# MICAS902

# Introduction to Probabilities and Statistics

# Part on “Detection and Estimation”

Philippe Ciblat

Telecom Paris, Institut Polytechnique de Paris, Palaiseau, France

# Outline

1. Motivation and preliminaries
2. Detection Theory (Bayesian approach)
   - Optimal detector: Maximum A Posteriori (MAP)
   - Optimal performance
3. Hypothesis Testing (deterministic approach)
   - Optimal test: Neyman-Pearson test (NPT)
   - Optimal Asymptotic performance
4. Estimation Theory (deterministic approach)
   - Optimal performance: Cramer-Rao bound (CRB),
   - Algorithms: Maximum likelihood (ML), Moments, Least Square (LS)
   - Asymptotic performance
5. Estimation Theory (Bayesian approach)
   - Optimal estimator: Mean A Posteriori (MeAP)
   - Optimal performance: Bayesian Cramer-Rao bound (BCRB)

# Part 1 : Motivation and Preliminaries

# Toy example: symbol detection

$$\underbrace{y}_{\text{observation}} = \underbrace{s}_{\text{information symbol: } 1 \text{ or } -1} + \underbrace{w}_{\text{noise}}$$

**Goal:** given $y$, recovering $s$ in the best way.

**Remarks:**

- Symbols are modeled as random: $\Pr\{s = 1\} = p$ with $p$ known
- Figure of merit: average error probability

$$\begin{aligned} P_e &:= \Pr\{\hat{s} \neq s\} \\ &= \Pr\{\hat{s} = 1 | s = -1\}\Pr\{s = -1\} + \Pr\{\hat{s} = -1 | s = 1\}\Pr\{s = 1\} \end{aligned}$$

- Conclusion: discrete-valued and random parameter $s$

# Toy example: signal detection

$$\begin{cases} \text{Hypothesis } \mathcal{H} = \mathcal{H}_0: & \underbrace{y}_{\text{observation}} = \underbrace{}_{\text{no signal}} + \underbrace{w}_{\text{noise}} \\ \text{Hypothesis } \mathcal{H} = \mathcal{H}_1: & \underbrace{y}_{\text{observation}} = \underbrace{x}_{\text{signal}} + \underbrace{w}_{\text{noise}} \end{cases}$$

**Goal:** given $y$, say if transmitted signal is active or not in the best way.

**Remarks:**

- Hypothesis parameter is not random usually
- Figure of merit:
  - maximizing signal detection probability $P_D = \Pr\{\mathcal{H}_1|\mathcal{H}_1\}$
  - given a maximum false alarm probability $P_{FA} = \Pr\{\mathcal{H}_1|\mathcal{H}_0\}$
- Conclusion: discrete-valued and deterministic parameter $\mathcal{H}$

**Applications:**

- radar (intrusion detection, missile detection),
- interweave cognitive radio

# Toy example: channel estimation

$$\underbrace{y_n}_{\text{set of observations}} = \underbrace{\sum_{\ell=0}^{L} h_\ell s_{n-\ell}}_{\text{unknown channel impulse response}} + \underbrace{w_n}_{\text{noise}}$$

**Goal:** given $\{y_n\}_{n=0,\cdots,N-1}$, recovering $\{h_\ell\}_{\ell=0,\cdots,L}$ in the best way.

**Remarks:**

- Symbols are known and channel modeled as unknown deterministic
- Figure of merit: mean square error

$$\mathrm{MSE} := \mathbb{E}[\|\hat{\mathbf{h}} - \mathbf{h}\|^2] = \sum_{\ell=0}^{L} \mathbb{E}[|\hat{h}_\ell - h_\ell|^2]$$

  with $\mathbf{h} = [h_0, \cdots, h_L]^{\mathrm{T}}$.

- Conclusion: continuous-valued and deterministic parameter $\mathbf{h}$

# Toy example: coin tossing parameter

Let $\mathcal{X} = \{x_0, \cdots, x_Q\}$ be a set of values

$$y_n = x_\ell \text{ with probability } p_\ell \text{ s.t. } \sum_{\ell=0}^{Q} p_\ell = 1.$$

**Goal:** given $\{y_n\}_{n=0,\cdots,N-1}$, recovering $\{p_\ell\}_{\ell=0,\cdots,Q}$ in the best way.
**Remarks:**

- Coin tossing parameter $\mathbf{p} = [p_0, \cdots, p_Q]^{\mathrm{T}}$ may be modeled as random parameter with a priori distribution (e.g. fluctuation around a predetermined value $p_\ell = p + \varepsilon_\ell$ with known $p$
- Figure of merit: mean square error

$$\mathrm{MSE} := \mathbb{E}[\|\hat{\mathbf{p}} - \mathbf{p}\|^2]$$

  Warning : expectation is over all the random variables (so averaging over the distributions of the noise and the parameter)
- Conclusion: continuous-valued and random parameter $\mathbf{p}$

**Applications:**

- Heads or tails, Loaded dice

# Problem classification

- Let $\theta_0$ be the true value of the parameter
- Let $\hat{\theta}_{(N)}$ be the estimated/guessed/decoded parameter (through the help of $N$ observations)
- Let $\theta$ be a generic variable of any function helping to estimate/guess/decode $\theta_0$.

| $\theta_0$ | random | deterministic |
|---|---|---|
| discrete | Detection (Part 2) | Hypothesis testing (Part 3) |
| continuous | Bayesian estimation (Part 5) | Estimation (Part 4) |

# Figures of merit for discrete-valued parameters

**Special Case:** binary parameter (0/1) leads to four probabilities

- $\Pr\{1|0\}$ (false alarm), $\Pr\{0|0\}$ (with $\Pr\{1|0\} + \Pr\{0|0\} = 1$)
- $\Pr\{1|1\}$ (correct detection), $\Pr\{0|1\}$ (with $\Pr\{1|1\} + \Pr\{0|1\} = 1$)

**Figures of merit:**

- If random, a priori distribution $\pi_0 = \Pr\{0\}$ and $\pi_1 = \Pr\{1\}$

  $$P = C_{0,0}\pi_0\Pr\{0|0\}+C_{1,0}\pi_0\Pr\{1|0\}+C_{0,1}\pi_1\Pr\{0|1\}+C_{1,1}\pi_1\Pr\{1|1\}$$

  with $C_{i,j}$ cost related to the configuration $i|j$

  Example: $P_e = \pi_0\Pr\{1|0\} + \pi_1\Pr\{0|1\}$
- If deterministic, tradeoff between both metrics (optimization for function output in $\mathbb{R}^2$ unfeasible)
  - Constant false alarm rate (CFAR): $\max \Pr\{1|1\}$ s.t. $\Pr\{1|0\} \leq C_{FA}$
  - Constant detection rate (CDR): $\min \Pr\{1|0\}$ s.t. $\Pr\{1|1\} \geq C_D$

# Figures of merit for continuous-valued parameters

**Remark:** $P_e$ usually no meaningful (except in some pathological cases)

**Goal:** find metric measuring the closeness of $\hat{\theta}$ to $\theta_0$. Typically Mean Square Error (MSE)

$$\begin{aligned} \text{MSE} &:= \mathbb{E}[\|\hat{\theta}-\theta_0\|^2] \\ \text{MSE}(\theta_0) &= \int \|v-\theta_0\|^2 p_{\hat{\theta}}(v)dv \text{ (if deterministic)} \\ \text{MSE} &= \iint \|v-u\|^2 p_{\hat{\theta},\theta_0}(v,u)dvdu \text{ (if random)} \end{aligned}$$

where the expectation is over all the random variables!

# Main results (take-home messages)

| $\theta_0$ | random | deterministic |
|---|---|---|
| discrete | Error probability | CFAR |
| | Max A Posteriori (MAP),<br>Max Likelihood (ML)<br>if equilikely | Likelihood Ratio Test<br>(LRT) |
| | Theoretical performance | Asymptotic performance<br>($N \to \infty$) |
| continuous | MSE | MSE |
| | Mean A Posteriori (MMSE) | Asymptotically ML<br>under some conditions |
| | Theoretical performance | Asymptotic performance |

# Generalities

- Let $\mathbf{X}_N = \{X_1, \dots, X_N\}$ be a random process
- The probability density function (pdf) $p_{\mathbf{X}}(\mathbf{x})$ depends on $\theta_0$, e.g., Gaussian process with unknown mean and variance ($\theta_0 = [\text{mean, variance}]$)

## Goal

Given a realization of the process (an event) $\mathbf{x}_N = \{x_1, \dots, x_N\}$, find out an estimated value, $\hat{\theta}_N$, of $\theta_0$, i.e., information on the pdf

**Notations:**

- If $\theta_0$ is random:
  - $p_{X,\theta}(\mathbf{x}_N, \theta)$: joint distribution between data and parameter
  - equivalently, $\underbrace{p_{X|\theta}(\mathbf{x}_N|\theta)}_{\text{likelihood}} . \underbrace{p_\theta(\theta)}_{\text{a priori distribution}}$
  - equivalently, $\underbrace{p_{\theta|X}(\theta|\mathbf{x}_N)}_{\text{a posteriori distribution}} . p_X(\mathbf{x}_N)$
- If $\theta_0$ is deterministic: $\underbrace{p_X(\mathbf{x}_N; \theta)}_{\text{pdf depending on } \theta}$ , equivalently, $\underbrace{p_{X|\theta}(\mathbf{x}_N|\theta)}_{\text{likelihood}}$

# Review of Matrix Algebra

**Non-singular square matrix:** $\mathbf{H} \in \mathbb{C}^{n\times n}$ is non-singular iff all its eigenvalues are non-zero

**Inverse of square matrix:** Let $\mathbf{H}^{-1} \in \mathbb{C}^{n\times n}$ be the matrix inverse of $\mathbf{H} \in \mathbb{C}^{n\times n}$.

- Then, $\mathbf{H}\mathbf{H}^{-1} = \mathbf{H}^{-1}\mathbf{H} = \mathbf{Id}$
- Moreover, $\mathbf{H}^{-1}$ exists iff $\mathbf{H}$ is non-singular

**Moore-Penrose pseudo-inverse of non-square matrix:** Let $\mathbf{H} \in \mathbb{C}^{M_R\times M_T}$ be a non-square full rank matrix.

- Right Pseudo-inverse: if $M_R < M_T$ then $\mathbf{H}$ admits a right pseudo-inverse, $\mathbf{H}^{\#} = \mathbf{H}^{\mathrm{H}}(\mathbf{H}\mathbf{H}^{\mathrm{H}})^{-1}$, such that $\mathbf{H}\mathbf{H}^{\#} = \mathbf{Id}$
- Left Pseudo-inverse: if $M_R > M_T$ then $\mathbf{H}$ admits a left pseudo-inverse, $\mathbf{H}^{\#} = (\mathbf{H}^{\mathrm{H}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{H}}$, such that $\mathbf{H}^{\#}\mathbf{H} = \mathbf{Id}$

# Review of Matrix Algebra (cont'd)

- Let $\mathbf{x}$, $\mathbf{y}$ be two vectors in $\mathbb{C}^n$
- Let (canonical) inner product : $<\mathbf{x}|\mathbf{y}>=\mathbf{x}^{\mathrm{H}}\mathbf{y}$ (bilinear sesqui-symmetric definite-positive)
- Norm: $\|\mathbf{x}\| = \sqrt{<\mathbf{x}|\mathbf{x}>} = \sqrt{\sum_{\ell=12}^{n}|x_\ell|^2}$; Euclidean distance: $\|\mathbf{x}-\mathbf{y}\|$
- Quadratic form (bilinear sesqui-symmetric form) : $\mathbf{x}^{\mathrm{H}}\mathbf{A}\mathbf{y}$ with $\mathbf{A}$ Hermitian matrix ($\mathbf{A}=\mathbf{A}^{\mathrm{H}}$)

Properties of quadratic form (and related matrix $\mathbf{A}$)

- **Positive Definite Quadratic form/matrix:** $\forall\mathbf{x}$, $\mathbf{x}^{\mathrm{H}}\mathbf{A}\mathbf{x}>0 \Leftrightarrow$ eigenvalues of $\mathbf{A}$ strictly positive (notation: $\mathbf{A}>0$)
- **Positive Semi-definite Quadratic form/matrix:** $\forall\mathbf{x}\neq\mathbf{0}$, $\mathbf{x}^{\mathrm{H}}\mathbf{A}\mathbf{x}\geq 0 \Leftrightarrow$ eigenvalues of $\mathbf{A}$ positive (notation: $\mathbf{A}\geq 0$)
- **Inequalities for positive semi-definite matrix:** partial order $\geq$ for two matrices $\mathbf{A}\geq 0$, $\mathbf{B}\geq 0$;

$$\mathbf{A}\geq\mathbf{B}\Leftrightarrow\mathbf{A}-\mathbf{B}\geq 0$$

# Part 2 : Detection Theory

# Introduction

Let

- $\Theta \in \mathbb{K}^n$ be the finite set of possible values for parameter $\theta$ ($\mathbb{K}$ any field)
- $\mathbf{y}$ be the observation depending on the parameter, let say, $\theta_0$.

