Intelligent distributed systems

Daniele Fontanelli

Department of Industrial Engineering
University of Trento
E-mail address: *daniele.fontanelli@unitn.it*

2022/2023

UNIVERSITÀ DI TRENTO | Dipartimento di Ingegneria Industriale

Outline

Outline

Stochastic process

Definition (Stochastic process)

A *stochastic process*, aka *random process*, is a collection of random variables representing the evolution of some system of random values.

In practice a *stochastic process* S is a collection of evolving random variables, i.e., $\{s_t, t \in T\}$, where T is an ordered set, in particular T can be the time.
For example, meteorological phenomena such as the random fluctuations in air temperature and air pressure are functions of time.

Stochastic process

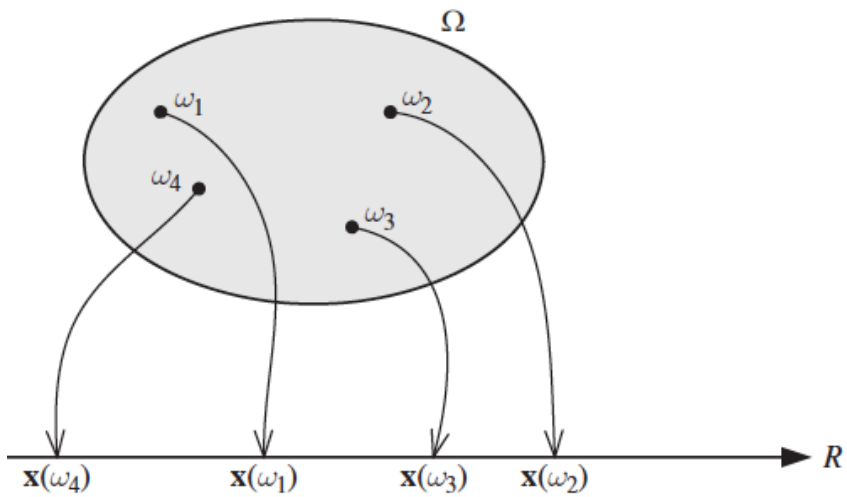

Figure: A random variable can be seen as a mapping from a sample space Ω to a continuous (discrete) set, e.g., the set $\mathbb{R}$ ($\mathbb{Z}$) of real (integer) numbers (Courtesy of Ha H. Nguyen Ed Shwedyk - A First Course in Digital Communications, Cambridge University Press, 2009).

Stochastic process

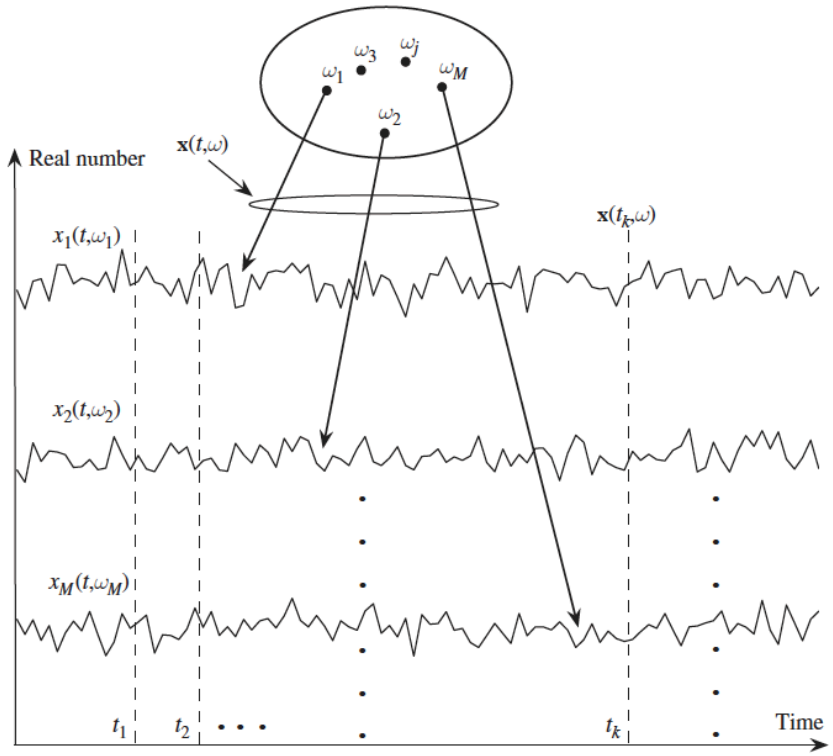

Figure: A stochastic process can be seen as a mapping from a sample space Ω to a set of time varying functions (Courtesy of Ha H. Nguyen Ed Shwedyk - A First Course in Digital Communications, Cambridge University Press, 2009).

Stochastic process

When dealing with *stochastic processes* is not of interest the knowledge of the precise mapping between the event set Ω and the function set.
What instead is of interest is the *ensemble of functions*.
Indeed, let $x(t)$ be the set of functions $x(\omega_1, t), x(\omega_2, t), \ldots x(\omega_n, t)$.
Fixing a specific time, say t_0, we have that $x(t_0)$ is a **rv**: therefore a *stochastic process can be interpreted as a time varying random variable*.

Stochastic process

In particular, given the process $x(\omega, t)$, we have that:

- $x(\omega_j, t)$ is the *sample function*, i.e., one of the possible time varying functions (i.e. a *realisation*);
- $x(\omega, t_i)$ is a **rv**;
- $x(\omega_j, t_i)$ is a *number*.

Stochastic process

Moreover:

- A *discrete stochastic process* describes a process whose **rv** $x(\omega, t_i)$ are discrete;
- A *continuous stochastic process* describes a process whose **rv** $x(\omega, t_i)$ are continuous;
- A *discrete time process* describes a process whose sample functions $x(\omega_j, t)$ are discrete-time;
- A *continuous time process* describes a process whose sample functions $x(\omega_j, t)$ are continuous-time.

Stochastic process

By selecting a set m of different time instants, we have that the **rv**s can be described by a *joint* pdf:

$$x(t_0), x(t_1), \ldots, x(t_m) \sim p_{x(t_0),x(t_1),\ldots,x(t_m)}(x).$$

Usually, the most important pdfs are given by $m = 1$ and 2.
Whatever is the order of the joint pdf we want to compute, we can get an estimate of it using the following frequency:

$$\frac{\text{Number of times the event is verified}}{\text{Number of trials}},$$

i.e., the pdf is estimated with a *histogram*.

Stochastic process

The *mean* of a *continuous stochastic process* for a given t_i is then computed as

$$\mu_{x(t_i)} = \mathsf{E}\left\{x(t_i)\right\} = \int_{-\infty}^{+\infty} \xi p_{x(t_i)}(\xi) d\xi,$$

which is usually referred to as the *ensemble average*.
Of course, for a *discrete stochastic process* for a given t_i is then computed as

$$\mu_{x(t_i)} = \mathsf{E}\left\{x(t_i)\right\} = \sum_j x_j \mathsf{Pr}\left[x(t_i) = x_j\right] = \sum_j x_j \pi_j(t_i).$$

Stochastic process

The *autocorrelation* of a *scalar real-valued continuous stochastic process* for a t_i and t_j is then computed as

$$\mathsf{R}\{t_i, t_j\} = \mathsf{E}\{x(t_i)x(t_j)\} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \xi\eta p_{x(t_i),x(t_j)}(\xi,\eta)d\xi d\eta.$$

The *autocovariance* of a *scalar real-valued continuous stochastic process* for a t_i and t_j is then computed as

$$\mathsf{V}\{t_i, t_j\} \triangleq \mathsf{E}\left\{(x(t_i) - \mu_{x(t_i)})(x(t_j) - \mu_{x(t_j)})\right\} = \mathsf{R}\{t_i, t_j\} - \mu_{x(t_i)}\mu_{x(t_j)}.$$

The extension to *vector-valued stochastic processes* follows the same steps carried out previously for the covariance.

Stochastic process

Notice that the *autocovariance* corresponds to the *variance* if $t_i = t_j = t$. Indeed, for the *autocorrelation*

$$\mathsf{R}\{t,t\} = \mathsf{E}\left\{x(t)^2\right\},$$

while for the standard variance of a **rv** y, we have:

$$\begin{aligned}\mathsf{V}\{y\} = \mathsf{E}\left\{(y-\mu_y)^2\right\} = \mathsf{E}\left\{y^2\right\} + \mathsf{E}\left\{\mu_y^2\right\} - \mathsf{E}\{2y\mu_y\} = \\ = \mathsf{E}\left\{y^2\right\} + \mu_y^2 - 2\mu_y\mathsf{E}\{y\} = \mathsf{E}\left\{y^2\right\} - \mu_y^2.\end{aligned}$$

Therefore

$$\mathsf{V}\{t,t\} = \mathsf{R}\{t,t\} - \mu_{x(t)}\mu_{x(t)} = \mathsf{E}\left\{x(t)^2\right\} - \mu_{x(t)}^2.$$

The term $\mathsf{E}\left\{x(t)^2\right\}$ is called the *mean squared value*.

Stochastic process

Stationarity

Definition (Wide sense stationarity)

A *wide sense stationary process* (often called *stationary process*) is a stochastic process having a *time invariant mean*, i.e., $\mu_{x(t)} = \mu_x$, and the *autocorrelation* $\mathsf{R}\{t_i, t_j\} = \mathsf{R}\{\tau\}$, where $\tau = t_j - t_i$ (hence, $\mathsf{V}\{t_i, t_j\} = \mathsf{R}\{\tau\} - \mu_x = \mathsf{V}\{\tau\}$).

Definition (Strict stationarity)

A *strict stationary process* is a stochastic process having all the pdfs time-invariant with respect to time shift, i.e.,

$$p_{x(t_0),x(t_1),\ldots,x(t_m)}(x) = p_{x(t_0+\tau),x(t_1+\tau),\ldots,x(t_m+\tau)}(x), \forall \tau.$$

In practice, *strict stationarity* (which is almost impossible to be verified in reality) requires that *all the moments* are stationary (not only the first two).

