}. \end{aligned}$$

Since $\mathbf{h} = [1,\ 1, \ldots,\ 1]^t$ we have:

- $$\hat{\theta}_{LS} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{27}$$

Least Squares Estimator (6)

- Consider now $s_k = \theta h_k$. Here $\mathbf{H}$ is a $n \times 1$ column vector.

$$\mathcal{J}(\theta) = \sum_{k=1}^{n} (x_k - \theta h_k)^2 .$$

  - Minimizing $\mathcal{J}(\theta)$ yields:

$$\hat{\theta}_{LS} = \frac{\sum_{k=1}^{n} h_k x_k}{\sum_{k=1}^{n} |h_k|^2}. \tag{28}$$

Least Squares Estimator (7)

- Consider the general case where:

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}.$$

  - where $\mathbf{H}$ is a $n \times K$ matrix, $\boldsymbol{\theta}$ a vector of K unknown parameters and $\mathbf{w}$ noise. The minimum of:

$$\mathcal{J}(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}),$$

    is obtained for :

$$\hat{\boldsymbol{\theta}}_{LS} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1} \mathbf{H}^T\mathbf{x}.$$

Least Squares Estimator (8)

- **Exercise 13: Fitting curve to pairs of measures**
- Consider an independent sequence of pairs of values $z_i = (u_i,\ x_i)$ which are related physically like $(T,\ V)$ the temperature and the volume of a gaz. We wish to fit a curve $x = f(u)$ to that sequence of pairs. We consider a polynomial function of order K. Find the coefficients of that polynomial curve that reduce least square error.

Least Squares Estimator (9)

- **Solution to exercise 13:**
- The polynomial function is $x = \theta_0 + \theta_1 u + \theta_2 u^2 + ... + \theta_K u^K$. This equation must be satisfied for all u_i, however the value of x are not exactly $f(u)$ and corrupted by errors $\mathbf{w}$. Let $\boldsymbol{H}$ be the matrix below:H

$$\mathbf{H} = \begin{bmatrix} 1 & u_1 & u_1^2 & \cdots & u_1^K \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & u_n & u_n^2 & \cdots & u_n^K \end{bmatrix}. \tag{29}$$

- We can write:

$$\mathbf{x} = \begin{bmatrix} 1 & u_1 & u_1^2 & \cdots & u_1^K \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & u_n & u_n^2 & \cdots & u_n^K \end{bmatrix} \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_K \end{bmatrix} + \mathbf{w}$$

- The least square solution for $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}_{LS} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}.$$

Sequential LSE (1)

- In many applications the original sequence is on going on time and after indices n more data are available.
  - In order to obtain a real time estimation of θ and avoid waiting for the entire sequence to be observed, it is important to perform a sequential estimation, where the estimator at time n can be updated to obtain an estimate at time $n+1$.

Sequential LSE (2)

- We will not develop the general case but give the results for the previous example. Let $\hat{\theta}_{n+1}$ be the estimate at time index $n+1$ and E_n and E_{n+1} the energy of the sequence (h_k) at indices n and $n+1$ respectively, $E_{n+1} \geq E_n$.

$$\hat{\theta}_{n+1} = \frac{\sum_{k=1}^{n+1} h_k x_k}{\sum_{k=1}^{n+1} |h_k|^2}.$$

$$\hat{\theta}_{n+1} = \frac{\sum_{k=1}^{n} h_k x_k + h_{n+1} x_{n+1}}{E_{n+1}}$$

Sequential LSE (3)

- With some simple computations we will have:

$$\begin{aligned}
\hat{\theta}_{n+1} &= \frac{E_n\hat{\theta}_n + h_{n+1}x_{n+1}}{E_{n+1}}, \\
&= \frac{E_n\hat{\theta}_n + h_{n+1}x_{n+1} + h_{n+1}^2\hat{\theta}_n - h_{n+1}^2\hat{\theta}_n}{E_{n+1}}, \\
&= \frac{E_{n+1}\hat{\theta}_n + h_{n+1}x_{n+1} - h_{n+1}^2\hat{\theta}_n}{E_{n+1}}, \\
&= \hat{\theta}_n + \frac{h_{n+1}}{E_{n+1}}\left(x_{n+1} - h_{n+1}\hat{\theta}_n\right).
\end{aligned}$$

  - We don't need to compute all the observed data. The term at the right hand side is a correction term. Further investigations show that the correction term decreases with n since E_n is increasing with n.

MLSE

To go further, MLSE, Bayesian Estimation

- We will not develop MLSE but sketch only its principle. MLSE, for Maximum Length Sequence Estimation, is generally resolved by an algorithm introduced by A. Viterbi in 1967.
  - Consider the linear model where $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$. In many applications we have a long sequence θ_1, θ_2, .., and we consider the vector of length K such that:

$$\boldsymbol{\theta_n} = \begin{bmatrix} \theta_{n-K+1} \\ \vdots \\ \theta_n \end{bmatrix} \tag{30}$$

    The linear model is then $\mathbf{x_n} = \mathbf{H}\boldsymbol{\theta_n} + \mathbf{w_n}$ with a sliding window to encompass all the sequence.
  - One may notice that $\boldsymbol{\theta}_{n-1}$ and $\boldsymbol{\theta}_n$ are dependent and the sequence of vectors is a Markov Chain.
  - The Viterbi Algorithm uses this Markovian characteristic to estimate the sequence θ_1, θ_2, ..,and performs a Maximum Likelihood Sequence Estimation.