**Goal :** make a decision on $\theta$ based on the observation. The decision is denoted by $\hat{\theta}$.

**Figure of merit :** average error probability

$$P_e = \Pr\{\hat{\theta} \neq \theta\}$$

# Decision regions

- The value of **y** leads to one deterministic decision
- The value of **y** can be viewed as a position in $\mathbb{K}^n$

Let *decision region associated with* $\theta_0$ be as follows

$$\Omega_{\theta_0} := \{\mathbf{y} \in \mathbb{K}^n : \hat{\theta}(\mathbf{y}) = \theta_0\}, \qquad \forall \theta_0 \in \Theta,$$

i.e., the set of observations **y** leading the decoder to decide the value $\theta_0$ for the parameter

**Remark:**
We have a partition of $\mathbb{K}^n$

$$\Omega_\theta \ \cap \ \Omega_{\theta'} = \emptyset, \qquad \forall \theta, \theta' \in \Theta, \ \theta \neq \theta'$$

and

$$\bigcup_{\theta \in \Theta} \Omega_\theta = \mathbb{K}^n.$$

# Main results

## Result 1

Minimizing $P_e$ leads to make the following decision

$$\hat{\theta} = \arg\max_{\theta\in\Theta} p_{\theta|Y}(\theta|\mathbf{y})$$

i.e., $\Omega_\theta = \{\mathbf{y} \in \mathbb{K}^n : p_{\theta|Y}(\theta|\mathbf{y}) \geq p_{\theta|Y}(\theta'|\mathbf{y}),\ \forall\theta' \neq \theta\}$
Optimal decoder: Maximum A Posteriori (MAP)

## Result 2 (special case)

Minimizing $P_e$ leads to make the following decision if $\theta$ equilikely

$$\hat{\theta} = \arg\max_{\theta\in\Theta} p_{Y|\theta}(\mathbf{y}|\theta)$$

i.e., $\Omega_\theta = \{\mathbf{y} \in \mathbb{K}^n : p_{Y|\theta}(\mathbf{y}|\theta) \geq p_{Y|\theta}(\mathbf{y}|\theta'),\ \forall\theta' \neq \theta\}$
Optimal decoder: Maximum Likelihood (ML)

# Main questions

- Description of $\Omega_\theta$ (region borders?)

  or equivalenty

- Derivations of $p_{\theta|Y}$ or $p_{Y|\theta}$ ?
- Finding out $\arg\max$ ?

# Sketch of proof

$$P_e = 1 - P_d \text{ with } P_d := \Pr\{\hat{\theta} = \theta\}.$$

We get

$$\begin{aligned} P_d &= \sum_{\theta_0 \in \Theta} \Pr\{\hat{\theta} = \theta_0 | \theta = \theta_0\} \cdot \Pr\{\theta = \theta_0\} \\ &= \sum_{\theta_0 \in \Theta} \int_{\mathbf{y} \in \Omega_\theta} p_{Y|\theta}(\mathbf{y}|\theta = \theta_0) \cdot \Pr\{\theta = \theta_0\}, \\ &= \int_{\mathbf{y} \in \mathbb{K}^n} \sum_{\theta_0 \in \Theta} \mathbf{1}\{\mathbf{y} \in \Omega_\theta\} p_{Y|\theta}(\mathbf{y}|\theta = \theta_0) \cdot \Pr\{\theta = \theta_0\} d\mathbf{y}, \\ &= \int_{\mathbf{y} \in \mathbb{K}^n} \left( \sum_{\theta_0 \in \Theta} \mathbf{1}\{\mathbf{y} \in \Omega_\theta\} p_{\theta|Y}(\theta = \theta_0 | \mathbf{y}) \right) p_Y(\mathbf{y}) d\mathbf{y}. \end{aligned}$$

For each $\mathbf{y}$, we select (and we need to select at most one) $\theta_0$ maximizing $\theta_0 \mapsto p_{\theta|Y}(\theta = \theta_0 | \mathbf{y})$

# Example 1: SISO case

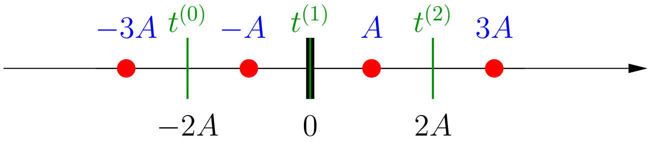
Let Single-Input-Single Output (SISO) case

$$y = s + w$$

with $s \in$ 4PAM and $w$ a zero-mean Gaussian noise with variance $\sigma^2$
ML can be written as follows

$$\hat{s} = \arg \min_{s \in \text{4PAM}} |y - s|^2$$

which leads to the following decision region




**Remark:** decision regions are described by the bisector between admissible points. We call this decoder as **threshold detector**.

# Example 2: MIMO

**Multiple Input - Multiple Ouput (MIMO):** $N_r$ receive antennas and $N_t$ transmit antennas

- increase the data rate significantly,
- better reliability for communications links.

$$y^{(r)} = \sum_{t=1}^{N_t} h_{r,t} s^{(t)} + w^{(r)} \Leftrightarrow \mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{w}$$

with $\mathbf{y} = [y^{(1)}, \cdots, y^{(N_r)}]^{\mathrm{T}}$, $\mathbf{H} = [h_{r,t}]_{1\leq r\leq N_r, 1\leq t\leq N_t}$, $\mathbf{s} = [s^{(1)}, \cdots, s^{(N_t)}]^{\mathrm{T}}$, and $\mathbf{w} = [w^{(1)}, \cdots, w^{(N_r)}]^{\mathrm{T}}$.

**Remark:** very generic model (actually any linear operator)

**Goal**

Carrying out the optimal decoder $\Leftrightarrow$ derive $p_{Y|S}(\mathbf{y}|\mathbf{s})$.

# Example 2: MIMO (cont'd)

As the noise is independent on each antenna, we have

$$p(\mathbf{y}|\mathbf{s}) = p(y^{(1)}|\mathbf{s}) \cdots p(y^{(N_r)}|\mathbf{s}).$$

As $w^{(r)}$ is a zero-mean Gaussian variable with variance $\sigma^2$, we obtain

$$p(y^{(r)}|\mathbf{s}) \propto e^{-\frac{\left(y^{(r)} - \sum_{t=1}^{N_t} h_{r,t} s_\ell^{(t)}\right)^2}{2\sigma^2}}$$

which leads to

$$p(\mathbf{y}|\mathbf{s}) \propto e^{-\frac{\sum_{r=1}^{N_r}\left(y^{(r)} - \sum_{e=1}^{N_e} h_{r,e} s_\ell^{(e)}\right)^2}{2\sigma^2}} = e^{-\frac{\|\mathbf{y}-\mathbf{H}\mathbf{s}\|^2}{2\sigma^2}}.$$

with the norm $L^2$ s.t. $\|\mathbf{x}\|^2 = \sum_r x_r^2$.

**Result**

$$\hat{\mathbf{s}} = \arg \min_{\mathbf{s}\in\mathcal{M}^{N_t}} \underbrace{\|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2}_{:=f(\mathbf{s})}.$$

**Remark:** discrete optimization in high dimension (*massive MIMO :* $N_t = 256$)

# Example 3: MIMO with Laplacian noise

We replace the Gaussian noise by a Laplacian noise (per antenna)

$$p_W(w) = \frac{1}{\sqrt{2\sigma^2}} e^{-\frac{2|w|}{\sqrt{2\sigma^2}}}.$$

Typically

- noise composed by some other users (collisions)
- more impulsive noise

Same approach as in previous slides, we have

$$\hat{s} = \arg \min_{s \in \mathcal{M}^{N_t}} \|\mathbf{y} - \mathbf{H}\mathbf{s}\|_1$$

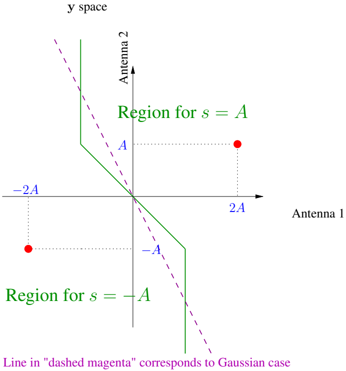
with the norm $L^1$ s.t. $\|\mathbf{x}\|_1 = \sum_r |x_r|$.

**Remarks:**

- **distance** $L^1 =$, Manhattan distance.
- Noise distribution (which provides the statistical link between input and output) plays a great role and strongly modifies the decoder through the involved distance!

# Example 3: special case (SITO)

$N_t = 1$, $N_r = 2$, $h_{1,1} = 2$ et $h_{2,1} = 1$, 2PAM.




Decision regions' border

- bisector in Gaussian case
- piecewise linear function (angles: 0, $90^o$, $45^o$, $-45^o$) in Laplacian case (counter-intuitive)

# Part 3 : Hypothesis Testing

# Introduction

$$\begin{cases} \text{Hypothesis } \mathcal{H}_0: & y \sim p_{Y|\mathcal{H}_0} \\ \text{Hypothesis } \mathcal{H}_1: & y \sim p_{Y|\mathcal{H}_1} \end{cases}$$

**Remark:** $p_{Y|\mathcal{H}_0} \neq p_{Y|\mathcal{H}_1}$. If not, problem unfeasible since we can not distinghuish between both hypotheses based on the statistical properties.

**Figure of merit :** maximizing probability of detection (power of the test)

$$P_D = \Pr\{\mathcal{H}_1|\mathcal{H}_1\}$$

or equivalenty minimizing probability of miss detection (probability of Type-II error)

$$P_M = \Pr\{\mathcal{H}_0|\mathcal{H}_1\}$$

s.t. probability of false alarm (probability of Type-I error) below a predefined threshold

$$P_{FA} = \Pr\{\mathcal{H}_1|\mathcal{H}_0\} \leq P_{FA}^{\text{target}}$$

# Main results

## Result

Minimizing the miss detection probability s.t. false alarm probability is below a threshold leads to the so-called Neyman-Pearson test, also called Likelihood Ratio Test (LRT), defined as follows

$$\Lambda(y) = \log\left(\frac{p_{Y|\mathcal{H}_1}(y)}{p_{Y|\mathcal{H}_0}(y)}\right) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \mu,$$

with

- $\Lambda$ the *Log Likelihood Ratio* (LLR)
- $\mu$ the threshold enabling to satisfy the target false alarm probability $P_{FA}^{\text{target}}$

# Main questions

- Derivations of $p_{Y|\mathcal{H}_0}$ ?
- Derivations of $p_{Y|\mathcal{H}_1}$ ?
- Derivations of $\mu$ ?