Stochastic process

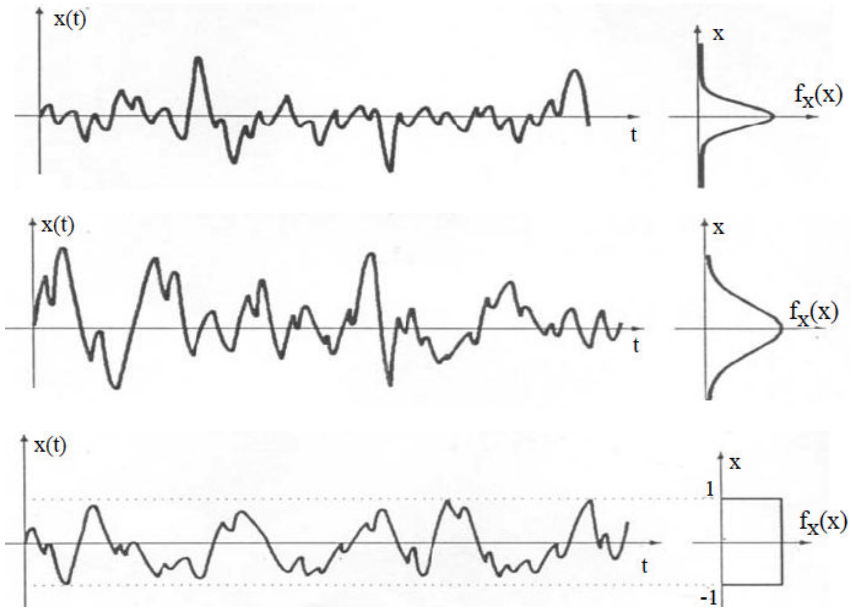

Figure: Three different pdfs $p_{x(t)}$ for three different zero-mean wide sense stationary processes.

Stochastic process

Autocorrelation function properties

The *autocorrelation function* $\mathsf{R}\{\tau\}$ of a *wide-sense stationary process* has several properties:

- $\mathsf{R}\{\tau\} = \mathsf{R}\{-\tau\}$: It is an even function of τ because the same set of product values is averaged across the ensemble, regardless of the direction of translation;
- $|\mathsf{R}\{\tau\}| \leq \mathsf{R}\{0\}$: The maximum always occurs at $\tau = 0$. Further, $\mathsf{R}\{0\}$ is the *mean squared value* of the random process;
- If $\mu_x \neq 0$, then $\mathsf{R}\{\tau\}$ tends asymptotically to μ_x^2;
- The *power spectral density* is the Fourier transform of its *autocorrelation*

$$\mathsf{R}\{\tau\} = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \mathsf{S}(\omega)\, e^{j\omega t} d\omega \leftrightarrow \mathsf{S}(\omega) = \int_{-\infty}^{+\infty} \mathsf{R}\{\tau\}\, e^{-j\omega\tau} d\tau.$$

Stochastic process

Autocorrelation function properties

The *power spectral density*

$$\frac{1}{2\pi}\int_{-\infty}^{+\infty} \mathsf{S}\left(\omega\right) e^{j\omega t} d\omega.$$

has the dimension of a power/hertz, that's why it is called like that.

Noise characterisation

Autocorrelation functions

Let $f(t)$ be a generic function of time, it is possible to define the *autocorrelation function* as

$$\mathrm{R}\{\tau\} = \int_{-\infty}^{+\infty} f(t+\tau) f^*(t) dt = \int_{-\infty}^{+\infty} f(t) f^*(t-\tau) dt,$$

where f^* stands for the complex conjugate of $f(\cdot)$.
In practice it is the *convolution* of a function by itself.
$\mathrm{R}\{\tau\}$ measures the *similarity* of a function by itself in order to detect repeating patterns or similarities.

Noise characterisation

Autocorrelation functions

If $f(k)$ is a generic discrete time function, it is possible to redefine the *autocorrelation function* with

$$\mathrm{R}\{k\} = \sum_{i \in \mathbb{Z}} f(i) f^*(i-k).$$

This definition applies to *finite energy* and *square summable* signals.

Noise characterisation

Cross-correlation functions

Assuming that $f(t)$ and $g(t)$ are two real functions of time, it is possible to generalise the *autocorrelation function* to the *cross-correlation*:

$$\mathrm{R}\{\tau\}_{fg} = \int_{-\infty}^{+\infty} f(t+\tau)g(t)dt = \int_{-\infty}^{+\infty} f(t)g(t+\tau)dt,$$

which is a measure of the similarity of the two functions $f(t)$ and $g(t)$.

Stochastic process

Ergodicity

Theorem (Ergodicity)

For an ergodic stationary process, the time average of some sample functions of the process is the same as the ensemble average.

Hence, for any given realisation of the stochastic process, the *time average* are equal to the *ensemble average with probability 1*.
Therefore, any realisation of the stochastic process gives a good estimate of the process *statistical description*.

Stochastic process

Average

It follows that for an *ergodic process*

$$\mathsf{E}\left\{x^r(t)\right\} = \int_{-\infty}^{+\infty} x^r(t) p_{x(t)}(x) dx = \lim_{\tau \to +\infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} x^r(\tau, \omega_i) d\tau, \forall i$$

where the factor $1/2\tau$ is adopted for normalisation.
Again the process is *wide sense ergodic* if the previous relation holds only for the *mean* and *autocorrelation* $(r = 1, 2)$.
If it holds $\forall r$, the process is *strictly ergodic*.

Stochastic process

Ergodicity

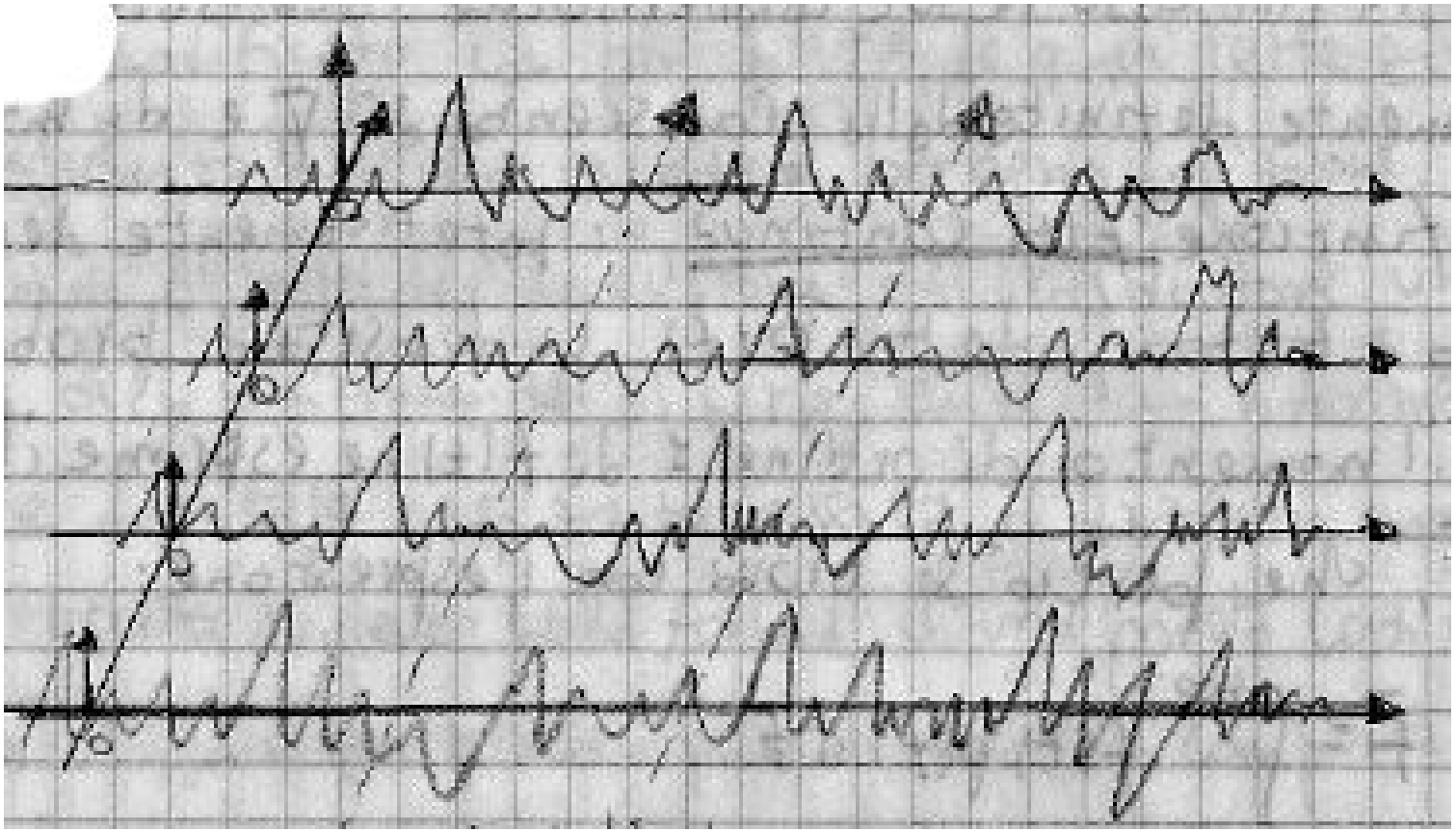

Figure: Ergodicity property.

Outline

Noise characterisation

Definition (White noise)

A (not necessarily stationary) stochastic process having the *autocovariance* null for any two different times is called a *white noise*.

In such a case:

$$\mathrm{V}\{t_i, t_j\} = \sigma^2(t_i)\delta(t_i - t_j)$$

where $\sigma^2(t_i)$ is the *instantaneous variance*.
Recall that if the *covariance* of two **rv**s is null, than the **rv**s are *uncorrelated*. It then follows that the time *uncorrelatedness* defines a *wide-sense whiteness* property.
There is also a *strict-sense whiteness* property that requests for the time *independence* rather than the time *uncorrelatedness* of the stochastic process.

Noise characterisation

By the definition of the *autocovariance* it follows that a *stationary zero-mean white* process has

$$\mathsf{R}\{\tau\} = \mathsf{E}\{x(t+\tau)x(t)\} = \sigma^2\delta(\tau)$$

It then follows that for a *stationary zero-mean white* process, the *power spectral density* is

$$\mathsf{S}(\omega) = \sigma^2,$$

hence constant across the frequency spectrum.
For *non-stationary zero-mean white* processes we have

$$\mathsf{V}\{t_i, t_j\} = \mathsf{R}\{t_i, t_j\} = \mathsf{E}\{x(t_i)x(t_j)\} = \sigma^2(t_i)\delta(t_i - t_j)$$

where the meaning of $\sigma^2(t_i)$ *is not* instantaneous variance but *time-varying spectral density*.

White noise

Remark

In general, the results related to the theory of estimators requests the strict-sense stationarity. However, this condition is hardly verified in practice, since only the first two moments of a stochastic process are available. Therefore, the estimators are applied to wide-sense stationarity, hence the results hold only approximately.

Definition

An *independent and identically distributed* (i.i.d.) random sequence corresponds to a stationary and white stochastic process.