Bayesian Estimation

To go further, MLSE, Bayesian Estimation

- We need to estimate the parameter π of a Binomial random variable. Consider this example with $n = 3$, suppose the outcome is $1,\ 1,\ 1$.
  - The ML estimator will give $\hat{\pi}_{ML} = 1$.
  - But if we expose π to be random with distribution $p(\pi)$, to estimate π we will maximize the a posterior probability $p(\pi|\mathbf{x})$ (MAP). From Bayes rule we have:

$$\begin{aligned} p(\pi|\mathbf{x}) &= \frac{p(\mathbf{x},\ \pi)}{p(\mathbf{x})} \\ &= \frac{p(\mathbf{x},|\pi)\,p(\pi)}{p(\mathbf{x})} \end{aligned}$$

Maximizing $p(\pi|\mathbf{x})$is equivalent to optimizing $p(\mathbf{x},|\pi)\,p(\pi)$. The problem above maximize then $\pi^3 p(\pi)$ and is different from the ML solution unless $p(\pi)$ is uniform. Suppose $p(\pi)$ is a discrete distribution with $p(0.25) = 0.6, p(0.5) = 0.35, p(0.75) = 0.05$. Then:

$$\hat{\pi}_{MAP} = arg\left\{max\left(0.25^3 \times 0.6,\ 0.5^3 \times 0.35,\ 0.75^3 \times 0.05\right)\right\} = 0.5$$

2. Hypothesis Testing

Hypothesis testing

Introduction (1)

- Hypothesis testing is a decision problem.
  - Consider a received outcome $\mathbf{x} = (x_1, \ldots, x_n)$ from a sample. In the simplest case, this sample is generated from one of two probability models.
    - we refer to these models as hypotheses and label them as H_0 and H_1 .
    - when observing $\mathbf{x}$ we don't know which hypothesis is responsible of the observed data and we aim at guessing which one it is .

Introduction (2)

- **Formulation and modeling.** We consider a source generating points according to one of two known probability densities:

$$\begin{aligned} H_0 &\iff f_{X|H_0}(x\,|H_0) \\ H_1 &\iff f_{X|H_1}(x\,|H_1) \end{aligned}$$

  - In many situations the two hypotheses obey to the same probability model but differ in the values of their parameters:

$$\begin{aligned} H_0 &\iff f_{\theta_0}(x) \\ H_1 &\iff f_{\theta_1}(x) \end{aligned}$$

    - the decision problem in this case is to choose between θ_0 and θ_1 after observing the outcome $\mathbf{x}$ of a sample.

The Likelihood Ratio Test (LRT) (1)

- One solution to perform hypothesis testing is the Likelihood Ratio Test (LRT). It is expressed as follows:

$$\Lambda(\mathbf{x}) = \frac{f_{\mathbf{X}|H_1}(\mathbf{x}\,|H_1)}{f_{\mathbf{X}|H_0}(\mathbf{x}\,|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda \tag{31}$$

  - λ is a threshold and we will discuss later the best strategy to fix it.
  - For an observed sample (the received points are independent and identically distributed (iid)), $m \in \{0, 1\}$:

$$f_{\mathbf{X}|H_m}(\mathbf{x}\,|H_m) = \prod_{i=1}^{n} f_{X|H_m}(x\,|H_m) \tag{32}$$

The Likelihood Ratio Test (LRT) (2)

- For computation simplicity we often use the log of the LRT and we have:

$$\ln \Lambda(\mathbf{x}) = \sum_{i=1}^{n} \ln f_{X|H_1}(x|H_1) - \sum_{i=1}^{n} \ln f_{X|H_0}(x|H_0) \underset{H_0}{\overset{H_1}{\gtrless}} \ln \lambda. \tag{33}$$

  - Example: Suppose a source has a Gaussian density function with $\mu_0 = 0$ or $\mu_1 = A > 0$ with the same variance σ^2 for both hypotheses. The test will resume in:

$$\ell(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} x_i \underset{H_0}{\overset{H_1}{\gtrless}} \gamma, \tag{34}$$

$$\textit{with} \quad \gamma = \frac{\sigma^2 \ln \lambda}{nA} + \frac{A}{2}. \tag{35}$$

The Likelihood Ratio Test (LRT) (3)