# Sketch of proof

$$P_D = \int_{\Omega_1} p_{Y|\mathcal{H}_1}(y)dy$$
$$= \int_{\mathbb{K}^n} \mathbf{1}\{y \in \Omega_1\} p_{Y|\mathcal{H}_1}(y)dy$$

and

$$P_{FA} = \int_{\Omega_1} p_{Y|\mathcal{H}_0}(y)dy$$
$$= \int_{\mathbb{K}^n} \mathbf{1}\{y \in \Omega_1\} p_{Y|\mathcal{H}_0}(y)dy$$

- Let $T$ be the Neyman-Parson test (written in terms of probablity of selecting $\mathcal{H}_1$)

$$T : \begin{cases} T(y) = 1 & \text{if} \quad p_{Y|\mathcal{H}_1}(y) > \mu p_{Y|\mathcal{H}_0}(y) \\ T(y) = t & \text{if} \quad p_{Y|\mathcal{H}_1}(y) = \mu p_{Y|\mathcal{H}_0}(y) \\ T(y) = 0 & \text{if} \quad p_{Y|\mathcal{H}_1}(y) < \mu p_{Y|\mathcal{H}_0}(y) \end{cases}$$

- Let $T'$ be any other test s.t. $P_{FA} \leq P_{FA}^{\text{target}}$

# Sketch of proof (cont'd)

We have

$$\forall y, \quad (T(y) - T'(y))(p_{Y|\mathcal{H}_1}(y) - \mu p_{Y|\mathcal{H}_0}(y)) \geq 0$$

$$\Rightarrow \quad \int_{\mathbb{K}^n} (T(y) - T'(y))(p_{Y|\mathcal{H}_1}(y) - \mu p_{Y|\mathcal{H}_0}(y))dy \geq 0$$

$$\Rightarrow \quad \int_{\mathbb{K}^n} (T(y) - T'(y))p_{Y|\mathcal{H}_1}(y)dy \geq \mu \int_{\mathbb{K}^n} (T(y) - T'(y))p_{Y|\mathcal{H}_0}(y)dy$$

$$\Rightarrow \quad P_D - P'_D \geq \mu(P_{FA} - P'_{FA})$$

$$\Rightarrow \quad P_D - P'_D \geq \mu(P_{FA}^{\text{target}} - P'_{FA})$$

$$\Rightarrow \quad P_D - P'_D \geq 0$$

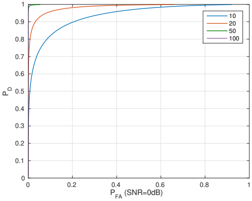
# ROC curve

**Remarks:**

- If $T(y) = 1, \forall y$, then $P_D = 1$ and $P_{FA} = 1$ (in military context: launch always a missile!)
- So $P_D$ strongly depends on $P_{FA}$, and $P_D$ should be plotted versus $P_{FA}$

**Definition:**

Given a configuration (SNR, number of samples, etc), function $P_{FA} \mapsto P_D$ is called *Receiver Operating Characteristics (ROC)* curve




How to draw it? plot the pair $(P_{FA}(\mu), P_D(\mu))$ for any $\mu$

# Example: Gaussian signal in Gaussian noise

$$\left\{\begin{array}{ll}\mathcal{H}_0 & : y(n) = w(n) \\ \mathcal{H}_1 & : y(n) = x(n) + w(n)\end{array}\right. , n = 1, \cdots, N$$

with

- $w(n)$ iid zero-mean Gaussian noise with known $\sigma_w^2 = \mathbb{E}[|w(n)|^2]$,
- $x(n)$ also iid zero-mean Gaussian with known variance $\sigma_x^2 = \mathbb{E}[|x(n)|^2]$

We have

$$\left\{\begin{array}{lll} p_{Y|\mathcal{H}_0}(\mathbf{y}) & = & \prod_{n=1}^{N} p_{Y|\mathcal{H}_0}(y_n) \text{ with } p_{Y|\mathcal{H}_0}(y_n) = \frac{1}{\pi\sigma_w^2} e^{-\frac{|y_n|^2}{\sigma_w^2}} \\ p_{Y|\mathcal{H}_1}(\mathbf{y}) & = & \prod_{n=1}^{N} p_{Y|\mathcal{H}_1}(y_n) \text{ with } p_{Y|\mathcal{H}_1}(y_n) = \frac{1}{\pi(\sigma_x^2+\sigma_w^2)} e^{-\frac{|y_n|^2}{\sigma_x^2+\sigma_w^2}} \end{array}\right.$$

with $\mathbf{y} = [y(1), \cdots, y(N)]^{\mathrm{T}}$

# Example: LRT

$$\begin{aligned}
\Lambda(\mathbf{y}) &= \log\left(\frac{\frac{1}{\pi(\sigma_x^2+\sigma_w^2)}e^{-\frac{\sum_{n=1}^{N}|y_n|^2}{\sigma_x^2+\sigma_w^2}}}{\frac{1}{\pi\sigma_w^2}e^{-\frac{\sum_{n=1}^{N}|y_n|^2}{\sigma_w^2}}}\right)\\
&= \log\left(\frac{\sigma_w^2}{\sigma_x^2+\sigma_w^2}e^{-\left(\frac{1}{\sigma_x^2+\sigma_w^2}-\frac{1}{\sigma_w^2}\right)\sum_{n=1}^{N}|y_n|^2}\right)\\
&= \text{positive constant}\times\sum_{n=1}^{N}|y_n|^2+\text{constant}
\end{aligned}$$

**LRT = energy test is optimal!**

$$T(\mathbf{y}) = \frac{1}{\sigma_x^2+\sigma_w^2}\sum_{n=1}^{N}|y(n)|^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta$$

# Example: performances derivations

- Under $\mathcal{H}_1$, $T(\mathbf{y})$ follows a $\chi_2$-distribution with $2N$ degrees of freedom with pdf

$$p_{\chi_2,2N}(x) = \frac{1}{\Gamma_c(N)} x^{N-1} e^{-x}, \ x \geq 0$$

- Under $\mathcal{H}_0$, $T(\mathbf{y})$ follows a $\chi_2$-distribution with $2N$ degrees of freedom with pdf

$$p_{\chi_2,2N}(x) = \frac{1}{(\sigma_w^2/(\sigma_x^2+\sigma_w^2))^N \Gamma_c(N)} x^{N-1} e^{-\frac{(\sigma_x^2+\sigma_w^2)x}{\sigma_w^2}}, \ x \geq 0$$

with complete and incomplete Gamma function

$$\Gamma_c(s) = \int_0^\infty x^{s-1} e^{-x} dx$$

and

$$\Gamma_{\text{inc}}(s,u) = \int_u^\infty x^{s-1} e^{-x} dx$$

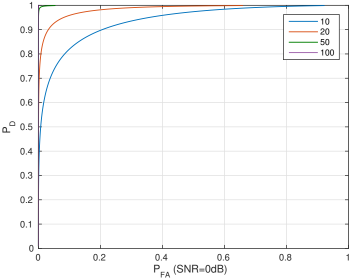
# Example: performances derivations (cont'd)

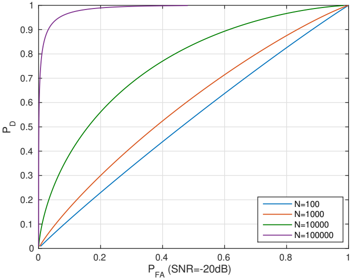
$$\begin{aligned} P_{FA} &= \Pr(T(\mathbf{y}) > \eta | \mathcal{H}_0) \\ &= \int_{\eta}^{\infty} \frac{1}{(\sigma_w^2/(\sigma_x^2+\sigma_w^2))^N \Gamma_c(N)} x^{N-1} e^{-\frac{(\sigma_x^2+\sigma_w^2)x}{\sigma_w^2}} dx \\ &= \frac{1}{\Gamma_c(N)} \cdot \frac{1}{(\sigma_w^2/(\sigma_x^2+\sigma_w^2))^N} \cdot \int_{\eta}^{\infty} x^{N-1} e^{-\frac{(\sigma_x^2+\sigma_w^2)x}{\sigma_w^2}} dx \\ &= \frac{\Gamma_{\text{inc}}\left(N, \eta\frac{\sigma_x^2+\sigma_w^2}{\sigma_w^2}\right)}{\Gamma_c(N)} \end{aligned}$$

Similarly

$$P_D = \frac{\Gamma_{\text{inc}}(N,\eta)}{\Gamma_c(N)}$$

# Example: ROC


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P_D
0
0.2
0.4
0.6
0.8
1
P_FA (SNR=0dB)
10
20
50
100


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P_D
0
0.2
0.4
0.6
0.8
1
P_FA (SNR=-20dB)
N=100
N=1000
N=10000
N=100000

# Asymptotic regime for generic case

- In general, very difficult to obtain $P_D$ and $P_{FA}$ in closed-form (the previous example is a counter-case)
- To overcome this issue, asymptotic regime ($N \to \infty$)

## Stein's lemma

Under iid assumption for $y_1, \cdots y_N$, we denote $P_1 = P_{\mathcal{H}_1}(y_1)$ and $P_0 = P_{\mathcal{H}_0}(y_1)$. For any $\varepsilon$,

- it exists a sequence of tests $T_N$, s.t.,
  - $P_D(T_N) > 1 - \varepsilon$ for $N$ large enough,
  - and $P_{FA}(T_N) < e^{-N(D(P_1 \| P_0) - \varepsilon)}$
- Let $T'_N$ be a sequence of tests s.t. $P_D(T'_N) > 1 - \varepsilon$. Then $P_{FA}(T'_N) > (1 - 2\varepsilon)e^{-N(D(P_1 \| P_0) + \varepsilon)}$

with $D(P_1 \| P_0)$ the Kullback-Leibler distance defined as

$$D(P_1 \| P_0) := \int P_1(y) \log \left( \frac{P_1(y)}{P_0(y)} \right) dy = \mathbb{E}_{y \sim P_1} \left[ \log \left( \frac{P_1}{P_0} \right) \right]$$

# Sketch of proof: achievability

Let $T_N$ be the following test:

$$\Omega_1 = \left\{ \mathbf{y} | D(P_1 \| P_0) - \varepsilon \leq \frac{1}{N} \log\left( \frac{P_{Y|\mathcal{H}_1}(\mathbf{y})}{P_{Y|\mathcal{H}_0}(\mathbf{y})} \right) \leq D(P_1 \| P_0) + \varepsilon \right\}$$

We have

1. If $\mathbf{y} \in \mathcal{H}_1$, $\lim_{N \to \infty} \frac{1}{N} \log\left( \frac{P_{Y|\mathcal{H}_1}(\mathbf{y})}{P_{Y|\mathcal{H}_0}(\mathbf{y})} \right) \overset{probability}{=} D(P_1 \| P_0)$
2. $P_D(T_N) > 1 - \varepsilon$, for $N$ large enough
3. $\forall \mathbf{y} \in \Omega_1$,
   $P_{Y|\mathcal{H}_1}(\mathbf{y}) e^{-N(D(P_1 \| P_0) + \varepsilon)} \leq P_{Y|\mathcal{H}_0}(\mathbf{y}) \leq P_{Y|\mathcal{H}_1}(\mathbf{y}) e^{-N(D(P_1 \| P_0) - \varepsilon)}$
4. $P_{FA}(T_N) \leq e^{-N(D(P_1 \| P_0) - \varepsilon)}$

# Sketch of proof: achievability (cont'd)

(1.)

$$\begin{aligned}\frac{1}{N}\log\left(\frac{P_{Y|\mathcal{H}_1}(\mathbf{y})}{P_{Y|\mathcal{H}_0}(\mathbf{y})}\right) &\overset{iid}{=} \frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{P_1(y_n)}{P_0(y_n)}\right)\\ &\overset{WLLN}{\rightarrow} \mathbb{E}_{y\sim P_1}\left[\log\left(\frac{P_1}{P_0}\right)\right] \text{ in probability}\\ &= D(P_1\|P_0)\end{aligned}$$

(2.) $\lim_{N\to\infty}\Pr\{|T_N(\mathbf{y})-D(P_1\|P_0)|>\varepsilon\}=0 \Rightarrow, \exists N_0(\varepsilon), N > N_0(\varepsilon), \text{ s.t. } 1-P_D(T_N)=\Pr\{|T_N(\mathbf{y})-D(P_1\|P_0)|>\varepsilon\}\leq\varepsilon$

(3.) Just manipulating the inequalities in $\Omega_1$

(4.)

$$\begin{aligned}P_{FA}(T_N) &= \int_{\mathbf{y}\in\Omega_1} P_{Y|\mathcal{H}_0}(\mathbf{y})d\mathbf{y} \overset{(3.)}{\leq} \int_{\mathbf{y}\in\Omega_1} P_{Y|\mathcal{H}_1}(\mathbf{y})e^{-N(D(P_1\|P_0)-\varepsilon)}d\mathbf{y}\\ &\leq e^{-N(D(P_1\|P_0)-\varepsilon)}\int_{\mathbf{y}\in\Omega_1} P_{Y|\mathcal{H}_1}(\mathbf{y})d\mathbf{y}\\ &\leq e^{-N(D(P_1\|P_0)-\varepsilon)}P_D(T_N)\leq e^{-N(D(P_1\|P_0)-\varepsilon)}\end{aligned}$$

# Sketch of proof: converse

Let $T_N \cap T'_N$ be the composite test ($\mathcal{H}_1$ is decided iff both decode $\mathcal{H}_1$)

