2. Estimation Theory

2.1. Introduction to Estimation Theory

1. Introduction to Estimation Theory

Given an $N-$point data set $\{x[1], x[2], \ldots, x[N]\}$ which **depends on an unknown parameter** θ (or set of parameters $\underline{\theta}$), we wish to determine θ based on the data, through the definition of an estimator:

$$\hat{\theta} = g(x[1], x[2], \ldots, x[N]) = g(\underline{x}),$$

where $g(\cdot)$ is some function.

The dependence of the available data $\underline{x}$ with respect to the parameters $\underline{\theta}$ is captured by the **model** that is proposed. As data is random in nature, we represent it by its **probability density function** (pdf):

$$f_{\underline{x}}(x[1], x[2], \ldots, x[N]; \underline{\theta}) = f_{\underline{x}}(\underline{x}; \underline{\theta}).$$

The pdf is **parametrized** by the unknown vector of parameters $\underline{\theta}$.

- **Case 1:** We are given a pdf. For instance, $N = 1$ ($x[1] = x$) and θ is the mean, the pdf could be

$$f_x(x; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right].$$

- **Case 2:** Usually, we are given data and we have to **choose a model**:
    1. Models should be **consistent** with the **problem** and **previous knowledge**.
    2. Models should be **mathematically tractable**.

$$x[n] = A + Bn + w[n], \quad f_{\underline{w}}(\underline{x}; \underline{\theta}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x[n] - A - Bn)^2\right].$$

- **Case 3**: *Bayesian approach.* We can assume that the parameters to be estimated are random variables (instead of deterministic but unknown). The knowledge about its pdf can be included.

$$f(\underline{x}, \underline{\theta}) = f(\underline{x}|\underline{\theta}) f(\underline{\theta})$$

In several situations, we want to estimate the mean value of a random process that can be modeled as a constant value θ embedded in stationary white noise $W[n]$:

$$\boxed{X[n] = \theta + W[n]}$$

$-$ **White noise:** each sample has a probability distribution with zero mean and finite variance, and samples are statistically independent and $r_W[n, l] = \sigma_W^2[n]\delta[l]$.

$-$ **Stationary white noise:** all variance samples have the same value and the autocorrelation function is $r_W[l] = \sigma_W^2 \cdot \delta[l]$.

How can we estimate the mean value of a random process given a set of observations (N) of a single realization? We can propose different estimators:

$$\begin{aligned}
\hat{m}_X^{(1)} &= \frac{1}{N}\sum_{n=1}^{N} x[n] \\
\hat{m}_X^{(2)} &= \text{median}(x[1], x[2], \ldots, x[n]) \\
\hat{m}_X^{(3)} &= \frac{\max(x[1], \ldots, x[N]) + \min(x[1], \ldots, x[N])}{2}
\end{aligned}$$

We need to assess the **performance of the estimators** to decide which one should be used.

1. Assessing Estimator Performance

How can we estimate the mean value of a random process given a set of samples (N) of a simple realization?

Let us assume that we select the average of the available samples (**sample mean**) as estimate of the mean value of the process. For this selection to be correct, we have to assume:

- **Stationarity:** the parameter to be estimated does not change through time.
- **Ergodicity:** any realization of the process ($X[n,i]$) assumes the statistical properties of the whole process,

$$m_X = \mathbb{E}[X[n]] = \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} X[n,i].$$

Estimators operate on the **samples of a given realization**. The estimated value depends on:

- The **available realization** $X[n,i]$.
- The **selected window** (n, N).

Thus, **any estimator is a random variable**.

Features of an estimator

The **bias of an estimator** is the difference between the expected value of the estimator and the true value of the parameter being estimated:

$$\boxed{B(\hat{\theta}) = |\theta - \mathbb{E}[\hat{\theta}]|}$$

- Estimations delivered by a biased estimator are **consistently different** from the parameter to be estimated.
- An estimator without bias is called **unbiased**.

Exercise: Given the signal model $X[n] = \theta + W[n]$, where $W[n]$ is a stationary white noise, calculate the bias of the estimator:

$$\hat{\theta}_N = \frac{1}{N}\sum_{n=1}^{N} x[n].$$

Solution:

$$\begin{aligned}
B(\hat{\theta}_N) = \left|\theta - \mathbb{E}\left[\hat{\theta}_N\right]\right| = \left|\theta - \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} x[n]\right]\right| = \left|\theta - \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(\theta + W[n])\right]\right| = \\
\left|\theta - \frac{1}{N}\sum_{n=1}^{N}(\theta + \mathbb{E}[W[n]])\right| = \left|\theta - \frac{1}{N}\sum_{n=1}^{N}\theta\right| = 0.
\end{aligned}$$

So, this estimator is unbiased.

