ESTIMATION THEORY

Michele TARAGNA

Dipartimento di Elettronica e Telecomunicazioni

Politecnico di Torino

michele.taragna@polito.it

III Level Course 02LCPRV / 01LCPRV / 01LCPIU

“Experimental modeling: model building from experimental data”

Estimation problem

The estimation problem refers to the empirical evaluation of an uncertain variable, like an unknown characteristic parameter or a remote signal, on the basis of observations and experimental measurements of the phenomenon under investigation.

An estimation problem always assumes a suitable mathematical description (*model*) of the phenomenon:

- in the classical statistics, the investigated problems usually involve *static models*, characterized by instantaneous (or algebraic) relationships among variables;
- in this course, estimation methods are introduced also for phenomena that are adequately described by *discrete-time dynamic models*, characterized by relationships among variables that can be represented by means of difference equations (i.e., for simplicity, the time variable is assumed to be discrete).

Estimation problem

$\theta(t)$: real variable to be estimated, scalar or vector, constant or time-varying;

$d(t)$: available data, acquired at N time instants $t_1, t_2, \ldots, t_N$;

$T = \{t_1, t_2, \ldots, t_N\}$: set of time instants used for observations, distributed with regularity (in this case, $T = \{1, 2, \ldots, N\}$) or non-uniformly;

$d = \{d(t_1), d(t_2), \ldots, d(t_N)\}$: observation set.

An **estimator** (or **estimation algorithm**) is a *function* $f(\cdot)$ that, starting from data, associates a value to the variable to be estimated:

$$\theta(t) = f(d)$$

The **estimate** term refers to the particular *value* given by the estimator when applied to the particular observed data.

Estimation problem classification

1) $\theta(t)$ is constant => **parametric identification** problem:
   - the estimator is denoted by $\hat{\theta}$ or by $\hat{\theta}_T$;
   - the true value of the unknown variable (if makes sense) is denoted by θ_o;
2) $\theta(t)$ is a time-varying function:
   - the estimator is denoted by $\hat{\theta}\left(t|T\right)$, or by $\hat{\theta}\left(t|N\right)$ if the time instants for observations are uniformly distributed;
   - according to the temporal relationship between t and the last time instant t_N:

   2.a) if $t > t_N$ => **prediction** problem;

   2.b) if $t = t_N$ => **filtering** problem;

   2.c) if $t_1 < t < t_N$ => **regularization** or **interpolation** or **smoothing** problem.

Example of prediction problem: time series analysis

Given a sequence of observations (time series or historical data set) of a variable y:

$$y(1), y(2), \ldots, y(t)$$

the goal is to evaluate the next value $y(t+1)$ of this variable

$$\Downarrow$$

it is necessary to find a good **predictor** $\hat{y}(t+1|t)$, i.e., a function of available data that provides the most accurate evaluation of the next value of the variable y:

$$\hat{y}(t+1|t) = f\left(y(t), y(t-1), \ldots, y(1)\right) \cong y(t+1)$$

A predictor is said to be *linear* if it is a linear function of data:

$$\hat{y}(t+1|t) = a_1(t)y(t) + a_2(t)y(t-1) + \ldots + a_t(t)y(1) = \sum_{k=1}^{t} a_k(t)y(t-k+1)$$

A linear predictor has *a finite memory* n if it is a linear function of the last n data only:

$$\hat{y}(t+1|t) = a_1(t)y(t) + a_2(t)y(t-1) + \ldots + a_n(t)y(t-n+1) = \sum_{k=1}^{n} a_k(t)y(t-k+1)$$

If all the parameters $a_i(t)$ are constant, the predictor is also *time-invariant*:

$$\hat{y}(t+1|t) = a_1 y(t) + a_2 y(t-1) + \ldots + a_n y(t-n+1) = \sum_{k=1}^{n} a_k y(t-k+1)$$

and it is characterized by the vector of constant parameters

$$\theta = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}^T \in \mathbb{R}^n$$

$$\Downarrow$$

The prediction problem becomes a parametric identification problem.

Questions:

- how to measure the predictor quality?
- how to derive the “best” predictor?

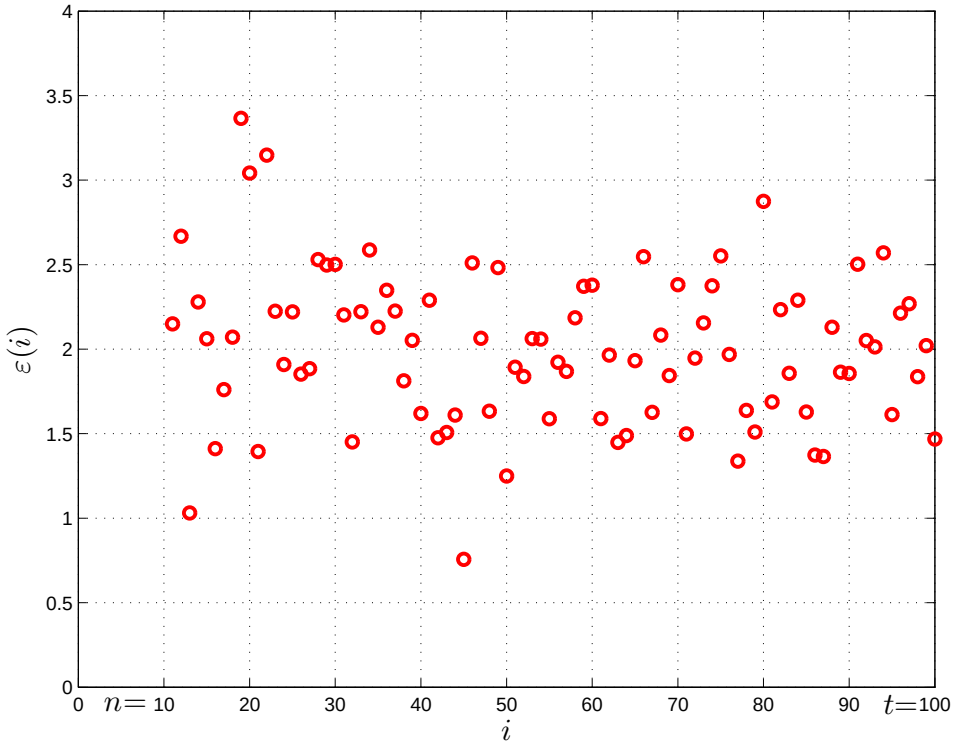
If the predictive model is linear, time-invariant, with finite memory n much shorter than the total number of data measured up to time instant t, its predictive capability over the available data $y(i)$, $i = 1, 2, \ldots, t$, can be evaluated in the following way:

- at each instant $i \geq n$, the prediction $\hat{y}(i+1|i)$ of the next value is computed:
  $\hat{y}(i+1|i) = a_1 y(i) + a_2 y(i-1) + \ldots + a_n y(i-n+1) = \sum_{k=1}^{n} a_k y(i-k+1)$
  and its *prediction error* $\varepsilon(i+1)$ with respect to $y(i+1)$ is evaluated:
  $$\varepsilon(i+1) = y(i+1) - \hat{y}(i+1|i)$$
- the model described by θ is a good predictive model if the error ε is "small" over all the available data $\Rightarrow$ the following figure of merit is introduced:
  $$J(\theta) = \sum_{k=n+1}^{t} \varepsilon(k)^2$$ (sum of squares of prediction errors)
- the best predictor is the one that minimizes J and the value of its parameters is:
  $$\theta^* = \arg\min_{\theta \in \mathbb{R}^n} J(\theta)$$

For example, if $t = 100$ and $n = 10 \ll t$, for a given $\theta = [a_1 \cdots a_{10}]^T$ it results:

$$\begin{cases} \hat{y}(11|10) = a_1 y(10) + \ldots + a_{10} y(1) & \Rightarrow \quad \varepsilon(11) = y(11) - \hat{y}(11|10) \\ \hat{y}(12|11) = a_1 y(11) + \ldots + a_{10} y(2) & \Rightarrow \quad \varepsilon(12) = y(12) - \hat{y}(12|11) \\ \vdots & \vdots \\ \hat{y}(100|99) = a_1 y(99) + \ldots + a_{10} y(90) & \Rightarrow \quad \varepsilon(100) = y(100) - \hat{y}(100|99) \end{cases}$$

and then the behaviour of the prediction error sequence $\varepsilon(\cdot)$ is plotted:



Fundamental question: is the predictor minimizing J necessarily a “good” model?

The predictor quality depends on the fact that the temporal behaviour of the prediction error sequence $\varepsilon(\cdot)$ has the following characteristics:

- its mean value is zero, i.e., it does not show a systematic error;
- it is “fully random”, i.e., it does not contain any regularity element.

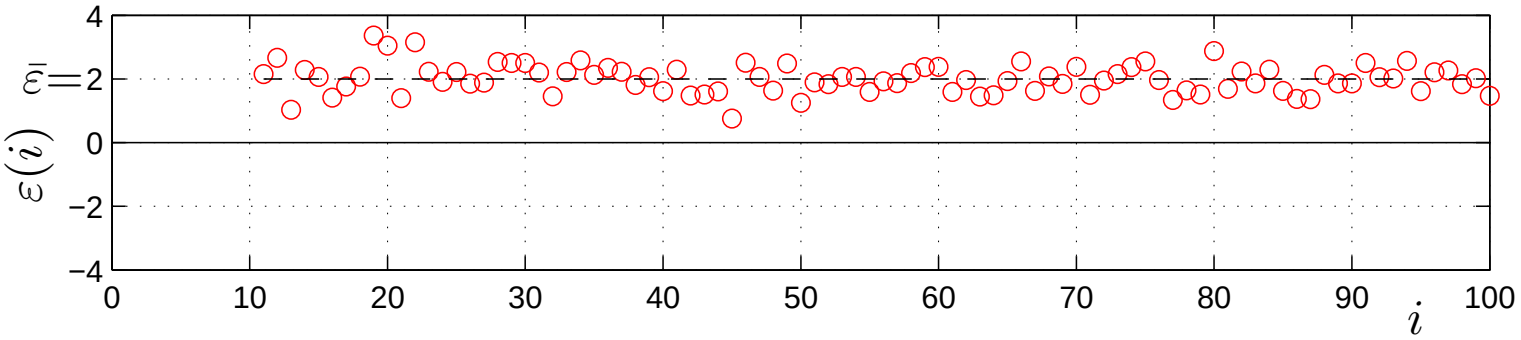
In probabilistic terms, this corresponds to require that the behaviour of the error $\varepsilon(\cdot)$ is that of a **white noise** (WN) process, i.e., a sequence of independent random variables with zero mean value and constant variance σ^2:

$$\varepsilon(\cdot) = WN\left(0, \sigma^2\right)$$

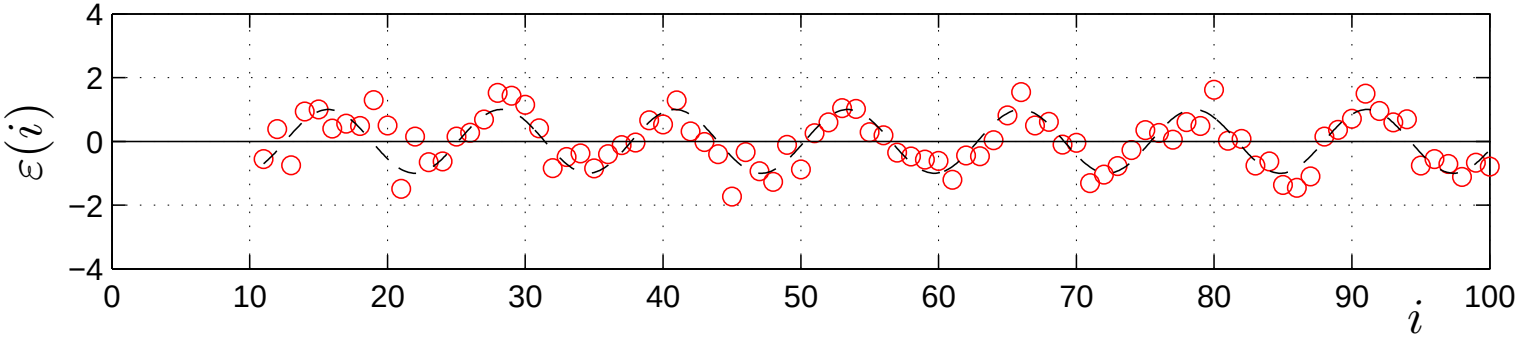
$$\Downarrow$$

A predictor is a “good” model if $\varepsilon(\cdot)$ has the white noise probabilistic characteristics.

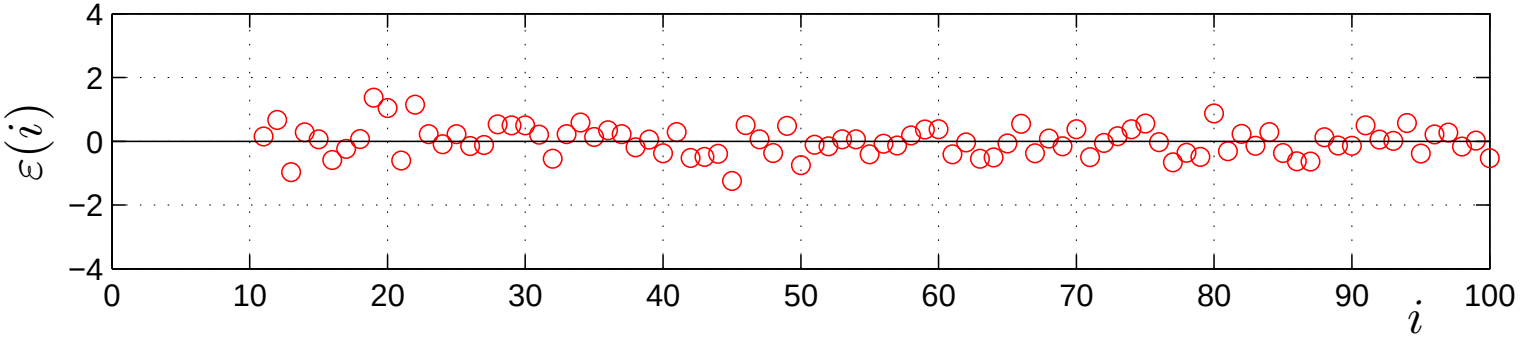
Example #1: prediction error with constant systematic error