As $P_D(T_N) > 1-\varepsilon$ and $P_D(T'_N) > 1-\varepsilon$, we have

$$P_D(T_N \cap T'_N) > 1-2\varepsilon$$

Moreover

$$\begin{aligned}
P_{FA}(T'_N) &\geq P_{FA}(T_N \cap T'_N) \\
&= \int_{\mathbf{y}\in\Omega_1(T_N)\cap\Omega_1(T'_N)} P_{Y|\mathcal{H}_0}(\mathbf{y})d\mathbf{y} \\
&\overset{(3.)}{\geq} \int_{\mathbf{y}\in\Omega_1(T_N)\cap\Omega_1(T'_N)} P_{Y|\mathcal{H}_1}(\mathbf{y})e^{-N(D(P_1\|P_0)+\varepsilon)}d\mathbf{y} \\
&= e^{-N(D(P_1\|P_0)+\varepsilon)}P_D(T_N \cap T'_N) \\
&\geq (1-2\varepsilon)e^{-N(D(P_1\|P_0)+\varepsilon)}
\end{aligned}$$

# Extension: Generalized LRT (GLRT)

**Problem:** in many applications, some parameters of the pdf are unknown (e.g. the variance)

**Goal:** testing the hypotheses but the hypotheses are partially unknown (through some parameters of nuisance)

- Let $\nu$ be the nuisance parameters
- Let $P_{Y|\mathcal{H}_1}(\mathbf{y};\nu)$ be the pdf under $\mathcal{H}_1$ for one value of $\nu$
- Let $P_{Y|\mathcal{H}_0}(\mathbf{y};\nu)$ be the pdf under $\mathcal{H}_0$ for one value of $\nu$

$$T(\mathbf{y}) = \frac{\max_\nu P_{Y|\mathcal{H}_1}(\mathbf{y};\nu)}{\max_\nu P_{Y|\mathcal{H}_0}(\mathbf{y};\nu)}$$

- No optimality result
- No asymptotic result

# Extension: Bayesian LRT (BLRT)

- We have a priori distribution on parameters of nuisance $\nu$
- Let $q$ be the known a priori distribution of $\nu$ (typically offset)

$$T(\mathbf{y}) = \frac{\int P_{Y|\mathcal{H}_1}(\mathbf{y};\nu)q(\nu)d\nu}{\int P_{Y|\mathcal{H}_0}(\mathbf{y};\nu)q(\nu)d\nu}$$

- No optimality result
- No asymptotic result

# Part 4 : Estimation for deterministic parameters

# Statistics

- Let $\mathbf{y}_N = \{y_1, \ldots, y_N\}$ be a (multi-variate) observation of the process $\mathbf{Y}_N$
- A *statistic* is any function $T$ only depending on the observation

$$T(\mathbf{y}_N)$$

- Any statistic is a random variable (and will be studied as it)

**but** few questions before

- How characterizing $T$ s.t. $T$ provides on $\theta$ information enough?
- In other words, how representing $\mathbf{y}_N$ in a compact form through $T$ without loosing information on $\theta$?

$\Rightarrow$ Fundamental concept of *sufficient* statistics

- If $T$ is a sufficient statistic, is it close to $\theta$?

$\Rightarrow$ Rao-Blackwell theorem

# Sufficient statistics

**Reminder**

$\mathbf{y}$ provides information on $\theta$ iff the pdf of $\mathbf{y}_N$, denoted by

$$p(\mathbf{y}_N;\theta) \text{ or } p(\mathbf{y}_N|\theta),$$

depends on $\theta$

$T$ is said sufficient statistics iff given the random variable $T(\mathbf{Y}_N)$, pdf on the whole observation is useless. Consequently, the random variable $\mathbf{Y}_N|T(\mathbf{Y}_N)$ has a pdf independent of $\theta$

$$p_{Y|T}(\mathbf{y}_N|T(\mathbf{Y}_N);\theta) \text{ does not depend on } \theta$$

**Remark:** in practice difficult to check that $T$ is a sufficient statistic by using this definition

# Sufficient statistics: properties

## Fisher factorization theorem

$T$ is a sufficient statistic of $\theta$ iff it exists two functions $g_\theta(.)$ (depending on $\theta$) and $h(.)$ (independent of $\theta$) s.t.

$$p(\mathbf{y}_N;\theta) = g_\theta(T(\mathbf{y}_N))h(\mathbf{y}_N)$$

**Remark:** The Likelihood Ratio (between two values: $\theta$ and $\theta'$) depends only on $T(\mathbf{y}_N)$

$$\frac{p(\mathbf{y}_N;\theta)}{p(\mathbf{y}_N;\theta')} = \frac{g_\theta(T(\mathbf{y}_N))}{g_{\theta'}(T(\mathbf{y}_N))}.$$

So, to distinguish $\theta$ from $\theta'$, evaluating $T(\mathbf{y}_N)$ is enough

# Sketch of proof

If $T$ is sufficient statistic, then

$$\begin{aligned} p_Y(\mathbf{y}_N;\theta) &= \int p_{Y|T}(\mathbf{y}_N|t;\theta)p_T(t;\theta)dt \\ &\stackrel{(a)}{=} p_{Y|T}(\mathbf{y}_N|T(\mathbf{y}_N);\theta)p_T(T(\mathbf{y}_N);\theta) \\ &\stackrel{(b)}{=} \underbrace{p_{Y|T}(\mathbf{y}_N|T(\mathbf{y}_N))}_{h(\mathbf{y}_N)}\underbrace{p_T(T(\mathbf{y}_N);\theta)}_{g_\theta(T(\mathbf{y}_N))} \end{aligned}$$

$(a)$ if $t' \neq T(\mathbf{y}_N)$, then $p_{Y|T}(\mathbf{y}_N|t';\theta) = 0$

$(b)$ $T$ sufficient statistic

# Sketch of proof (cont'd)

If $p(\mathbf{y}_N;\theta) = g_\theta(T(\mathbf{y}_N))h(\mathbf{y}_N)$, we have

- If $t \neq T(\mathbf{y}_N)$,

$$p_{Y|T}(\mathbf{y}_N|T(\mathbf{Y}_N) = t;\theta) \quad = \quad 0$$

- If $t = T(\mathbf{y}_N)$,

$$\begin{aligned}
p_{Y|T}(\mathbf{y}_N|T(\mathbf{Y}_N) = t;\theta) &\overset{\text{Bayes}}{=} \frac{p_{Y,T}(\mathbf{y}_N, T(\mathbf{Y}_N) = t;\theta)}{p_T(T(\mathbf{Y}_N) = t;\theta)} \\
&\overset{(c)}{=} \frac{p_Y(\mathbf{y}_N;\theta)}{p_T(T(\mathbf{Y}_N) = t;\theta)} \\
&\overset{(d)}{=} \frac{p_Y(\mathbf{y}_N;\theta)}{\int_{y|T(y)=t} p_Y(y;\theta)dy} \\
&= \frac{g_\theta(t)h(\mathbf{y}_N)}{\int_{y|T(y)=t} g_\theta(t)h(y)dy} = \frac{h(\mathbf{y}_N)}{\int_{y|T(y)=t} h(y)dy}
\end{aligned}$$

(c) $p_{Y,T}(\mathbf{y}_N, T(\mathbf{Y}_N) = t;\theta) = p_Y(\mathbf{y}_N;\theta)$

(d) $p_T(T(\mathbf{Y}_N) = t;\theta) = \int_y p_{Y,T}(y, T(\mathbf{Y}_N) = t;\theta)dy \overset{(c)}{=} \int_{y|T(y)=t} p_Y(y;\theta)dy$

# Application

As an estimate of $\theta$, we may have

$$\hat{\theta}_N = \arg\max_{\theta} p(\mathbf{y}_N; \theta)$$

If $T$ is a sufficient statistic, then

$$\begin{aligned} \hat{\theta}_N &= \arg\max_{\theta} g_\theta(T(\mathbf{y}_N)) \\ &= \mathrm{fct}(T(\mathbf{y})) \end{aligned}$$

and only the knowledge of $T(\mathbf{y}_N)$ is enough to estimate $\theta$.

**Questions:**

- What is the function $\mathrm{fct}$?
- Is $\hat{\theta}_N = T(\mathbf{y}_N)$ a reasonnable choice?

# Figures of merit for $\hat{\theta}_N$

**Remarks:**

- An *estimate* $\hat{\theta}_N$ of $\theta$ is just a statistic “close” to $\theta$

$$\hat{\theta}_N = \hat{\theta}(\mathbf{y}_N)$$

- “Close” implies we need a cost function $C(\hat{\theta}_N, \theta)$ to measure the gap between $\hat{\theta}_N$ and $\theta$.
  - $\mathbf{1}(\|\hat{\theta}_N - \theta\| \geq \varepsilon)$ : uniform cost
  - $\|\hat{\theta}_N - \theta\|_{L^1}$ : Manhattan cost ($L^1$ norm)
  - $\|\hat{\theta}_N - \theta\|^2$ : quadratic/Euclidian cost ($L^2$ norm)

**Risk**

We average the cost function over all the values of $\mathbf{y}_N$

$$\begin{aligned} R(\hat{\theta}_N, \theta) &= \mathbb{E}[C(\hat{\theta}_N, \theta)] \\ &= \int C(\hat{\theta}_N(\mathbf{y}_N), \theta) p(\mathbf{y}_N; \theta) d\mathbf{y}_N \end{aligned}$$

# Biais and Mean Square Error (MSE)

**Bias:**

$$b(\theta, \hat{\theta}_N) = \mathbb{E}[\hat{\theta}(\mathbf{y}_N)] - \theta$$

**Variance:**

$$\mathrm{var}(\theta, \hat{\theta}_N) = \mathbb{E}[\|\hat{\theta}(\mathbf{y}_N) - \mathbb{E}[\hat{\theta}(\mathbf{y}_N)]\|^2]$$

**Mean Square Error:**

$$\begin{aligned} \mathrm{MSE}(\theta, \hat{\theta}_N) &= \mathbb{E}[\|\hat{\theta}(\mathbf{y}_N) - \theta\|^2] \\ &= \|b(\theta, \hat{\theta}_N)\|^2 + \mathrm{var}(\theta, \hat{\theta}_N) \end{aligned}$$

## Remarks

- Bias and variance are the mean and variance of the random variable $\hat{\theta}_N$ respectively
- An estimate is called *unbiased/biasfree* iff $b(\theta, \hat{\theta}_N) = 0$
- Warning: the quality of the estimate depends on the considered figures of merit

# Sufficient statistics and estimate's design

## Rao-Blackwell theorem

- Let $T$ be a sufficient statistic for $\theta$
- Let $T'$ be an unbiased estimate for $\theta$
- Let $T'' = \mathbb{E}[T'|T]$

Then

- $T''$ is an unbiased estimate of $\theta$

$$\mathbb{E}[T''(\mathbf{y}_N)] = \theta$$

- $T''$ does not offer a worse MSE than $T'$

$$\mathbb{E}[\|T''(\mathbf{y}_N) - \theta\|^2] \leq \mathbb{E}[\|T'(\mathbf{y}_N) - \theta\|^2]$$

# Sketch of proof

- As $T$ sufficient statistic, $T''$ does not depend on $\theta$

$$T'' = \mathbb{E}[T'|T] = \int t'(y)p_{Y|T}(y|t;\theta)dy = \int t'(y)p_{Y|T}(y|t)dy,$$

can be evaluated by knowing $\mathbf{y}_N$ only. So, $T''$ is a statistic for $\theta$

- In addition, we get

$$\begin{aligned}
\mathbb{E}[T''] &= \mathbb{E}[\mathbb{E}[T'|T]] \\
&= \iint t' p_{T'|T}(t';\theta)dt' p_T(t)dt \\
&= \iint t' p_{T',T}(t',t;\theta)dt'dt \\
&= \int t' \left( \int p_{T',T}(t',t;\theta)dt \right) dt' \\
&= \int t' p_{T'}(t';\theta)dt' \\
&= \mathbb{E}[T'] = \theta
\end{aligned}$$

# Sketch of proof (cont'd)

- If $\tilde{T}$ unbiased $\mathbb{E}[(\tilde{T}-\theta)^2] = \mathbb{E}[\tilde{T}^2] + \theta^2$, then

$$\mathbb{E}[\| T''(\mathbf{y}_N) - \theta \|^2] \leq \mathbb{E}[\| T'(\mathbf{y}_N) - \theta \|^2] \Leftrightarrow \mathbb{E}[\| T''(\mathbf{y}_N) \|^2] \leq \mathbb{E}[\| T'(\mathbf{y}_N) \|^2]$$