The unbiased constrain is desirable and, among all unbiased estimators, that of **minimum variance** is preferred; it is called the **Minimum Variance Unbiased** (MVU) estimator. The variance of the estimator is calculated as

$$\sigma_{\hat{\theta}}^2 = \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2\right]$$

An estimator is **consistent** if, as the number of samples (N) increases, the resulting sequence of estimates converges to θ, and the variance of the estimates converges to zero:

$$\lim_{N\to\infty} \mathbb{E}[\hat{\theta}] \to \theta, \quad \lim_{N\to\infty} \sigma_{\hat{\theta}}^2 \to 0$$

Exercise: Given the signal model $X[n] = \theta + W[n]$, where $W[n]$ is a stationary white noise, calculate the variance of the estimator:

$$\hat{\theta}_N = \frac{1}{N}\sum_{n=1}^{N} x[n].$$

Solution:

$$\sigma_{\hat{\theta}_N}^2 = \mathbb{E}\left[\left(\hat{\theta}_N - \mathbb{E}[\hat{\theta}_N]\right)^2\right] = \mathbb{E}\left[(\hat{\theta}_N - \theta)^2\right] =$$
$$\mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N} x[n] - \theta\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N}(\theta + W[n]) - \theta\right)^2\right] =$$
$$\mathbb{E}\left[\left(\frac{1}{N}N\theta + \frac{1}{N}\sum_{n=1}^{N} W[n] - \theta\right)^2\right] = \mathbb{E}\left[\frac{1}{N^2}\left(\sum_{n=1}^{N} W[n]\right)^2\right] = \frac{1}{N^2}\mathbb{E}\left[\sum_{n=1}^{N} W[n]\sum_{m=1}^{N} W[m]\right] =$$
$$\frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbb{E}[W[n]W[m]] = \left[\, r_W[l] = \sigma_W^2\delta[l]\,\right] = \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\delta[n-m]\sigma_W^2 =$$
$$\frac{1}{N^2}N\sigma_W^2 \implies \boxed{\sigma_{\hat{\theta}_N}^2 = \frac{\sigma_W^2}{N}.}$$

This last equality implies that the estimator is consistent.

If the estimator is biased, the dispersion of the estimations with respect to the actual value to be estimated (θ) is not the variance but the **Mean Square Error** of the estimator ($\mathrm{MSE}(\hat{\theta})$). The MSE can be a measure of assessment for a given estimator, but to define an estimator **optimizing the MSE usually leads to unrealizable estimators**.

$$\sigma_{\hat{\theta}}^2 = \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2\right] = [\text{Biased}] \neq \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \mathrm{MSE}(\hat{\theta}).$$

Exercise: Prove that, for a given estimator $\hat{\theta}$,

$$\boxed{\mathrm{MSE}(\hat{\theta}) = \sigma_{\hat{\theta}}^2 + B^2(\hat{\theta}).}$$

Solution:

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta}-\theta)^2\right] = \mathbb{E}\left[(\hat{\theta}-\mathbb{E}[\hat{\theta}]+\mathbb{E}[\hat{\theta}]-\theta)^2\right] = \mathbb{E}\left[((\hat{\theta}-\mathbb{E}[\hat{\theta}])-(\theta-\mathbb{E}[\hat{\theta}]))^2\right] =$$
$$\mathbb{E}\left[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2 - 2(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\theta-\mathbb{E}[\hat{\theta}]) + (\theta-\mathbb{E}[\hat{\theta}])^2\right] =$$
$$\mathbb{E}\left[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\right] - 2\mathbb{E}\left[(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\theta-\mathbb{E}[\hat{\theta}])\right] + \mathbb{E}\left[(\theta-\mathbb{E}[\hat{\theta}])^2\right] =$$
$$\sigma^2_{\hat{\theta}} + B^2(\hat{\theta}) - 2\mathbb{E}\left[(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\theta-\mathbb{E}[\hat{\theta}])\right] = \sigma^2_{\hat{\theta}} + B^2(\hat{\theta}) - 2\mathbb{E}\left[(\hat{\theta}-\mathbb{E}[\hat{\theta}])\right](\theta-\mathbb{E}[\hat{\theta}]) =$$
$$\sigma^2_{\hat{\theta}} + B^2(\hat{\theta}) - 2\left(\mathbb{E}[\hat{\theta}] - \mathbb{E}\left[\mathbb{E}[\hat{\theta}]\right]\right)(\theta-\mathbb{E}[\hat{\theta}]) = \boxed{\sigma^2_{\hat{\theta}} + B^2(\hat{\theta}).}$$

Vector notation

The previous **sample mean** estimator can be interpreted as a **filter** and, this way, we can generalize the study of its properties:

$$\hat{\theta}_N = \frac{1}{N}\sum_{n=1}^{N} x[n] = \frac{1}{N}\underline{1}^T\underline{x} \implies \boxed{\hat{\theta}_N = \underline{h}^T\underline{x}}$$

This estimator is **linear** in the (N) data.

Exercise: Given the signal model $X[n] = \theta + W[n]$, analyze the **bias** of the estimator $\hat{\theta}_N$.