Example #2: prediction error with sinusoidal systematic error



Example #3: "fully random" prediction error, with no systematic error



Then, the prediction problem can be recast as the study of a **stochastic system**, i.e., a dynamic system whose inputs are probabilistic signals; in fact:

$$\begin{cases} \hat{y}(t|t-1) = a_1 y(t-1) + a_2 y(t-2) + \ldots + a_n y(t-n) \\ \varepsilon(t) = y(t) - \hat{y}(t|t-1) \end{cases} \Rightarrow$$

$$y(t) = \hat{y}(t|t-1) + \varepsilon(t) = a_1 y(t-1) + a_2 y(t-2) + \ldots + a_n y(t-n) + \varepsilon(t)$$

represents a discrete-time LTI dynamic system with output $y(t)$ and input $\varepsilon(t)$

$$\Downarrow$$

$\mathcal{Z}$-transforming, with $\mathcal{Z}[y(t-k)] = z^{-k}Y(z)$ and z^{-1} the unitary delay operator:

$$Y(z) = a_1 z^{-1} Y(z) + a_2 z^{-2} Y(z) + \ldots + a_n z^{-n} Y(z) + \varepsilon(z)$$

$$\Downarrow$$

$$H(z) = \frac{Y(z)}{\varepsilon(z)} = \frac{1}{1 - a_1 z^{-1} - a_2 z^{-2} - \ldots - a_n z^{-n}} = \frac{z^n}{z^n - a_1 z^{n-1} - a_2 z^{n-2} - \ldots - a_n}$$

represents the transfer function of a LTI dynamic system $\Rightarrow$ in order to be a "good" model, its input $\varepsilon(\cdot)$ shall have the white noise probabilistic characteristics.

Classification of data descriptions

- The actually available information is always:
  - bounded $\Rightarrow$ the measurement number N is necessarily finite;
  - corrupted by different kinds of uncertainty (e.g., measurement noise).

- The uncertainty affecting the data can be described:
  - in probabilistic terms $\Rightarrow$ we talk about **statistical** or **classical estimation**;
  - in terms of set theory, as a member of some bounded set $\Rightarrow$ we talk about **Set Membership** or **Unknown But Bounded** (UBB) **estimation**.

Probabilistic description of data

In the *probabilistic* (or *classical* or *statistical*) framework, data d are assumed to be produced by a random source of data $\mathcal{S}$, influenced by:

- the outcome s of a random experiment $\mathcal{E}$
- the "true" value θ_o of the unknown variable to be estimated

$$d = d\left(s, \theta_o\right)$$

$$\Downarrow$$

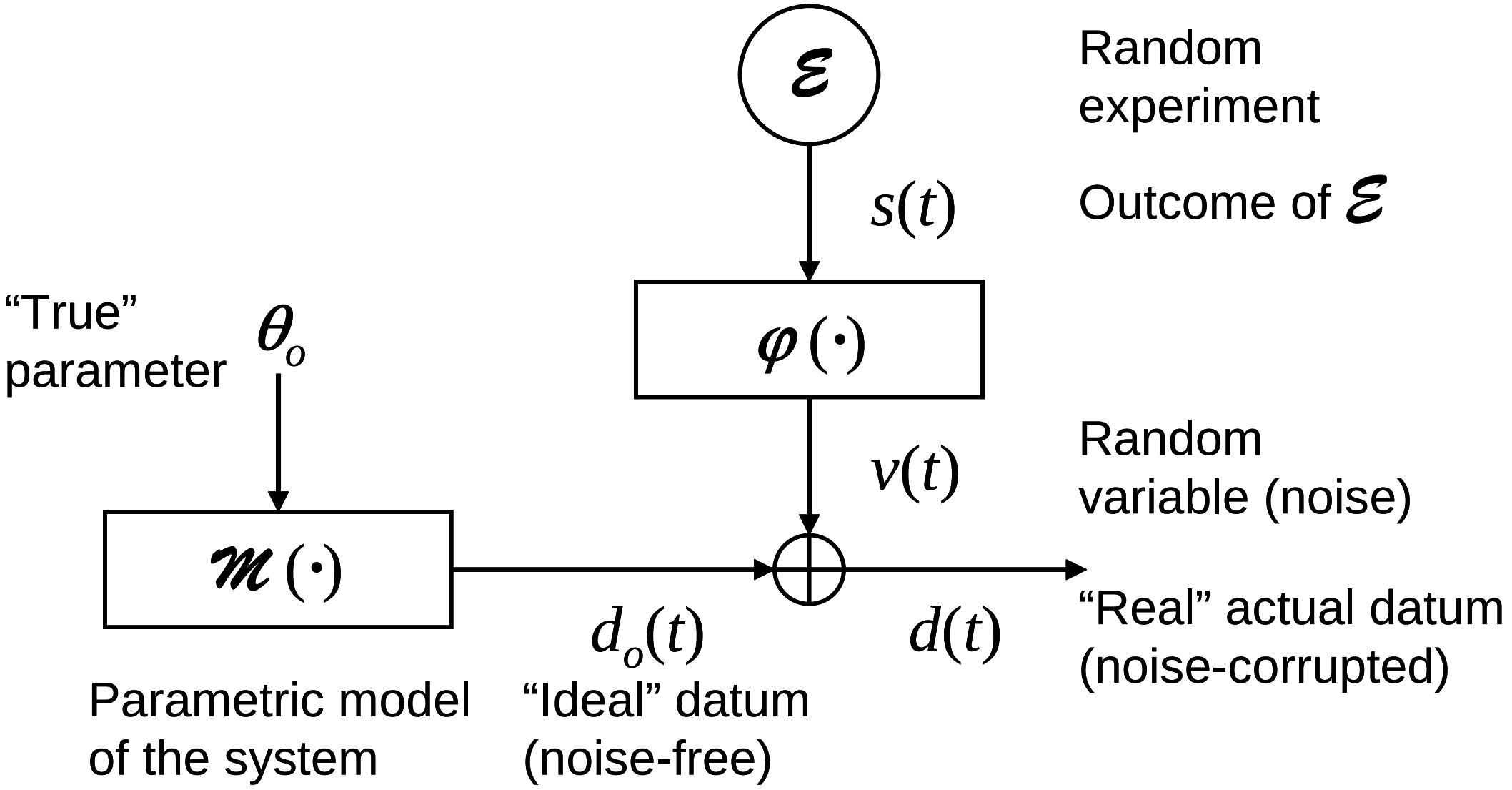
data d are random variables, since they are functions of the outcome s

$$\Downarrow$$

A full probabilistic description of data is constituted by

- its **probability distribution** $F(q) = Prob\left\{d\left(s, \theta_o\right) \leq q\right\}$ or
- its **probability density function** $f(q) = \dfrac{dF(q)}{dq}$, often denoted by p.d.f.