- Then

$$\begin{aligned} \mathbb{E}[\| T''(\mathbf{y}_N) \|^2] &\overset{(a)}{=} \mathbb{E}[\| \mathbb{E}[T'(\mathbf{y}_N) | T] \|^2] \\ &\overset{(b)}{\leq} \mathbb{E}[\mathbb{E}[\| T'(\mathbf{y}_N) \|^2 | T]] \\ &\overset{(c)}{=} \mathbb{E}[\| T'(\mathbf{y}_N) \|^2] \end{aligned}$$

(a) replace $T''$ by its definition
(b) Jensen inequality: let $\phi$ be a convex function, then $\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]$
(c) similar to previous slide with $\| T' \|^2$ instead of $T'$

# Consequences

## Minimum-Variance Unbiased Estimator (MVUE)

- Let $T$ be a sufficient statistic for $\theta$
- Assume that it exists an unique function $f$ s.t. $\mathbb{E}[f(T)] = \theta$

Then $f(T)$ is a Minimum Variance Unbiased Estimate of $\theta$

## Notion of completeness

A sufficient statistic $T$ is said *complete* iff

$$\mathbb{E}[h(T)] = 0 \Rightarrow h(T) = 0, \ \forall\theta$$

As a consequence, $f(T)$ is MVUE

**Remarks:**

- Easy to find $f$ ? no
- Completeness is easier to check

# Sketch of proof

- Let $T'$ be an unbiased estimate of $\theta$. We know that $T'' = \mathbb{E}[T'|T]$ is a function of $T$ and also an unbiased estimate ($\mathbb{E}[T''] = \theta$). So $T'' = f(T)$. Consequently, $\forall T'$, we have

$$\mathbb{E}[\|T'' - \theta\|^2] \leq \mathbb{E}[\|T' - \theta\|^2]$$

- Assume $T$ complete. Consider $f_1$ and $f_2$ s.t. $\mathbb{E}[f_1(T)] = \theta$ and $\mathbb{E}[f_2(T)] = \theta$.

$$\begin{aligned} \mathbb{E}[f_1(T)] &= \mathbb{E}[f_2(T)] \\ \mathbb{E}[(f_1 - f_2)(T)] &= 0 \\ f_1 - f_2 &= 0 \\ f_1 &= f_2 \end{aligned}$$

# Example

- Let $\mathbf{Y}_N$ be a iid Gaussian vector with unknown mean $m$ and unit-variance.
- Let $\theta = m$

We have

$$\begin{aligned} p_Y(\mathbf{y}_N;\theta) &= \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}\sum_{n=1}^N (y_n - m)^2} \\ &= \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}\sum_{n=1}^N \left(y_n^2 + m^2 - 2y_n m\right)} \\ &= \underbrace{\frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}\sum_{n=1}^N y_n^2}}_{h(\mathbf{y}_N)} \underbrace{e^{-\frac{1}{2}(Nm^2 - 2m\sum_{n=1}^N y_n)}}_{g_\theta(T(\mathbf{y}_N))} \end{aligned}$$

with

- $\hat{m}_N = \sum_{n=1}^N y_n/N$ : empirical mean
- $T(\mathbf{y}_N) = \hat{m}_N$

$T$ is a sufficient statistic for $\theta$

# Example (cont'd)

$\hat{m}_N$ is

- unbiased
- MSE $\mathrm{var}(m, \hat{m}_N) = \frac{1}{N}$
- complete statistic

$$
\begin{aligned}
\mathbb{E}[\phi(T(\mathbf{y}_N))] & \overset{(a)}{\propto} \int h(t) e^{-\frac{N}{2}(t-\theta)^2} dt = 0 \\
& \overset{(b)}{=} h \star g = 0 \\
& \overset{(c)}{=} H.G = 0 \\
& \overset{(d)}{=} H = 0
\end{aligned}
$$

*(a)* : $\hat{m}_N$ is Gaussian with mean $\theta$ and variance $1/N$
*(b)* : convolution with a Gaussian function $g$.
*(c)* : $H$ and $G$ Fourier transform of $h$ and $g$ respectively
*(d)* : $G$ is still a Gaussian function

- MVUE

# Counter-example

Consider $T(\mathbf{y}_N) = y_1$

$$
\begin{aligned}
p_{Y|Y_1}(\mathbf{y}_N|y_1;\theta) &= \frac{p_{Y,Y_1}(\mathbf{y}_N, y_1;\theta)}{p_{Y_1}(y_1;\theta)} \\
&= \mathbf{1}_{\mathbf{y}_N(1)=y_1} \frac{p_Y(\mathbf{y}_N;\theta)}{p_{Y_1}(y_1;\theta)} \\
&\propto \frac{e^{-\frac{1}{2v}\sum_{n=1}^{N}(y_n-\theta)^2}}{e^{-\frac{1}{2v}(y_1-\theta)^2}} \\
&\propto e^{-\frac{1}{2v}\sum_{n=2}^{N}(y_n-\theta)^2}
\end{aligned}
$$

- $p_{Y|Y_1}(\mathbf{y}_N|y_1;\theta)$ still depends on $\theta$
- $T = Y_1$ is not a sufficient statistic

# Performances

What have we seen?

- Sufficient statistic
- If some additional properties (difficult to satisfy), MVUE

**Still open questions**

- Fair comparison between two estimates: $\hat{\theta}_1$ is better then $\hat{\theta}_2$ wrt the risk $R$ iff

$$R(\hat{\theta}_1, \theta) \leq R(\hat{\theta}_2, \theta) \quad \forall\, \theta$$

- Is there a minimum value for the risk ?
  - if the risk is quadratic
  - if the problem is smooth enough
  - if we reduce the class of considered estimates
- the answer is **yes**
  - Cramer-Rao bound (CRB)
  - achievable sometimes (more often when $N \to \infty$)

# Smoothness

Problem is said smooth if

- $$\frac{\partial p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial \theta}_{|\theta=\theta_0}$$
  exists for any $\mathbf{y}_N$ and any $\theta_0$.
- $\mathbf{y}_N \mapsto p_{Y|\theta}(\mathbf{y}_N|\theta)$ has the same support for any $\theta$
- $$\int \frac{\partial p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial \theta} d\mathbf{x}_N = \frac{\partial}{\partial \theta} \int p_{Y|\theta}(\mathbf{y}_N|\theta) d\mathbf{y}_N = 0$$

# Example

Let $\mathbf{Y}_N$ be a iid Gaussian vector with unknown mean $\theta = m$ and unit-variance

$$p_{Y|\theta}(\mathbf{y}_N|\theta) = \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}\sum_{n=1}^{N}(y_n-\theta)^2}$$

- $\frac{\partial p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial \theta}_{|\theta=\theta_0} = (\sum_{n=1}^{N} y_n - \theta) p_{Y|\theta}(\mathbf{y}_N|\theta)$
- the support is $\mathbb{R}^N$ for any $\theta$
- $\int_{\mathbb{R}^N} p_{Y|\theta}(\mathbf{y}_N|\theta) d\mathbf{y}_N = 1$

# Cramer-Rao bound

## Result

Any unbiased estimate satisfies

$$\mathbb{E}\left[\left(\hat{\theta}-\theta_0\right)\left(\hat{\theta}-\theta_0\right)^{\mathrm{T}}\right] \geq F(\theta_0)^{-1} = \mathrm{CRB}(\theta_0)$$

with

- $F(\theta_0)$ the so-called Fisher Information Matrix (FIM) defined as

$$F(\theta_0) = \mathbb{E}\left[\left(\frac{\partial \log p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial \theta}_{|\theta=\theta_0}\right)\left(\frac{\partial \log p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial \theta}_{|\theta=\theta_0}\right)^{\mathrm{T}}\right]$$

- $\geq$: order for quadratic semi-definite matrix

# *Sketch of proof*

Let us first consider the scalar case

$$
\begin{aligned}
\mathbb{E}\left[(\hat{\theta}-\theta_0)\frac{\partial \log p_{Y|\theta}(y|\theta)}{\partial\theta}_{|\theta_0}\right] &= \int (\hat{\theta}-\theta_0)\frac{\partial \log p_{Y|\theta}(y|\theta)}{\partial\theta}_{|\theta_0} p_{Y|\theta}(y|\theta_0)dy \\
&= \int (\hat{\theta}-\theta_0)\frac{\partial p_{Y|\theta}(y|\theta)}{\partial\theta}_{|\theta_0} dy \\
&= \int \hat{\theta}\frac{\partial p_{Y|\theta}(y|\theta)}{\partial\theta}_{|\theta_0} dy - \theta_0 \int \frac{\partial p_{Y|\theta}(y|\theta)}{\partial\theta}_{|\theta_0} d \\
&= \frac{\partial}{\partial\theta}\mathbb{E}[\hat{\theta}] = \frac{\partial}{\partial\theta}\theta \\
&= 1
\end{aligned}
$$

Then Cauchy-Schwartz inequality

$$\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$$

# Sketch of proof (cont'd)

By setting $D(y,\theta_0) = \frac{\partial \log p_{Y|\theta}(y|\theta)}{\partial \theta}_{|\theta_0}$ a column vector, we have by construction

$$M := \mathbb{E}\left[\left(F(\theta_0)^{-1}D(y,\theta_0) - (\hat{\theta} - \theta_0)\right)\left(F(\theta_0)^{-1}D(y,\theta_0) - (\hat{\theta} - \theta_0)\right)^{\mathrm{T}}\right] \geq 0$$

So

$$\begin{aligned} M &= \mathbb{E}\left[F(\theta_0)^{-1}D(y,\theta_0)D(y,\theta_0)^{\mathrm{T}}F(\theta_0)^{-1}\right] + \mathbb{E}[(\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)^{\mathrm{T}}] \\ &- \mathbb{E}\left[F(\theta_0)^{-1}D(y,\theta_0)(\hat{\theta}-\theta_0)^{\mathrm{T}}\right] - ()^{\mathrm{T}} \\ &= F(\theta_0)^{-1} + \mathbb{E}[(\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)^{\mathrm{T}}] - F(\theta_0)^{-1}\mathbb{E}[D(y,\theta_0)(\hat{\theta}-\theta_0)^{\mathrm{T}}] - ()^{\mathrm{T}} \end{aligned}$$

In addition

$$\mathbb{E}[D(y,\theta_0)(\hat{\theta}-\theta_0)^{\mathrm{T}}] = \mathbf{Id}$$

# Application

$$\mathrm{MSE}(\hat{\theta}, \theta_0) \geq \mathrm{trace}(F(\theta_0)^{-1})$$

since

$$\mathrm{MSE} = \sum_{n=1}^{N_\theta} \mathbb{E}[|\hat{\theta}(n) - \theta_0(n)|^2] = \mathrm{trace}(\mathbb{E}[(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^{\mathrm{T}}])$$

and

$$\mathbf{A} \geq \mathbf{B} \Rightarrow \mathrm{trace}(\mathbf{A}) \geq \mathrm{trace}(\mathbf{B})$$

- CRB exists also for biased case
- An unbiased estimate achieving the CRB is said *efficient*

# Cramer-Rao bound (cont'd)

## Result

If $\theta \mapsto \log p_{Y\theta}(\mathbf{y}_N|\theta)$ has a second-order derivative, then

$$F(\theta_0) = -\mathbb{E}\left[\frac{\partial^2 \log p_{Y|\theta}(\mathbf{y}_N|\theta)}{(\partial\theta)^2}_{|\theta=\theta_0}\right]$$

where $\mathbb{E}[\partial^2 \log p_{Y|\theta}(\mathbf{y}_N|\theta)/(\partial\theta)^2_{|\theta=\theta_0}]$ is the Hessian matrix whose components $(\ell, m)$ are

$$\mathbb{E}\left[\frac{\partial^2 \log p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial\theta(\ell)\theta(m)}_{|\theta=\theta_0}\right]$$