Solution:

$$B(\hat{\theta}_N) = |\theta - \mathbb{E}[\hat{\theta}_N]|.$$
$$\mathbb{E}[\hat{\theta}_N] = \mathbb{E}\left[\underline{h}^T\underline{x}\right] = \underline{h}^T\mathbb{E}[\underline{x}] = [\underline{x} = \underline{\theta} + \underline{W}] = \underline{h}^T\mathbb{E}[\underline{\theta}+\underline{W}] = [\theta \text{ is deterministic}] =$$
$$\underline{h}^T\left[\underline{\theta} + \mathbb{E}[\underline{W}]\right] = \left[\mathbb{E}[\underline{W}] = 0\right] = \underline{h}^T\underline{\theta} = \theta\underline{h}^T\underline{1}.$$
$$\implies B(\hat{\theta}_N) = |\theta - \theta\underline{h}^T\underline{1}| = |\theta(1-\underline{h}^T\underline{1})|.$$
$$\text{As the estimator is unbiased (as we saw earlier), } \boxed{\underline{h}^T\underline{1} = 1.}$$

For instance, $\underline{h}^T = \frac{1}{N}\underline{1}$.

Exercise: Given the signal model $X[n] = \theta + W[n]$, analyze the **variance** of the estimator $\hat{\theta}_N$.

Solution:

$$\hat{\sigma}^2_{\hat{\theta}_N} = \mathbb{E}\left[(\hat{\theta}_N - \mathbb{E}[\hat{\theta}_N])^2\right] = \mathbb{E}\left[(\hat{\theta}_N - \theta)^2\right] = \mathbb{E}\left[(\underline{h}^T\underline{x}-\theta)^2\right] = \mathbb{E}\left[(\underline{h}^T(\underline{\theta}+\underline{w})-\theta)^2\right] =$$
$$\mathbb{E}\left[(\underline{h}^T\underline{\theta}+\underline{h}^T\underline{w}-\theta)^2\right] = \left[\underline{h}^T\underline{\theta} = \theta\cdot\underline{h}^T\underline{1} = \theta\right] = \mathbb{E}\left[(\theta+\underline{h}^T\underline{w}-\theta)^2\right] = \mathbb{E}\left[(\underline{h}^T\underline{w})^2\right] =$$
$$\mathbb{E}\left[\underline{h}^T\underline{w}\cdot\underline{h}^T\underline{w}\right] = \mathbb{E}\left[\underline{h}^T\underline{w}\cdot\underline{w}^T\underline{h}\right] = \underline{h}^T\mathbb{E}\left[\underline{w}\underline{w}^T\right]\underline{h} = \boxed{\underline{h}^T\underline{\underline{R}}_W\underline{h}.}$$

---

2. Minimum Variance Unbiased Estimator

After generalizing the **sample mean** as a **filter**, we have obtained a **family of unbiased linear estimators** of the mean of a random process, for which we have the expression of their variance:

$$\hat{\theta}_N = \underline{h}^T\underline{x} \implies \text{Unbiased if: } \underline{h}^T\underline{1} = 1, \quad \sigma^2_{\hat{\theta}_N} = \underline{h}^T\underline{\underline{R}}_W\underline{h}.$$

Note: we imposed zero-mean noise and the use of the unbiased estimator.

To obtain the **Minimum Variance Unbiased (MVU)** estimator, we should solve the following problem of optimization with constraints:

$$\min_{\underline{h}} \left(\underline{h}^T \underline{\underline{R}}_W \underline{h}\right)$$
$$\text{subject to } \underline{h}^T \underline{1} = 1.$$

This optimization problem is formulated through **Lagrange multipliers**. This method allows an optimization problem with constraints to be solved **without explicit parametrization** in terms of the constraints.

Given a function $f(\underline{x})$ that we want to optimize subject to a constraint (described by another function) $g(\underline{x})$, we can define a **Lagrange function** (or **Lagrangian**) $\mathcal{L}(\underline{x}, \lambda)$ whose first derivatives are zero at the solutions of the original constrained problem.

Note: the theory of Lagrange multipliers will be studied in the Mathematical Optimization course.

$$\left.\begin{array}{r} \underset{\underline{x}}{\text{optimize}}\, f(\underline{x}) \\ \text{subject to } g(\underline{x}) = 0 \end{array}\right\} \implies \mathcal{L}(\underline{x}, \lambda) := f(\underline{x}) - \lambda g(\underline{x}) \implies \left\{\begin{array}{l} \nabla_{\underline{x}} \mathcal{L}(\underline{x}, \lambda) = 0 \\ \dfrac{\partial \mathcal{L}(\underline{x}, \lambda)}{\partial \lambda} = 0 \end{array}\right.$$

It is necessary to **derivate a scalar function with respect to a vector**.

Rules to derivate a scalar with respect to a vector.