Random source of data:



Estimator characteristics

A random source of data $\mathcal{S}$, influenced by the outcome s of a random experiment $\mathcal{E}$ and by the "true" value θ_o of the unknown variable to be estimated, produces data d:

$$d = d\left(s, \theta_o\right)$$

$$\Downarrow$$

data d are random variables, since they are functions of the outcome s

$$\Downarrow$$

the estimator $f(\cdot)$ and the estimate $\hat{\theta}$ are random variables too, being functions of d:

$$\hat{\theta} = f(d) = f\left(d\left(s, \theta_o\right)\right)$$

$$\Downarrow$$

the quality of $f(\cdot)$ and $\hat{\theta}$ depends on their probabilistic characteristics.

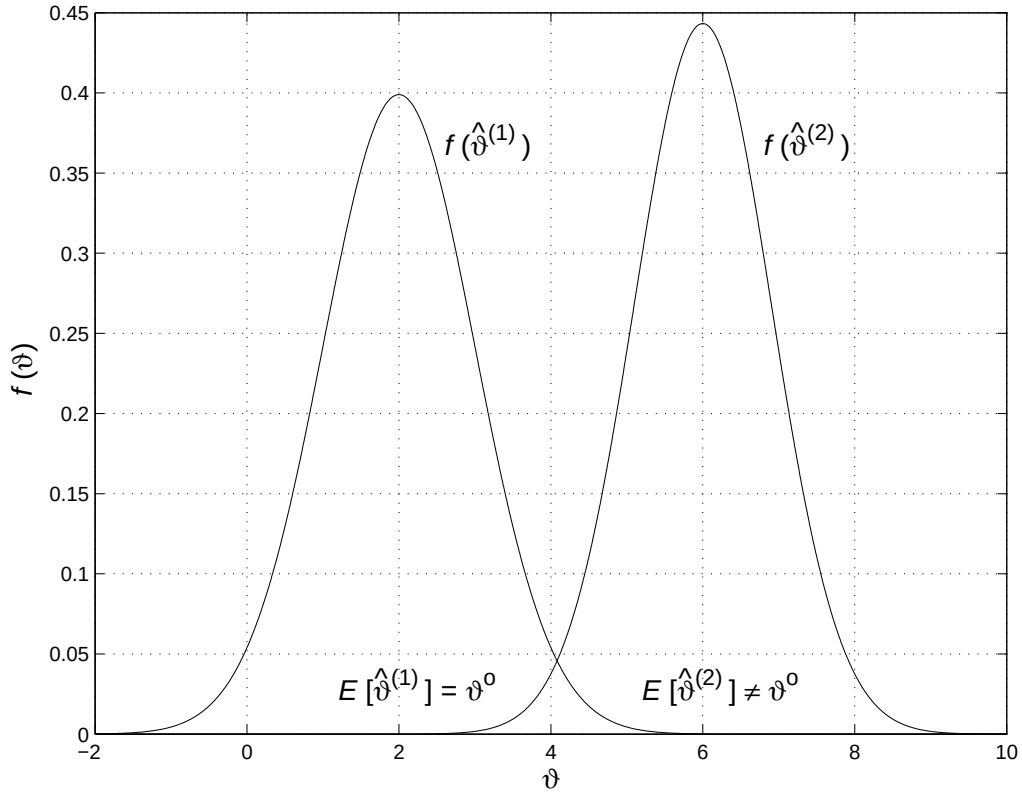
Estimator probabilistic characteristics

- No bias (in order to avoid to introduce any systematic estimation error)
- Minimum variance (smaller scattering around the mean value guarantees higher probability of obtaining values close to the “true” value θ_o)
- Asymptotic characteristics (for $N \to \infty$):
  - quadratic-mean convergence
  - almost-sure convergence
  - consistency

Estimator probabilistic characteristics

An estimator is said to be **unbiased** (or **correct**) if

$$E\left[\hat{\theta}\right] = \theta_o$$

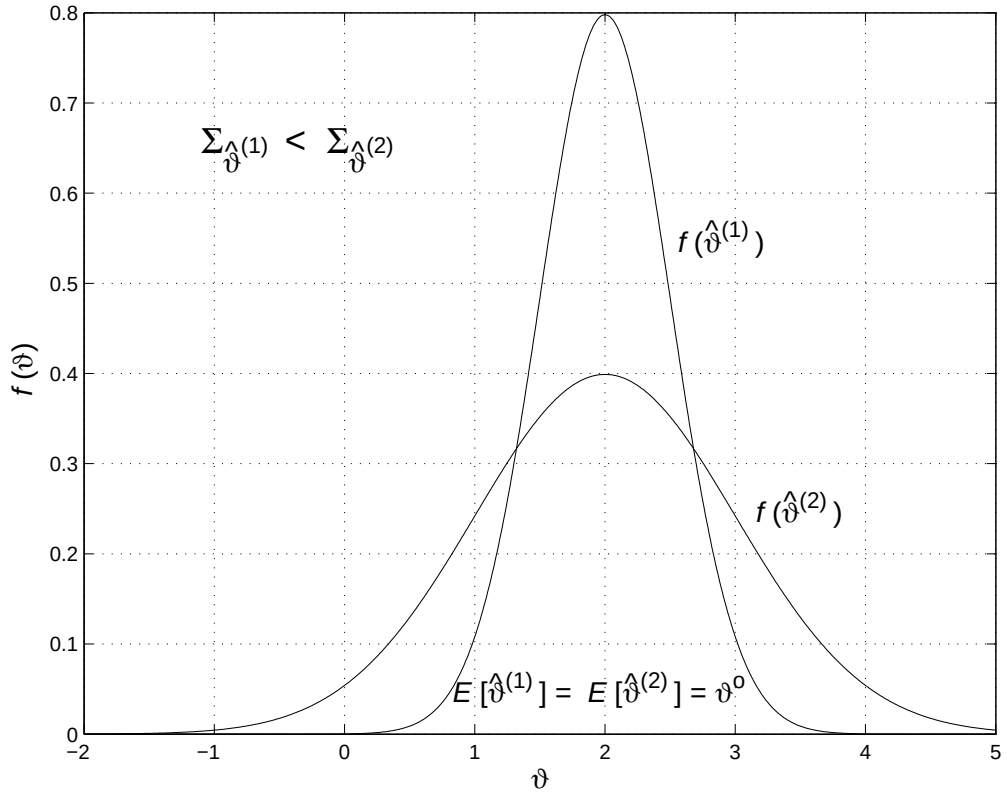


An unbiased estimator does not introduce any systematic estimation error.