# Sketch of proof

Let us consider the scalar case

$$\mathbb{E}\left[\frac{\partial^2 \log p_{Y|\theta}(y|\theta)}{(\partial\theta)^2}_{|\theta=\theta_0}\right] = -\mathbb{E}\left[\frac{1}{p_{Y|\theta}(y|\theta_0)^2}\left(\frac{\partial p_{Y|\theta}(y|\theta)}{\partial\theta}_{|\theta=\theta_0}\right)^2\right]$$
$$+ \mathbb{E}\left[\frac{1}{p_{Y|\theta}(y|\theta_0)}\frac{\partial^2 p_{Y|\theta}(y|\theta)}{(\partial\theta)^2}_{|\theta=\theta_0}\right]$$
$$= -\mathbb{E}\left[\left(\frac{1}{p_{Y|\theta}(y|\theta_0)}\frac{\partial p_{Y|\theta}(y|\theta)}{\partial\theta}_{|\theta=\theta_0}\right)^2\right]$$
$$= -\mathbb{E}\left[\left(\frac{\partial \log p_{Y|\theta}(y|\theta)}{\partial\theta}_{\theta=\theta_0}\right)^2\right]$$

# Example 1

Let $\mathbf{Y}_N$ be a iid Gaussian vector with unknown mean $\theta = m$ and unit-variance

$$p_{Y|\theta}(\mathbf{y}_N|\theta) = \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}\sum_{n=1}^{N}(y_n-\theta)^2}$$

## Result

Fisher Information Matrix is s.t.

$$F^{-1}(\theta_0) = \frac{1}{N}$$

**Remarks:** the empirical mean estimate

- unbiased, MVUE with variance $1/N$
- efficient (rational since MVUE and CRB evaluation)

# Example 2

- Let $\mathbf{Y}_N$ be a process depending on two multi-variate parameters $\theta_1$ and $\theta_2$.
- Let $\theta = [\theta_1^{\mathrm{T}}, \theta_2^{\mathrm{T}}]^{\mathrm{T}}$

$$F(\theta) = \begin{bmatrix} \mathbb{E}\left[\frac{\partial \log p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial \theta_1}\frac{\partial \log p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial \theta_1^{\mathrm{T}}}\right] & \mathbb{E}\left[\frac{\partial \log p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial \theta_1}\frac{\partial \log p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial \theta_2^{\mathrm{T}}}\right] \\ \mathbb{E}\left[\frac{\partial \log p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial \theta_2}\frac{\partial \log p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial \theta_1^{\mathrm{T}}}\right] & \mathbb{E}\left[\frac{\partial \log p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial \theta_2}\frac{\partial \log p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial \theta_2^{\mathrm{T}}}\right] \end{bmatrix}$$

## Matrix inversion lemma

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \Rightarrow M^{-1} = \begin{bmatrix} S^{-1} & -S^{-1}BD^{-1} \\ -D^{-1}CS^{-1} & D^{-1} + D^{-1}CS^{-1}BD^{-1} \end{bmatrix}$$

with the so-called Schur complement

$$S = A - BD^{-1}C$$

# Example 2 (cont'd)

- If $B = C = 0$, then performance for joint optimization (both $\theta_1$ and $\theta_2$ are unknown) are the same as only one of them is unknown
- If $B \neq 0$ and $C \neq 0$ (actually $C = B^{\mathrm{T}}$), then
  - Schur complement is definite-positive (take $\tilde{x} = [x^{\mathrm{T}}, -x^{\mathrm{T}}BD^{-1}]^{\mathrm{T}}$)
  - $D^{-1}B^{\mathrm{T}}S^{-1}BD^{-1}$ is positive
  - joint estimation for $\theta_2$ is worse
  - $BS^{-1}B^{\mathrm{T}}$ is positive and as $A - S = BS^{-1}B^{\mathrm{T}}$, then $A \geq S$ and $S^{-1} \geq A^{-1}$
  - joint estimation for $\theta_1$ is worse

# Asymptotic analysis

- In many estimation problems, very difficult to obtain performance at fixed $N$ for the variance
- Consequently difficult to know the distance to CRB
- Extremely difficult to design an (almost)-efficient algorithm at fixed $N$ (see the characterization of the complete sufficient statistic)
- To overcome these issues, $N \to \infty$ is useful

**Goal**

Analyze the performance (bias, variance, ...) of $\hat{\theta}_N$ when $N \to \infty$

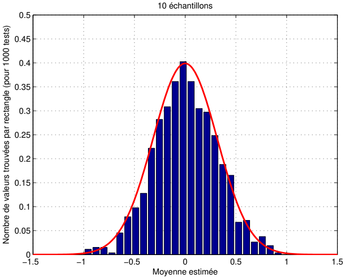
# Example

Let $\mathbf{Y}_N$ be a iid vector with unknown mean $\theta = m$ and unit-variance

$$p_{Y|\theta}(\mathbf{y}_N|\theta) = \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}\sum_{n=1}^{N}(y_n-\theta)^2}$$

Let

$$\hat{m}_N = \frac{1}{N}\sum_{n=1}^{N} y_n$$




- Convergence?
- Distribution?
  - Shape
  - Mean (value of convergence necessary)
  - Variance

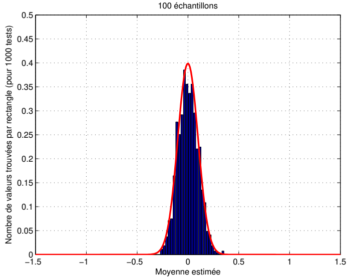
# Example

Let $\mathbf{Y}_N$ be a iid vector with unknown mean $\theta = m$ and unit-variance

$$p_{Y|\theta}(\mathbf{y}_N|\theta) = \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}\sum_{n=1}^{N}(y_n-\theta)^2}$$

Let

$$\hat{m}_N = \frac{1}{N}\sum_{n=1}^{N} y_n$$




- Convergence?
- Distribution?
  - Shape
  - Mean (value of convergence necessary)
  - Variance

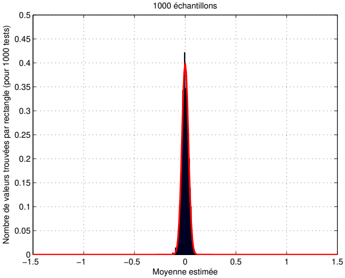
# Example

Let $\mathbf{Y}_N$ be a iid vector with unknown mean $\theta = m$ and unit-variance

$$p_{Y|\theta}(\mathbf{y}_N|\theta) = \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}\sum_{n=1}^{N}(y_n-\theta)^2}$$

Let

$$\hat{m}_N = \frac{1}{N}\sum_{n=1}^{N} y_n$$




- Convergence?
- Distribution?
  - Shape
  - Mean (value of convergence necessary)
  - Variance

# Example (cont'd)

In any case

$$\mathbb{E}[\hat{m}_N] = m \text{ and } \lim_{N\to\infty} \hat{m}_N \stackrel{a.s.}{=} m$$

but

- If $y_n$ Gaussian,

$$\sqrt{N}(\hat{m}_N - m) \stackrel{\mathcal{D}}{=} \mathcal{N}(0,1)$$

- If $y_n$ non-Gaussian, Central-Limit Theorem

$$\lim_{N\to\infty} \sqrt{N}(\hat{m}_N - m) \stackrel{\mathcal{D}}{=} \mathcal{N}(0,1)$$

**Goal**

Extend similar results to other cases

# Consistency

## Definition

$$\lim_{N\to\infty} \hat{\theta}_N \overset{a.s.}{=} \theta_0$$

with the almost surely convergence

$$\Pr\left(\omega : \lim_{N\to\infty} \hat{\theta}_N(\omega) = \theta_0\right) = 1$$

**Standard approaches for proving it:**

- Strong law of large number (SLLN)
- Weak law of large numbers (WLLN) if convergence in probability
- Other way:
  - Borel-Cantelli lemma

$$\forall\varepsilon > 0,\ \sum_{n\in\mathbb{N}} \Pr(\|\hat{\theta}_n - \theta_0\| > \varepsilon) < +\infty \Rightarrow \Pr(\lim_{N\to\infty} \hat{\theta}_N = \theta_0) = 1$$

  - Markov/Tchebitchev inequality and Doob trick

$$\Pr(\|\hat{\theta}_N - \theta_0\| > \varepsilon) \leq \frac{\mathbb{E}[\|\hat{\theta}_N - \theta_0\|^2]}{\varepsilon^2}$$

# Asymptotic normality

## Definition

An estimate is said *asymptotically normal* iff $\exists p$ s.t.

$$\lim_{N\to\infty} N^{p/2}(\hat{\theta}_N - \theta_0) \stackrel{\mathcal{D}}{=} \mathcal{N}(0, \Gamma)$$

where

- $p$ is the so-called *speed of convergence*

$$\mathrm{MSE} = \mathbb{E}[\|\hat{\theta}_N - \theta_0\|^2] \sim \frac{\mathrm{trace}(\Gamma)}{N^p}$$

- $\Gamma$ is the so-called *asymptotic covariance matrix*

**Standard approaches for proving it:**

- Central-Limit Theorem
- Standard proof by using the characteristic function of the second-kind

$$t \mapsto \log \mathbb{E}[e^{iXt}]$$

# Definitions

- An estimate is said *asymptotically unbiased* iff

$$\lim_{N \to \infty} \mathbb{E}[\hat{\theta}_N] = \theta_0$$

- An estimate is said *asymptotically efficient* iff

$$\lim_{N \to \infty} \frac{\mathrm{MSE}(N)}{\mathrm{trace}(\mathrm{CRB}(N))} = 1$$

# Algorithms

- Contrast-based estimate
- Maximum-Likelihood (ML) estimate
- Least-Square (LS) estimate
- Moments-matching (MM) estimate

# Definition for contrast estimate

- Let $J$ be a bivariate function. It is called a *contrast function* iff

$$\theta \mapsto J(\theta, \theta_0)$$

is minimum in $\theta = \theta_0$

- Let $J_N$ a statistic of $\mathbf{y}_N$ depending on generic parameter $\theta$

$$\theta \mapsto J_N(\mathbf{y}_N, \theta)$$

$J_N$ is called a *contrast process* iff

$$\lim_{N \to \infty} J_N(\mathbf{y}_N, \theta) \stackrel{p.}{=} J(\theta, \theta_0)$$

- The so-called *minimum constrast estimate* $\hat{\theta}_N$ is obtained by

$$\hat{\theta}_N = \arg\min_{\theta} J_N(\mathbf{y}_N, \theta)$$

# Example

Let $\mathbf{Y}_N$ be a iid Gaussian vector with unknown mean $\theta = m$ and unit-variance

$$p_{Y|\theta}(\mathbf{y}_N|\theta) = \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}\sum_{n=1}^{N}(y_n-\theta)^2}$$

We have

$$J(\theta, \theta_0) = 1 + (\theta_0 - \theta)^2$$

$$J_N(\mathbf{y}_N, \theta) = \frac{1}{N}\sum_{n=1}^{N}(y_n - \theta)^2$$

$$\hat{\theta}_N = \frac{1}{N}\sum_{n=1}^{N} y_n$$

The empirical mean is a minimum contrast estimate (unbiased, efficient, asymptotically normal with $p = 1$)

# Main results

## Consistency

If $\theta \mapsto J(\theta, \theta_0)$ and $\theta \mapsto J_N(\mathbf{y}_N, \theta)$ are continuous in $\theta$ (and other mild regularity conditions on $J_N$), then minimum contrast estimate $\hat{\theta}_N$ consistent

## Asymptotic normality

- $\theta \mapsto J_N(\mathbf{y}_N, \theta)$ twice-differentiable in an open neighborhood of $\theta_0$
- $\sqrt{N}\frac{\partial J_N(\mathbf{y}_N,\theta)}{\partial\theta}_{|\theta=\theta_0}$ converges in distribution to a zero-mean Gaussian distribution with covariance matrix $\Delta(\theta_0)$
- the Hessian matrix $\frac{\partial^2 J_N(\mathbf{y}_N,\theta)}{(\partial\theta)^2}_{|\theta=\theta_0}$ converges in probability to the definite-positive matrix $H(\theta_0)$
- and mild regularity technical conditions on $J_N$

then minimum contrast estimate $\hat{\theta}_N$ asymptotically normal with $p = 1$ and asymptotic covariance matrix

$$\Gamma(\theta_0) = H^{-1}(\theta_0)\Delta(\theta_0)H^{-1}(\theta_0)$$

# Sketch of proof

By applying second-order Taylor series expansion around $\theta_0$, we get

$$\frac{\partial J_N(\mathbf{y}_N,\theta)}{\partial\theta}_{|\theta=\hat{\theta}_N} = 0 = \frac{\partial J_N(\mathbf{y}_N,\theta)}{\partial\theta}_{|\theta=\theta_0} + \frac{\partial^2 J_N(\mathbf{y}_N,\theta)}{(\partial\theta)^2}_{|\theta=\theta_0}(\hat{\theta}_N-\theta_0)$$