Definition. *Gradient.* Given a scalar function $f(\underline{x}) \in \mathbb{R}$, with $\underline{x} \in \mathbb{R}^N$, we define its gradient with respect to $\underline{x}$ as

$$\nabla_{\underline{x}} f(\underline{x}) = \left(\frac{\partial f(\underline{x})}{\partial x_1}, \frac{\partial f(\underline{x})}{\partial x_2}, \ldots, \frac{\partial f(\underline{x})}{\partial x_N}\right)^T \in \mathbb{R}^N$$

Given this definition, the most common cases that we will work with are:

$$\nabla_{\underline{x}} \left(\underline{h}^T \underline{x}\right) = \nabla_{\underline{x}} \left(\sum_{i=1}^N h_i x_i\right) = \left(\frac{\partial \sum_{i=1}^N h_i x_i}{\partial x_1}, \frac{\partial \sum_{i=1}^N h_i x_i}{\partial x_2}, \ldots, \frac{\partial \sum_{i=1}^N h_i x_i}{\partial x_N}\right)^T = (h_1, h_2, \ldots, h_N)^T = \underline{h}.$$

In the same way, we can obtain $\nabla_{\underline{x}} \left(\underline{x}^T \underline{h}\right) = \underline{h}$.

$$\nabla_{\underline{x}} \left(\underline{z}^T \underline{\underline{A}} \underline{x}\right) = \left[\underline{z}^T \underline{\underline{A}} = \underline{v}^T\right] = \nabla_{\underline{x}} \left(\underline{v}^T \underline{x}\right) = \underline{v} = \left(\underline{v}^T\right)^T = \left(\underline{z}^T \underline{\underline{A}}\right)^T = \underline{\underline{A}}^T \underline{z}.$$

In the same way, $\nabla_{\underline{x}} \left(\underline{x}^T \underline{\underline{A}} \underline{z}\right) = \underline{\underline{A}} \underline{z}$. If we have a symmetric matrix, such as a correlation matrix, it can be shown that $\nabla_{\underline{x}} \left(\underline{x}^T \underline{\underline{A}} \underline{x}\right) = 2\underline{\underline{A}} \underline{x}$.

Obtaining MVU through Lagrange optimization

To obtain the MVU estimator, we should solve the following problem of optimization with constraints:

$$\min_{\underline{h}} \left(\sigma^2_{\hat{\theta}_N}\right) = \min_{\underline{h}} \left(\underline{h}^T \underline{\underline{R}}_W \underline{h}\right)$$
$$\text{subject to } \underline{h}^T \underline{1} = 1.$$

Note: only unbiased estimator and zero-mean noise were imposed to obtain these results.

Exercise: Given the signal model $X[n] = \theta + W[n]$, find the **MVU estimator** for the parameter θ.

Solution:

$$\left.\begin{array}{r}\min_{\underline{h}}\left(\underline{h}^T\underline{\underline{R}}_W\underline{h}\right)\\ \underline{h}^T\underline{1}=1\end{array}\right\}\mathcal{L}(\underline{h},\lambda)=\underline{h}^T\underline{\underline{R}}_W\underline{h}-\lambda\left(\underline{h}^T\underline{1}-1\right).$$

$$\nabla\mathcal{L}=0\iff\begin{cases}\nabla_{\underline{h}}\mathcal{L}(\underline{h},\lambda)=0\\ \frac{\partial\mathcal{L}}{\partial\lambda}=0\end{cases}.$$

$$\nabla_{\underline{h}}\left[\underline{h}^T\underline{\underline{R}}_W\underline{h}-\lambda\left(\underline{h}^T\underline{1}-1\right)\right]=2\underline{\underline{R}}_W\underline{h}-\lambda\underline{1}=0. \qquad (1)$$

$$\text{Previous constraint: }\frac{\partial\mathcal{L}}{\partial\lambda}=\underline{h}^T\underline{1}-1=0. \qquad (2)$$

$$\text{From (1) we have that }2\underline{\underline{R}}_W\underline{h}-\lambda\underline{1}=0\text{ and this holds }\iff\underline{h}=\frac{\lambda}{2}\underline{\underline{R}}_W^{-1}\underline{1}. \qquad (3)$$

$$\text{Using (3) on (2), we get to }\left[\frac{\lambda}{2}\underline{\underline{R}}_W^{-1}\underline{1}\right]^T\underline{1}=1\iff\lambda=\frac{2}{\underline{1}^T\underline{\underline{R}}_W^{-1}\underline{1}}. \qquad (4)$$

$$\text{Now, using (4), we have an expression for the filter }\underline{h}\text{: }\boxed{\underline{h}=\frac{\underline{\underline{R}}_W^{-1}\underline{1}}{\underline{1}^T\underline{\underline{R}}_W^{-1}\underline{1}}.}$$

In the case of stationary white noise, the correlation matrix is the identity and, as such, the filter $\underline{h}$ is $\underline{h}=\frac{1}{N}\underline{1}$. In either case, the parameter θ is $\hat{\theta}_N=\underline{h}^T\underline{x}$. We can see that it is unbiased, as

$$\underline{h}^T\underline{1}=\left[\frac{\underline{\underline{R}}_W^{-1}\underline{1}}{\underline{1}^T\underline{\underline{R}}_W^{-1}\underline{1}}\right]^T\underline{1}=\left[\left(\underline{\underline{R}}_W^{-1}\right)^T=\underline{\underline{R}}_W^{-1}\right]=\frac{\underline{1}^T\underline{\underline{R}}_W^{-1}\underline{1}}{\underline{1}^T\underline{\underline{R}}_W^{-1}\underline{1}}=1.$$

---

3. Function Estimation

In some cases, we want to **estimate a function** rather than a single parameter. Common cases are:

- The **self-correlation** function of a process.
- The **spectral density** function of a process.