Estimator probabilistic characteristics

An unbiased estimator $\hat{\theta}^{(1)}$ is said to be **efficient** (or with **minimum variance**) if

$$Var[\hat{\theta}^{(1)}] \leq Var[\hat{\theta}^{(2)}], \quad \forall \text{ unbiased } \hat{\theta}^{(2)} \neq \hat{\theta}^{(1)}$$



Smaller scattering around the mean value $\Rightarrow$ higher probability of approaching θ_o.

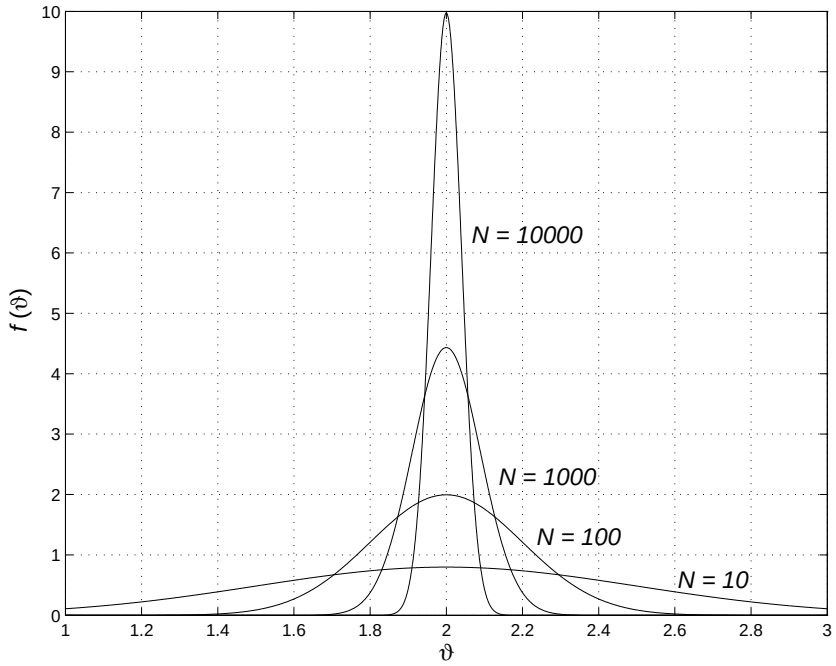
Estimator probabilistic characteristics

An unbiased estimator **converges in quadratic mean to θ_o**, i.e., $\underset{N\to\infty}{l.i.m.}\,\hat{\theta}_N = \theta_o$, if

$$\lim_{N\to\infty} E\left[\|\hat{\theta}_N - \theta_o\|^2\right] = 0$$

where $\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}, \quad \forall x \in \mathbb{R}^n$, is the Euclidean norm.

An unbiased estimator such that $\lim_{N\to\infty} Var\left[\hat{\theta}_N\right] = 0$ converges in quadratic mean:



Sure and almost-sure convergence, consistency

An estimator is function of both the outcome s of a random experiment $\mathcal{E}$ and θ_o:

$$\hat{\theta} = f(d) = f(d(s, \theta_o)) \quad \Rightarrow \quad \hat{\theta} = \hat{\theta}(s, \theta_o)$$

If a particular outcome $\overline{s} \in S$ is considered and the sequence of estimates $\hat{\theta}_N(\overline{s},\theta_o)$ is evaluated for increasing N, a numerical series $\hat{\theta}_1(\overline{s},\theta_o)$, $\hat{\theta}_2(\overline{s},\theta_o)$, ..., is derived that may converge to θ_o for some $\overline{s}$, and may not converge for some other $\overline{s}$.

Let A be the set of outcomes $\overline{s}$ guaranteeing the convergence to θ_o:

- if $A \equiv S$, then we have **sure convergence**, since it holds $\forall \overline{s} \in S$;
- if $A \subset S$, considering A like an event, the probability $P(A)$ may be defined; if A is such that $P(A) = 1$, we say that $\hat{\theta}_N$ converges to θ_o *with probability* 1:

$$\lim_{N \to \infty} \hat{\theta}_N = \theta_o \qquad w.p.1$$

we have **almost-sure convergence** $\Rightarrow$ the algorithm is said to be **consistent**.

Example

Problem: N scalar data d_i with the same mean value $E[d_i] = \theta_o$, with variances $Var[d_i]$ possibly different but bounded ($\exists \sigma \in \mathbb{R}_+ : Var[d_i] \leq \sigma^2 < \infty, \forall i$); data are uncorrelated, i.e.:

$$E[\{d_i - E[d_i]\}\{d_j - E[d_j]\}] = 0, \quad \forall i \neq j$$

Estimator #1 (sample mean):

$$\boxed{\hat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} d_i}$$

- it is an unbiased estimator:

$$E\left[\hat{\theta}_N\right] = E\left[\tfrac{1}{N}\textstyle\sum_{i=1}^{N} d_i\right] = \tfrac{1}{N}\textstyle\sum_{i=1}^{N} E[d_i] = \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \theta_o = \theta_o$$

- it converges in quadratic mean:

$$
\begin{aligned}
Var\left[\hat{\theta}_N\right] &= E\left[\left(\hat{\theta}_N - E\left[\hat{\theta}_N\right]\right)^2\right] = E\left[\left(\tfrac{1}{N}\sum_{i=1}^N d_i - \theta_o\right)^2\right] = \\
&= E\left[\left(\tfrac{1}{N}\sum_{i=1}^N d_i - \tfrac{1}{N}\sum_{i=1}^N \theta_o\right)^2\right] = E\left[\left(\tfrac{1}{N}\sum_{i=1}^N (d_i - \theta_o)\right)^2\right] = \\
&= E\left[\tfrac{1}{N^2}\left(\sum_{i=1}^N (d_i - \theta_o)\right)^2\right] = \tfrac{1}{N^2} E\left[\left(\sum_{i=1}^N (d_i - \theta_o)\right)^2\right] = \\
&= \tfrac{1}{N^2} E\left[\sum_{i=1}^N (d_i - \theta_o)^2 + \sum_{i=1}^N (d_i - \theta_o)\sum_{j=1, j\neq i}^N (d_j - \theta_o)\right] = \\
&= \tfrac{1}{N^2}\left\{\sum_{i=1}^N E\left[(d_i - \theta_o)^2\right] + \sum_{i=1}^N E\left[(d_i - \theta_o)\sum_{j=1, j\neq i}^N (d_j - \theta_o)\right]\right\} = \\
&= \tfrac{1}{N^2}\sum_{i=1}^N Var\left[d_i\right] \leq \tfrac{1}{N^2}\sum_{i=1}^N \sigma^2 = \sigma^2 / N
\end{aligned}
$$

$$\Downarrow$$

$$\lim_{N\to\infty} Var\left[\hat{\theta}_N\right] \leq \lim_{N\to\infty} \frac{\sigma^2}{N} = 0$$

$$\Downarrow$$

the algorithm converges in quadratic mean, since it is unbiased and with $\lim_{N\to\infty} Var\left[\hat{\theta}_N\right] = 0$.

Estimator #2:

$$\boxed{\hat{\theta}_N = d_j}$$

- it is an unbiased estimator:

$$E\left[\hat{\theta}_N\right] = E\left[d_j\right] = \theta_o$$

- it does not converge in quadratic mean:

$$Var\left[\hat{\theta}_N\right] = E\left[\left(\hat{\theta}_N - E\left[\hat{\theta}_N\right]\right)^2\right] = E\left[\left(d_j - \theta_o\right)^2\right] = Var\left[d_j\right] \leq \sigma^2$$

and then it does not vary with the number N of data

$$\Downarrow$$

the estimation uncertainty is constant and, in particular, it does not decrease when the number of data grows.

Estimator #3 (weighted sample mean):

$$\boxed{\hat{\theta}_N = \sum_{i=1}^{N} \alpha_i d_i}$$

- it is an unbiased estimator if and only if $\sum_{i=1}^{N} \alpha_i = 1$, because

$$E\left[\hat{\theta}_N\right] = E\left[\sum_{i=1}^{N} \alpha_i d_i\right] = \sum_{i=1}^{N} \alpha_i E\left[d_i\right] = \theta_o \sum_{i=1}^{N} \alpha_i = \theta_o \;\Leftrightarrow\; \sum_{i=1}^{N} \alpha_i = 1$$

  Note: the algorithm #1 corresponds to the case $\alpha_i = \frac{1}{N}, \forall i$;

  the algorithm #2 corresponds to the case $\alpha_j = 1$ and $\alpha_i = 0, \forall i \neq j$

- it can be proven that the minimum variance unbiased estimator has weights

$$\alpha_i = \frac{\alpha}{Var\left[d_i\right]}, \quad \alpha = \left[\sum_{i=1}^{N} \frac{1}{Var\left[d_i\right]}\right]^{-1}$$

  intuitively, more uncertain data are considered as less trusted, with lower weights

- the variance of the minimum variance unbiased estimator is

$$
\begin{aligned}
Var\left[\hat{\theta}_N\right] &= E\left[\left(\hat{\theta}_N - E\left[\hat{\theta}_N\right]\right)^2\right] = E\left[\left(\sum_{i=1}^N \alpha_i d_i - \theta_o\right)^2\right] = \\
&= E\left[\left(\sum_{i=1}^N \alpha_i d_i - \sum_{i=1}^N \alpha_i \theta_o\right)^2\right] = E\left[\left(\sum_{i=1}^N \alpha_i \left(d_i - \theta_o\right)\right)^2\right] = \\
&= E\left[\sum_{i=1}^N \alpha_i^2 (d_i - \theta_o)^2 + \sum_{i=1}^N \alpha_i (d_i - \theta_o) \sum_{j=1, j\neq i}^N \alpha_j (d_j - \theta_o)\right] = \\
&= \sum_{i=1}^N \alpha_i^2 E\left[(d_i - \theta_o)^2\right] + \sum_{i=1}^N \alpha_i E\left[(d_i - \theta_o) \sum_{j=1, j\neq i}^N \alpha_j (d_j - \theta_o)\right] = \\
&= \sum_{i=1}^N \alpha_i^2 Var\left[d_i\right] = \sum_{i=1}^N \frac{\alpha^2}{Var[d_i]^2} Var[d_i] = \alpha^2 \sum_{i=1}^N \frac{1}{Var[d_i]} = \\
&= \alpha = \left[\sum_{i=1}^N \frac{1}{Var\left[d_i\right]}\right]^{-1} \leq \left[\sum_{i=1}^N \frac{1}{\sigma^2}\right]^{-1} = \frac{\sigma^2}{N}
\end{aligned}
$$

- the minimum variance unbiased algorithm converges in quadratic mean, since

$$
\lim_{N\to\infty} Var\left[\hat{\theta}_N\right] \leq \lim_{N\to\infty} \frac{\sigma^2}{N} = 0
$$

Maximum Likelihood estimators

The actual data are generated by a random source, which depends on the outcome s of a random experiment and on the "true" value θ_o of the unknown to be estimated. However, if a generic value θ of the unknown parameter is considered, the data can be seen as function of both the value θ and the outcome s $\Rightarrow$

the data can be denoted by $d^{(\theta)}(s)$, with p.d.f. $f(q,\theta)$ that is function of θ too.

Let δ be the particular data observation that corresponds to a particular outcome $\overline{s}$ of the random experiment:

$$\delta = d^{(\theta)}(\overline{s})$$

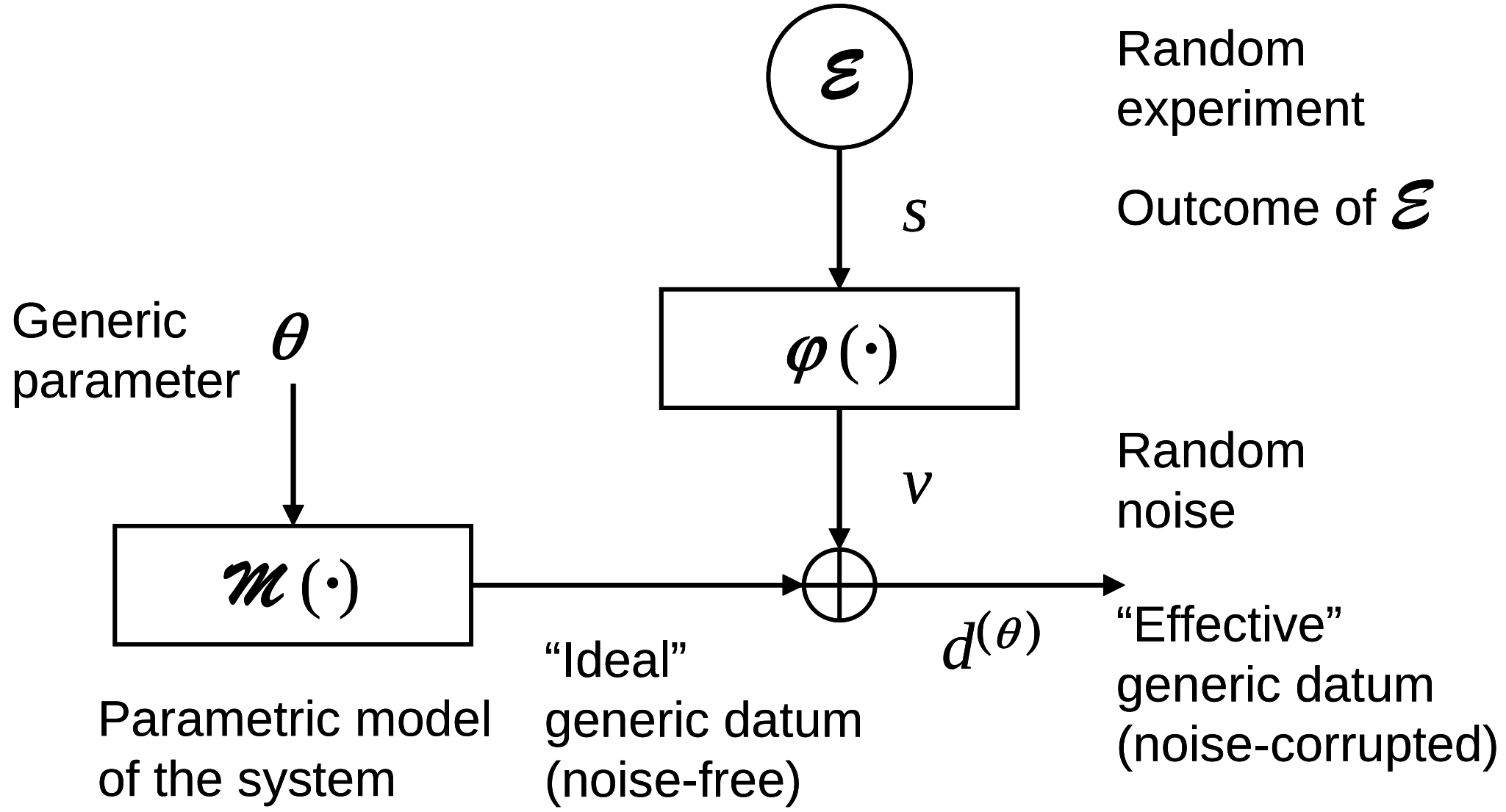
The so-called **likelihood function** is given by the p.d.f. of the data evaluated in δ:

$$L(\theta) = f(q,\theta)\big|_{q=\delta}$$

The **Maximum Likelihood (ML) estimate** is defined as:

$$\hat{\theta}_{ML} = \arg\max_{\theta\in\mathbb{R}^n} L(\theta)$$

Random source of data for a generic value θ of the unknown parameter:



Example: a scalar parameter $\theta_o \in \mathbb{R}$ is estimated using a unique measurement (i.e., $N = 1$), corrupted by a zero-mean Gaussian disturbance with variance σ_v^2

$\Rightarrow$ the random source of data has the following structure:

$$y = \theta_o + v$$

where the noise v is a scalar zero-mean Gaussian random variable with p.d.f.

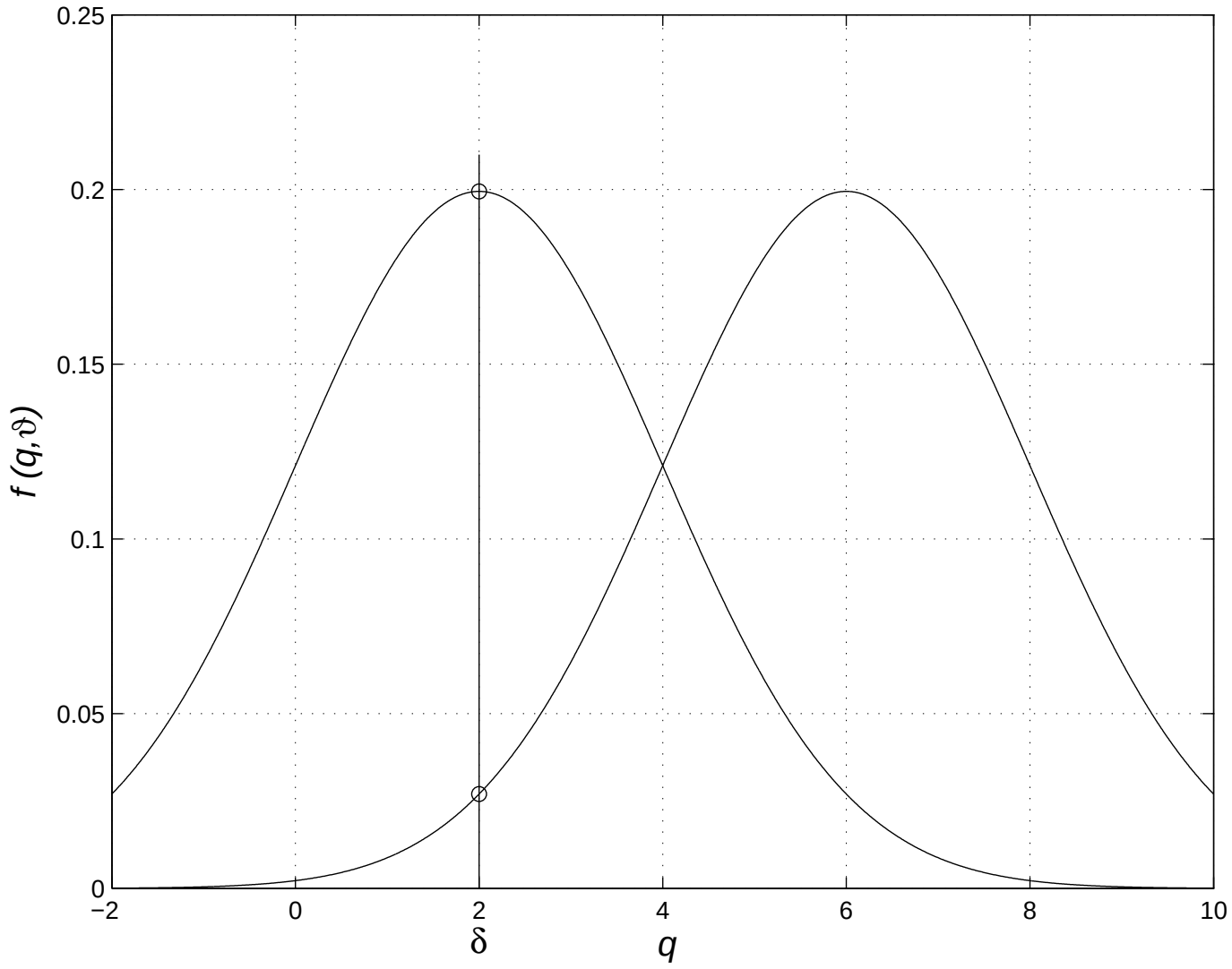
$$f(q) = \mathcal{N}\left(0, \sigma_v^2\right) = \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left(\frac{-q^2}{2\sigma_v^2}\right)$$

Since $v = y - \theta_o \Rightarrow$ the p.d.f. of data y generated by a random source where a generic value θ is considered instead of θ_o is then given by

$$f(q, \theta) = \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left(\frac{-(q-\theta)^2}{2\sigma_v^2}\right) = \mathcal{N}\left(\theta, \sigma_v^2\right) \quad \Rightarrow$$

$$L(\theta) = f(q, \theta)\big|_{q=\delta} = \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left(\frac{-(\delta-\theta)^2}{2\sigma_v^2}\right) = \mathcal{N}\left(\delta, \sigma_v^2\right)$$

$f(q, \theta)$ translates when the value of θ changes $\Rightarrow L(\theta) = f(q, \theta)\big|_{q=\delta}$ varies too.



$$f(q,\theta) = \mathcal{N}\left(\theta, \sigma_v^2\right) \quad \Rightarrow \quad L(\theta) = f(q,\theta)\big|_{q=\delta} = \mathcal{N}\left(\delta, \sigma_v^2\right)$$

Maximum Likelihood estimator properties

The estimate $\hat{\theta}_{ML}$ is:

- asymptotically unbiased: $E\left(\hat{\theta}_{ML}\right) \xrightarrow[N \to \infty]{} \theta_o$
- asymptotically efficient: $\Sigma_{\hat{\theta}_{ML}} \leq \Sigma_{\hat{\theta}}$, $\forall$ unbiased $\hat{\theta} \neq \hat{\theta}_{ML}$, if $N \to \infty$
- consistent: $\lim_{N \to \infty} \Sigma_{\hat{\theta}_{ML}} = 0$
- asymptotically Gaussian (for $N \to \infty$)

Example: let us assume that the random source of data has the following structure:

$$y(t) = \psi(t, \theta_o) + v(t)\,, \quad t = 1, 2, \ldots, N \quad \Leftrightarrow \quad y = \Psi(\theta_o) + v$$

where $\psi(t, \theta_o)$ is a generic *nonlinear* function of θ_o and the disturbance v is a vector of zero-mean Gaussian random variables with variance Σ_v and p.d.f.

$$f(q) = \mathcal{N}(0, \Sigma_v) = \frac{1}{\sqrt{(2\pi)^N \det \Sigma_v}} \exp\left(-\tfrac{1}{2} q^T \Sigma_v^{-1} q\right)$$

Since $v = y - \Psi(\theta_o) \Rightarrow$ the p.d.f. of data generated by a random source where a generic value θ is considered instead of θ_o is then given by

$$f(q, \theta) = \frac{1}{\sqrt{(2\pi)^N \det \Sigma_v}} \exp\left(-\tfrac{1}{2} \left[q - \Psi(\theta)\right]^T \Sigma_v^{-1} \left[q - \Psi(\theta)\right]\right)$$

$$\Downarrow$$

$$L(\theta) = f(q, \theta)\big|_{q=\delta} = \frac{1}{\sqrt{(2\pi)^N \det \Sigma_v}} \exp\left(-\tfrac{1}{2} \left[\delta - \Psi(\theta)\right]^T \Sigma_v^{-1} \left[\delta - \Psi(\theta)\right]\right)$$

$$L(\theta) = f(q,\theta)\big|_{q=\delta} = \frac{1}{\sqrt{(2\pi)^N \det \Sigma_v}} \exp\left(-\tfrac{1}{2}\left[\delta - \Psi(\theta)\right]^T \Sigma_v^{-1} \left[\delta - \Psi(\theta)\right]\right)$$

$$\Downarrow$$

$f(q,\theta)\big|_{q=\delta}$ is an exponential function of θ

$$\Downarrow$$

$$\hat{\theta}_{ML} = \arg\max_{\theta\in\mathbb{R}^n} L(\theta) = \arg\min_{\theta\in\mathbb{R}^n} \underbrace{\left\{\left[\delta - \Psi(\theta)\right]^T \Sigma_v^{-1} \left[\delta - \Psi(\theta)\right]\right\}}_{R(\theta)}$$

Problem: the global minimum of $R(\theta)$ has to be found with respect to θ, but $R(\theta)$ may have many local minima if $\Psi(\theta)$ is a generic nonlinear function of the unknown variable; the standard nonlinear optimization algorithms do not guarantee to find always the global minimum.

Particular case: $\Psi(\theta) =$ *linear* function of the unknown parameters $= \Phi\theta$

$$\Downarrow$$

$R(\theta)$ is a quadratic function of θ : $R(\theta) = [\delta - \Phi\theta]^T \Sigma_v^{-1} [\delta - \Phi\theta]$

$$\Downarrow$$

there exists a unique minimum of $R(\theta)$, if $\det\left(\Phi^T \Sigma_v^{-1} \Phi\right) \neq 0$

$$\Downarrow$$

$$\hat{\theta}_{ML} = \left(\Phi^T \Sigma_v^{-1} \Phi\right)^{-1} \Phi^T \Sigma_v^{-1} \delta = \textbf{Gauss-Markov estimate} = \hat{\theta}_{GM} =$$

$=$ *Weighted Least Squares estimate* using the disturbance variance Σ_v

If $\Sigma_v = \sigma_v^2 I_N$, i.e., independent identically distributed (*i.i.d.)* disturbance:

$$\hat{\theta}_{ML} = \hat{\theta}_{GM} = \left(\Phi^T \Phi\right)^{-1} \Phi^T \delta = \textbf{Least Squares estimate}$$

Gauss-Markov estimate properties

If the disturbance v is Gaussian and $\Psi(\theta)$ is linear, then the estimate $\hat{\theta}_{GM}$ is:

- unbiased: $E\left(\hat{\theta}_{GM}\right) = \theta_o$
- efficient: $\Sigma_{\hat{\theta}_{GM}} = [\Phi^T \Sigma_v^{-1} \Phi]^{-1} \leq \Sigma_{\hat{\theta}}, \quad \forall$ unbiased $\hat{\theta} \neq \hat{\theta}_{GM}$
- consistent: $\lim_{N \to \infty} \Sigma_{\hat{\theta}_{GM}} = 0$
- Gaussian

If the disturbance v is not Gaussian and $\Psi(\theta)$ is linear, then the estimate $\hat{\theta}_{GM}$ is the minimum variance estimator among all unbiased and linear estimators.

- Note that the variance σ_v^2 of the disturbance v is usually unknown $\Rightarrow$

  if the random source of data has the following linear structure

$$y(t) = \varphi(t)^T \theta_o + v(t)\,, \quad t = 1, 2, \ldots, N \quad \Leftrightarrow \quad y = \Phi\theta_o + v$$

  where $v \in \mathbb{R}^N$ is a vector of zero-mean random variables that are uncorrelated and with the same variance σ_v^2 (i.e., $Var[v] = E\left[vv^T\right] = \sigma_v^2 I_N$),

  as in the case of disturbance $v(\cdot)$ given by a white noise $WN(0, \sigma_v^2)$,

  then a "reasonable" unbiased estimate $\hat{\sigma}_v^2$ (such that $E[\hat{\sigma}_v^2] = \sigma_v^2$) can be directly derived from data as

$$\hat{\sigma}_v^2 = \frac{J(\hat{\theta})}{N - n}$$

  where $N =$ measurement number, $n =$ number of unknown parameters of θ,

$$J(\hat{\theta}) = \sum_{t=1}^{N} \varepsilon(t)^2 \Big|_{\theta=\hat{\theta}} = \sum_{t=1}^{N} \left[y(t) - \varphi(t)^T \hat{\theta}\right]^2 = [y - \Phi\hat{\theta}]^T [y - \Phi\hat{\theta}]$$