So

$$-\underbrace{\sqrt{N}\frac{\partial J_N(\mathbf{y}_N\theta)}{\partial\theta}_{|\theta=\theta_0}}_{\text{cv. in distribution to } \mathcal{N}(0,\Delta(\theta_0))} = \underbrace{\frac{\partial^2 J_N(\mathbf{y}_N,\theta)}{(\partial\theta)^2}_{|\theta=\theta_0}}_{\text{cv. in probability to } H(\theta_0)}\sqrt{N}(\hat{\theta}_N-\theta_0)$$

Then

$$\lim_{N\to\infty} H(\theta_0).\sqrt{N}(\hat{\theta}_N-\theta_0) \overset{\mathcal{D}}{=} \mathcal{N}(0,\Delta(\theta_0))$$

# Maximum-Likelihood (ML) estimate

## Definition

Let $p_{Y|\theta}(.|\theta_0)$ be a probability density parametrized by $\theta_0$
The Maximum-Likelihood (ML) estimate for $\theta_0$ is defined as follows

$$\hat{\theta}_{\mathrm{ML},N} = \arg\max_{\theta} p_{Y|\theta}(\mathbf{y}_N|\theta)$$

**Likelihood equations:** If $\theta \mapsto p_{Y|\theta}(\mathbf{y}_N|\theta)$ is differentiable, then

$$\left.\frac{\partial p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}_{\mathrm{ML},N}} = 0$$

Warning: the ML estimate is not necessary unique (if more than one global maximum)

# Link with minimum contrast estimate

## Fundamental assumption

$\mathbf{Y}_N$ iid vector

We have

$$\hat{\theta}_{\mathrm{ML},N} = \arg\min_{\theta} J_N(\mathbf{y}_N, \theta)$$

with the following contrast process

$$J_N(\mathbf{y}_N, \theta) := -\frac{1}{N} \log p_{Y|\theta}(\mathbf{y}_N|\theta) = -\frac{1}{N} \sum_{n=1}^{N} \log p_{Y|\theta}(y_n|\theta)$$

One can prove

$$\lim_{N\to\infty} J_N(\mathbf{y}_N, \theta) \overset{p.}{=} J(\theta, \theta_0)$$

with the contrast function (maximum in $\theta = \theta_0$)

$$\begin{aligned} J(\theta, \theta_0) &= -\mathbb{E}[\log p_{Y|\theta}(\mathbf{y}_N|\theta)] \\ &= -\int \log(p_{Y|\theta}(\mathbf{y}_N|\theta)) p_{Y|\theta}(\mathbf{y}_N|\theta_0) d\mathbf{y}_N \\ &= D(p_{Y|\theta}(.|\theta_0) || p_{Y|\theta}(.|\theta)) + H(p_{Y|\theta}(.|\theta_0)) \end{aligned}$$

# Asymptotic analysis

**Result**

If $\mathbf{Y}_N$ iid vector, and the ML-related constrast function and process satisfy standard conditions, then ML estimate is

- consistent
- asymptotically unbiased
- asymptotically normal with $p = 1$
- asymptotically efficient
  $(\lim_{N\to\infty} \mathrm{trace}(\Gamma(\theta_0))/(N\mathrm{trace}(\mathrm{CRB}(N))) = 1)$

**General case (non-iid):**

- no general result
- should be analyzed case by case
- nevertheless iid result often valid

# Sketch of proof

Let $F(\theta_0)$ the FIM when $N$ samples are available.

$$\begin{aligned} F(\theta_0) &\overset{iid}{=} \sum_{n,n'=1}^{N} \mathbb{E}\left[ \frac{\partial \log p_{Y|\theta}(y_n|\theta)}{\partial \theta}_{|\theta=\theta_0} \cdot \frac{\partial \log p_{Y|\theta}(y_{n'}|\theta)}{\partial \theta}^{\mathrm{T}}_{|\theta=\theta_0} \right] \\ &= NF_1(\theta_0) + \sum_{n \neq n'} \mathbb{E}\left[ \frac{\partial \log p_{Y|\theta}(y_n|\theta)}{\partial \theta}_{|\theta=\theta_0} \right] \mathbb{E}\left[ \frac{\partial \log p_{Y|\theta}(y_{n'}|\theta)}{\partial \theta}^{\mathrm{T}}_{|\theta=\theta_0} \right] \\ &= NF_1(\theta_0) \end{aligned}$$

with $F_1(\theta_0)$ the FIM for one sample

# Sketch of proof (cont'd)

$$-\sqrt{N}\frac{\partial \frac{1}{N}\log p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial\theta}_{|\theta=\theta_0} = -\sqrt{N}\left(\frac{1}{N}\sum_{n=0}^{N}\frac{\partial \log p_{Y|\theta}(y_n|\theta)}{\partial\theta}_{|\theta=\theta_0} - 0\right)$$
$$\xrightarrow{\mathcal{D}} \mathcal{N}(0,\Delta(\theta_0))$$

since $\mathbb{E}[\frac{\partial \log p_{Y|\theta}(y_n|\theta)}{\partial\theta}_{|\theta=\theta_0}] = 0$ and with

$$\begin{aligned}\Delta(\theta_0) &= \lim_{N\to\infty} N\mathbb{E}\left[\frac{\partial \frac{1}{N}\log p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial\theta}_{|\theta=\theta_0}\frac{\partial \frac{1}{N}\log p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial\theta}^{\mathrm{T}}_{|\theta=\theta_0}\right]\\ &= \frac{1}{N}\sum_{n,n'=1}^{N}\mathbb{E}\left[\frac{\partial \log p_{Y|\theta}(y_n|\theta)}{\partial\theta}_{|\theta=\theta_0}.\frac{\partial \log p_{Y|\theta}(y_{n'}|\theta)}{\partial\theta}^{\mathrm{T}}_{|\theta=\theta_0}\right]\\ &= F_1(\theta_0)\end{aligned}$$

# Sketch of proof (cont'd)

$$
\begin{aligned}
-\frac{\partial^2 \frac{1}{N}\log p_{Y|\theta}(\mathbf{y}_N|\theta)}{(\partial\theta)^2}_{|\theta=\theta_0} &= -\frac{1}{N}\sum_{n=1}^{N}\frac{\partial^2 \log p_{Y|\theta}(y_n|\theta)}{(\partial\theta)^2}_{|\theta=\theta_0} \\
&\overset{p.\ N\to\infty}{=} -\mathbb{E}\left[\frac{\partial^2 \log p_{Y|\theta}(y_n|\theta)}{(\partial\theta)^2}_{|\theta=\theta_0}\right] \\
&= F_1(\theta_0)
\end{aligned}
$$

Consequently,

$$H(\theta_0) = F_1(\theta_0)$$

# Sketch of proof (cont'd)

We remind

$$\sqrt{N}(\hat{\theta}_{\mathrm{ML},N} - \theta_0) \stackrel{\mathcal{D}}{\rightarrow} \mathcal{N}(0, \Gamma(\theta_0))$$

with

$$\Gamma(\theta_0) := H^{-1}(\theta_0)\Delta(\theta_0)H^{-1}(\theta_0) = F_1^{-1}(\theta_0)$$

Consequently

$$\lim_{N\to\infty} N\mathbb{E}\left[(\hat{\theta}_{\mathrm{ML},N} - \theta_0)(\hat{\theta}_{\mathrm{ML},N} - \theta_0)^{\mathrm{T}}\right] = F_1^{-1}(\theta_0)$$

$$\mathbb{E}\left[(\hat{\theta}_{\mathrm{ML},N} - \theta_0)(\hat{\theta}_{\mathrm{ML},N} - \theta_0)^{\mathrm{T}}\right] \approx \frac{1}{N}F_1^{-1}(\theta_0) = F^{-1}(\theta_0) = \mathrm{CRB}(N)$$

# Example 1

Let $\mathbf{Y}_N$ be a iid Gaussian vector with unknown mean $\theta = m$ and unit-variance

$$p_{Y|\theta}(\mathbf{y}_N|\theta) = \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}\sum_{n=1}^{N}(y_n-\theta)^2}$$

We can see that $\hat{\theta}_{\mathrm{ML},N}$ is the empirical mean, and

$$\sqrt{N}(\hat{\theta}_{\mathrm{ML},N} - m) \stackrel{\mathcal{D}}{\to} \mathcal{N}(0,1)$$

and

$$\mathrm{CRB}(N) = \frac{1}{N}$$

# Example 2

Let a pure harmonic with additive noise

$$y_n = e^{2i\pi f_0 n} + w_n, \; n = 1, \cdots, N$$

with $w_n$ iid zero-mean (circularly) Gaussian noise with variance $\sigma_w^2$

**Remarks:**

- independent sample but not identically distributed (non-id)
- none of previous results applies!

## Results

$$\hat{f}_{\mathrm{ML},N} = \arg\max_f \Re\left\{ \frac{1}{N} \sum_{n=1}^{N} y_n e^{-2i\pi f n} \right\}$$

$$N^{3/2}(\hat{f}_{\mathrm{ML},N} - f_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \frac{3\sigma_w^2}{8\pi^2})$$

Much faster than standard case since $\mathbb{E}[(\hat{f}_{\mathrm{ML},N} - f_0)^2] \approx \frac{3\sigma_w^2}{8\pi^2 N^3}$

*Proof:* see Exercises' session

# Least-Square (LS) estimate

Let

$$y_n = f_n(\theta_0) + w_n, \; n = 1, \cdots, N$$

with

- $f_n(.)$ deterministic function
- $w_n$ zero-mean process

## Least-Square (LS) estimate

We fit the model with the data wrt the Euclidian distance

$$\hat{\theta}_{\mathrm{LS},N} = \arg\min_{\theta} \sum_{n=1}^{N} |y_n - f_n(\theta)|^2$$

Related to the closest vector problem in a (non-discrete) vector space

## Result

If $w_n$ is iid Gaussian noise (with variance $\sigma_w^2$), LS is identical to ML

**Example :** empirical mean is both LS and ML (with Gaussian noise)

# Example

**Linear model:**

$$\mathbf{y}_N = \mathbf{H}\theta + \mathbf{w}_N$$

Then

$$\hat{\theta}_{\mathrm{LS},N} = \arg\min_{\theta} \|\mathbf{y}_N - \mathbf{H}\theta\|^2$$

If **H** column full rank, then

$$\hat{\theta}_{\mathrm{LS},N} = \mathbf{H}^{\#}\mathbf{y}_N$$

# Moments-matching (MM) estimate

**$q$-order moments:**

- statistical terms with the following form

$$\mathbb{E}[f(Y_1, \cdots, Y_p)]$$

  with $f$ a monomial function of degree $q$
- related to the Taylor-series expansion of the function

$$\omega \mapsto \mathbb{E}[e^{iY\omega}]$$

  if we consider only one variable $Y$

**Notations:**

- Let $S(\theta)$ a vector of moments depending on $\theta$
- Let $\hat{S}_N$ the empirical estimate of $S(\theta_0)$ with $N$ samples