When estimating a parameter, the used **estimator** becomes a **random variable**. Therefore, when estimating a function (an ordered set of parameters) the **estimator** becomes a **random process** too (an ordered set of random variables).

Given N samples $\{x[0],\ldots,x[N-1]\}$ of a realization of an ergodic process $X[n]$, we want to estimate the self-correlation of that process; let us analyze how to estimate each lag l of the self-correlation function $r_x[l]=\mathbb{E}[X[n+l]\cdot X[n]]$. We will first assess the following estimator $\check{r}_x$:

$$\check{r}_x[l]=\begin{cases}\dfrac{1}{N-l}\displaystyle\sum_{n=0}^{N-l-1}x[n+l]x[n], & 0\le l\le N-1,\\ \dfrac{1}{N-|l|}\displaystyle\sum_{n=|l|}^{N-1}x[n+l]x[n], & -N+1\le l\le 0.\end{cases}$$

As the correlation function is symmetric ($r_x[l]=r_x[-l]$) the second expression (for negative lags) is not computed. The $\check{r}_x$ estimator is **unbiased**, and, terefore, $\mathrm{MSE}\left(\check{r}_x[l]\right)=\sigma^2\left(\check{r}_x[l]\right)$. However, the value of $\sigma^2\left(\check{r}_x[l]\right)$ is not known. It has only been approximated for specific cases of random processes.

Let's see that the estimator is unbiased: we will only check for positive lags, as we know the function is symmetric.

$$\mathbb{E}[\check{r}_x[l]]=\mathbb{E}\left[\frac{1}{N-l}\sum_{n=0}^{N-l-1}x[n+l]x[n]\right]=\frac{1}{N-l}\sum_{n=0}^{N-l-1}\mathbb{E}[x[n+l]x[n]]=\frac{1}{N-l}\sum_{n=0}^{N-l-1}r_x[l]=r_x[l].$$

Therefore, $B(\check{r}_x) = 0$ and the estimator is unbiased.

❖ The value of its variance has only been rpoven for the Gaussian case and $N >> l$, and it's equal to

$$\sigma^2(\check{r}_x[l]) = \frac{N}{(N-|l|)^2} \sum_{k=-\infty}^{\infty} \left(r_x^2[k] + r_x[k+l] + r_x[k-l]\right).$$

However unknown their value, it is known that the $\check{r}_x$ estimator behaves commonly for all probability distributions:

- Its variance increases with the absolute value of the lag $|l|$.
- The estimator is consistent, meaning that $\lim_{N\to\infty} \sigma^2(\check{r}_x[l]) = 0$.

How to improve the variance behavior?

To remove the dependency of l from the variance, a new estimator for the self-correlation is proposed:

$$\hat{r}_x[l] = \begin{cases} \frac{1}{N} \sum_{n=0}^{N-l-1} x[n+l]x[n], & 0 \le l \le N-1 \\ \frac{1}{N} \sum_{n=|l|}^{N-1} x[n+l]x[n], & -N+1 \le l \le 0 \end{cases}$$

Both estimators are clearly related: $\hat{r}_x[l] = \frac{N-|l|}{N}\check{r}_x[l]$. We will now see that the new estimator is **biased**, and that it reduces the variance and the MSE.

- As the two estimators are linearly related, we can see that the expected value of $\hat{r}_x[l]$ is

$$\mathbb{E}[\hat{r}_x[l]] = \frac{N-|l|}{N}\mathbb{E}[\check{r}_x[l]] = \frac{N-|l|}{N} r_x[l].$$

- The new variance is independent of l, has decreased, and it still makes the estimator be consistent:

$$\sigma^2(\hat{r}_x[l]) = \frac{1}{N} \sum_{k=-\infty}^{\infty} \left(r_x^2[k] + r_x\left[k+l\right] + r_x[k-l]\right)$$

- It can be shown that the MSE has decreased, $\mathrm{MSE}(\hat{r}_x) < \mathrm{MSE}(\check{r}_x)$.

The available N samples can be modeled as having a **whole realization** of the process that has been **windowed**. A (consistent) **square window** upon the data samples $v[n]$ produces a **triangular window** $w[l]$ upon the mean of the correlation samples: $w[l] = \frac{1}{N}v[k] * v[-k]$.

2.2. Cramer-Rao bound and Efficient Estimator

In the previous unit we have been able to find the MVU estimator for the estimation of the mean value of a signal $X[n]$ that can be modeled as a constant value embedded in zero-mean noise, $\theta + W[n]$. To obtain the estimator, we have used the method of Lagrange multipliers to minimize a given criterion subject to an unbiased constraint.