- If we fix $\lambda = 1$, we simply test which of the two possible densities is greater at point $\mathbf{x}$ (see (31)).
  - From (34) the test is resumed by testing if the sample mean exceeds half of $[0;\ A]$ or not.
    - $\ell(\mathbf{x}) \in \mathbb{R}$ and this quantity in $\mathbb{R}$ resume the test previously formulated in $\mathbb{R}^n$. But this is not always the case !
    - The real axis is then divided into two regions $]-\infty,\ \gamma[\cup [\gamma,\ \infty[$.
    - The distribution of $\ell(\mathbf{x})$ is easily extracted as it is the sum on n independent Gaussian:

$$Y\,|H_0 = \ell\,(\boldsymbol{X}\,|H_0\,) \quad \sim \quad \mathcal{N}(0;\ \frac{\sigma^2}{n}),$$

$$Y\,|H_1 = \ell\,(\boldsymbol{X}\,|H_0\,) \quad \sim \quad \mathcal{N}(A;\ \frac{\sigma^2}{n}).$$

Hypothesis testing

Error probabilities of the first and the second kind (1)

- Let $y = \ell(\mathbf{x})$ and the real line $\mathbb{R}$ separated into the two decision regions R_0 and R_1:

$$R_0 = \{y \in \mathbb{R} : y < \gamma\} =]-\infty, \gamma[\,,$$

$$R_1 = \{y \in \mathbb{R} : y \geq \gamma\} = [\gamma, \infty[\,.$$

  - $R_0 \cup R_1 = \mathbb{R}$ and $R_0 \cap R_1 = \emptyset$.

  - We define then the two error probabilities:
    - error of the first kind:

$$P_F = Pr\{y \in R_1 \,|H_0\} = \int_{\gamma}^{\infty} f_{Y|H_0}(y\,|H_0)\, dy \tag{36}$$

    - error of the second kind:

$$P_M = Pr\{y \in R_0 \,|H_1\} = \int_{-\infty}^{\gamma} f_{Y|H_1}(y\,|H_1)\, dy \tag{37}$$

Error probabilities of the first and the second kind (2)

For the example above:

$$P_F = \sqrt{\frac{n}{2\pi\sigma^2}} \int_{\gamma}^{\infty} \exp\left(-\frac{n}{2\sigma^2} y^2\right) dy = Q\left(\frac{\sqrt{n}\gamma}{\sigma}\right),$$

$$P_M = P_M = \sqrt{\frac{n}{2\pi\sigma^2}} \int_{-\infty}^{\gamma} \exp\left(-\frac{n}{2\sigma^2} (y-A)^2\right) dy = 1 - Q\left(\frac{\sqrt{n}(\gamma - A)}{\sigma}\right).$$

- where $Q(x) = \int_{x}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2}) du$
  - the values of $Q(x)$ can be obtained from Gaussian tables in the literature or from software.

Error probabilities of the first and the second kind (3)

- The subscripts F and M are chosen from the traditional radar problem where H_0 corresponds to the absence of a target and H_1 to the presence of a target.
  - P_F is the a *false alarm*, i.e. we decide that a target is present when it is not.
    - P_M is the *miss detection*, i.e. we decide that the target is absent when it is present.
    - We also use the conditional probability $P_D = 1 - P_M$ = the probability of *detection*.

Error probabilities of the first and the second kind (4)

More generally we say that each decision rule partitions the observation space $\mathbb{R}^n$ into **two disjoint** regions $\mathcal{R}_0$ and $\mathcal{R}_1$ such that $\mathcal{R}_0 \cup \mathcal{R}_1 = \mathbb{R}^n$:

$$\begin{aligned} \mathcal{R}_0 &= \{\mathbf{x} \in \mathbb{R}^n : \ell(\mathbf{x}) < \ln \lambda.\} \\ \mathcal{R}_1 &= \{\mathbf{x} \in \mathbb{R}^n : \ell(\mathbf{x}) \geq \ln \lambda.\}. \end{aligned}$$

- the probability of error of the first kind (historically noted as α) or false alarm defined as:

$$\alpha = P_F = Pr\{\mathbf{x} \in \mathcal{R}_1 \,|H_0\} = \int_{\mathcal{R}_1} f_{\mathbf{X}|H_0}(\mathbf{x}\,|H_0)\, d\mathbf{x} \tag{38}$$

Error probabilities of the first and the second kind (5)

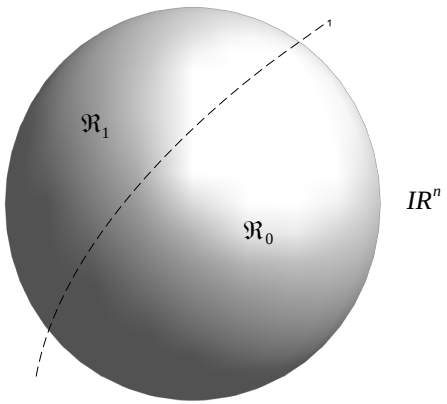
- the probability of error of the second kind (historically noted as β) or miss detection defined as:

$$\beta = P_M = Pr\{\mathbf{x} \in \mathcal{R}_0 \,|H_1\} = \int_{\mathcal{R}_0} f_{\mathbf{X}|H_1}(\mathbf{x}\,|H_1)\, d\mathbf{x} \tag{39}$$

  - we also use P_{ND} for *non detection* instead of P_M.
  - Note: The sign $\int$ for α and β should be understood as a multiple integral.

Hypothesis testing

- This can be illustrated by the figure below:
  - partition the observation space $\mathbb{R}^n$ into two disjoint regions $\mathcal{R}_0$ and $\mathcal{R}_1$ corresponding to decisions in favor of hypotheses H_0 and H_1 respectively.



Hypothesis testing

The Neyman-Pearson test (1)

- For the LRT test the two error probabilities depend on the choice of λ.
  - As the value of the threshold changes, α and β change but **in an opposite way.**
- we should like to make α and β as small as possible but these are conflicting objectives.

- To chose a threshold, the Neyman-Pearson criterion suggests the following:
  - constrain $\alpha \leq \alpha_{max}$ and choose the threshold such that β is minimized:

$$\begin{aligned} \min_{\lambda} \quad & \beta(\lambda) \\ \textit{such that} \quad & \alpha(\lambda) \quad \leq \alpha_{max} \end{aligned}$$

- Clearly, choosing λ^* such that $\alpha(\lambda^*) = \alpha_{max}$ will minimize β .

The Neyman-Pearson test (2)

- The problem is then a constrained optimization problem which involves a Lagrange multiplier λ:

$$L(\mathbf{x}, \lambda) = \int f_{\mathbf{X}|H_1}(\mathbf{x}\,|H_1)\, d\mathbf{x} + \lambda \left(\int f_{\mathbf{X}|H_0}(\mathbf{x}\,|H_0)\, d\mathbf{x} - \alpha_{max} \right) \quad (40)$$

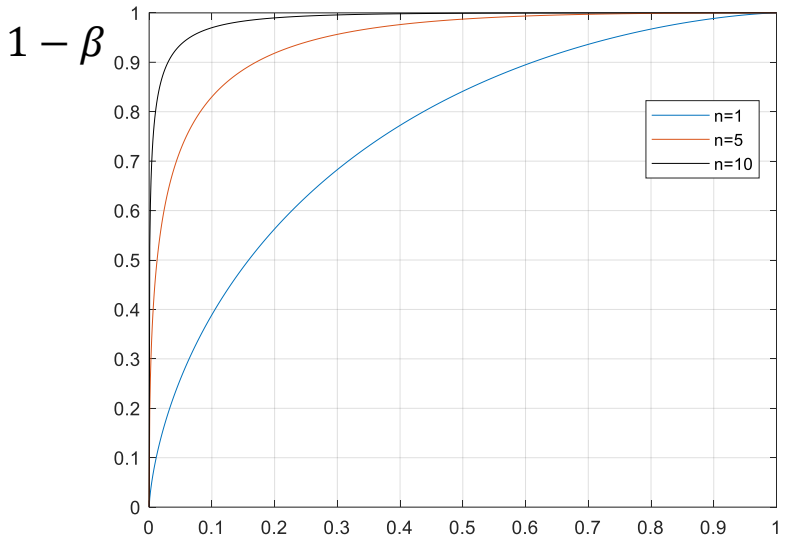
- where the first term is to minimize and with $\lambda \geq 0$. As λ is generally often active we choose λ such that $P_F = \alpha_{max}$ (the border of the constraint) .
- Example for an outcome of a Gaussian sample: Suppose $\alpha = P_F = Q\left(\frac{\sqrt{n}\gamma}{\sigma}\right)$. Fixing $\alpha = \alpha_{max}$ the threshold is then

$$\gamma^* = \frac{\sigma}{\sqrt{n}} Q^{-1}(\alpha_{max}) .$$

Hypothesis testing

The ROC curve

- The performance of the Neyman-Pearson test may be represented in a curve called the ROC curve (Receiver Operating Characteristics) .
  - It draws the evolution of β (or 1-β) with respect to α.



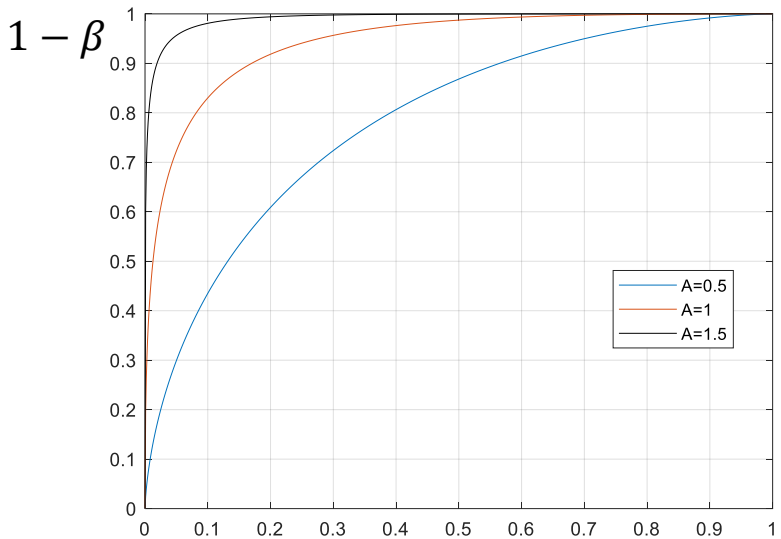


Figure: ROC curves for the example above. To the left, ROC for different number of points in a sample. To the right ROC for larger A . The higher the curve the better the test.

The Bayes test (1)

- Bayes test is also an LRT test. To compute the threshold we need two additional information:
  - the proportions of hypotheses H_0 and H_1 , i.e the prior probabilities $\mathbb{P}(H_0)$ and $\mathbb{P}(H_1)$.
    - the costs of each decision with respect to the true hypothesis, namely: $C_{0|0}$, $C_{0|1}$,$C_{1|0}$,$C_{1|1}$ ($C_{decide\ H_i | H_j\ is\ true}$).
    - the test and its threshold are then (without proof):

$$\Lambda(\mathbf{x}) = \frac{f_{\mathbf{X}|H_1}(\mathbf{x}|H_1)}{f_{\mathbf{X}|H_0}(\mathbf{x}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{C_{1|0} - C_{0|0}}{C_{0|1} - C_{1|1}} \times \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)} \tag{41}$$

The Bayes test (2)

- The main idea of Bayes test is to minimize the average risk:

$$\begin{aligned} \mathfrak{R} &= \mathbb{P}(H_0)P_F C_{1|0} + \mathbb{P}(H_0)(1-P_F)C_{0|0} \\ &+ \mathbb{P}(H_1)P_M C_{0|1} + \mathbb{P}(H_1)(1-P_M)C_{1|1} \end{aligned} \tag{42}$$

- In binary communication receivers we set $C_{0|0}$ and $C_{1|1}$ to zero and $C_{0|1}$ and $C_{1|0}$ to one.
  - The threshold is then $\lambda = \frac{P(H_0)}{P(H_1)}$ and for this application the two hypotheses are generally equally likely so $\lambda = 1$.
- Setting $\mu_0 = -1$ and $\mu_1 = +1$, the threshold of the log of the LRT is zero.
  - recalling that the error probabilities are $\alpha = P_F$ and $\beta = P_M$, minimizing the risk will be equivalent to minimizing the total error: $\mathbb{P}(H_0)P_F + \mathbb{P}(H_1)P_M$.

Hypothesis testing

Nomenclature

- Hypothesis H_0 is generally called the $null - hypothesis$.
  - hypothesis H_1 the *alternative hypothesis*.
    - region $\mathcal{R}_1$ for deciding H_1 is usually called the *critical region* and noted $\mathcal{R}_c$.
- We discussed about *binary* hypotheses.
  - we may have more than two hypotheses: $M > 2$ hypotheses. We talk then about M-ary hypotheses or Multiple Hypotheses
- We discussed above *simple* hypotheses where the values of the parameters are known.
  - in many cases of interest, one or all of the parameters are not defined exactly, but known to belong to a subspace of the space of parameters.
  - Example: hypothesis $H_0 : \mu < \mu_0$ and hypothesis H_1: $\mu \geq \mu_0$
    - generally $\theta \in \Omega_0$ or $\theta \in \Omega_1$. We have then *composite* hypotheses.

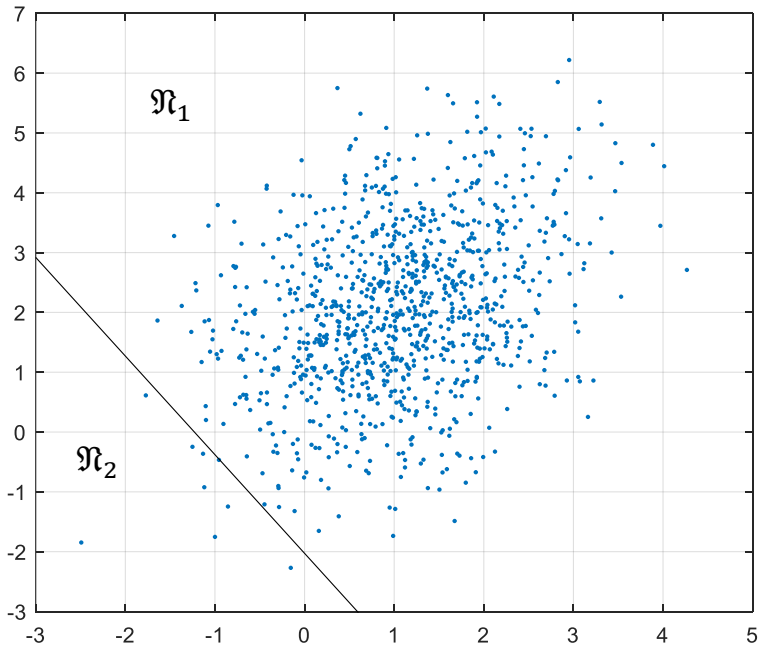
Hypothesis testing

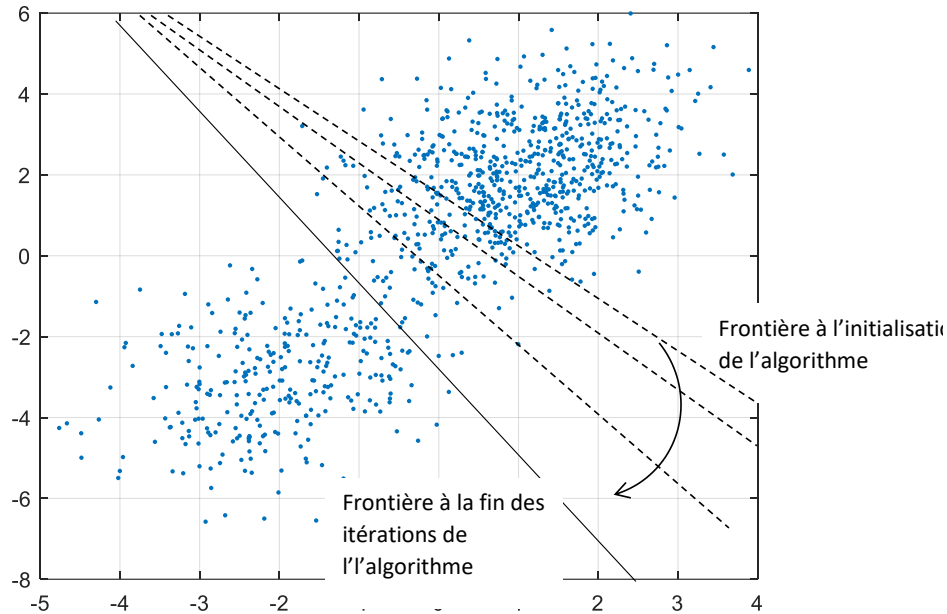
Hypothesis testing and Classification

- In hypothesis testing, we observe a realization of an sample,and want to assign one of two possible distributions to this observed sequence. The test is performed with respect to a **pre-calculated boundary** between two regions of space $\mathbb{R}^n$

- During classification, we may receive a realization of a sample containing two modes (a mixture), and we want **to draw a boundary** that clearly separates the two modes. We assign then to each element of the observed sequence the corresponding class or mode.

Hypothesis testing





- To the left Hypothesis Testing, and to the right classification problem.

Hypothesis testing

- **Exercise 14:**
  - Light bulbs are manufactured by two companies, A and B. Their lifetime is a random variable whose distribution depends on their origin.
  - The light bulbs manufactured by company A have a lifetime distributed according to an exponential distribution with parameter a, while those manufactured by company B have a lifetime distributed according to an exponential distribution with parameter b.
  - We observe the lifetime of a light bulb, which we denote by Y, and wish to deduce its origin.
  - To this end, we construct a hypothesis test where the null hypothesis corresponds to company A and the alternative hypothesis to company B:

$$H_0 : \quad f_{Y|H_0}(y) \quad = \quad a \exp(-ay)\, I_{[0,\,\infty[}(y), \tag{43}$$

$$H_1 : \quad f_{Y|H_1}(y) \quad = \quad b \exp(-by)\, I_{[0,\,\infty[}(y). \tag{44}$$

  where $b > a > 0$.

Hypothesis testing

1. Compute the Likelihood ratio $\Lambda(y)$.
   1. Show that the test $\Lambda(y) \underset{H_0}{\overset{H_1}{\gtrless}} \eta$ can be resumed by the test:

$$y \underset{H_1}{\overset{H_0}{\lessgtr}} \gamma \tag{45}$$

      and precise the expression of γ with respect to η, a,and b.
   2. The light bulbs manufactured by B have a poor environmental record and require special treatment before destruction. Using the assumptions defined above, calculate the probability of detection and the probability of false alarm.
   3. Plot them graphically.
   4. Deduce the probabilities $P(\text{accept } H_0 \mid H_0)$ and $P(\text{accept } H_0 \mid H_1)$.
   5. Specify the threshold if the chosen test is the Neyman-Pearson test.

- **Exercise 15: Application to Radar systems**
  - The observed sequence is a collection of n values which can either be a noisy reflected wave from a target, or in case of no target noise only is observed. We then assume that the observed sequence has a constant DC component A under hypothesis H_1(the presence of a target) or is a zero DC value under hypothesis H_0 (no target). These observations are corrupted by an additive zero mean white and Gaussian noise W_i with known variance σ^2. So that:

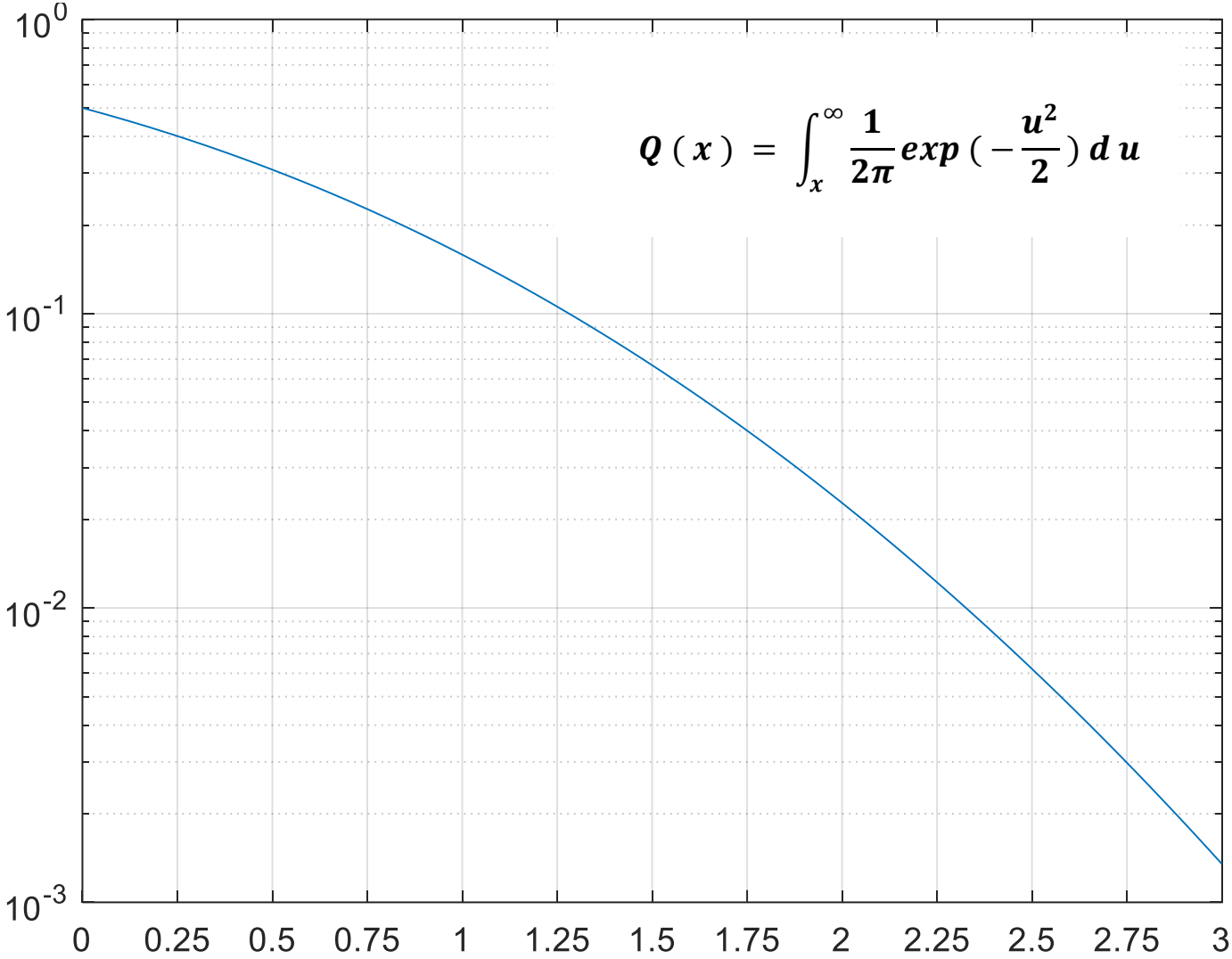
$$\begin{aligned} H_0 : \quad x_i &= w_i \qquad 1 \leq i \leq n, \\ H_1 : \quad x_i &= A + w_i \quad 1 \leq i \leq n. \end{aligned}$$

Hypothesis testing

- **Exercise 15: Application to Radar systems**

1. Write the LLRT (Log Likelihood Ratio Test) to decide whenever there is a target or not.
2. Write the probability of False Alarm (FA, the error of the first kind) and the probability of non detection (Pnd the error of the second kind).
3. Let $\sigma^2 = 1$, $A = 1$ and $n = 25$. What is the Neyman-Pearson test such that the FA is lower than 10^{-2}(use the curve of the complementary of the cumulative function $Q(x)$ of the zero mean unit variance of the Gaussian distribution given below).
4. What will be in this case the Pnd ?
5. We consider a lower FA upper limit, say $1.5 * 10^{-3}$. In what direction will the Pnd evolve: will it increase or decrease? Compute it.

Hypothesis testing



- **Exercise 15: Solution**

(1) Let us write the two probability (density) functions corresponding to the two hypotheses:

$$H_0: \quad p_{H_0}(x_1, \ldots, x_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right),$$

$$H_1: \quad p_{H_1}(x_1, \ldots, x_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} (x_i - A)^2\right).$$

Hypothesis testing

- The LLRT is:

$$\ln \Lambda(\mathbf{x}) = \ln \left[\frac{\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - A)^2\right)}{\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right)} \right] \underset{H_0}{\overset{H_1}{\gtrless}} \lambda, \quad (46)$$

$$= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - A)^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 \underset{H_0}{\overset{H_1}{\gtrless}} \lambda, \quad (47)$$

$$= \frac{A\sum_{i=1}^{n} x_i}{\sigma^2} - \frac{nA^2}{2\sigma^2} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda, \quad (48)$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i \underset{H_0}{\overset{H_1}{\gtrless}} \frac{A}{2} + \frac{\lambda\sigma^2}{nA}. \quad (49)$$

Hypothesis testing

- (2) The LLRT test may be written as:

$$\ell(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i \underset{H_0}{\overset{H_1}{\gtrless}} \gamma. \tag{50}$$

- The FA probability α is $Pr\{$decide "a target is present" given that "there is no target"$\}$:

$$\begin{aligned} \alpha &= Pr\left\{\text{decide } H_1 \,|H_0 \text{is true}\right\} \\ &= Pr\left\{\frac{1}{n}\sum_{i=1}^{n} X_i > \gamma \,|H_0 \text{is true}\right\}. \end{aligned}$$

- We need to derive the distribution of $\frac{1}{n}\sum_{i=1}^{n} X_i$. The probability α is conditioned with the hypothesis H_0. Each X_i is then $\sim \mathcal{N}(0;\, \sigma^2)$ and:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$$

Hypothesis testing

- Let $V = \frac{1}{n} \sum_{i=1}^{n} X_i$:

$$\begin{aligned} \alpha &= Pr\{V > \gamma\}, \\ &= Pr\left\{ \frac{V}{\sqrt{\frac{\sigma^2}{n}}} > \frac{\gamma}{\sqrt{\frac{\sigma^2}{n}}} \right\}, \\ &= Pr\left\{ U > \frac{\gamma\sqrt{n}}{\sigma} \right\}. \end{aligned}$$

- where $U \sim \mathcal{N}(0,1)$ and :

$$\alpha = \int_{\frac{\gamma\sqrt{n}}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du. \tag{51}$$

Hypothesis testing

- The non detection probability β is: $Pr\{$decide "there is no target" given that "there is truly a target"$\}$:

$$\begin{aligned} \beta &= Pr\left\{\text{decide } H_0 \,|H_1 \text{is true}\right\} \\ &= Pr\left\{\frac{1}{n}\sum_{i=1}^{n} X_i < \gamma \,|H_1 \text{is true}\right\} \end{aligned}$$

Hypothesis testing

- The probability β is conditioned by the hypothesis H_1 so each $X_i \sim \mathcal{N}(A;\, \sigma^2)$ and $V = \frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathcal{N}(A;\, \frac{\sigma^2}{n})$:

$$\begin{aligned} \beta &= Pr\left\{V < \gamma\right\}, \\ &= Pr\left\{\left(\frac{V-A}{\sqrt{\frac{\sigma^2}{n}}}\right) < \left(\frac{\gamma-A}{\sqrt{\frac{\sigma^2}{n}}}\right)\right\}, \\ &= Pr\left\{U < \left(\frac{\gamma-A}{\sqrt{\frac{\sigma^2}{n}}}\right)\right\}. \end{aligned}$$

Hypothesis testing

- $U \sim \mathcal{N}(0,1)$ and :

$$\beta = \int_{-\infty}^{\frac{(\gamma - A)\sqrt{n}}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du. \tag{52}$$

- To conclude:

$$\begin{aligned} \alpha &= Q\left(\frac{\gamma\sqrt{n}}{\sigma}\right), \\ \beta &= 1 - Q\left(\frac{(\gamma - A)\sqrt{n}}{\sigma}\right) \end{aligned}$$

- where $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} \exp\left(-\frac{u^2}{2}\right) du.$

Hypothesis testing

- (3) We seek in the Neyman-Pearson test to find γ_s that minimize β such that $\alpha \leq \alpha_{max}$.
- From (51), α is a decreasing function of γ. Let γ^* be the value of γ such that $\alpha = \alpha_{max} = 10^{-2}$. In consequence the set of γ such that $\alpha \leq \alpha_{max}$ is:

$$\Upsilon_{\alpha_{max}} = \{\gamma \in \mathbb{R} : \gamma \geq \gamma^*\}. \tag{53}$$

- From (52) β is an increasing function with γ . β is minimum for $\gamma \in \Upsilon_{\alpha_{max}}$ for the smallest value of γ in this set which is γ^*.
- The solution of the Neyman-Pearson test is then to choose the threshold $\gamma_s = \gamma^*$:

$$10^{-2} = Q\left(\frac{\gamma_s\sqrt{25}}{1}\right).$$

- From the graph of $Q(x)$ $\gamma_s = \frac{2.3}{5} = 0.46$.

Hypothesis testing

- (4) $\frac{(\gamma_s - A)\sqrt{n}}{\sigma} = \frac{(0.46-1)\sqrt{25}}{1} = -2.7$
- $\beta = 1 - Q(-2.7) = Q(2.7) \cong 3.10^{-3}$.
- α and β evolve in opposite direction. So when α decreases β will increase:
  - for $\alpha_{max} = 1.5 * 10^{-3}$, $\gamma_s = \frac{3}{5} = 0.6$
  - $\frac{(\gamma_s - A)\sqrt{n}}{\sigma} = \frac{(0.6-1)\sqrt{25}}{1} = -2.$
  - $\beta = 1 - Q(-2) = Q(2) \cong 2.10^{-2}$

References

1. Van Trees, Harry L. Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory. John Wiley & Sons, 2004.
2. KAY, Steven M. Fundamentals of statistical signal processing: estimation theory. Prentice Hall, 1993.
3. KAILATH, Thomas et SAYED, Ali H. Linear estimation. IEEE Control Systems, 2001,