# Moments-matching (MM) estimate (cont'd)

## Algorithm

If

- $S(\theta) = S(\theta_0) \Rightarrow \theta = \theta_0$,
- and $\theta \mapsto S(\theta)$ is continuous,

we define the contrast process

$$J_N(\mathbf{y}_N, \theta) = \|\hat{S}_N - S(\theta)\|^2$$

and the MM estimate as

$$\hat{\theta}_{\mathrm{MM},N} = \arg\min_{\theta} \|\hat{S}_N - S(\theta)\|^2$$

**Vocabulary:** MM estimate is also called Method of Moments (MoM)

# Moments-matching (MM) estimate (cont'd)

## Result

If

- $\theta \mapsto S(\theta)$ is twice differentiable in $\theta_0$
- $\sqrt{N}(\hat{S}_N - S(\theta_0)) \stackrel{\mathcal{D}}{\to} \mathcal{N}(0, R(\theta_0))$

then

$$\sqrt{N}(\hat{\theta}_{\mathrm{MM},N} - \theta_0) \stackrel{\mathcal{D}}{\to} \mathcal{N}(0, \Gamma(\theta_0))$$

with

$$\Gamma(\theta_0) = \left(D(\theta_0)^{\mathrm{T}} D(\theta_0)\right)^{-1} D(\theta_0)^{\mathrm{T}} R(\theta_0) D(\theta_0) \left(D(\theta_0)^{\mathrm{T}} D(\theta_0)\right)^{-1}$$

with $D(\theta_0) = \partial S(\theta)/\partial\theta_{|\theta=\theta_0}$

**Remark:** second bullet is often satisfied (if iid, straightforward).

# Sketch of proof

We have

$$\sqrt{N}\frac{\partial J_N(\mathbf{y}_N,\theta)}{\partial\theta}_{|\theta=\theta_0} = -2D(\theta_0)^{\mathrm{T}}\sqrt{N}(\hat{S}_N - S(\theta_0)) \stackrel{\mathcal{D}}{\to} \mathcal{N}(0,\Delta(\theta_0))$$

with $\Delta(\theta_0) = 4D(\theta_0)^{\mathrm{T}}R(\theta_0)D(\theta_0)$

$$\frac{\partial^2 J_N(\mathbf{y}_N,\theta)}{(\partial\theta)^2}_{|\theta=\theta_0} = 2D(\theta_0)^{\mathrm{T}}D(\theta_0) - 2\frac{\partial^2 S^{\mathrm{T}}}{(\partial\theta)^2}_{|\theta=\theta_0}\underbrace{(\hat{S}_N - S(\theta_0))}_{\mathrm{cv}\ \stackrel{p.}{\to} 0}$$

# Example

Let $\mathbf{Y}_N$ be a iid Gaussian vector with unknown mean $\theta_0 = m$ and unit-variance

We consider

$$J_N(\mathbf{y}_N, \theta) = \|\hat{m}_N - \theta\|^2$$

and the MM estimate as

$$\begin{aligned} \hat{\theta}_{\mathrm{MM},N} &= \arg\min_{\theta} J_N(\mathbf{y}_N, \theta) \\ &= \hat{m}_N \end{aligned}$$

Then

$$\sqrt{N}(\hat{m}_N - m) \stackrel{\mathcal{D}}{\to} \mathcal{N}(0, 1)$$

since

- $D(\theta_0) = 1$
- $R(\theta_0) = \lim_{N\to\infty} N\mathbb{E}[(\hat{m}_N - m)^2] = 1$

# Extension to complex-valued case

Let us now consider that $\theta \in \mathbb{C}^K$

- Can apply all previous results by working on

$$\tilde{\theta} = [\Re\{\theta\}^{\mathrm{T}}, \Im\{\theta\}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{2K}$$

- But result not easy to interpret: use

$$\tilde{\theta} = \underbrace{\frac{1}{2}\begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix}}_{M} \cdot \underbrace{\begin{bmatrix} \theta \\ \overline{\theta} \end{bmatrix}}_{\underline{\theta}} \quad \text{and} \quad \underline{\underline{\theta}} = \begin{bmatrix} \overline{\theta} \\ \theta \end{bmatrix}$$

- Remind

$$\frac{\partial .}{\partial \theta} = \frac{1}{2}\left(\frac{\partial .}{\partial \Re\{\theta\}} - i\frac{\partial .}{\partial \Im\{\theta\}}\right) \text{ and } \frac{\partial .}{\partial \underline{\theta}} = \frac{1}{2}\left(\frac{\partial .}{\partial \Re\{\theta\}} + i\frac{\partial .}{\partial \Im\{\theta\}}\right)$$

(see $f(\theta) = \theta$)

Apply previous results with changes of variables ($\underline{\theta}$ and $\underline{\underline{\theta}}$)

# Main result

We have

$$\mathbb{E}[(\underline{\hat{\theta}} - \underline{\theta}_0)(\underline{\hat{\theta}} - \underline{\theta}_0)^{\mathrm{H}}] \geq \underline{F}^{-1}(\underline{\theta}_0)$$

with

$$F(\underline{\theta}_0) = \mathbb{E}\left[\left(\frac{\partial \log p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial \underline{\underline{\theta}}}_{|\theta=\theta_0}\right)\left(\frac{\partial \log p_{Y|\theta}(\mathbf{y}_N|\theta)}{\partial \underline{\underline{\theta}}}_{|\theta=\theta_0}\right)^{\mathrm{H}}\right]$$

**Remarks:**

- we can use "real-valued" CRB expression with $^{\mathrm{H}}$ and $\overline{\theta}$ instead of $^{\mathrm{T}}$ and $\theta$ iff cross term vanishes in $F(\underline{\theta}_0)$
- many examples in telecommunications (as working in baseband with on complex enveloppe)

# Sketch of proof

We have

$$\begin{aligned}
\mathbb{E}[(\hat{\tilde{\theta}} - \tilde{\theta}_0)(\hat{\tilde{\theta}} - \tilde{\theta}_0)^{\mathrm{T}}] &\geq \mathrm{CRB}(\tilde{\theta}_0) \\
\mathbb{E}[(\hat{\tilde{\theta}} - \tilde{\theta}_0)(\hat{\tilde{\theta}} - \tilde{\theta}_0)^{\mathrm{H}}] &\geq \mathrm{CRB}(\tilde{\theta}_0) \\
M\mathbb{E}[(\hat{\underline{\theta}} - \underline{\theta}_0)(\hat{\underline{\theta}} - \underline{\theta}_0)^{\mathrm{H}}]M^{\mathrm{H}} &\geq M\mathrm{CRB}(\underline{\theta}_0)M^{\mathrm{H}} \\
\mathbb{E}[(\hat{\underline{\theta}} - \underline{\theta}_0)(\hat{\underline{\theta}} - \underline{\theta}_0)^{\mathrm{H}}] &\geq \mathrm{CRB}(\underline{\theta}_0)
\end{aligned}$$

since

$$\begin{aligned}
F(\underline{\theta}_0) = M^{\mathrm{H}} F(\tilde{\theta}_0) M \quad &\Rightarrow \quad \mathrm{CRB}(\underline{\theta}_0) = M^{-1}\mathrm{CRB}(\tilde{\theta}_0){M^{-1}}^{\mathrm{H}} \\
&\Rightarrow \quad \mathrm{CRB}(\tilde{\theta}_0) = M\mathrm{CRB}(\underline{\theta}_0)M^{\mathrm{H}}
\end{aligned}$$

with $\partial./\partial\underline{\underline{\theta}} = M^{\mathrm{H}}\partial/\partial\tilde{\theta}$

# Part 5 : Bayesian estimation (for random parameters)

# Principle of Bayesian approach

- Let us consider $\theta_0$ as a random variable with a known *a priori* probability density function $p_\theta(\theta)$.
- Let us consider the joint pdf between observations and unknown parameter $\theta_0$. Bayes' rule leads to

$$p_{Y,\theta}(\mathbf{y}_N, \theta) = p_{Y|\theta}(\mathbf{y}_N|\theta)p_\theta(\theta)$$

## Quadratic risk

$$\begin{aligned} \mathrm{MSE} &= \mathbb{E}_{[}\|\hat{\theta}_N - \theta_0\|^2] \\ &= \int \|\hat{\theta}_N - \theta\|^2 p_{Y,\theta}(\mathbf{y}_N, \theta) d\mathbf{y}_N d\theta \\ &= \mathbb{E}[\mathbb{E}_{.|\theta}[\|\hat{\theta}_N - \theta\|^2]] \\ &= \mathbb{E}[\mathrm{MSE}_{\mathrm{det.}}(\theta)] \end{aligned}$$

**Remark:** the risk is averaged over all the values of $\theta_0$. It is not evaluated for a specific value of $\theta_0$.

# Optimal estimate

## Result

The optimal unbiased estimate (wrt MSE) exists and is given by

$$\hat{\theta}_{\mathrm{MMSE},N} = \mathbb{E}_{\theta|Y}[\theta] = \int \theta p_{\theta|Y}(\theta|\mathbf{y}_N)d\theta$$

This estimate is called MMSE and corresponds to the mean of the *a posteriori* pdf of $\theta$

**Remarks:**

- The optimal estimate is the Mean A Posteriori (MeAP) insterad of the Maximum A Posteriori (MAP) defined as follows

$$\hat{\theta}_{\mathrm{MAP},N} = \arg\max_{\theta} p_{\theta|Y}(\theta|\mathbf{y}_N)$$

- In deterministic approach, the optimal unbiased estimate does not exist in general. But often exists asymptotically (through ML)

# Sketch of proof

Let us consider the scalar case

$$\mathrm{MSE}(\hat{\theta}_N) = \int \left( \int (\hat{\theta}_N - \theta)^2 p_{\theta|Y}(\theta|\mathbf{y}_N) d\theta \right) p_Y(\mathbf{y}_N) d\mathbf{y}_N$$

As inner integral and $p_Y(\mathbf{y}_N)$ are positive, $\mathrm{MSE}(\hat{\theta}_N)$ is minimum if for each observation $\mathbf{y}_N$, the inner integral is minimum itself.
So we are looking for $\hat{\theta}_N$ s.t.

$$\frac{d}{d\hat{\theta}_N} \int (\hat{\theta}_N - \theta)^2 p_{\theta|Y}(\theta|\mathbf{y}_N) d\theta = 0$$

which implies

$$\hat{\theta}_N \underbrace{\int p_{\theta|Y}(\theta|\mathbf{y}_N) d\theta}_{1} = \int \theta p_{\theta|Y}(\theta|\mathbf{y}_N) d\theta$$

# Example

$$y_n = m + w_n, \quad \text{pour } n = 1, \ldots, N$$

with

- $m$ zero-mean Gaussian variable with known variance $\sigma_m^2$
- $w_n$ iid zero-mean Gaussian process wth known variance $\sigma_w^2$

$$\begin{aligned} p_{m,Y}(m|\mathbf{y}_N) &= p_{Y|m}(\mathbf{y}_N|m)p_m(m)/p_Y(\mathbf{y}_N) \\ &\propto e^{-\left(m - \frac{\sigma_m^2}{\sigma_m^2+\sigma_w^2/N}\frac{1}{N}\sum_{n=1}^N y_n\right)^2/2\sigma_w^2} \end{aligned}$$

$$\hat{m}_{\mathrm{MMSE},N}(=\hat{m}_{\mathrm{MAP},N}) = \frac{\sigma_m^2}{\sigma_m^2+\sigma_b^2/N}\left(\frac{1}{N}\sum_{n=1}^N y_n\right)$$

**Remarks:**

- If $\sigma_m^2 \ll \sigma_w^2$, $\hat{m}_{\mathrm{MMSE},N}$ close to *a priori* mean (0)
- If $\sigma_m^2 \gg \sigma_w^2$, $\hat{m}_{\mathrm{MMSE},N}$ close to empirical mean

# Bayesian Cramer-Rao Bound

## Result

Let $\hat{\theta}$ be an unbiased estimate of $\theta_0$, then

$$\mathrm{MSE}(\hat{\theta}) \geq F^{-1} = \mathrm{BCRB}$$

with

$$F = \mathbb{E}\left[\left(\frac{\partial \log p_{Y,\theta}(\mathbf{y}_N, \theta)}{\partial \theta}\right)\left(\frac{\partial \log p_{Y,\theta}(\mathbf{y}_N, \theta)}{\partial \theta}\right)^{\mathrm{T}}\right]$$

**Remarks:**

- We have

$$F = -\mathbb{E}\left[\frac{\partial^2 \log p_{Y,\theta}(\mathbf{y}_N, \theta)}{(\partial \theta)^2}\right]$$

- No link between $\mathrm{BCRB}$ and $\mathbb{E}[\mathrm{CRB}_\theta(\theta)]$

# General conclusion

- Rich topic with four main configurations
- In deterministic approach: mainly asymptotic results and Maximum Likelihood plays a great role
- In Bayesian approach: optimal estimate fully characterized and finite-data analysis possible

# References


- H. Cramer, “Mathematical Methods of Statistics”, 1946.
- V. Kotelnikov, “The Theory of Optimum Noise Immunity”, 1947 (1959 in English).
- H.L. Van Trees, “Detection, Estimation, and Modulation Theory”, Part 1, 1968.
- H.V. Poor, “An introduction to signal detection and estimation”, 1988.
- S.M. Kay, “Fundamentals of Statistical Signal Processing”, 1993.
- B. Porat, “Digital Processing of Random Signals : Theory and Methods”, 1994.