However, if a MVU estimator exists, there is no method that ensures that we are able to find it. Nevertheless, the **Cramer-Rao Lower Bound** (CRLB or CRB):

- Determines the minimum possible variance for any unbiased estimator. This bound, then, provides a benchmark for assessing any estimator performance.
- Provides, in some cases, the expression for the MVU estimator.
- Can be used to estimate the (non-linear) function of a parameter.

Definition. *Efficient estimator.* We say that an estimator is efficient if it attains the CRLB.

Cramer-Rao bound for parameters

There exists a **lower bound** for the variance of the whole set of unbiased estimators of a parameter θ. the bound is related to the **probability density function** of the data: when the pdf is viewed as a function of the unknown parameters (with $\underline{x}$ fixed), it is known as the **likelihood function**:

$$f_{\underline{x}}(x[0], \ldots, x[N-1]; \theta) = f_{\underline{x}}(\underline{x}; \theta)$$

Then, while we won't prove it in this course, we state the **Cramer-Rao Lower Bound**:

Proposition. *Cramer-Rao Lower Bound.* The variance of any unbiased estimator $\hat{\theta}$ must satisfy

$$\boxed{\operatorname{Var}(\hat{\theta}) \geq \frac{1}{-\mathbb{E}\left[\frac{\partial^2 \ln f_{\underline{x}}(\underline{x};\theta)}{\partial \theta^2}\right]},}$$

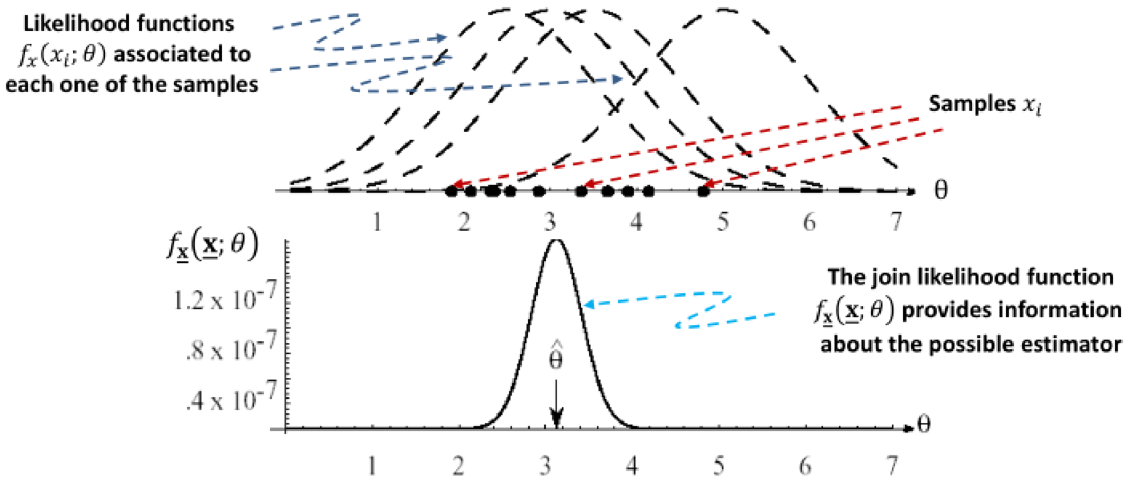
and equality holds when, for some function of the parameter $k(\theta)$,

$$\boxed{\frac{\partial \ln f_{\underline{x}}(\underline{x}; \theta)}{\partial \theta} = k(\theta)(\hat{\theta}_{\text{opt}}(\underline{x}) - \theta).}$$

Let us analyze the case of the likelihood function $f_{\underline{x}}(\underline{x}; \theta)$ of a set of N Gaussian, independent samples:

$$f_{\underline{x}}(\underline{x}; \theta) = \prod_{i=1}^{N} f_x(x_i; \theta)$$

each sample has a likelihood function $f_x(x_i; \theta_i)$ associated to it, and the joint likelihood function $f_{\underline{x}}(\underline{x}; \theta)$ provides information about the possible estimator. Looking for a maximum (we will see this later) in the joint likelihood function will provide a Maximum Likelihood Estimator for the parameter.



The more informative the set of samples $\underline{x}$, the sharper the likelihood function $f_{\underline{x}}(\underline{x}; \theta)$: a measure of sharpness is the **curvature**.

Definition. *Curvature.* The curvature of a likelihood function $f_{\underline{x}}(\underline{x}; \theta)$ is

$$-\mathbb{E}\left[\frac{\partial^2 \ln f}{\partial \theta^2}\right]$$

The larger the curvature, the smaller the Cramer-Rao bound on the variance. We can easily see this as the Cramer-Rao bound is nothing more than

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{-\mathbb{E}\left[\frac{\partial^2 \ln f_{\underline{x}}(\underline{x};\theta)}{\partial\theta^2}\right]}.$$

The curvature depends on both the number of samples N and the likelihood function $f_{\underline{x}}(\underline{x};\theta)$.