Noise characterisation

Extension to vectors

If the stochastic process defines vectors, i.e. $x(t) \in \mathbb{R}^n$, we have immediately for a generic *stochastic process* that

$$\mathsf{R}\{t_i, t_j\} = \mathsf{E}\left\{x(t_i)x(t_j)^T\right\} = Q(t_i, t_j)$$

and

$$\mathsf{V}\{t_i, t_j\} = \mathsf{R}\{t_i, t_j\} = \mathsf{E}\left\{(x(t_i) - \mu_{x(t_i)})(x(t_j) - \mu_{x(t_j)})^T\right\} = Q(t_i, t_j).$$

It then follows that for a *zero-mean white* stochastic process

$$\mathsf{R}\{t_i, t_j\} = \mathsf{V}\{t_i, t_j\} = Q(t_i)\delta(t_i - t_j),$$

and, for a *stationary zero-mean white* process

$$\mathsf{R}\{t_i, t_j\} = \mathsf{V}\{t_i, t_j\} = Q\delta(t_i - t_j) = Q\delta(\tau).$$

Outline

Markovian processes

The *Markov processes* are defined by the following *Markov property*:

$$p(x(t)|x(\tau), \tau \leq t_1) = p(x(t)|x(t_1)), \forall t > t_1$$

So, the past up to any t_1 is *fully characterised* by the value of the process at time t_1.
To state it with the words of Bar-Shalom: *the future is independent from the past if the present is known*.
This is very important, since the state of a possibly *time varying dynamic system* driven by *white noise* $\nu(t)$, i.e.,

$$\dot{x}(t) = f(x(t), \nu(t), t),$$

is indeed a *Markov process*.

Markovian processes

Example: Wiener stochastic process

As an example, let us consider the *Wiener stochastic process*.
Being $\nu(t)$ a *zero-mean white noise*, the *Wiener stochastic process* is given by

$$w(t) = \int_0^t \nu(\tau) d\tau,$$

which is a form of the *random walk*.
An alternative way of writing it is

$$dw(t) = \nu(t) dt,$$

hence an *independent increment process*.

Markovian processes

Example: Wiener stochastic process

Formally, the *white noise* is the *Wiener stochastic process* derivative.
However, this is not rigorous, since the Wiener process is *nowhere differentiable*, otherwise the variance of the derivative would be unlimited.
Nevertheless, the *Wiener stochastic process* is Markovian, indeed

$$w(t) = \int_0^t \nu(\tau)d\tau = w(t_1) + \int_{t_1}^t \nu(\tau)d\tau,$$

and $\nu(\tau)$ is *uncorrelated* from $w(t_1)$.
The *Wiener stochastic process* and the *white noise* are used to *model unknown inputs* for linear systems.

Markovian processes

Prewhitening

The *autocorrelation function* of the input is in general a matrix given by

$$\mathsf{R}\{t_1, t_2\} = \mathsf{E}\left\{\nu(t_1)\nu(t_2)^T\right\} = Q\delta(t_1 - t_2),$$

where Q is the *power spectral density* of the white noise input, usually (and improperly) referred to as the *covariance matrix* of $\nu(t)$.
It then follows that:

Remark

Every Markov process with a rational spectrum can be represented as a linear time-invariant system excited by white noise.

Markovian processes

Prewhitening

This is of paramount importance in our set-up. Indeed, given a system that is driven by a *stationary and not white* noise $\eta(t)$

$$\dot{x}(t) = Ax(t) + B\eta(t).$$

Since $\eta(t)$ is not white, the state $x(t)$ is *not* a *Markov process*. However, if $\eta(t)$ has a *rational spectrum*, it can be considered as the output of a linear system driven by *white noise* $\nu(t)$, i.e.,

$$\begin{aligned}\dot{z}(t) &= A_z z(t) + B_z \nu(t)\\ \eta(t) &= C_z z(t)\end{aligned}$$

Basically in this operation we consider the *stationary and not white* noise $\eta(t)$ as a Markov process. This operation is called the *prewhitening*.

Markovian processes

Prewhitening

The system A_z, B_z, C_z is called the *prewhitening system* or *shaping filter*.
In the *prewhitening* we consider an overall system driven by the *white noise* $\nu(t)$ and comprising the series of two linear systems.
Notice that the overall system state

$$y(t) = \begin{bmatrix} x(t) \\ z(t) \end{bmatrix}$$

is now a *Markovian state*.

Remark

A non white noise is also called an autocorrelated noise or a coloured noise.

Markovian processes

Discrete time

A *discrete-time stochastic process* is a time indexed sequence of random variables:

$$X_k = \{x(i)\}_{i=1}^{k}$$

Similarly to the continuous time, a *random sequence is Markovian* if

$$\Pr\left[x(k)|X_j\right] = \Pr\left[x(k)|x(j)\right], \forall k > j.$$

The *zero-mean sequence* $\nu(j)$, $j = 1, \ldots$ is a *discrete-time white noise sequence* if

$$\mathrm{E}\left\{v(k)v(j)\right\} = Q(k)\delta_{jk},$$

where δ_{jk} is the *Kronecker delta function* and $Q(k)$ is the *covariance matrix*.

Markovian processes

Discrete time

If the discrete-time process generates a *stationary white sequence*, $Q(k) = Q$ is the time invariant covariance matrix.
A *discrete-time Markov process* is given by a time invariant system as

$$x(k+1) = f(k, x(k), \nu(k)).$$

The state of a linear time-invariant system excited by a *white Gaussian noise*

$$x(k+1) = Ax(k) + \nu(k),$$

is a *Gauss-Markov sequence*.
If the system is scalar with $A = 1$

$$x(k+1) = x(k) + \nu(k),$$

hence $x(k)$ is the integral of the $\nu(k)$ and, hence, it is called a *discrete-time Wiener process*.

Markovian processes

Discrete time

A *Markov chain* is a special case of the *Markov sequence* in which the state is *discrete and finite*.
A *Markov chain* is fully characterised by the *transition probabilities*

$$\Pr\left[x(k) = x_i | x(k-1) = x_j\right] \triangleq p_{ij}$$

Markovian processes

Discrete time

Therefore the probability of being in the state x_i at time k, i.e., π_i, given the probabilities of being in the state x_j at time $k-1$ is then given by

$$\pi_i = \sum_{j=1}^{n} p_{ij}\pi_j.$$

By collecting all the probabilities in a single row vector $\pi = [\pi_1, \pi_2, \ldots, \pi_n]$, one has

$$\pi = \pi P,$$

where $P = (p_{ij})_{ij}$ is the *transition probability matrix*.

Outline

Fundaments of estimation

Let us recall the already introduced concepts about *estimators*.
Let us assume that we want to *estimate* (that is to *increase our knowledge*) about a certain quantity x.
A *measurement process* can be formally defined as

$$z = h(x, w, t),$$

where $h(\cdot)$ can represent either a *direct* measurement process, e.g. the thermometer measuring the temperature, or an *indirect* measurement, e.g. the velocity inferred from position measurements.
The data $z \in \mathbb{R}^m$ are the *measurements*.
The vector $w \in \mathbb{R}^m$ is the *measurement noise*.
t expresses the *time variability* of the measurement process.

Fundaments of estimation

In more strict terms, the problem of estimation of x given the set of measures

$$z(j) = h(j, x, w(j)), \quad j = 1, \ldots, k,$$

where $w(j)$ is a sequence of measurement noises (or *disturbances*) amounts to find the *estimates*

$$\hat{x}(k) = \hat{x}[k, Z^k],$$

being Z^k the set of all measures up to time k, i.e.,

$$Z^k = \{z(j)\}_{j=1,\ldots,k}.$$

The set of measurements is usually referred to as a *time series*, which is a *discrete time process*.
While $\hat{x}(k)$ are the *estimates* of x, the function $\hat{x}[k, Z^k]$ is the *estimator*.

Fundaments of estimation

For the *measurement model*, we usually adopt a *forward map*

$$z(k) = h(x(k)).$$

The model in the easiest case is linear, i.e.

$$z(k) = Hx(k).$$

If this is not the case, usually a *Taylor approximation* is employed.

Fundaments of estimation

Usually in the linear model $H \in \mathbb{R}^{m \times n}$. The objective of the *estimator* is to retrieve $x \in \mathbb{R}^n$ from $z \in \mathbb{R}^m$.

- If $n < m$ the problem is *over-determined* and the problem is usually solvable in the linear and nonlinear cases;
- If $n = m$ the problem is more involved since it may happens that it does not have solutions for the linear and the nonlinear cases;
- If $n > m$ the problem is not solvable unless *prior knowledge* is available.

Fundaments of estimation

In a very simple linear case, the *measurement error* w is the error affecting the measures at time k

$$z(k) = x(k) + w(k),$$

where $w(k)$ is a *random variable*.
The *estimation error* at time k is instead given by

$$\tilde{x}(k) = x(k) - \hat{x}(k),$$

where $\hat{x}(k)$ is a function of the measurements $z(k)$, $\forall k$.

Fundaments of estimation

To design an *estimator* we resort to the following description:

- The role of an estimator is to retrieve a *correct estimate* $\hat{x}(k)$, possibly the *best estimate*, i.e. the one with the smallest *estimation error* $\tilde{x}$;
- The information available are: a) the *measurements* $z(k)$, which are *noisy*; b) an idea about the *system output model*, i.e. the function $h(\cdot)$; c) maybe, some *additional knowledge*, e.g. the quantity is always positive;
- *All* the available knowledge *must be exploited*, so the estimator needs to incorporate it properly;
- We can treat or not the quantity to estimate as a *random variable*.

Fundaments of estimation

Example

To design an estimator, let us first model the data, i.e. the *measurements*. Since, the *time series*

$$Z^k = \{z(j)\}_{j=1,\dots,k}$$

is *inherently random*, it can be described by a *joint pdf*, which is *parametrised* by the unknown parameter x, i.e.

$$p(z(1), \dots, z(k); x).$$

For a single measure

$$p(z(1); x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z(1)-x)^2}{2\sigma^2}}.$$

The choice of the modelling pdf turns to be *very critical* for the estimator design.
Let us make an example: three possible values of $x \in \{-10, 10, 20\}$, with $\sigma = 2$. What is the *most probable value* if $z(1) < 0$?

Fundaments of estimation

Example

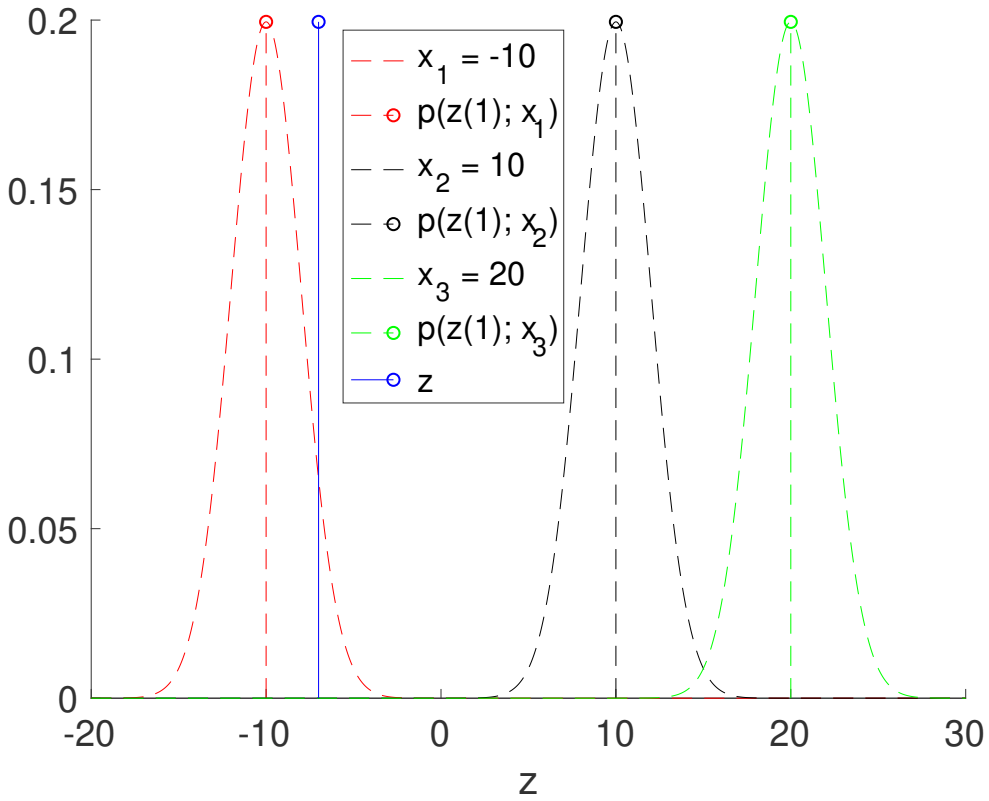

Figure: Three different pdfs for three different possible values of x, i.e. $x \in \{-10, 10, 20\}$, with $\sigma = 2$ and Gaussian pdf. In this case, $z(1) = \overline{z} = -7$. What is the most probable value of x?

Fundaments of estimation

The likelihood

But there is also an alternative representation of this information.
The pdf $\Pr[z(1) < z < z(1) + \delta_z] \approx p(z(1); x)dz(1)$ gives the *probability of observing* $z(1)$ in a small area for a given x.
As an example, we can plot the pdfs $p(z(1); x)$ for a *given* $z(1) = \overline{z} < 0$ considering x *variable*, i.e. $x \in \mathbb{R}$.
In practice, the value of $p(z(1) = \overline{z}; x)dz(1)$ for each possible x tell us the *probability of observing* $z(1)$ *in a region* $\mathbb{R}$ *centred around* $\overline{z}$ *with are* $dz(1)$ *assuming the given value of* x.
This is the *likelihood function*.

Fundaments of estimation

The Likelihood

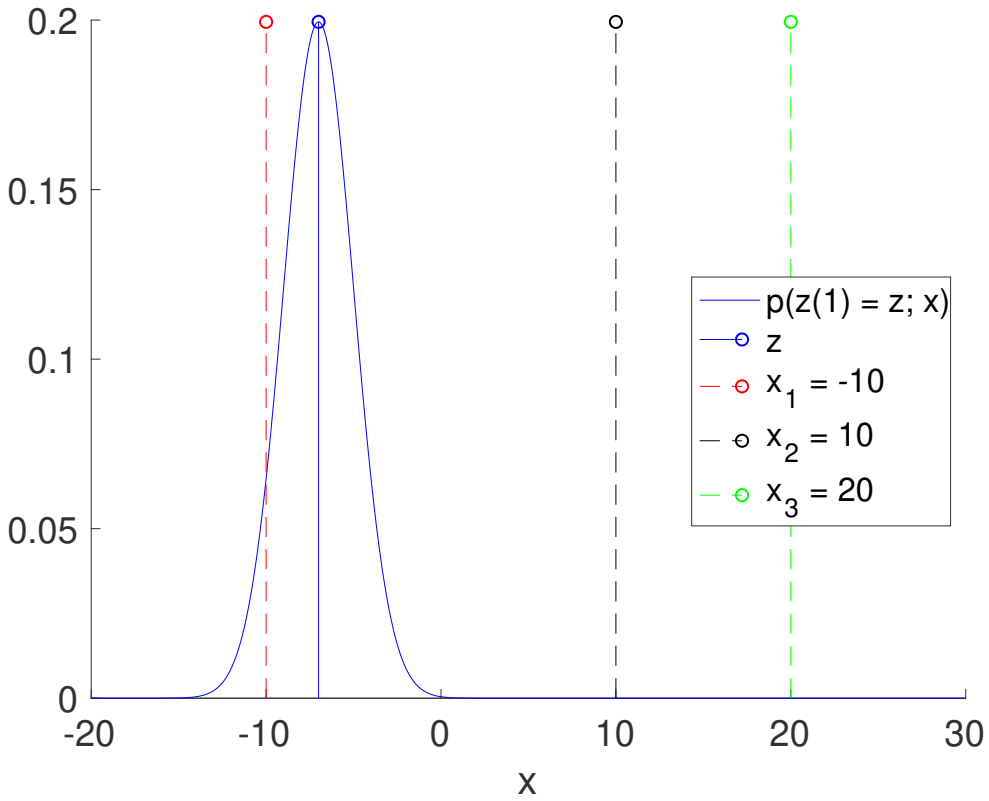

Figure: A likelihood function for the problem at hand, with $x \in \{-10, 10, 20\}$, $\sigma = 2$ and Gaussian pdf. In this case, $z(1) = \overline{z} = -7$. What is the most probable value of x?

Fundaments of estimation

The likelihood

Observing $z(1) = \overline{z}$ it is quite unlikely that $x = 10$ or $x = 20$. Indeed, in those cases *the probability of observing that measurement would be very low*!

On the other hand, it is more *likely* that we have observed $x = -10$, i.e. *assuming that $x = -10$, we have a higher probability that we actually observe $z(1) = \overline{z}$*.

As a consequence, our estimate would be $\hat{x} = -10$!

This quite immediate idea carries a lot of information with it. Indeed, we may actually try to estimate x by *maximising the likelihood* function, i.e. by choosing the estimate $\hat{x}$ that give the *highest probability of having a measurement $z(1) = \overline{z}$*.

This estimator is the *Maximum Likelihood* estimator that we will see in a while in a more general framework.

Unbiasedness

Definition

An estimator is *unbiased* if, on average, it converges to the true value x, i.e. if $\mathsf{E}\{\hat{x}\} = x$, *for all* the possible values of x in its domain.

Alternatively, we can say that since $\hat{x} = \hat{x}[k, Z^k]$, we have

$$\mathsf{E}\{\hat{x}\} = \int \hat{x}[k, Z^k] p(Z^k; x) dZ^k = x.$$

Remark

A biased estimator is affected by a systematic error, i.e. a bias $E\{\hat{x}\} - x$.

Mean Square Error

A natural way to synthesise an estimator is to use an *optimal criterium*. For example, one may try to minimise the *Mean Square Error* (MSE) that we have already introduced:

$$\mathsf{mse}(\hat{x}) = \mathsf{E}\left\{(\hat{x} - x)^2\right\}.$$

However, for a *biased estimator* with bias $b(x) = \mathsf{E}\{\hat{x}\} - x$, we have

$$\mathsf{mse}(\hat{x}) = \mathsf{E}\left\{((\hat{x} - \mathsf{E}\{\hat{x}\}) + (\mathsf{E}\{\hat{x}\} - x))^2\right\} = \mathsf{V}\{\hat{x}\} + b(x)^2.$$

Mean Square Error

Hence, to minimise the MSE, it is necessary to use the *bias* $b(x)$, which is usually *unknown* since it depends on the *actual value* of x, which is *not available*.
Therefore, synthesise an estimator that *minimises the MSE* is in general *not possible*!

MVUE

If the estimator is instead *unbiased*, as we noticed for the arithmetic mean, the MSE turns to be the *variance*.
Therefore, minimising the MSE corresponds to minimise the *variance* and the estimator is then called *Minimum Variance Unbiased Estimator* (MVUE).
In general the MVUE *does not* always exists. Indeed, it needs to be based on an *unbiased* estimator with *minimum variance* but *for all the possible values* of x.

Remark

Notice that for multidimensional parameters $x \in \mathbb{R}^n$, the MVUE has the property that $V\{x_i\}$ is minimum.

Summary

- We are interested in *unbiased* estimators. this property *must* hold for any value of the parameter x;
- The *unbiased* estimator with the *least variance* is in general chosen;
- The *Minimum Mean Square Error* (MMSE) leads in general to *unrealisable estimators*;
- *Minimum variance unbiased estimators* (MVUE) do not in general exists;
- When an MVUE exists, it can be found using the *Cramer-Rao lower bound* (CRLB) or the *Rao-Blackwell-Lehmann-Scheffe* theorem;
- When an MVUE does not exist, an approximation of an estimator that is *linear in the measurements* can be derived.

CRLB

In other words,

MVUE

Linear case

Theorem ((MVUE for the Linear Models))

Assuming a linear model $Z = Hx + w$ for the observations, where $Z \in \mathbb{R}^k$, $x \in \mathbb{R}^n$, $H \in \mathbb{R}^{k \times n}$ and $w \in \mathbb{R}^k$ with $w \sim \mathcal{N}(0, R)$, then the MVUE is

$$\hat{x} = (H^T R^{-1} H)^{-1} H^T R^{-1} Z,$$

with covariance matrix

$$C\{\hat{x}\} = I^{-1}(x) = (H^T R^{-1} H)^{-1}.$$

For the linear model, the MVUE is efficient and attains the CRLB.

MVUE

Linear case

Remark

A measure of the information an estimator of a quantity x uses is called Fisher information $I(x)$.

The *Fisher information* has the properties of the *information measure*: a) it is non negative; b) it is additive for independent measures.

Remark

When the CRLB is attained, the Fisher information is the reciprocal of the variance: more information, less variance.

For the previous linear MVUE, the *covariance matrix*

$$\mathsf{C}\{\hat{x}\} = I^{-1}(x) = (H^T R^{-1} H)^{-1}.$$

Outline

BLUE

In general, even if the pdf is known and the CRLB available, it is not given for granted that an MVUE could be designed. In those cases, it is needed to use a *suboptimal estimator*.
Usually, we can restrict to estimators that are *linear in the measurements* and hence find the *Best Linear Unbiased Estimator* (BLUE).
The BLUE can be determined knowing only the *the first two moments* of the pdf.

BLUE

Gauss-Markov

Theorem

If the measurements are linear, i.e. $Z^k = Hx + w$ *where* w *is zero-mean and covariance* R*, the BLUE is given by*

$$\hat{x} = (H^T R^{-1} H)^{-1} H^T R^{-1} Z^k,$$

with minimum variance

$$C\{\hat{x}\} = (H^T R^{-1} H)^{-1}.$$

This is the Gauss-Markov theorem.

BLUE

Example

A few remarks are now in order:

- The *arithmetic mean* is the BLUE for *any* pdf;
- If the noise sequence w is Gaussian, than the BLUE is also the MVUE;
- If the noise sequence w *is not* Gaussian, this is no longer true and the BLUE is clearly a *sub-optimal* solution

Consistency

Definition

The estimator $\hat{x}[k, Z^k]$ of the quantity x is *consistent* if:

$$\lim_{k \to +\infty} \Pr\left[|\hat{x}[k, Z^k] - x| \geq \varepsilon\right] = 0, \forall \varepsilon > 0.$$

For example, the arithmetic mean is consistent since the MSE is $\frac{\sigma^2}{k}$.

Outline

Fundaments of estimation

Linear trend immersed in noise

Suppose we have a *time series* representing the time evolution of the price of the oil on the market, which shows a *linear trend on average*.
We can assume that a valid model is

$$z(j) = a + bj + w(j), \quad j = 1, \ldots, k.$$

To have a model tractable we assume that $w(j) \sim \mathcal{N}(0, \sigma^2)$ and that $\mathsf{E}\{w(j)w(i)\} = \sigma^2 \delta_{ji}$.
Defining $x = [a, b]^T$, we finally have

$$p(Z^k; x) = \frac{1}{\sqrt{(2\pi\sigma^2)^k}} e^{-\frac{\sum_{j=1}^{k}(z(j)-a-bj)^2}{2\sigma^2}}.$$

Fundaments of estimation

Of course, the choice of the pdf is *critical* for the estimator design, which relies exactly on the chosen model.
In general, we hope that the estimator is *robust*, i.e. changes in the pdf *do not affect* that much the estimates. Otherwise, *robust statistical procedures* should be adopted.
In the *classical estimation* theory the parameter(s) x is assumed *unknown but constant*.
If instead, we know for the previous example that $a \in [\underline{a}, \overline{a}]$, we can incorporate this knowledge with a *stochastic prior* $p(x)$, thus having a *joint pdf*

$$p(Z^k, x) = p(Z^k|x)p(x),$$

and a *Bayesian estimator*.
Usually, when the parameter(s) x is assumed *unknown but constant*, the approaches of the *classical estimation* are referred to as *non Bayesian*.

Fundaments of estimation

The philosophical meaning to the prior

The prior pdf assumed in the problem is the *subjective assessment of the phenomenon*.
If the prior is *uniform*, we assume that the Nature is *indifferent*.
In game theory, the Nature is assumed to *play against our purposes* (which is the basis of the H_∞ filter design).
None of the previous assumptions are strictly correct, however there is always the *principle of perversity of inanimate objects*. According to Richard Bellman, this fact is proved by a set of experiments.

Example

If you drop a piece of buttered toast on a rug, you have 79.3% possibilities that the toast fell buttered face down. There is also a mathematical proof...

Fundaments of estimation

The *prior knowledge* can be given by the pdf $p_x(x)$.
For example if x is in a certain interval, we can define

$$p_x(x) \sim \mathcal{U}(x_{\mathsf{min}}, x_{\mathsf{max}}).$$

If x has some typical values

$$p_x(x) \sim \mathcal{N}(\mu_x, \sigma_x^2).$$

Outline

Fundaments of estimation

Non-Bayesian approach

If no prior $p(x)$ is available, no *posterior* pdf can be defined.
However, we can define the *pdf of the measurements conditioned on the parameter* x, which is called the *likelihood function* (LF) of x and it is given by

$$\Lambda_k(x) = p(Z^k|x).$$

The *likelihood function*, that measures how "likely" a parameter value is given the observations, measure the *evidence from the data*.

Fundaments of estimation

Non-Bayesian estimator: Maximum Likelihood

The *Maximum Likelihood* (ML) estimator maximises the LF, i.e.,

$$\hat{x}^{ML}[k, Z^k] = \arg\max_{\hat{x}} \Lambda_k(\hat{x}) = \arg\max_{\hat{x}} p(Z^k|\hat{x}).$$

Notice that since $\hat{x}^{ML}[k, Z^k]$ is a function of the random variables Z^k, it is a **rv** even if the parameter x is not.
The solution of the ML is given by

$$\frac{d\Lambda_k(\hat{x})}{d\hat{x}} = \frac{dp(Z^k|\hat{x})}{d\hat{x}} = 0.$$

Fundaments of estimation

Bayesian approach

As pointed out in the previous slides, the *Bayesian* estimator computes the *posterior* $p(\hat{x}|Z^k)$ given the *prior* $p(\hat{x})$ and the *likelihood function* $p(Z^k|\hat{x})$, i.e.,

$$p(\hat{x}|Z^k) = \frac{p(Z^k|\hat{x})p(\hat{x})}{p(Z^k)} = cp(Z^k|\hat{x})p(\hat{x}),$$

where c is the normalisation constant computed in the denominator, which is not a function of $\hat{x}$.
In practice, the *Bayesian* estimators apply the *Bayes' Theorem*.
Once the *posterior* pdf is known, several different algorithms can be applied.

Fundaments of estimation

Bayesian estimator: Maximum A Posteriori

An example is the *Maximum A Posteriori* (MAP) that computes the maximum of the *posterior* pdf

$$\hat{x}^{MAP}[k, Z^k] = \arg\max_{\hat{x}} p(\hat{x}|Z^k) = \arg\max_{\hat{x}} p(Z^k|\hat{x})p(\hat{x}),$$

where the normalisation constant has been removed since it does not depend on $\hat{x}$.
Obviously, the MAP estimator $\hat{x}^{MAP}[k, Z^k]$ returns a value that is the realisation of a random variable.
WARNING: the MAP estimates are *point estimates*, whereas Bayesian methods are characterised by the use of distributions to summarise data and draw inferences, hence the MAP is *marginally representative* of the potentiality of the *Bayesian inference*. Take a look to the next figure...

Fundaments of estimation

Bayesian estimator: Maximum A Posteriori

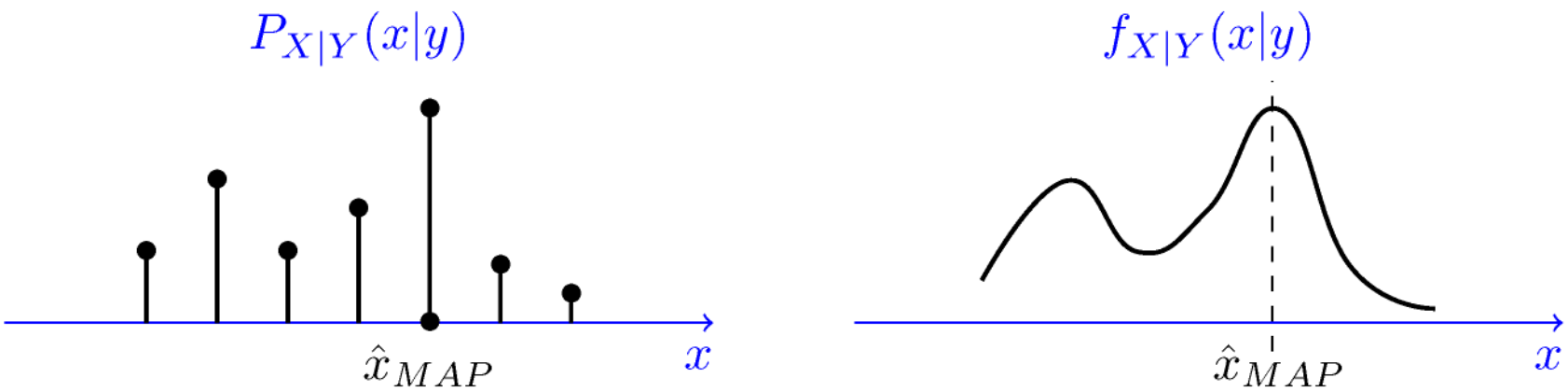

Figure: The work of the MAP estimator: the estimate is the *mode* of the posterior $p(x|z)$.

Fundaments of estimation

ML vs MAP

Let us make an example: consider a single measurement:

$$z = x + w,$$

and suppose that x is an unknown parameter and that $w \sim \mathcal{N}(0, \sigma^2)$. Therefore, we know that z is a Gaussian **rv**, characterised by knowing only the first two moments, i.e.

$$\mathsf{E}\{z\} = \mathsf{E}\{x + w\} = x + \mathsf{E}\{w\} = x,$$

and

$$\mathsf{V}\{z\} = \mathsf{E}\left\{(z - \mathsf{E}\{z\})^2\right\} = \mathsf{E}\left\{(x + w - x)^2\right\} = \mathsf{E}\left\{w^2\right\} = \sigma^2.$$

As a consequence $z \sim \mathcal{N}(x, \sigma^2)$.

Fundaments of estimation

ML vs MAP

First assume that no prior is given.
Hence, for the ML estimator, the *likelihood function* will be given by

$$\Lambda(\hat{x}) = p(z|\hat{x}) = \mathcal{N}(\hat{x}, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-\hat{x})^2}{2\sigma^2}}.$$

Therefore,

$$\hat{x}^{ML} = \arg\max_{\hat{x}} \Lambda(\hat{x}) = \arg\min_{x} (z - \hat{x})^2 = z.$$

Fundaments of estimation

ML vs MAP

This idea can be generalised to a set of *i.i.d.*, e.g. *independent and identically distributed*, measures Z^k, each normally distributed with the likelihood function $\Lambda(\hat{x}) = \mathcal{N}(\hat{x}, \sigma^2)$.
In this case, we have

$$p(Z^k|\hat{x}) = p(z(1), \ldots, z(k)|\hat{x}) = \\ = \prod_{j=1}^{k} \mathcal{N}(\hat{x}, \sigma^2) = c'' e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}(z(i)-\hat{x})^2}$$

Therefore,

$$\hat{x}^{ML} = \arg\max_{\hat{x}} \Lambda(\hat{x}) = \arg\min_{x} \sum_{i=1}^{k}(z(i) - \hat{x})^2 = \frac{1}{k}\sum_{i=1}^{k} z(i).$$

Fundaments of estimation

ML vs MAP

Moreover, it they are not *i.i.d.* but still *uncorrelated* (i.e. *independent*, since we are considering Gaussian **rv**s), measures Z^k, each normally distributed with likelihood function $\Lambda_i(\hat{x}) = \mathcal{N}(\hat{x}, \sigma_i^2)$.
In this case, we have

$$p(Z^k|\hat{x}) = p(z(1), \ldots, z(k)|\hat{x}) = \\ = \prod_{j=1}^{k} \mathcal{N}(\hat{x}, \sigma_j^2) = c'' e^{-\frac{1}{2}\sum_{i=1}^{k} \frac{(z(i)-\hat{x})^2}{\sigma_i^2}}$$

Therefore,

$$\hat{x}^{ML} = \arg\max_{\hat{x}} \Lambda(\hat{x}) = \arg\min_{x} \sum_{i=1}^{k} \frac{(z(i)-\hat{x})^2}{\sigma_i^2} = \frac{\sum_{i=1}^{k} \frac{z(i)}{\sigma_i^2}}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}},$$

i.e. a *weighted arithmetic mean*.

Fundaments of estimation

ML vs MAP

Suppose now that an a priori information is given, i.e., $\hat{x} \sim \mathcal{N}(\mu_x, \sigma_x^2)$, where $\hat{x}$ is *independent* form w.
Hence the posterior is given by

$$p(\hat{x}|z) = c' p(z|\hat{x}) p(\hat{x}) = c e^{-\frac{(z-\hat{x})^2}{2\sigma^2} - \frac{(\hat{x}-\mu_x)^2}{2\sigma_x^2}}.$$

Notice that

$$\hat{x}^{MAP}(z) = \arg\max_{\hat{x}} p(\hat{x}|z) = \arg\min_{\hat{x}} -\frac{(z-\hat{x})^2}{2\sigma^2} - \frac{(\hat{x}-\mu_x)^2}{2\sigma_x^2}.$$

It immediately follows that

$$\hat{x}^{MAP}(z) = \frac{\sigma^2}{\sigma^2+\sigma_x^2}\mu_x + \frac{\sigma_x^2}{\sigma^2+\sigma_x^2} z.$$

Fundaments of estimation

ML vs MAP

Since the conditional pdf of a Gaussian distribution *is* a Gaussian distribution, it follows immediately that

$$p(\hat{x}|z) = \mathcal{N}(\eta_{x|z}, \sigma^2_{x|z}) = \frac{1}{\sqrt{2\pi}\sigma_{x|z}} e^{-\frac{(\hat{x}-\eta_{x|z})^2}{2\sigma^2_{x|z}}}.$$

In this case, we have obviously that

$$\hat{x}^{MAP}(z) = \arg\max_{\hat{x}} p(\hat{x}|z) \Rightarrow \hat{x}^{MAP}(z) = \eta_{x|z}.$$

In other words, we have immediately that the posterior Gaussian is expressed in terms of the *conditional mean* and *conditional variance*!

Fundaments of estimation

ML vs MAP

From the previous analysis, $\eta_{x|z}$ is given by

$$\eta_{x|z} = \frac{\sigma^2}{\sigma^2 + \sigma_x^2}\mu_x + \frac{\sigma_x^2}{\sigma^2 + \sigma_x^2}z \text{ and } \sigma_{x|z}^2 = \frac{\sigma^2\sigma_x^2}{\sigma^2 + \sigma_x^2},$$

which is the formula of parallel resistors.
Notice how the component with the highest uncertainty, weights less in the estimation.
Moreover,

$$\eta_{x|z} = \mu_x + \frac{\sigma_x^2}{\sigma^2 + \sigma_x^2}(z - \mu_x).$$

Fundaments of estimation

ML vs MAP

This idea can be generalised to a set of *i.i.d.*, e.g. *independent and identically distributed*, measures Z^k, each normally distributed $\mathcal{N}(x, \sigma^2)$. In this case, we have

$$p(Z^k|x) = p(z(1), \ldots, z(k)|x) = \\ = \prod_{j=1}^{k} \mathcal{N}(x, \sigma^2) = c'' e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{k} (z(i)-x)^2}$$

Hence, the posterior is given by

$$p(\hat{x}|Z^k) = c' p(Z^k|\hat{x}) p(\hat{x}) = c e^{\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{k} (z(i)-\hat{x})^2\right] - \frac{(\hat{x}-\mu_x)^2}{2\sigma_x^2}}.$$

Fundaments of estimation

ML vs MAP

For the MAP we have

$$\hat{x}^{MAP}[k, Z^k] = \arg\max_{\hat{x}} p(\hat{x}|Z^k) =$$

$$= \arg\min_{\hat{x}} \left[-\frac{1}{2\sigma^2} \sum_{i=1}^{k} (z(i) - \hat{x})^2 \right] - \frac{(\hat{x} - \mu_x)^2}{2\sigma_x^2}.$$

It immediately follows that

$$\hat{x}^{MAP}(z) = \frac{\frac{\sigma^2}{k}}{\frac{\sigma^2}{k} + \sigma_x^2} \mu_x + \frac{\sigma_x^2}{\frac{\sigma^2}{k} + \sigma_x^2} \left(\frac{1}{k} \sum_{i=1}^{k} z(i) \right) = \eta_{x|z}.$$

Fundaments of estimation

ML vs MAP

Some remarks are now in order.

Remark

The MAP estimator is a weighted sum between: the ML estimator and μ_x, which is the peak of the prior.

Remark

Depending on the number of the available measurements Z^k, the MAP choses either the prior (when k is small), the measurements (when k is relatively high) or a combination thereof.

Fundaments of estimation

Non-Bayesian estimator: Least Squares

Another non-Bayesian estimator for nonrandom parameters is the *Least Squares* (LS) method.
Let us consider a set of measurements about a *nonrandom* variable x affected by random noise w as

$$z(j) = x + w(j), \quad j = 1, \ldots, k,$$

the *Least Squares Estimator* is given by

$$\hat{x}^{LS}[k, Z^k] = \arg\min_{\hat{x}} \sum_{j=1}^{k} [z(j) - \hat{x}]^2.$$

This is a *linear LS problem*, the simplest form of *linear regression*.

Fundaments of estimation

Non-Bayesian estimator: Least Squares

This solution *makes no assumption* about the noise.
If, by chance, this noise is *i.i.d.* (independent and identically distributed) and zero-mean Gaussian $w(j) \sim \mathcal{N}(0, \sigma^2)$, then it turns out that

$$z(j) \sim \mathcal{N}(x, \sigma^2), \quad j = 1, \ldots, k.$$

Fundaments of estimation

Non-Bayesian estimator: Least Squares

Since the *likelihood* function of x is

$$\Lambda_k(\hat{x}) = p(Z^k|\hat{x}) = p(z(1), \ldots, z(k)|\hat{x}) =$$
$$= \prod_{j=1}^{k} \mathcal{N}(\hat{x}, \sigma^2) = ce^{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}(z(j)-\hat{x})^2},$$

if follows that the maximisation of the ML is *equivalent* to the minimisation of the LS, that is the LS method is a *disguised* ML approach for Gaussian noises.

Fundaments of estimation

Non-Bayesian estimator: Least Squares

It has to be noted that the *Least Squares* (LS) method works also for *nonlinear* and possibly *time varying* output functions.
Given a nonlinear function of the parameter x affected by random noise w as

$$z(j) = h(j, x) + w(j), \quad j = 1, \ldots, k,$$

the *Least Squares Estimator* is given by

$$\hat{x}^{LS}(k) = \arg\min_{\hat{x}} \sum_{j=1}^{k} [z(j) - h(j, \hat{x})]^2.$$

This is a *nonlinear LS problem*.

Fundaments of estimation

Non-Bayesian estimator: Least Squares

If, by chance, this noise is *i.i.d.* (independent and identically distributed) and zero-mean Gaussian $w(j) \sim \mathcal{N}(0, \sigma^2)$, then it turns out that

$$z(j) \sim \mathcal{N}(h(j, x), \sigma^2), \;\; j = 1, \ldots, k.$$

Fundaments of estimation

Non-Bayesian estimator: Least Squares

The *likelihood* function of x is again given by

$$\Lambda_k(\hat{x}) = p(Z^k|\hat{x}) = p(z(1), \ldots, z(k)|\hat{x}) =$$

$$= \prod_{j=1}^{k} \mathcal{N}(h(j,\hat{x}), \sigma^2) = ce^{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}(z(j)-h(j,\hat{x}))^2}$$

which again reveals that the maximisation of the ML is equivalent to the minimisation of the LS for Gaussian noises.

Fundaments of estimation

Bayesian estimator: Minimum Mean Square Error

For random parameters, the "counterpart" of the LS is the *Minimum Mean Square Error* (MMSE) estimator.
In practice, we consider the *square error* of the estimator

$$(\hat{x} - x)^2.$$

Since we are dealing with a **rv**, we want to consider its *mean* assuming the knowledge of the measurements Z^k, i.e.

$$\mathsf{E}\left\{(\hat{x} - x)^2 | Z^k\right\}.$$

Finally, the MMSE is given by

$$\hat{x}^{MMSE}[k, Z^k] = \arg\min_{\hat{x}} \mathsf{E}\left\{(\hat{x} - x)^2 | Z^k\right\}.$$

Fundaments of estimation

Bayesian estimator: Minimum Mean Square Error

To find the solution to the MMSE estimator, i.e.,

$$\hat{x}^{MMSE}[k, Z^k] = \arg\min_{\hat{x}} \mathsf{E}\left\{(\hat{x} - x)^2 | Z^k\right\},$$

we set the derivative with respect to $\hat{x}$ to zero

$$\frac{d\mathsf{E}\left\{(\hat{x} - x)^2 | Z^k\right\}}{d\hat{x}} = \mathsf{E}\left\{2(\hat{x} - x) | Z^k\right\} = 2(\hat{x} - \mathsf{E}\left\{x | Z^k\right\}) = 0$$

It then follows that

$$\hat{x}^{MMSE}[k, Z^k] = \mathsf{E}\left\{\hat{x} | Z^k\right\} \triangleq \int_{-\infty}^{+\infty} \hat{x} p(\hat{x} | Z^k) d\hat{x},$$

i.e. the *conditional mean*!.

Fundaments of estimation

LS, MMSE and MAP

Remark

Both $\hat{x}^{LS}[k, Z^k]$ (with noise) and $\hat{x}^{MMSE}[k, Z^k]$ returns random variables.

Remark

The MMSE is a particular case of Bayesian estimator where the expected value of a quadratic function has to be minimised.

Remark

The difference between the MAP and the MMSE is that the MAP finds the most probable x, i.e. the mode of the posterior $p(\hat{x}|Z^k)$, while the MMSE finds the mean value of the posterior $p(\hat{x}|Z^k)$.

Fundaments of estimation

LS, MMSE and MAP

The MAP and the MMSE estimators for the Gaussian can be rewritten in terms of the *conditional mean* and the *conditional variance*, i.e.

$$\eta_{x|z} = \frac{\sigma^2}{\sigma^2 + \sigma_x^2}\mu_x + \frac{\sigma_x^2}{\sigma^2 + \sigma_x^2} z \text{ and } \sigma_{x|z}^2 = \frac{\sigma^2 \sigma_x^2}{\sigma^2 + \sigma_x^2}.$$

Notice that in the case of the Gaussian, the mean and the mode actually coincides, hence the equivalence between the two methods.

Fundaments of estimation

LS, MMSE and MAP

Remark

The MAP and the MMSE estimators for the Gaussian can be rewritten as

$$\hat{x}^{MMSE} = \sigma_x^{-2}(\sigma^{-2} + \sigma_x^{-2})^{-1}\mu_x + \sigma^{-2}(\sigma^{-2} + \sigma_x^{-2})^{-1}z =$$
$$= (\sigma^{-2} + \sigma_x^{-2})^{-1}\left(\frac{\mu_x}{\sigma_x^2} + \frac{z}{\sigma^2}\right),$$

which again expresses that the weightings of the prior and the measurements are inversely proportional to their variances.

Remark

Since the inverse of the variances is called the information and since $\sigma_{x|z}^{-2} = \sigma^{-2} + \sigma_x^{-2}$*, the inverse of the variances are additive. This additivity property holds in general when the pdfs are independent.*

Fundaments of estimation

Jointly gaussian **rv**s

Let us generalise the MMSE analysis thus introduced and let us consider two vectors of **rv**s $x \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$ that are jointly Gaussian, we have:

$$y = \begin{bmatrix} x \\ z \end{bmatrix} \sim \mathcal{N}(\mu_y, P_{yy}).$$

In other words:

$$p(y) = p(x, z) = \frac{1}{\sqrt{|2\pi P_{yy}|}} e^{-\frac{1}{2}(y-\mu_y)^T P_{yy}^{-1}(y-\mu_y)},$$

where

$$\mu_y = \begin{bmatrix} \mu_x \\ \mu_z \end{bmatrix} \quad \text{and} \quad P_{yy} = \begin{bmatrix} P_{xx} & P_{xz} \\ P_{zx} & P_{zz} \end{bmatrix},$$

where $P_{xz} = P_{zx}^T$ is the *cross covariance matrix*.
Let us first try derive the marginals, i.e. $p_x(x)$ and $p_z(z)$.

Fundaments of estimation

MMSE for jointly Gaussian **rv**s

Let us first define $a = x - \mu_x$ and $b = z - \mu_z$, i.e.

$$p(x,z) = \frac{1}{\sqrt{|2\pi P_{yy}|}} e^{-\frac{1}{2}\begin{bmatrix} a^T & b^T \end{bmatrix} P_{yy}^{-1} \begin{bmatrix} a \\ b \end{bmatrix}}.$$

Let us consider

$$P_{yy}^{-1} = \begin{bmatrix} P_{xx} & P_{xz} \\ P_{zx} & P_{zz} \end{bmatrix}^{-1} = \begin{bmatrix} Q_{xx} & Q_{xz} \\ Q_{zx} & Q_{zz} \end{bmatrix},$$

where by definition if the covariance matrix $P_{xx} = P_{xx}^T$, $P_{zz} = P_{zz}^T$ and $P_{xz} = P_{zx}^T$.

Fundaments of estimation

MMSE for jointly Gaussian **rv**s

By computing the matrix product at the exponent, we have the scalar value

$$g = a^T Q_{xx} a + a^T Q_{xz} b + b^T Q_{zx} a + b^T Q_{zz} b$$

By the matrix inversion lemma we have immediately that:

$$\begin{aligned}
Q_{xx} &= (P_{xx} - P_{xz}P_{zz}^{-1}P_{zx})^{-1} = P_{xx}^{-1} + P_{xx}^{-1}P_{xz}(P_{zz} - P_{zx}P_{xx}^{-1}P_{xz})^{-1}P_{zx}P_{xx}^{-1}, \\
Q_{zz} &= (P_{zz} - P_{zx}P_{xx}^{-1}P_{xz})^{-1} = P_{zz}^{-1} + P_{zz}^{-1}P_{zx}(P_{xx} - P_{xz}P_{zz}^{-1}P_{zx})^{-1}P_{xz}P_{zz}^{-1}, \\
Q_{xz} &= -P_{xx}^{-1}P_{xz}(P_{zz} - P_{zx}P_{xx}^{-1}P_{xz})^{-1} = Q_{zx}^T, \\
Q_{zx} &= -P_{zz}^{-1}P_{zx}(P_{xx} - P_{xz}P_{zz}^{-1}P_{zx})^{-1} = Q_{xz}^T,
\end{aligned}$$

which ensures the fact that the invers is symmetric as well, i.e. $Q_{xx} = Q_{xx}^T$, $Q_{zz} = Q_{zz}^T$ and $Q_{xz} = Q_{zx}^T$.

Fundaments of estimation

MMSE for jointly Gaussian **rv**s

Substituting the previous equations in g, we have

$$
\begin{aligned}
g &= a^T Q_{xx} a + a^T Q_{xz} b + b^T Q_{zx} a + b^T Q_{zz} b = a^T Q_{xx} a + 2 b^T Q_{zx} a + b^T Q_{zz} b = \\
&= a^T (P_{xx} - P_{xz} P_{zz}^{-1} P_{zx})^{-1} a - 2 b^T P_{zz}^{-1} P_{zx} (P_{xx} - P_{xz} P_{zz}^{-1} P_{zx})^{-1} a + \\
&+ b^T \left[P_{zz}^{-1} + P_{zz}^{-1} P_{zx} (P_{xx} - P_{xz} P_{zz}^{-1} P_{zx})^{-1} P_{xz} P_{zz}^{-1} \right] b = \\
&= a^T (P_{xx} - P_{xz} P_{zz}^{-1} P_{zx})^{-1} a - 2 b^T P_{zz}^{-1} P_{zx} (P_{xx} - P_{xz} P_{zz}^{-1} P_{zx})^{-1} a + \\
&+ b^T P_{zz}^{-1} b + b^T P_{zz}^{-1} P_{zx} (P_{xx} - P_{xz} P_{zz}^{-1} P_{zx})^{-1} P_{xz} P_{zz}^{-1} b = \\
&= [a - P_{xz} P_{zz}^{-1} b]^T (P_{xx} - P_{xz} P_{zz}^{-1} P_{zx})^{-1} [a - P_{xz} P_{zz}^{-1} b] + b^T P_{zz}^{-1} b
\end{aligned}
$$

Fundaments of estimation

MMSE for jointly Gaussian **rv**s

Recalling that $a = x - \mu_x$ and $b = z - \mu_z$, we can define

$$\gamma = \mu_x + P_{xz}P_{zz}^{-1}(z - \mu_z) \text{ and } \Gamma = P_{xx} - P_{xz}P_{zz}^{-1}P_{zx},$$

so hence to have

$$G_1(z) = (z - \mu_z)^T P_{zz}^{-1}(z - \mu_z) \text{ and } G_2(x) = (x - \gamma)^T \Gamma^{-1}(x - \gamma),$$

and finally

$$g = G_1(z) + G_2(x).$$

Fundaments of estimation

MMSE for jointly Gaussian **rv**s

Therefore, we have

$$p(x,z) = \frac{1}{\sqrt{|2\pi P_{yy}|}} e^{-\frac{1}{2}\begin{bmatrix} a^T & b^T \end{bmatrix} P_{yy}^{-1} \begin{bmatrix} a \\ b \end{bmatrix}} =$$
$$= \frac{1}{\sqrt{|2\pi P_{yy}|}} e^{-\frac{1}{2}(G_1(z)+G_2(x))}.$$

First notice that the term $|2\pi P_{yy}| = (2\pi)^{n+m}|P_{yy}|$ and then we notice that

$$|P_{yy}| = |P_{zz}||P_{xx} - P_{xz}P_{zz}^{-1}P_{zx}| = |P_{zz}||\Gamma|.$$

Fundaments of estimation

MMSE for jointly Gaussian **rv**s

Indeed, for any block partitioned matrix

$$P_{yy} = \begin{bmatrix} P_{xx} & P_{xz} \\ P_{zx} & P_{zz} \end{bmatrix},$$

it can be expressed as the product of two *triangular* matrices, i.e.

$$P_{yy} = \begin{bmatrix} P_{xx} & P_{xz} \\ P_{zx} & P_{zz} \end{bmatrix} = \begin{bmatrix} I & P_{xz} \\ 0 & P_{zz} \end{bmatrix} \begin{bmatrix} P_{xx} - P_{xz}P_{zz}^{-1}P_{zx} & 0 \\ P_{zz}^{-1}P_{zx} & I \end{bmatrix}.$$

Recalling that $|AB| = |A||B|$, the proof follows.

Fundaments of estimation

MMSE for jointly Gaussian **rv**s

As a consequence

$$p(x,z) = \frac{1}{\sqrt{|2\pi P_{yy}|}} e^{-\frac{1}{2}(G_1(z)+G_2(x))} = \\ = \frac{1}{\sqrt{(2\pi)^m |P_{zz}|}} e^{-\frac{1}{2}G_1(z)} \frac{1}{\sqrt{(2\pi)^n |\Gamma|}} e^{-\frac{1}{2}G_2(x)}.$$

It then follows that the marginal

$$p_z(z) = \int_{-\infty}^{+\infty} p(x,z)dx = \frac{1}{\sqrt{(2\pi)^m |P_{zz}|}} e^{-\frac{1}{2}G_1(z)} = \\ = \frac{1}{\sqrt{(2\pi)^m |P_{zz}|}} e^{-\frac{1}{2}(z-\mu_z)^T P_{zz}^{-1}(z-\mu_z)}.$$

hence we have shown that the marginal $p_z(z) \sim \mathcal{N}(\mu_z, P_{zz})$!

Fundaments of estimation

MMSE for jointly Gaussian **rv**s

In practice, given the joint Gaussian pdf defined as

$$p(y) = p(x,z) = \frac{1}{\sqrt{|2\pi P_{yy}|}} e^{-\frac{1}{2}(y-\mu_y)^T P_{yy}^{-1}(y-\mu_y)},$$

where

$$\mu_y = \begin{bmatrix} \mu_x \\ \mu_z \end{bmatrix} \text{ and } P_{yy} = \begin{bmatrix} P_{xx} & P_{xz} \\ P_{zx} & P_{zz} \end{bmatrix},$$

the marginal $p_z(z) \sim \mathcal{N}(\mu_z, P_{zz})$ is a Gaussian whose mean value and covariance matrix are *the correct portion of the mean values vector and the correct portion of the covariance matrix!*
Of course, by exchanging the role of z and x, we can compute similarly the marginal $p_x(x) \sim \mathcal{N}(\mu_x, P_{xx})$.

Fundaments of estimation

MMSE for jointly Gaussian **rv**s

The interesting thing is that if we compute the conditional $p_c(x|z)$, we have immediately by definition

$$p_c(x|z) = \frac{p(x,z)}{p_z(z)} = \frac{\frac{1}{\sqrt{(2\pi)^m|P_{zz}|}}e^{-\frac{1}{2}G_1(z)}\frac{1}{\sqrt{(2\pi)^n|\Gamma|}}e^{-\frac{1}{2}G_2(x)}}{\frac{1}{\sqrt{(2\pi)^m|P_{zz}|}}e^{-\frac{1}{2}G_1(z)}} =$$

$$= \frac{1}{\sqrt{(2\pi)^n|\Gamma|}}e^{-\frac{1}{2}G_2(x)},$$

which states that $p_c(x|z) \sim \mathcal{N}(\gamma, \Gamma)$, i.e. a Gaussian with mean value and covariance matrix

$$\gamma = \mu_x + P_{xz}P_{zz}^{-1}(z - \mu_z) \text{ and } \Gamma = P_{xx} - P_{xz}P_{zz}^{-1}P_{zx}.$$

Fundaments of estimation

MMSE for jointly Gaussian **rv**s

We know that if x is the unknown vector and z are the measurements, we have to compute $p(x|z)$ to have an *estimate*.
Moreover, we know that the *minimum mean square error* (MMSE) is a *Bayesian* estimator that is computed with the *conditional mean*, i.e.

$$\hat{x}^{MMSE} = \mathsf{E}\{x|z\}.$$

Hence, it follows immediately that

$$\hat{x}^{MMSE} = \mathsf{E}\{x|z\} = \gamma = \mu_x + P_{xz}P_{zz}^{-1}(z - \mu_z).$$

Fundaments of estimation

MMSE for jointly Gaussian **rv**s

Remark

The expressions of the conditional mean *and* conditional covariance *are then*

$$E\{x|z\} = \mu_x + P_{xz}P_{zz}^{-1}(z - \mu_z) \text{ and } C\{x|z\} = \Gamma = P_{xx} - P_{xz}P_{zz}^{-1}P_{zx},$$

that are referred to as the fundamental equations of linear estimation*.*

Indeed, the *conditional mean* depends *linearly* by the observations z, while the *conditional covariance* is *independent* from the observations.
Notice that, again, the results for the mean and the covariance are a weighted composition between the prior and the measures.
Of course it is possible to compute the *marginal mean* and the *marginal covariance* using the *Total Expectations* and *Total Variance* Laws.

Fundaments of estimation

Recap on the MMSE estimator

Remark

The linear MMSE is also called in the literature least mean squares (LMS), minimum variance (MV) or leat squares (LS).

Personally I prefer *minimum variance* since it immediately explains what is the job of the estimator.

Fundaments of estimation

Recap on the MMSE estimator

Facts on the MMSE:

- The *linear MMSE* estimator is identical to the expression of the *conditional mean* of Gaussian random vectors.
- The *linear MMSE* estimator is the *best estimator* if the random variables are *Gaussian*.
- The quite interesting thing is that the *linear MMSE* estimator is the *best linear estimator* if the random variables are *not Gaussian* and generically distributed.

Fundaments of estimation

LS Example

Let us make the same example. Consider a single measurement:

$$z = x + w,$$

and suppose that x is an unknown nonrandom parameter.
The *Least Squares* criterion leads to

$$\hat{x}^{LS} = \arg\max_{\hat{x}} (z - \hat{x})^2 = z.$$

Recall that if the noise is Gaussian, i.e. $w \sim \mathcal{N}(0, \sigma^2)$, $\hat{x}^{LS} = \hat{x}^{ML}$.

Fundaments of estimation

MAP and MMSE: the Gaussian case

Let us consider a single measurement:

$$z = x + w,$$

and suppose that x is an unknown random parameter.
Assuming both $\hat{x}$ and w normally distributed, the *conditional* pdf is given by

$$p(\hat{x}|z) = \mathcal{N}(\eta_{x|z}, \sigma^2_{x|z}) = \frac{1}{\sqrt{2\pi}\sigma_{x|z}} e^{-\frac{(\hat{x}-\eta_{x|z})^2}{2\sigma^2_{x|z}}}.$$

Notice that for the Gaussian, the peak (the *mode*) coincides with the *mean* of the conditional pdf, which is the value returned by the MMSE, hence

$$\hat{x}^{MMSE} = \mathsf{E}\{\hat{x}|z\} = \eta_{x|z} = \hat{x}^{MAP}.$$

Fundaments of estimation

Unbiased estimators

For a nonrandom parameter x_0, we say that an estimator is *unbiased* if

$$\mathsf{E}\left\{\hat{x}[k, Z^k]\right\} = x_0,$$

where x_0 is the true value of x.
For a random parameter x with prior $p(x)$, we say that an estimator is *unbiased* if

$$\mathsf{E}\left\{\hat{x}[k, Z^k]\right\} = \mathsf{E}\left\{x\right\},$$

where the left term is computed on the *joint* pdf $p(x, Z^k)$, while the term on the right is computed on the $p(x)$.

Fundaments of estimation

Unbiased estimators

In the general case, an estimator is *unbiased* if

$$\mathsf{E}\left\{\tilde{x}(k)\right\} = 0,$$

where $\tilde{x}(k) = x - \hat{x}(k)$ is the *estimation error* at the k–th iteration of the estimator.
An estimator is *asymptotically unbiased* is the previous holds only for $k \to +\infty$.

Fundaments of estimation

Example: ML is unbiased

Let us consider a single measurement:

$$z = x + w,$$

and suppose that x is an unknown parameter.
For the ML we have

$$\mathsf{E}\left\{\hat{x}^{ML}\right\} = \mathsf{E}\left\{z\right\} = \mathsf{E}\left\{x + w\right\} = x.$$

Equivalently:

$$\mathsf{E}\left\{\tilde{x}\right\} = \mathsf{E}\left\{x - \hat{x}^{ML}\right\} = x - \mathsf{E}\left\{\hat{x}^{ML}\right\} = 0.$$

Fundaments of estimation

Example: MAP is unbiased

Let us consider a single measurement:

$$z = x + w,$$

and suppose that x is an unknown random parameter.
For the MAP, we have

$$\begin{aligned}
\mathsf{E}\left\{\hat{x}^{MAP}\right\} &= \mathsf{E}\left\{\frac{\sigma^2}{\sigma^2+\sigma_x^2}\mu_x + \frac{\sigma_x^2}{\sigma^2+\sigma_x^2} z\right\} = \\
&= \frac{\sigma^2}{\sigma^2+\sigma_x^2}\mu_x + \frac{\sigma_x^2}{\sigma^2+\sigma_x^2}\mathsf{E}\left\{z\right\} = \\
&= \frac{\sigma^2}{\sigma^2+\sigma_x^2}\mu_x + \frac{\sigma_x^2}{\sigma^2+\sigma_x^2}(\mu_x + \mathsf{E}\left\{w\right\}) = \mu_x.
\end{aligned}$$

Fundaments of estimation

Example: MAP is unbiased

We always mentioned that the choice of the prior is crucial. Indeed, if an incorrect choice is made, we have:

$$\begin{aligned}\mathsf{E}\left\{\hat{x}^{MAP}\right\} &= \mathsf{E}\left\{\frac{\sigma^2}{\sigma^2+\bar{\sigma}_x^2}\bar{\mu}_x + \frac{\bar{\sigma}_x^2}{\sigma^2+\bar{\sigma}_x^2}z\right\} = \\ &= \frac{\sigma^2}{\sigma^2+\bar{\sigma}_x^2}\bar{\mu}_x + \frac{\bar{\sigma}_x^2}{\sigma^2+\bar{\sigma}_x^2}\mu_x \neq \mu_x.\end{aligned}$$

However, when a sufficiently large number of measurements are available, we have

$$\mathsf{E}\left\{\hat{x}^{MAP}\right\} = \frac{\frac{\sigma^2}{k}}{\frac{\sigma^2}{k}+\bar{\sigma}_x^2}\bar{\mu}_x + \frac{\bar{\sigma}_x^2}{\frac{\sigma^2}{k}+\bar{\sigma}_x^2}\hat{\mu}_x,$$

where $\hat{\mu}_x$ is the arithmetic mean. Notice that when $k \to +\infty$, we have that $\hat{\mu}_x \to \mu_x$ and the MAP rewards the measurements, hence

$$\lim_{k\to+\infty} \mathsf{E}\left\{\hat{x}^{MAP}\right\} = \mu_x.$$

Fundaments of estimation

Variance of an estimator

With the definition of the estimation error $\tilde{x} = x - \hat{x}$, we can say that:

$$\mathsf{E}\left\{\tilde{x}^2\right\} = \mathsf{V}\left\{\hat{x}\right\},$$

if $\hat{x}$ is *unbiased* and x is *nonrandom*.
On the contrary, if $\hat{x}$ is *biased* and/or x is *random*, we have

$$\mathsf{E}\left\{\tilde{x}^2\right\} = MSE(\hat{x}),$$

where MSE stands for the *mean squared error* .

Fundaments of estimation

Summary

There are basically two main approaches to estimation given the peculiarity of the time-invariant parameter x to be estimated.

Definition

If the parameter is *nonrandom*, then x has a true value x_0 that has to be estimated. To this end, *non-Bayesian* or *Fisher* approaches will be used.

Definition

If the parameter is *random* and if *a prior* on x in terms of a pdf $p(x)$ is known, a particular *realisation* is supposed to be available and, hence, the *Bayesian* approaches will be used.

Outline

Basic estimators

A *stochastic process* maps the outcome of an *event* onto a set of *functions* thus generating *time varying* **rv**s. Each sequence of **rv**s is called a *realisation*.
A *white process* is a *stochastic process* with *autocovariance* function equal to zero for any two time instants t_i and t_j.
A *Markovian process* is a *stochastic process* with *one step memory*.
The *BLUE* estimator is in general a *sub-optimal* estimator that turns to be optimal, i.e. am *MVUE*, if the *measurement noise* is Gaussian.
Two approaches to estimators: *non-Bayesian* (e.g., ML, LS) and *Bayesian* (e.g. MAP and MMSE).