The **optimal (efficient) estimator** can be obtained through the condition of minimum variance: that is, imposing that

$$\frac{\partial \ln f_{\underline{x}}(\underline{x};\theta)}{\partial\theta} = k(\theta)(\hat{\theta}_{\text{opt}}(\underline{x}) - \theta)$$

we can see that the optimal estimator $\hat{\theta}_{\text{opt}}$ is

$$\hat{\theta}_{\text{opt}}(\underline{x}) = \frac{1}{k(\theta)}\frac{\partial \ln f_{\underline{x}}(\underline{x};\theta)}{\partial\theta} + \theta.$$

For the estimator to be efficient, the dependence on θ should cancel out. We can see that the **achieved minimum variance** is given by

$$\mathrm{Var}_{\text{opt}}(\hat{\theta}) = \frac{1}{k(\theta)},$$

because if we calculate the curvature,

$$\begin{aligned} -\mathbb{E}_{\underline{x}}\left[\frac{\partial^2 \ln f}{\partial\theta^2}\right] = -\mathbb{E}_{\underline{x}}\left[\frac{\partial}{\partial\theta}(k(\theta)\hat{\theta}_{\text{opt}}(\underline{x}) - k(\theta)\theta)\right] = \\ -\mathbb{E}_{\underline{x}}\left[k'(\theta)\hat{\theta}_{\text{opt}}(\underline{x}) - k'(\theta)\theta - k(\theta)\right] = \left[\mathbb{E}[\hat{\theta}_{\text{opt}}(\underline{x})] = \theta\right] = \\ -\mathbb{E}_{\underline{x}}[-k(\theta)] = k(\theta). \end{aligned}$$

The denominator in the CRLB is referred to as the **Fisher Information** $I(\theta)$:

$$I(\theta) := -\mathbb{E}\left[\frac{\partial^2 \ln f_{\underline{x}}(\underline{x};\theta)}{\partial\theta^2}\right] = \mathbb{E}\left[\left(\frac{\partial \ln f_{\underline{x}}(\underline{x};\theta)}{\partial\theta}\right)^2\right]$$

Exercise: Given N samples of a process that can be modeled as $\underline{x} = \theta\underline{1} + \underline{w}$, compute an **efficient estimator** of its mean θ.

Note: $\underline{w}$ is a Gaussian stationary white noise.

Solution:

Exercise: Given N samples of a process that can be modeled as $\underline{x} = \theta\underline{1} + \underline{w}$, compute an **efficient estimator** of its mean θ.

Note: $\underline{w}$ is a Gaussian colored white noise.

Solution:

Cramer-Rao bound for parameter vectors

The extension to the case of a vector parameter $\underline{\theta}$ is as follows: the pdf is

$$f_{\underline{x}}(x[0],\ldots,x[N-1];\theta_1,\ldots,\theta_P) = f_{\underline{x}}(\underline{x};\underline{\theta}),$$

and the lower bound for estimator variance is the following:

Proposition. *Cramer-Rao Lower Bound for vector parameters.* The variance of any unbiased estimator $\hat{\theta}_i$ must satisfy

$$\boxed{\operatorname{Var}(\hat{\theta}_i) \geq \left[\underline{\underline{I}}^{-1}(\underline{\theta})\right]_{ii},}$$

where $\underline{\underline{I}}(\underline{\theta})$ is the $P \times P$ **Fisher Information Matrix**,

$$\left[\underline{\underline{I}}(\underline{\theta})\right]_{ij} = -\mathbb{E}\left[\frac{\partial^2 \ln f_{\underline{x}}(\underline{x};\underline{\theta})}{\partial\theta_i \partial\theta_j}\right].$$

Equality for the variance bound holds whenever the gradient of f with respect to $\underline{\theta}$ satisfies the following:

$$\boxed{\nabla_{\underline{\theta}}\left(f_{\underline{x}}(\underline{x};\underline{\theta})\right) = \underline{\underline{I}}^{-1}(\underline{\theta})\left(\underline{\theta}_{\text{opt}}(\underline{x}) - \underline{\theta}\right).}$$

Exercise: Given N samples of a process that can be modeled as $\underline{x} = A\underline{1} + \underline{w}$, compute an **efficient estimator** of its mean A and variance σ^2.

Note: $\underline{w}$ is a Gaussian stationary white noise.

Solution:

2.3. Maximum Likelihood & Maximum a Posteriori Estimator

The CRLB states that there exists a lower bound for the variance of the whole set of unbiased estimators of a parameter θ. It proposes a mechanism that, in some cases, allows obtaining this estimator; this particular estimator that attains the variance bound is termed **efficient**. Nevertheless, there is no feasible estimator that satisfies the Cramer-Rao Lower Bound.

Maximum Likelihood Estimator

Let us define the ML estimator:

Definition. *Maximum Likelihood Estimator.* The maximum likelihood estimator for a parameter θ is

$$\hat{\underline{\theta}}_{\text{ML}} = \underset{\underline{\theta}\in\Theta}{\operatorname{argmax}}\, f_{\underline{x}}(\underline{x};\underline{\theta}).$$

Properties. The ML estimator has the following properties:

- It is **asymptotically unbiased** (and in most cases, unbiased).
- It is **asymptotically efficient**: when N increases, its variance attains the Cramer-Rao bound.
- It is closely related to **efficiency**. In fact, whenever there exists an efficient estimator for a parameter, it is the ML estimator.
- It follows a **Gaussian distribution** for large N, characterized by its mean and variance.
- **Invariance through maps:** the ML estimator of a function of a parameter, $\alpha = g(\theta)$, can be obtained as $\hat{\alpha}_{\text{ML}} = g(\hat{\theta}_{\text{ML}})$.

Let's see why the efficient estimator is exactly the ML estimator: if there exists such an estimator, the following factorization

$$\frac{\partial \ln f_{\underline{x}}(\underline{x};\theta)}{\partial \theta} = k(\theta)(g(\underline{x}) - \theta)$$

has been possible. As $\ln(\cdot)$ is a monotonically increasing function, the positions of the extrema do not change. Mathematically speaking,

$$\frac{\partial \ln f_{\underline{x}}(\underline{x};\theta)}{\partial \theta} = 0 \iff \frac{\partial f_{\underline{x}}(\underline{x};\theta)}{\partial \theta} = 0.$$

Thus, if there is an efficient estimator, the Cramer-Rao and the Maximum Likelihood estimators are the same, as

$$\frac{\partial \ln f_{\underline{x}}(\underline{x};\theta)}{\partial \theta} = 0 \iff g(\underline{x}) - \theta = 0 \iff \boxed{\hat{\theta}_{\mathrm{ML}} = g(\underline{x}) = \hat{\theta}_{\mathrm{opt}}(\underline{x}) = \hat{\theta}_{\mathrm{CR}}.}$$

Exercise: Given N samples of a process that can be modeled as $\underline{x} = A\underline{1} + \underline{w}$, compute the **ML estimator** of its mean A and variance σ^2.

Note: $\underline{w}$ is a Gaussian stationary white noise.

Solution:

Exercise: Given N samples of a process that can be modeled as $\underline{x} = \theta\underline{1} + \underline{w}$, compute the **ML estimator** of its mean θ.

Note: $\underline{w}$ is a Gaussian stationary colored noise.

Solution:

Exercise: Given N independent samples of a Laplacian process $\underline{x} = m\underline{1} + \underline{w}$, we want to obtain the **ML estimator** of their mean m and diversity λ.

Note: $\underline{w}$ is a Laplacian stationary white noise. The parameter vector is $\underline{\theta} = (m, \lambda)$.

Solution:

Exercise: We have 2 measures of a magnitude $z_i = x + v_i$, with different errors. The errors are Gaussian, zero-mean, with variance σ_i^2 and independent. Compute the ML estimator of the magnitude to be measured.

Solution:

Maximum a Posteriori Estimator

A **Bayesian estimator** models the parameter we are attempting to estimate as a **realization of a random variable**, instead of as a constant unknown value. With this approach, we can include the **prior pdf of the parameter** $f_\theta(\theta)$, which summarizes our *a priori* knowledge about the parameter.

$$\hat{\theta}_{\mathrm{MAP}} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, f_{\underline{x},\theta}(\underline{x},\theta) = \underset{\theta\in\Theta}{\operatorname{argmax}}\big(f_{\underline{x}}(\underline{x}|\theta) f_\theta(\theta)\big)$$

Note: conceptually, $f_{\underline{x}}(\underline{x};\theta)$ is a family of pdf's and $f_{\underline{x}}(\underline{x}|\theta)$ is a conditional pdf.

It is called the **Maximum a Posteriori** (MAP) estimator, since it can be formulated as:

$$\hat{\theta}_{\mathrm{MAP}} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, f_\theta(\theta|\underline{x}) = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \frac{f_{\underline{x}}(\underline{x}|\theta) f_\theta(\theta)}{f_{\underline{x}}(\underline{x})} = \underset{\theta\in\Theta}{\operatorname{argmax}}\big(f_{\underline{x}}(\underline{x}|\theta) f_\theta(\theta)\big).$$

MAP and ML estimators: The conditional probability function $f_{\underline{x}}(\underline{x}|\theta)$ will be sharper around θ_0, as the number of samples N increases. In this case, if the information provided by $f_\theta(\theta)$ is correct, both estimators tend to be the same.

MAP with different priors: if we do not have any prior information about the parameter to be estimated, its pdf $f_\theta(\theta)$ is a constant and any possible value has the same likelihood. Then, the MAP estimator becomes the ML estimator.

Exercise: Given N samples of a process that can be modeled as $\underline{x} = \mu\underline{1} + \underline{w}$, compute the **MAP estimator** of its mean μ, knowing that it is a random variable with distribution $\mathcal{N}\left(\mu_m, \sigma_m^2\right)$.

Note: $\underline{w}$ is a Gaussian stationary colored noise.

Solution:
