Advanced Signal Processing

Lecture 1: Random Variables

Prof Danilo Mandic

room 813, ext: 46271

Department of Electrical and Electronic Engineering

Imperial College London, UK

d.mandic@imperial.ac.uk, URL: www.commsp.ee.ic.ac.uk/∼mandic

Introduction $\rightsquigarrow$ Recap

Discrete Random Signals:

discrete vs. digital $\looparrowright$ quantisation

- $\{x[n]\}_{n=0:N-1}$ is a sequence of indexed random variables $x[0], x[1], \ldots, x[N-1]$, and the symbol $'[\cdot]'$ indicates the random nature of signal x *(every sample is random too!)*

- The sequence is discrete with respect to sample index n (*discrete time or some other physical variable, such as spatial index in arrays of sensors*)

- $x[n]$ real or complex with discrete or continuous values

NB: signals can be continuous or discrete in *time* as well as *amplitude*.

Digital signal $=$ discrete in time and amplitude

Discrete–time signal $=$ discrete in time, amplitude either discrete or continuous

Standardisation and normalisation

(e.g. to be invariant of amplifier gain or the quality of sensor contact)

Some real-world applications require data of specific mean and variance, yet measured variables are usually of different natures and magnitudes. We refer to **standardisation** as the process of converting the data to an arbitrary mean $\bar{\mu}$ and variance $\bar{\sigma}^2$, and to **normalisation** as the particular case $\bar{\mu} = 0$, $\bar{\sigma}^2 = 1$. In practice, **raw data** $\{x[n]\}_{n=0:N-1}$ are normalised by subtracting the sample mean, μ, and dividing by the sample std. dev., σ

- **Compute statistics:** $\mu = \frac{1}{N}\sum_{n=0}^{N-1} x[n], \quad \sigma^2 = \frac{1}{N}\sum_{n=0}^{N-1}(x[n]-\mu)^2$
- **Centred data:** $x^{\mathsf{C}} = x - \mu$
- **Centred and scaled data (normalised):** $x^{\mathsf{CS}} = \frac{x^{\mathsf{C}}}{\sigma} \qquad (\mu = 0, \sigma = 1)$

Normalised data can be **standardised** to any mean $\bar{\mu}$ and variance $\bar{\sigma}^2$ by

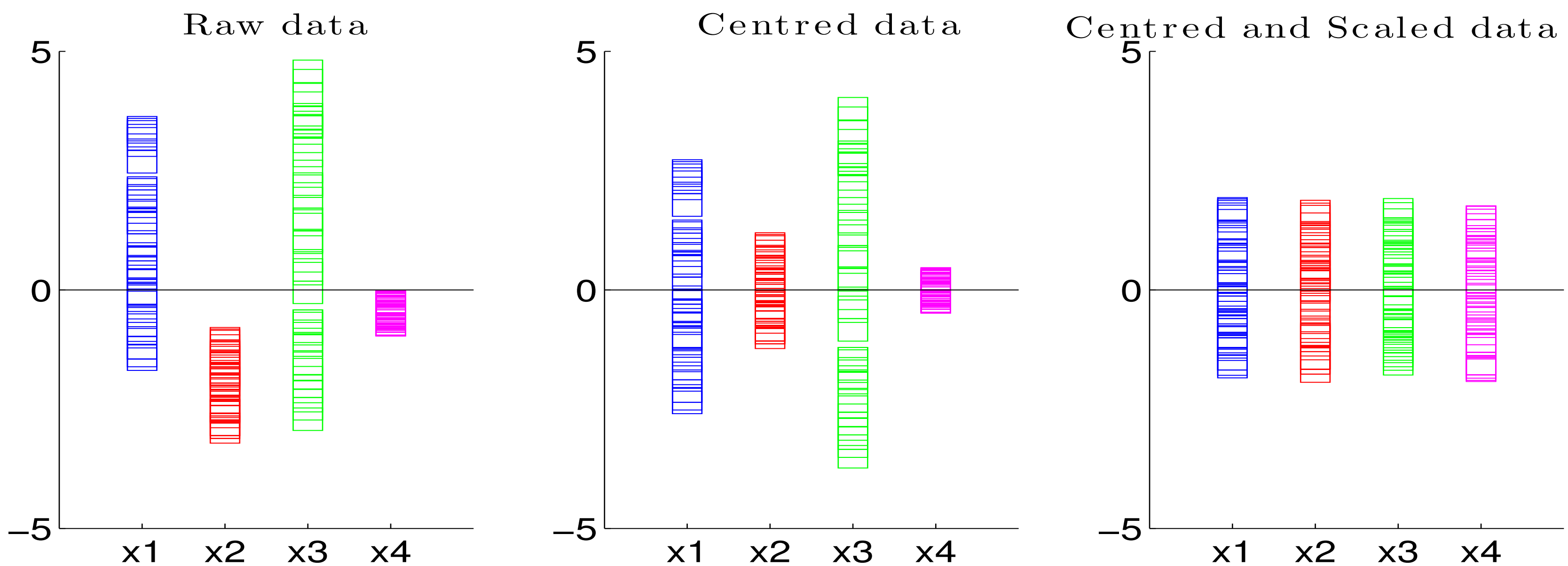
$$x^{\mathsf{ST}} = \frac{x^{\mathsf{CS}} - \bar{\mu}}{\bar{\sigma}}$$

Standardize to zero mean and range $[-1, 1]$ $\hookrightarrow \quad x(n) = \frac{x(n) - \big(x_{max} + x_{min}\big)/2}{\big(x_{max} - xmin\big)/2}$

Standardisation: Example

The bars denote the amplitudes of the samples of signals x_1-x_4

For the raw measurements: $\{x_1[n], x_2[n], x_3[n], x_4[n]\}_{n=1:N}$



◦ Standardisation allows for a *coherent and aligned* handling of different variables, as the amplitude plays a role in regression algorithms.

◦ Furthermore, input variable selection can be performed by assigning smaller or larger weighting to samples (confidence intervals).

How do we describe a signal? (statistically)

Probability distribution functions – very convenient!

- **Cumulative Density Function (CDF)** – probability of a random variable falling within a given range.

$$F_X\left(x[n]\right) = \text{Probability}\left(X[n] \leq x[n]\right) \tag{1}$$

  $X[n]$ – random quantity, $x[n]$ – particular fixed value.

- **Probability Density Function (pdf)** – relative likelihood for a random variable to occur at a given point in the observation space.

$$p\left(x[n]\right) = \frac{\partial F_X(x[n])}{\partial x[n]} \qquad \Leftrightarrow \qquad F(x) = \int_{-\infty}^{x} p(X)dX \tag{2}$$

NB: For random signals, for two time instants n_1 and n_2, the pdf of $x[n_1]$ need not be identical to that of $x[n_2]$. (e.g. $\sin(n) + w(n)$)

Statistical distributions: Uniform distribution

Important: Recall that probability densities sum up to unity

$$\int_{-\infty}^{\infty} p\big(x[n]\big)dx[n] = 1$$

and that the connection between pdf and its ***cumulative density function*** CDF is

$$F(x[n]) = \int_{-\infty}^{x[n]} p(z)dz, \quad \text{also} \quad \lim_{x[n]\rightarrow\infty} F\big(x[n]\big) = 1$$

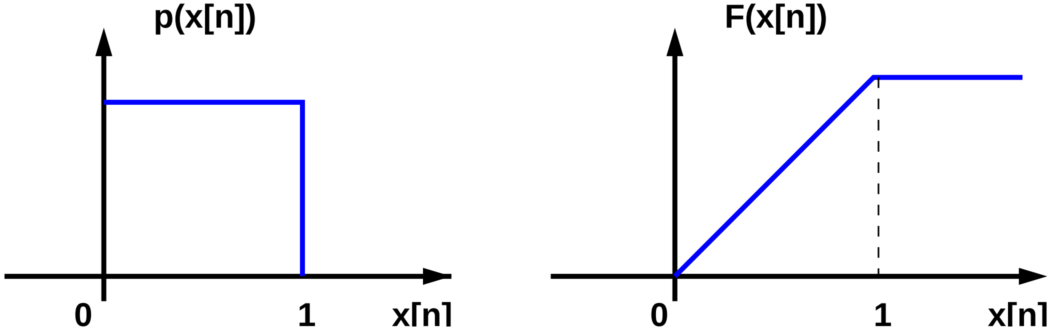
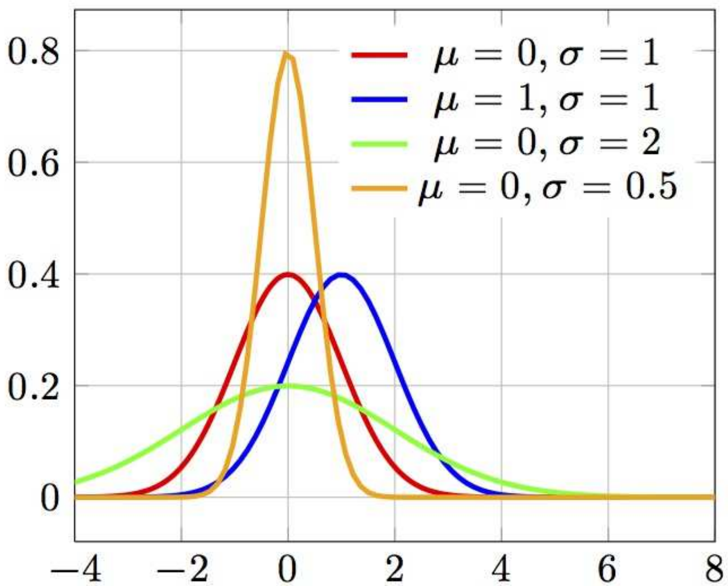


Figure: pdf and CDF for uniform distribution. **In MATLAB – function rand**

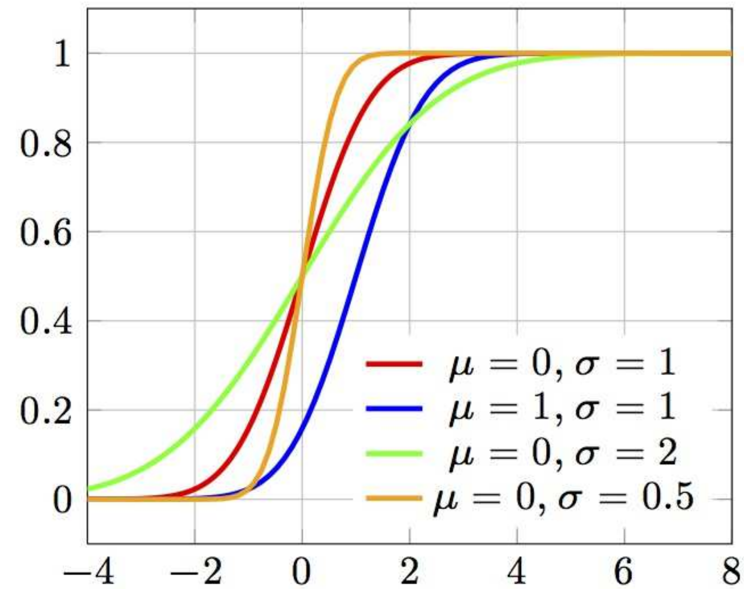
Gaussian probability and cumulative density functions

How does the variance σ^2 influence the shape of CDP and pdf?

Gaussian pdf



Gaussian CDF



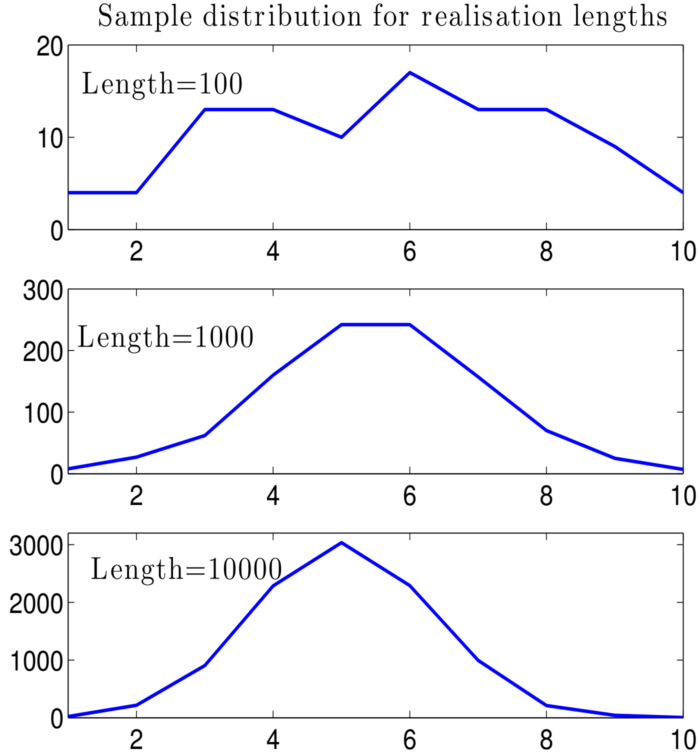
$$p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad P(x; \mu, \sigma) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right]$$

The **standard Gaussian** distribution ($\mu = 0, \sigma = 1$) is given by $p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$

Statistical distributions: Gaussian $\looparrowright$ randn in Matlab

Very convenient (mathematical tractability) - especially in terms of **log-likelihood** $\log p(x[n])$

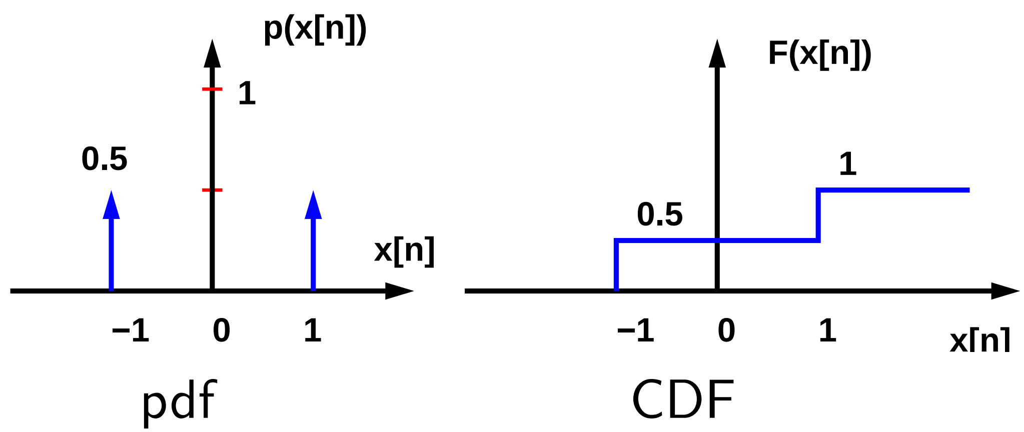
$$p(x[n]) = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{(x[n]-\mu_x)^2}{2\sigma_x^2}} \quad \Rightarrow \quad \log p(x[n]) = -\frac{(x[n]-\mu_x)^2}{2\sigma_x^2} - \frac{1}{2}\log\left(2\pi\sigma_x^2\right)$$



Sample densities for varying N

$$x[n] \sim \mathcal{N}\left(\mu_x, \sigma_x^2\right) \qquad \mu_x \to \text{mean}, \;\; \sigma_x^2 \to \text{variance}$$

Bipolar distribution



Multi-dimensionality versus multi-variability

Univariate vs. **Multivariate** vs. **Multidimensional**

- Single input single output (SISO) e.g. single-sensor system
- Multiple input multiple output (MIMO) (arrays of transmitters and receivers) ***can measure one source with many sensors***
- Multidimensional processes (3D inertial bodymotion sensors, radar, vector fields, wind anemometers) – ***intrinsically multidimensional***

Example: Multivariate function with single output (MISO)

$$stockvalue = f(stocks, oilprice, GNP, month, \ldots)$$

$\Rightarrow$ Complete probabilistic description of $\{x[n]\}$ is given by its pdf

$$p\,(x[n_1], \ldots, x[n_k]) \quad \text{for all } k \text{ and } n_1, \ldots, n_k.$$

Much research is being directed towards the reconstruction of the process history from observations of one variable only (Takens)

Joint distributions of delayed samples (temporal)

Joint distribution (bivariate CDF)

$$F\left(x[n_1], x[n_2]\right) = \text{Prob}\left(X[n_1] \leq x[n_1], X[n_2] \leq x[n_2]\right)$$

and its pdf

$$p\left(x[n_1], x[n_2]\right) = \frac{\partial^2 F(x[n_1], x[n_2])}{\partial x[n_1] \partial x[n_2]}$$

A k–th order multivariate CDF distribution

$$F\left(x[n_1], x[n_2], \ldots, x[n_k]\right) = \text{Prob}\left(X[n_1] \leq x[n_1], \ldots, X[n_k] \leq x[n_k]\right)$$

and its pdf

$$p\left(x[n_1], x[n_2], \ldots, x[n_k]\right) = \frac{\partial^k F\left(x[n_1], \ldots, x[n_k]\right)}{\partial x[n_1] \cdots \partial x[n_k]}$$

Mathematically simple, but complicated to evaluate in reality

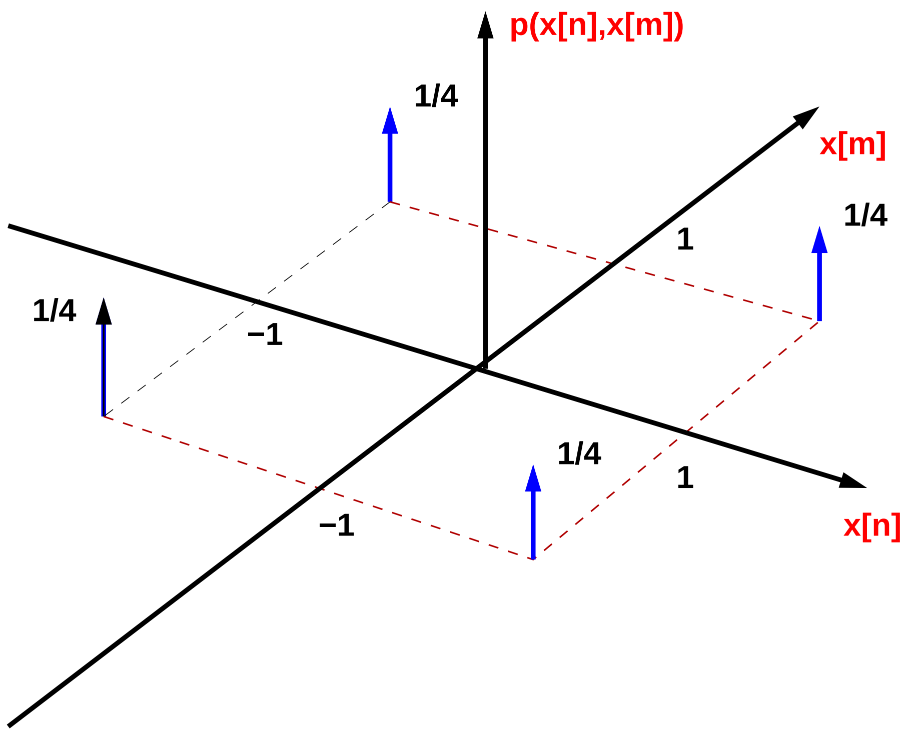
Luckily, real world time series often have "finite memory" (Markov)

Example 1.1. Bivariate pdf

Notice the change in indices (assuming **discrete time** signals)

$$CDF: \qquad F\left(x[n], x[m]\right) = \text{Prob}\left\{X[n] \le x[n], X[m] \le x[m]\right\}$$

$$PDF: \qquad p\left(x[n], x[m]\right) = \frac{\partial^2 F(x[n], x[m])}{\partial x[n] \partial x[m]}$$



Homework: Plot the CDF for this case. (What would happen in $\mathbb{C}$?)

Properties of the statistical expectation operator

P1: Linearity

$$E\{ax[n] + by[m]\} = aE\{x[n]\} + bE\{y[m]\}$$

P2: $E\{x[m]y[n]\} \neq E\{x[m]\}E\{y[n]\}$ unless $\{x[m]\}$ and $\{y[n]\}$ are independent random processes

P3: If $y[n] = g(x[n])$ and the pdf of $x[n]$ is $p(x[n])$ then

$$E\{y[n]\} = \int_{-\infty}^{\infty} g(x[n])p(x[n])dx[n]$$

that is, we DO NOT need to know the pdf of $\{y[n]\}$ to find its expected values (when $g(\cdot)$ is a deterministic function).

NB: Think of a saturation-type sensor (microphone)

Example 1.2. Mean for linear systems (use P1 & P2)

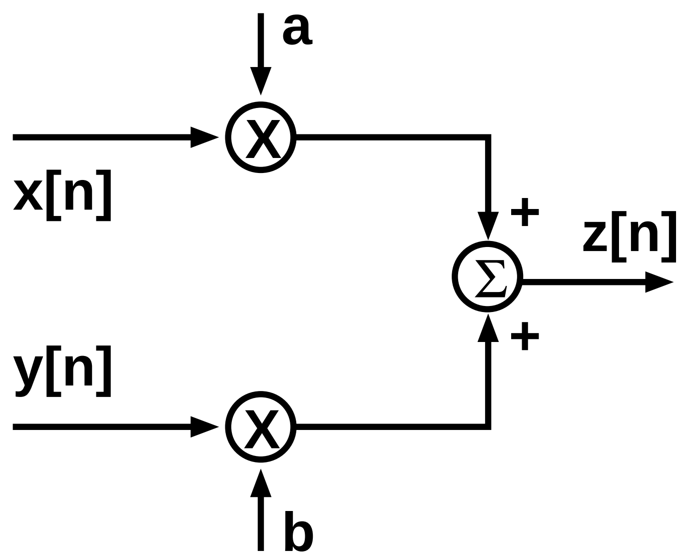
Consider a general linear system given by $z[n] = ax[n] + by[n]$. Find the mean ($E\{x[n]\} = \mu_x, E\{y[n]\} = \mu_y$, and $x \perp y$).

Solution:

$$E\{z[n]\} = E\{ax[n] + by[n]\} = aE\{x[n]\} + bE\{y[n]\}$$

that is

$$\mu_z = a\mu_x + b\mu_y$$



This is a consequence of the linearity of the $E\{\cdot\}$ operator.

Example 1.3. Mean for nonlinear systems (use P3)

For a nonlinear system, say the sensor nonlinearity is given by

$$z[n] = x^2[n]$$

using Property P3 of the statistical expectation operator, we have

$$\mu_z = E\{x^2[n]\} = \int_{-\infty}^{\infty} x^2[n] p(x[n]) dx[n]$$

This is extremely useful, since most of the real–world signals are observed through sensors, e.g.

microphones, geophones, various probes ...

which are **almost invariably nonlinear** (typically a saturation type nonlinearity)

Dealing with ensembles of random processes

Ensemble $\looparrowright$ collection of **all possible realisations** of a **random signal**

The Ensemble Mean

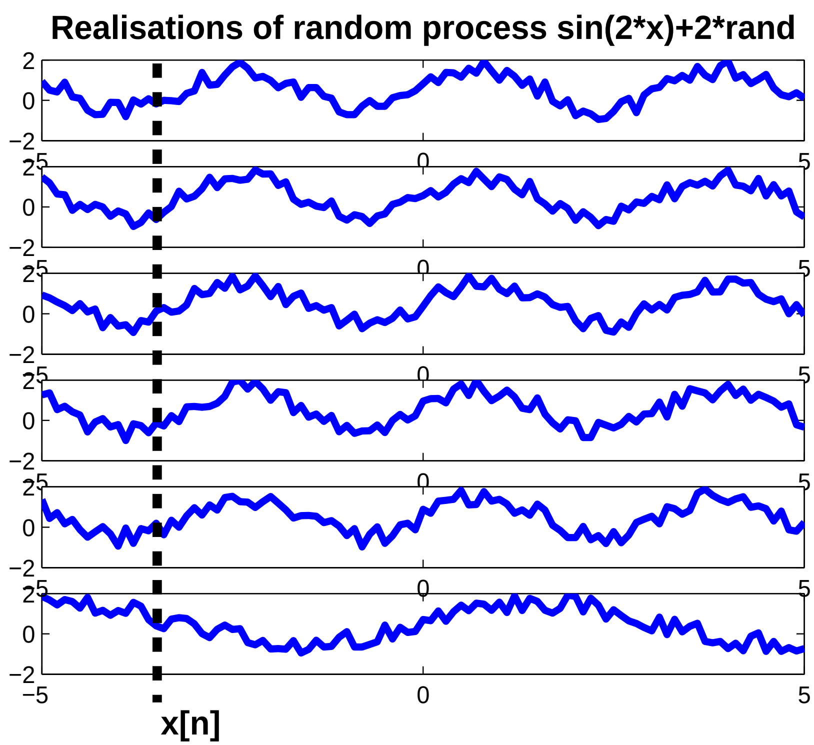
$$\mu_x(n) = \frac{1}{N}\sum_{i=1}^{N} x_i[n]$$

where $x_i[n]$ $\looparrowright$ outcome of i–th experiment at sample n.

For $N \rightarrow \infty$ we have

$$\mu_x(n) = \lim_{N\rightarrow\infty} \frac{1}{N}\sum_{i=1}^{N} x_i[n]$$

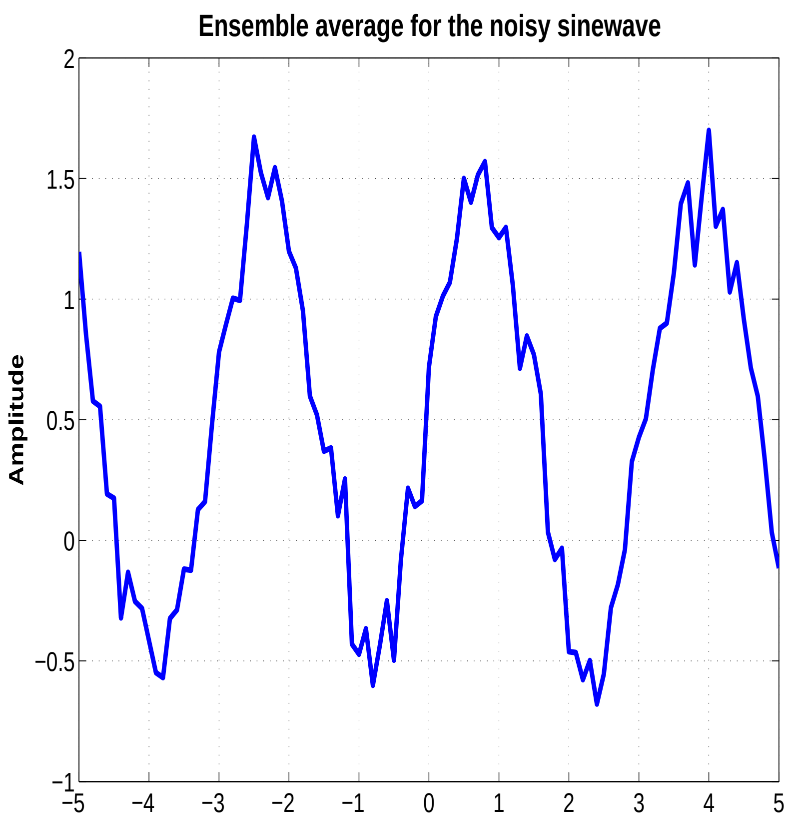
Average both **along** one and **across** all realisations?

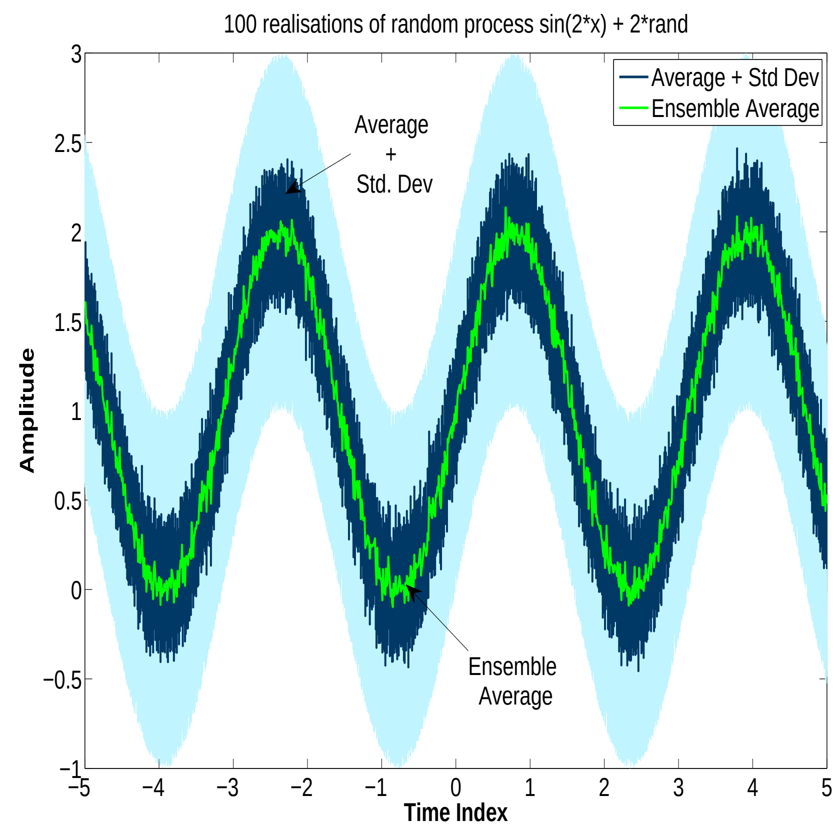


$$\text{Average Statistically} \qquad E\{x[n]\} \quad = \quad \mu_x = \int_{-\infty}^{\infty} x[n]p(x[n])dx[n]$$

Ensemble Average = Ensemble Mean

Our old noisy sine example
stochastic process is a collection of random variables





The pdf at time instant n is different from that at m, in particular:

$$\mu(n) \neq \mu(m) \qquad m \neq n$$

Left & Right: Ensemble average

$$\sin(2x) + 2 * rand$$

Left: 6 realisations, **Right:** 100 realisations (and the overlay plot)

Second order statistics: 1) Correlation

- **Correlation (also known as Autocorrelation Function (ACF))**

$$r(m,n) = E\{x[m]x[n]\}, \quad \text{that is}$$

$$r(m,n) = \int_{-\infty}^{\infty} x[m]x[n]p\big(x[m],x[n]\big)dx[m]dx[n]$$

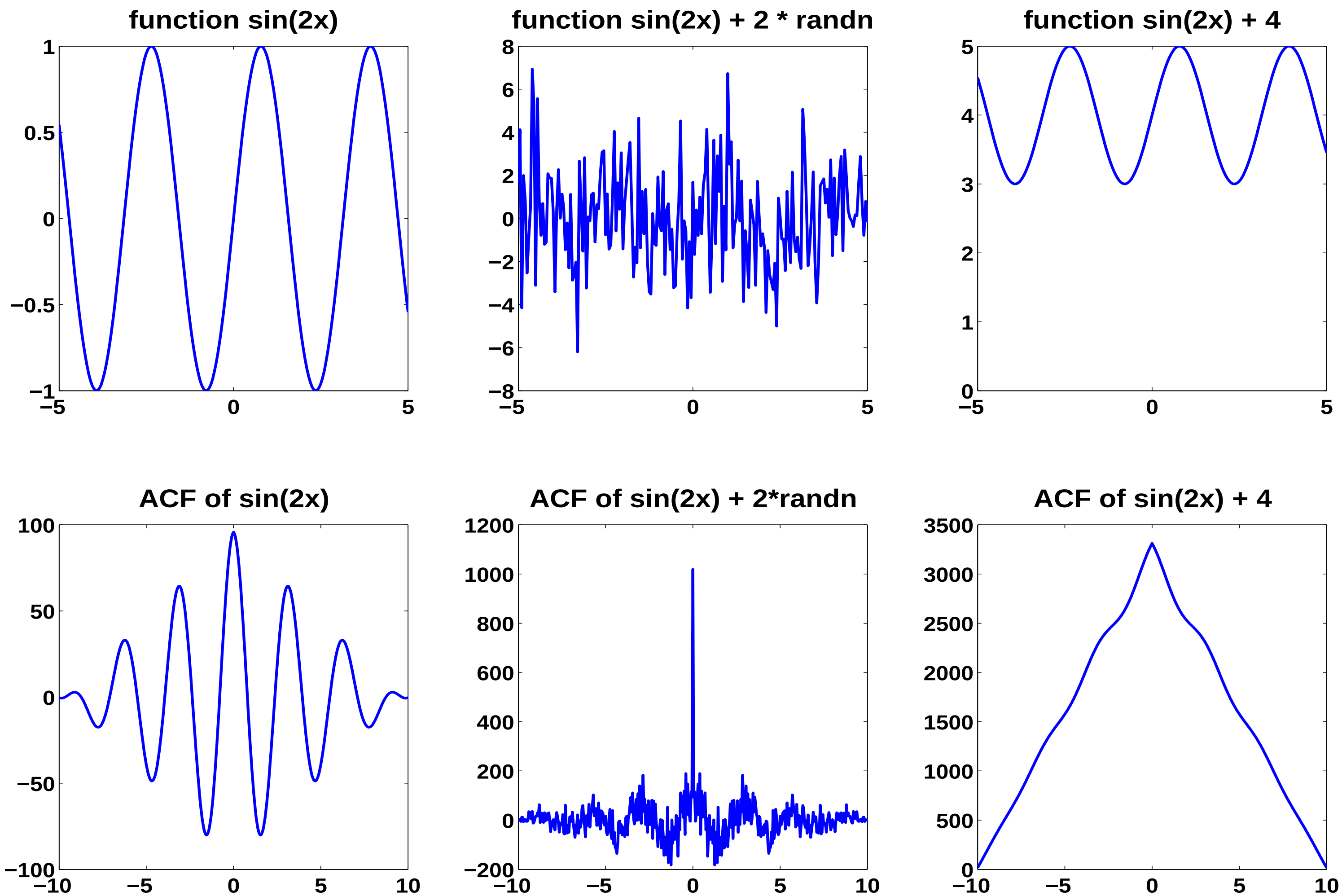
  - in practice, for **ergodic signals** we calculate correlations from the **relative frequency perspective**

$$r(m,n) = \lim_{N\to\infty} \left\{ \frac{1}{N}\sum_{i=1}^{N} x_i[m]x_i[n] \right\}, (i \text{ denotes the ensemble index})$$

  - $r(m,n)$ measures the **degree of similarity** between $x[n]$ and $x[m]$.
  - $r(n,n) = E\{x^2[n]\} \Rightarrow$ is the average "power" of a signal
  - $r(m,n) = r(n,m) \Rightarrow$ the autocorrelation matrix of all $\{r(m,n)\}$

$$\mathbf{R} = \{r(m,n)\} = E\big[\mathbf{x}\mathbf{x}^H\big] \text{ is } \textbf{symmetric}$$

Example 1.4. Autocorrelation of sinewaves
(the need for a covariance function)



$\rightsquigarrow$ Useful information gets obscured in noise or under a DC offset

Second order statistics: 2) Covariance

- **Covariance** is defined as

$$\begin{aligned} c(m,n) &= E\{(x[m]-\mu(m))(x[n]-\mu(n))\} \\ &= E\{x[m]x[n]\}-\mu(m)\mu(n) \\ c(n,n) &= \sigma_n^2 = E\{(x[n]-\mu(n))^2\} \qquad \text{for m = n} \end{aligned}$$

- Properties:
  - $c(m,n) = c(n,m) \Rightarrow$ the covariance matrix for $\mathbf{x} = \big[x[0],\ldots,x[N-1]\big]^T$ is **symmetric** and given by

$$\mathbf{C} = \{c(m,n)\} = E\big[\mathbf{x}\mathbf{x}^H\big], \text{where } \ \mathbf{x} = \{x-\mu\}$$

  - For zero mean signals, $c(m,n) = r(m,n)$

(see also the Standardisation slide and Example 1.4)

Higher order moments

For a zero-mean stochastic process $\{x[n]\}$:

- Third and fourth order moments

$$\text{Skewness}: R_3(l,m,n) = E\{x[l]x[m]x[n]\}$$

$$\text{Kurtosis}: R_4(l,m,n,p) = E\{x[l]x[m]x[n]x[p]\}$$

- In general, n–th order moment

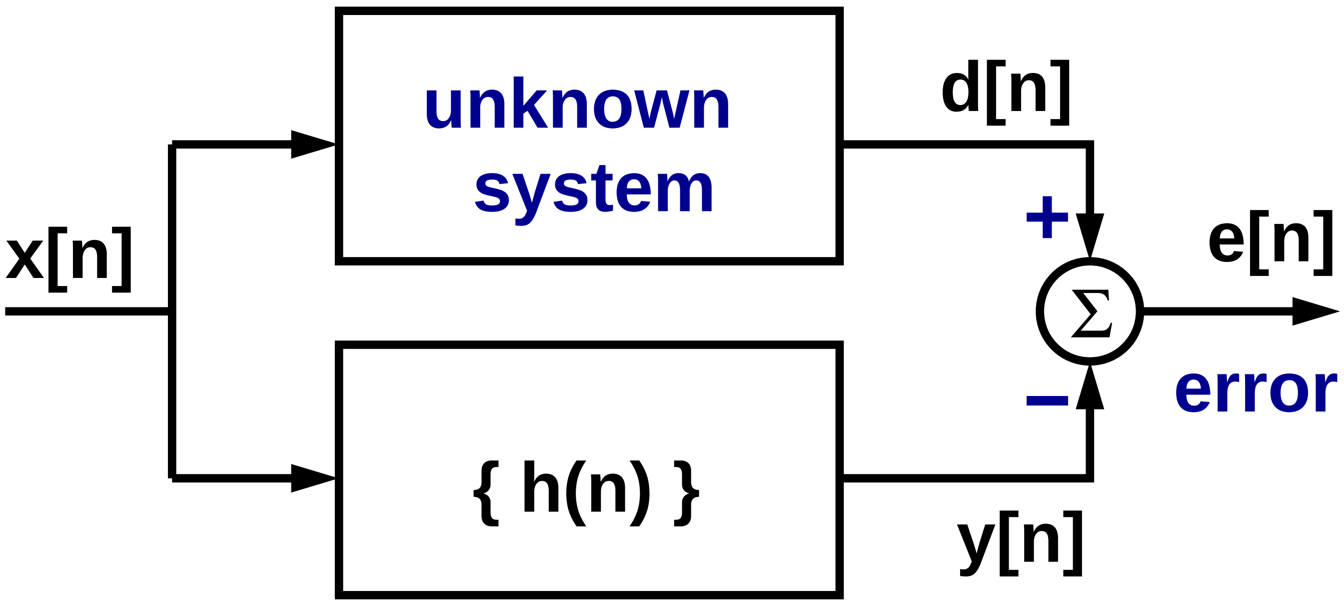
$$R_N(l_1, l_2, \ldots, l_n) = E\{x[l_1]x[l_2]\cdots x[l_n]\}$$

Higher order moments can be used to form noise-insensitive statistics (cumulants).

- Important in non-linear signal processing
- Applications: blind source separation

⊛ In many applications the signals are assumed to be, or are reduced to, zero-mean stochastic process.

Example 1.5. Use of statistics in system identification (statistical rather than transfer function based analysis)



Task: Select $\{h(n)\}$ such that $y[n]$ is as similar to $d[n]$ as possible.

Measure of "goodness" is the distribution of the error $\{e[n]\}$.

Ideally, the error should be zero mean, white, and uncorrelated with the output signal

Solution: Minimise error power $E\{e^2[n]\}$ by selecting suitable $\{h(k)\}$

- Cost function: $J = E\left\{\left(d[n] - \sum_k h(k)x[n-k]\right)^2\right\}$

- Setting $\nabla_h J = 0$ for $h = h(i)$, gives (you will see more detail later)

$$E\{d[n]x[n-i]\} - \sum_k h(k)E\{x[n-k]x[n-i]\} = 0$$

- The solution $r_{dx}(-i) = \sum_k h(k)r_{xx}(i-k)$ in vector form is

$$\mathbf{h} = \mathbf{R}^{-1}\mathbf{r}_{dx}$$

$\Rightarrow$ The optimal coefficients are **inversely proportional** to the autocorrelation matrix and **directly proportional** to the estimate of the crosscorrelation.

Independence, uncorrelatedness and orthogonality

- Two RV are **independent** if the realisation of one does not affect the distribution of the other, consequently, the joint density is separable:

  $$p(x, y) = p(x)p(y)$$

  Example: Sunspot numbers on 31 December and Argentinian debt

- Two RVs are **uncorrelated** if their cross-covariance is zero, that is

  $$c(x, y) = E[(x - \mu_x)(y - \mu_y)] = E[xy] - E[x]E[y] = 0$$

  Example: $x \sim \mathcal{N}(0, 1)$ and $y = x^2$ (impossible to relate through a linear relationship)

- Two RV are **orthogonal** if $r(x, y) = E[xy] = 0$

  Example: Two uncorrelated RVs with at least one of them zero-mean

Independence, uncorrelatedness and orthogonality - Properties

- **Independent** RVs are always uncorrelated

- **Uncorrelatedness** can be seen as a 'weaker' form of independence since only the expectation (rather than the density) needs to be separable.

- **Uncorrelatedness** is a measure of **linear** independence. For instance, $x \sim \mathcal{N}(0,1)$ and $y = x^2$ are clearly **dependent but uncorrelated**, meaning that there is no **linear** relationship between them

- Since $c_{xy} = r_{xy} - m_x m_y$ orthogonal RVs x and y need not be uncorrelated. Furthermore:

  – **uncorrelated** if they are independent and one them is zero mean
  – **orthogonal** if they are uncorrelated and one them is zero mean

- For uncorrelated random variables: $\text{var}\{x + y\} = \text{var}\{x\} + \text{var}\{y\}$

Stationarity: Strict and wide sense

- **Strict Sense Stationarity (SSS):** The process $\{x[n]\}$ is SSS if for all k the joint distribution $p(x[n_1], \ldots, x[n_k])$ ***is invariant under time shifts***, i.e. (all moments considered)

$$p\left(x[n_1+n_0], \ldots, x[n_k+n_0]\right) = p\left(x[n_1], \ldots, x[n_k]\right), \forall n_0$$

  As SSS is too strict for practical applications, we consider the more 'relaxed' stationarity condition: .

- **Wide-Sense Stationarity (WSS):** The process $\{x[n]\}$ is WSS if $\forall m, n$:
  - Mean: $E\{x[m]\} = E\{x[m+n]\}$,
  - Covariance: $c(m,n) = c(m-n,0) = c(m-n)$

  Note that only the first two moments are considered.

Example of WSS: $x[n] = \sin(2\pi f n + \phi)$, where ϕ is uniformly distributed on $[-\pi, \pi]$

Autocorrelation function $r(m)$ of WSS processes

i) **Time/shift invariant:** $r(m,n) = r(m-n,0) = r(m-n)$ (follows from the covariance WSS requirement)

ii) **Symmetric:** $r(-m) = r(m)$ (follows from the definition)

iii) $r(0) \geq |r(m)|$ (maximum at $m = 0$)

The signal power $= r(0) \looparrowright$ Parseval's relationship

Follows from $E\{(x[n] - \lambda x[n+m])^2\} \geq 0$, i.e.

$$E\{x^2[n]\} - 2\lambda E\{x[n]x[n+m]\} + \lambda^2 E\{x^2[n+m]\} \geq 0 \qquad \forall \lambda$$

$$r(0) - 2\lambda r(m) + \lambda^2 r(0) \geq 0 \qquad \forall \lambda$$

which is quadratic in λ and required to be positive for all λ, i.e. the equation determinant: $\Delta = r^2(m) - r(0)r(0) \leq 0 \Rightarrow r(0) \geq |r(m)|$.

Properties of ACF – continued

iv) The AC matrix for a stationary $\mathbf{x} = [x[0], \ldots, x[L-1]]^T$ is

$$\mathbf{R} = E\{\mathbf{x}\mathbf{x}^T\} = E\left\{ \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{L-1} \end{bmatrix} [x_0, x_1, \ldots, x_{L-1}] \right\} = \begin{bmatrix} r(0) & r(1) & \cdots & r(L-1) \\ r(1) & r(0) & \cdots & \vdots \\ \vdots & \vdots & \ddots & r(1) \\ r(L-1) & r(L-2) & \cdots & r(0) \end{bmatrix}$$

is **symmetric and Toeplitz**.

v) $\mathbf{R}$ is **positive semi–definite**, that is

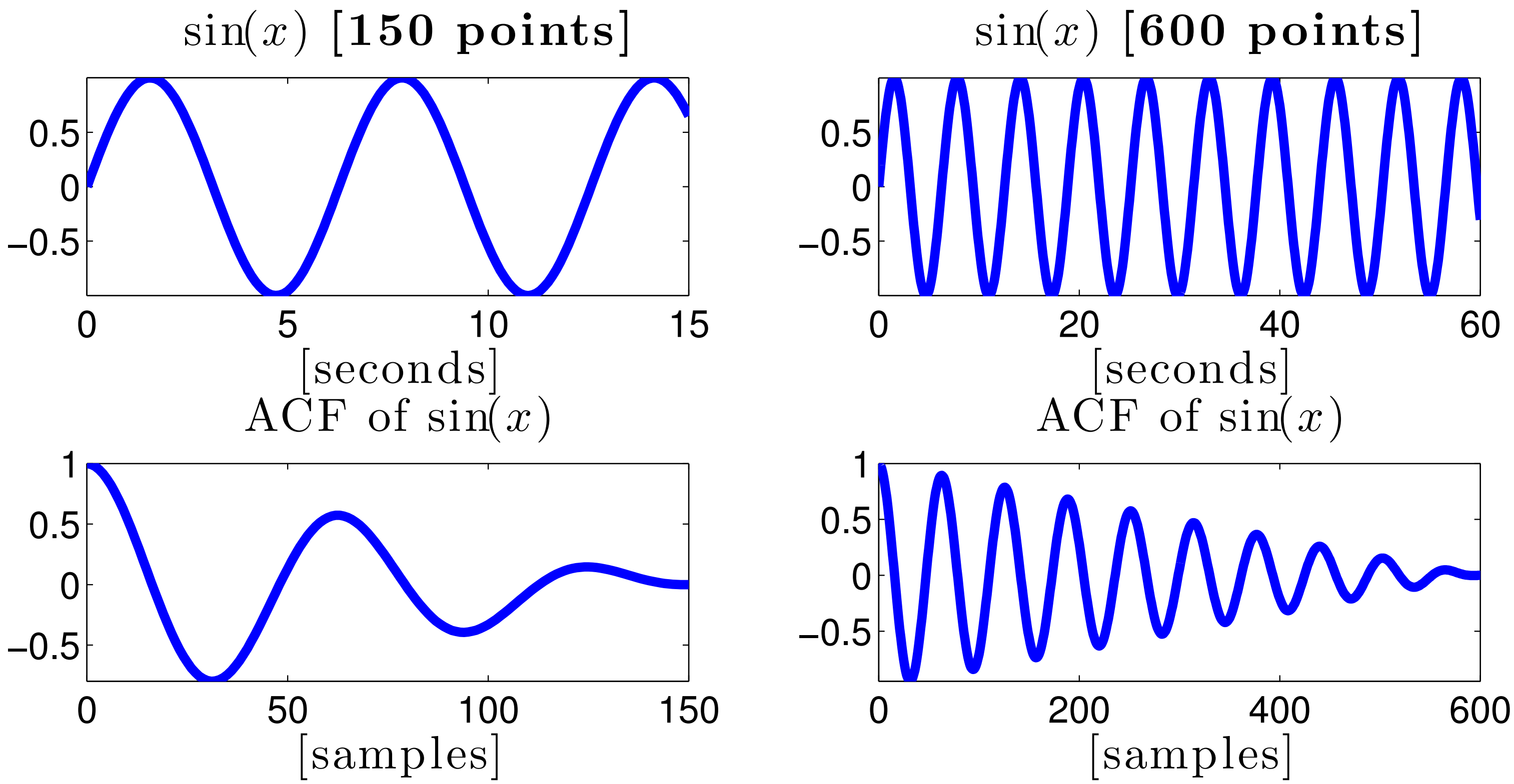
$$\mathbf{a}^T\mathbf{R}\mathbf{a} \geq 0 \qquad \forall \mathbf{a} \neq 0$$

which follows from $y = \mathbf{a}^T\mathbf{x}$ and $y^T = \mathbf{x}^T\mathbf{a}$ so that

$$E\{y^2[n]\} = E\{y[n]y^T[n]\} = E\{\mathbf{a}^T\mathbf{x}\mathbf{x}^T\mathbf{a}\} = \mathbf{a}^T E\{\mathbf{x}\mathbf{x}^T\}\mathbf{a} = \mathbf{a}^T\mathbf{R}\mathbf{a} \geq 0$$

Properties of $r(m)$ – contd II

vi) The autocorrelation function **reflects the basic shape of a signal**, for instance if the signal is periodic, the autocorrelation function will also be periodic



Sinewave and its ACF - Sampling rate=10Hz

Properties of the crosscorrelation

i) $r_{xy}(m) = E\{x[n]y[n+m]\} = r_{yx}(-m)$ (accounts for the lead/trail signal - see also the radar principle in Example 1.6)

ii) If $z[n] = x[n] + y[n]$ then

$$\begin{aligned} r_{zz}(m) &= E\left\{(x[n]+y[n])(x[n+m]+y[n+m]\right\} \\ &= r_{xx}(m) + r_{yy}(m) + r_{xy}(m) + r_{yx}(m) \end{aligned}$$

and if $x[n]$ and $y[n]$ are independent or uncorrelated

$$r_{zz}(m) = r_{xx}(m) + r_{yy}(m)$$

(therefore for $m = 0$ we have var(z) = var(x) + var y)

iii) $r_{xy}^2(m) \leq r_{xx}(0)r_{yy}(0)$ (Same as ACF P(iii) when $x = y$)

Example 1.6. The use of (cross-)correlation

Detection of Tones in Noise:

Consider a noisy tone $x = A\cos(\omega n + \theta)$

$$y[n] = A\cos(\omega n + \theta) + w[n]$$

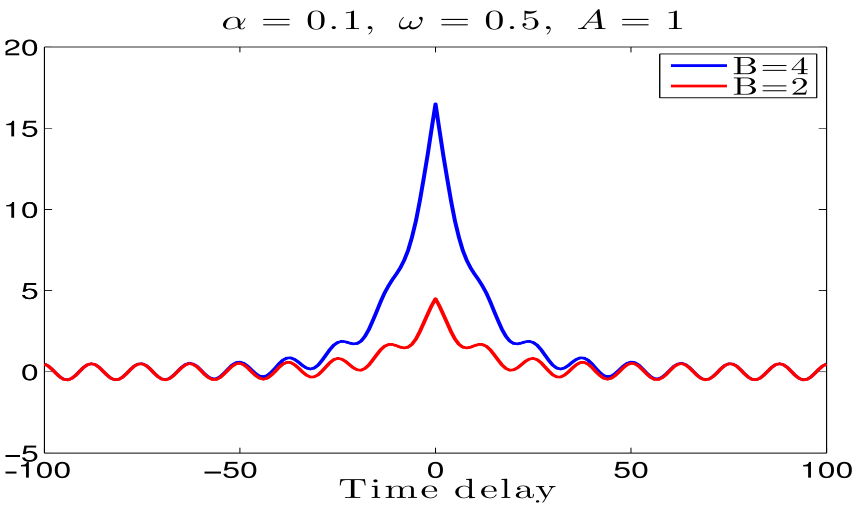
$$\textbf{ACF}: \; R(m) = E\big[y[n]y[n+m]\big] =$$

$$= R_x(m) + R_w(m) + R_{xw}(m) + R_{wx}(m)$$

For $R_w = B^2 \exp(-\alpha|m|)$ & $x \perp w$, then

$$R_y(m) = \frac{1}{2}A^2\cos(\omega m) + B^2\exp(-\alpha|m|)$$

- for large m, the ACF $\propto$ the signal
- $\exists$ extract tiny signal from large noise



Principle of Radar:

The received signal (see previous slide)

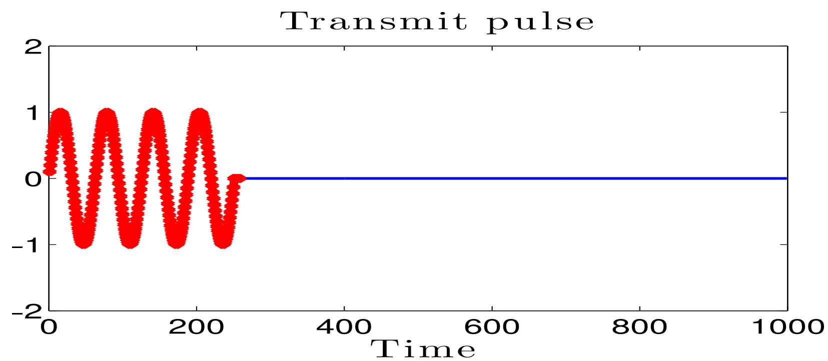
$$y[n] = ax[n - T_0] + w[n], \quad \text{so that}$$

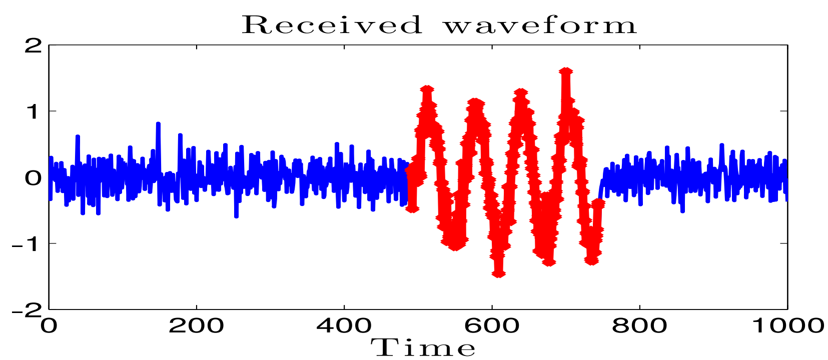
$$R_{xy}(\tau) = E\{x(n)y(n+\tau)\}$$

$$= aR_x(\tau - T_0) + R_{xw}(\tau)$$

Since

$$\boldsymbol{x \perp w \rightsquigarrow R_{xy}(\tau) = aR_x(\tau - T_0)}$$





Example 1.7. Range of a radar

Unbiased estimate of a true radar delay δ_0 that has distribution $\delta \sim \mathcal{N}(\delta_0, \sigma_0^2)$

Q: What is the distribution of the range of the radar, and how should the radar be designed (i.e. what should σ_0 be) so that the range estimate is within 100 of the actual range with a probability of 99%?

A: The range is given by $R = \delta \frac{C}{2}$, therefore, $R \sim \mathcal{N}(\delta_0 \frac{C}{2}, \sigma_0^2 \frac{C^2}{4})$, where $R_0 = \delta_0 \frac{C}{2}$ is the actual true range.

To fulfil the radar design requirement, we need, $\mathbb{P}\{|R - R_0| < 100\} = 0.99$, or equivalently (due to the symmetry of the RV R)

$$\mathbb{P}\left\{\frac{(R - R_0)}{\sigma_0^2 C/2} < \frac{100}{\sigma_0^2 C/2}\right\} = 0.995,$$

and as $\frac{(R-R_0)}{\sigma_0^2 C/2} \sim \mathcal{N}(0, 1)$, we have $P\left(\frac{100}{\sigma_0^2 C/2}; 1, 0\right) = 0.995$. Evaluating this from the expression of the Gaussian CDF in an earlier slide we have

$$\frac{100}{\sigma_0^2 C/2} = 2.58 \Rightarrow \sigma_0 = \sqrt{\frac{200}{2.58 \times 3 \times 10^8}} = 0.51\text{milliseconds}$$

NB: By dividing $\mathcal{N}(0, \sigma)$ with σ we standardise pdf to unit variance $\mathcal{N}(0, 1)$.

Power spectral density (PSD)

The **power spectrum** or **power spectral density** $S(f)$ of a process $\{x[n]\}$ is the Fourier transform of its ACF (Wiener–Khintchine Theorem)

$$S(f) = \mathcal{F}\{r_{xx}(m)\} = \sum_{m=-\infty}^{\infty} r_{xx}(m) e^{-\jmath 2\pi n f} \qquad f \in (-1/2, 1/2], \omega \in (-\pi, \pi]$$

The sampling period T is assumed to be unity, thus f is the *normalised frequency*.

From the inversion formula (Fourier), we can write

$$r_{xx}(m) = \int_{-1/2}^{1/2} S(f) e^{\jmath 2\pi m f} df$$

$\Rightarrow$ ACF tells us about the power within the signal (Average)
$\Rightarrow$ PSD tell us about the distribution of power across frequencies (Density)

PSD properties

i) $S(f)$ is a **positive real** function **(it is a distribution)** $\Rightarrow S(f) = S^*(f)$.
Since $r(-m) = r(m)$ we can write

$$S(f) = \sum_{m=-\infty}^{\infty} r_{xx}(-m) e^{\jmath 2\pi m f} = \sum_{m=-\infty}^{\infty} r_{xx}(m) e^{-\jmath 2\pi m f}$$

and hence

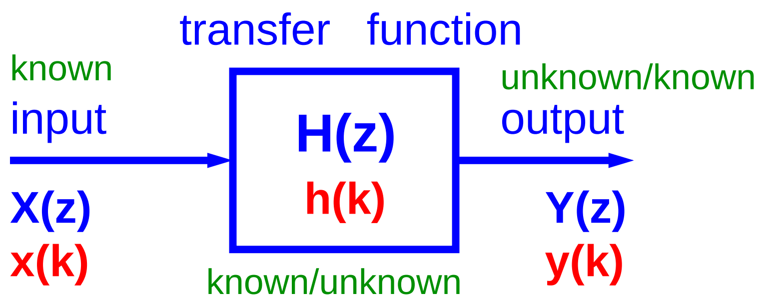
$$S(f) = \sum_{m=-\infty}^{\infty} r_{xx}(m) \cos(2\pi m f) = r_{xx}(0) + 2 \sum_{m=1}^{\infty} r_{xx}(m) \cos(2\pi m f)$$

ii) $S(f)$ is a **symmetric** function, $S(-f) = S(f)$. This follows from the last expression.

iii) $r(0) = \int_{-1/2}^{1/2} S(f) df = E\{x^2[n]\} \geq 0$.

$\Rightarrow$ **the area below the PSD (power spectral density) curve = Signal Power.**

Linear systems



$$H(z) = \frac{Y(z)}{X(z)}$$

Described by their impulse response $h(n)$ or the transfer function $H(z)$

In the frequency domain (remember that $z = e^{\jmath\theta}$) the transfer function is

$$H(\theta) = \sum_{n=-\infty}^{\infty} h(n)e^{-\jmath n\theta} \qquad \{x[n]\} \rightarrow \left| \begin{array}{c} \{h(n)\} \\ H(\theta) \end{array} \right| \rightarrow \{y[n]\}$$

that is $\qquad y[n] = \sum_{r=-\infty}^{\infty} h(r)x[n-r] = h * x$

Example 1.8. Linear systems – statistical properties ↬ mean and variance

i) **Mean**

$$E\{y[n]\} = E\left\{\sum_{r=-\infty}^{\infty} h(r)x[n-r]\right\} = \sum_{r=-\infty}^{\infty} h(r)E\{x[n-r]\}$$

$$\Rightarrow \qquad \mu_y = \mu_x \sum_{r=-\infty}^{\infty} h(r) = \mu_x H(0)$$

[NB: $H(\theta) = \sum_{r=-\infty}^{\infty} h(r)e^{-jr\theta}$. For $\theta = 0$, then $H(0) = \sum_{r=-\infty}^{\infty} h(r)$]

ii) **Cross–correlation**

$$r_{yx}(m) = E\{y[n]x[n+m]\} = \sum_{r=-\infty}^{\infty} h(r)E\{x[n-r]x[n+m]\}$$

$$= \sum_{r=-\infty}^{\infty} h(r)r_{xx}(m+r) \quad \textbf{convolution of input ACF and } \{\mathbf{h}\}$$

$\Rightarrow$ **Cross-power spectrum** $S_{yx}(f) = \mathcal{F}(r_{yx}) = S_{xx}(f)H(f)$

Example 1.9. Linear systems – statistical properties $\looparrowright$ crosscorrelation (this will be used in AR spectrum)

From $r_{xy}(m) = r_{yx}(-m)$ we have
$r_{xy}(m) = \sum_{r=-\infty}^{\infty} h(r) r_{xx}(m-r)$. Now we write

$$r_{yy}(m) = E\{y[n]y[n+m]\} = \sum_{r=-\infty}^{\infty} h(r)E\{x[n-r]y[n+m]\}$$

$$= \sum_{r=-\infty}^{\infty} h(r) r_{xy}(m+r) = \sum_{r=-\infty}^{\infty} h(-r) r_{xy}(m-r)$$

by taking Fourier transforms we have

$$\begin{aligned} S_{xy}(f) &= S_{xx}(f)H(f) \\ S_{yy}(f) &= S_{xy}(f)H(-f) = \mathcal{F}(r_{xx}) \end{aligned}$$

or

$$\mathbf{S_{yy}(f) = H(f)H(-f)S_{xx}(f) = |H(f)|^2 S_{xx}(f)}$$

Output power spectrum = input power spectrum $\times$ squared transfer function

Crosscorrelation and cross–PSD (recap)

- CC of two jointly WSS discrete time signals **(this is not symmetric)**

$$r_{xy}(m) = E\{x[n]y[n+m]\} = r_{yx}(-m)$$

- For $z[n] = x[n] + y[n]$ where $x[n]$ and $y[n]$ are zero mean and independent, we have $r_{xy}(m) = r_{yx}(m) = 0$, therefore

$$\begin{aligned} r_{zz}(m) &= r_{xx}(m) + r_{yy}(m) + r_{xy}(m) + r_{yx}(m) \\ &= r_{xx}(m) + r_{yy}(m) \end{aligned}$$

- Cross Power Spectral Density

$$P_{xy}(f) = \mathcal{F}\{r_{xy}(m)\}$$

Generally a complex quantity and so will contain both the **magnitude** and **phase** information.

Special signals: a) White noise

If the joint pdf is separable

$$p(x[0], x[1], \ldots x[n]) = p(x[0])p(x[1]) \cdots p(x[n]) \quad \forall n$$

where the pdf's $p(x[r])$ are identical $\forall r$, then all the pairs $x[n], x[m]$ are **independent** and $\{x[n]\}$ is said to be an **independent identically distributed (iid) signal.**

Since independent samples $x[n]$ are also uncorrelated, then for a zero–mean signal we have

$$r(n-m) = E\{x[m]x[n]\} = \sigma^2 \delta(n-m)$$

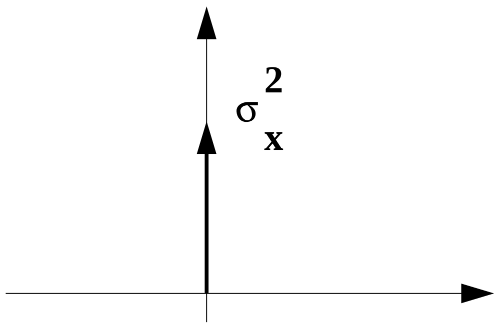
where the variance (signal power) $\sigma^2 = E\{x^2[n]\}$ and

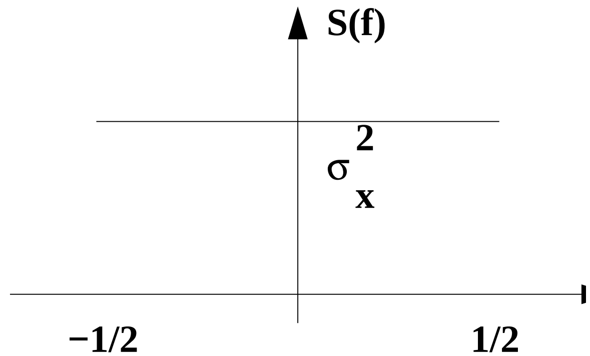
$$\delta(n-m) = \begin{cases} 1 & n = m \\ 0 & \text{elsewhere} \end{cases}$$

where $\delta(n)$ is the Kronecker delta operator

Example 1.10. ACF and power spectrum of white noise

The Fourier transform of WN is constant for all frequencies, hence "white".





- The autocorrelation matrix

$$\mathbf{R} = \sigma^2 \mathbf{I} \qquad r(m) = \sigma_x^2 \delta(m)$$

  Since $E\{x[n]x[n-1]\} = 0$, the variance $r(0) = \sigma_x^2$ is the power of WN.

- The shape of the pdf $p(x[n])$ determines whether the white noise is called Gaussian (WGN), uniform (UWN), Poisson, Laplacian, etc.

  From the Wiener–KhinchineTheorem:

PSD(White Noise) = FT(ACF(WN)) = FT($\delta(t)$ function) = constant

b) First order Markov signals: autoregressive modelling

(finite memory in the description of a random signal)

If instead of the iid condition, we have the **first order conditional expectation**, then

$$p(x[n], x[n-1], x[n-2], \ldots, x[0]) = p(x[n]|x[n-1])$$

where $p(a|b)$ is defined as the pdf of "a" *conditioned* upon the (possibly correlated) observation "b"

$\Rightarrow$ the signal above is the **first order Markov signal**.

Example: Examine the statistical properties of the signal given by

$$y[n] = ay[n-1] + w[n]$$

where $a = 0.8$ and $w[n] \sim \mathcal{N}(0, 1)$ (see your coursework).

c) Minimum phase signals

Let $\{x[n]\}$ be observed for $n = 0, 1, \ldots, N-1$.

$$X(z) = x[0] + x[1]z^{-1} + \cdots + x[N-1]z^{-(N-1)} =$$

$$A \prod_{i=1}^{N} \left(1 - z_i z^{-1}\right), \quad A(0) = x[0]$$

- $|z_i| \leq 1, \quad \forall i$ then $X(z)$ is said to be *minimum phase*
- $|z_i| \geq 1, \quad \forall i$, then $X(z)$ is said to be *maximum phase*
- $|z_i| \geq 0$ for some i while for others $|z_i| \leq 1$ then $X(z)$ is said to be of *mixed phase*.

In DSP, the algorithms often rely on the minimum phase property of a signal for stability (of the inverse system) and to be able to have real-time implementation (causality).

d) Gaussian random signals

A signal in which each of the L samples is Gaussian distributed

$$p(x[i]) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{(x[i]-\mu(i))^2}{2\sigma_i^2}} \qquad i = 0, \ldots, L-1$$

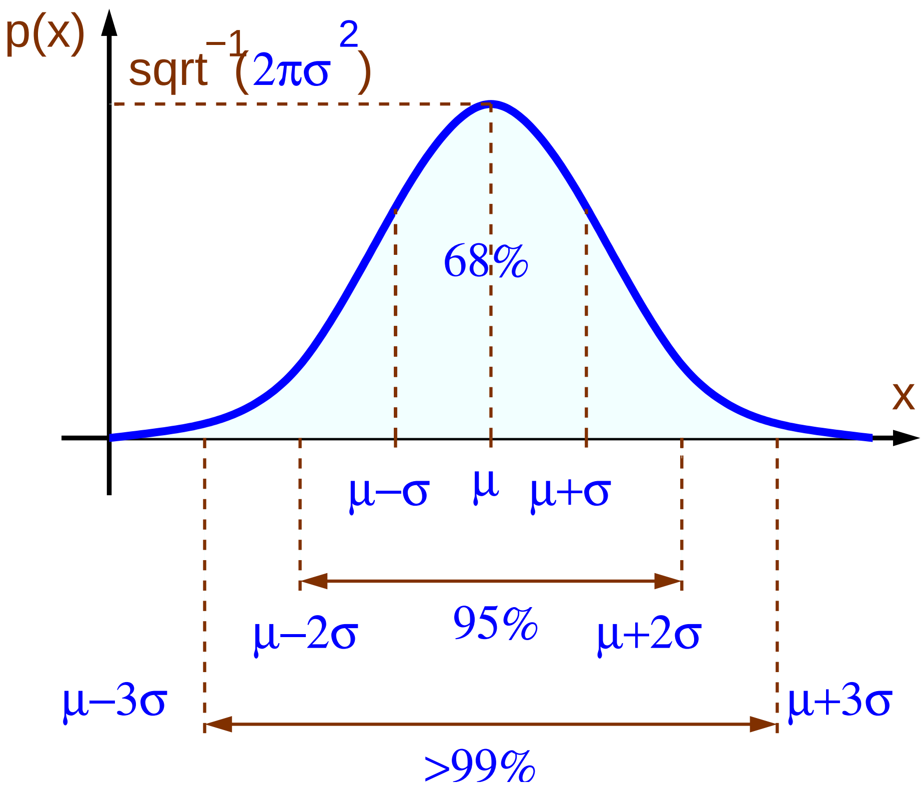
denoted by $\mathcal{N}(\mu(i), \sigma_i^2)$.

The joint pdf of L samples $x[n_0], x[n_1], \ldots, x[n_{L-1}]$ is then

$$\begin{aligned} p(\mathbf{x}) &= p(x[n_0], x[n_1], \ldots, x[n_{L-1}]) \\ p(\mathbf{x}) &= \frac{1}{[2\pi]^{L/2} det(\mathbf{C})^{1/2}} e^{\frac{(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}} = \frac{1}{(2\pi\sigma^2)^{L/2}} e^{\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{L-1}(x[n]-\mu)^2\right]} \end{aligned}$$

where $\mathbf{x} = [x[n_0], x[n_1], \ldots, x[n_{L-1}]]$, $\boldsymbol{\mu} = [\mu[n_0], \mu[n_1], \ldots, \mu[n_{L-1}]]$ and $\mathbf{C}$ is a covariance matrix with determinant Δ.

e) Properties of a Gaussian distribution



1) If x and y are jointly Gaussian, then for any constants a and b the random variable

$$z = ax + by$$

is Gaussian with mean

$$m_z = am_x + bm_y$$

and variance

$$\sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\sigma_x\sigma_y\rho_{xy}$$

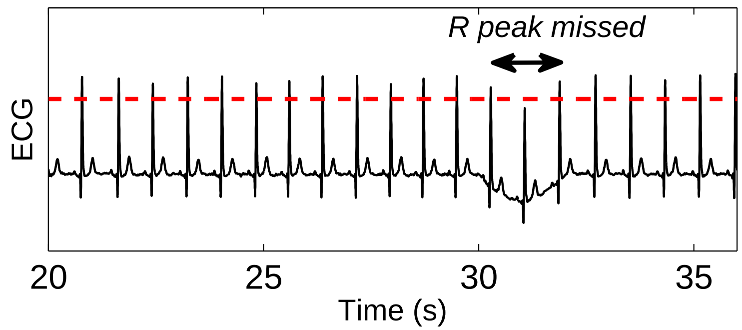
2) If two jointly Gaussian random variables are *uncorrelated* ($\rho_{xy} = 0$) then they are statistically independent,

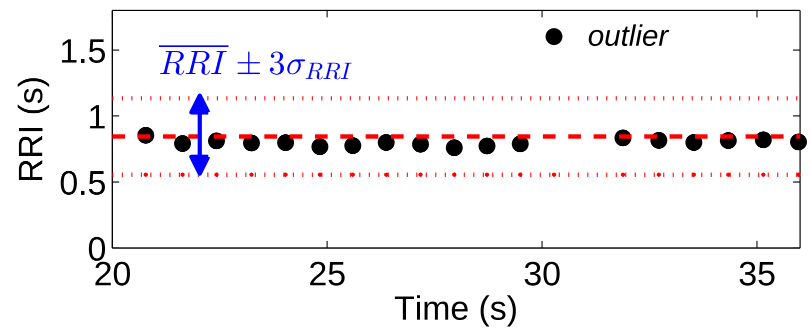
$$f_{x,y} = f(x)f(y)$$

For $\mu = 0$, $\sigma = 1$, the inflection points are ± 1

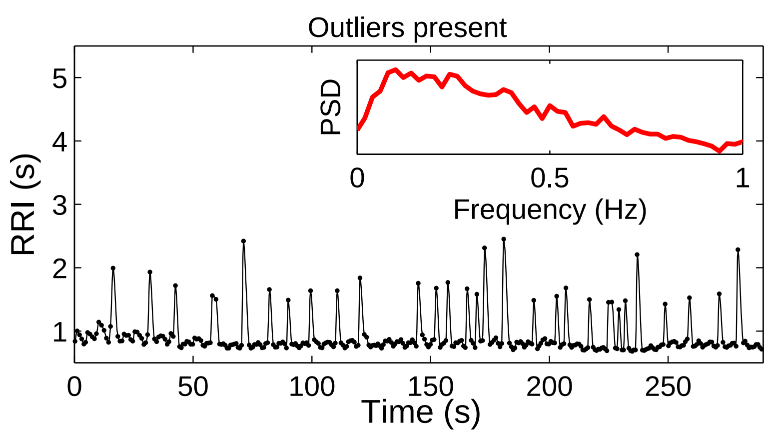
Example 1.11. Rejecting outliers from cardiac data

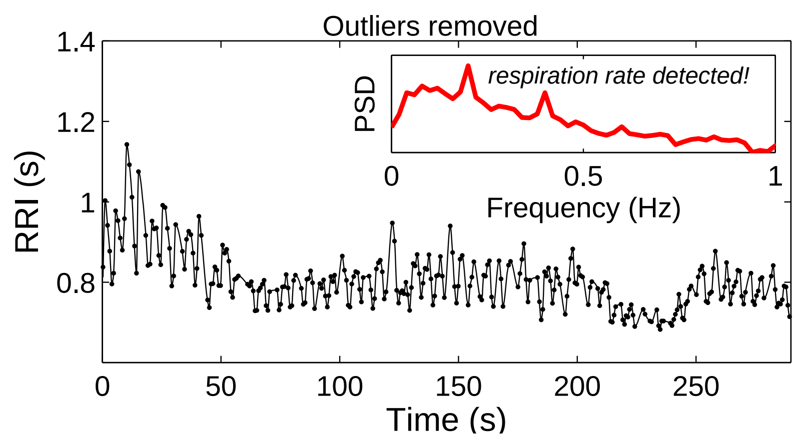
○ Failed detection of R peaks in ECG [left] causes outliers in R-R interval (RRI, time difference between consecutive R peaks) [right]





○ No clear outcome from PSD analysis of outlier-compromised RRI [left], but PSD of RRI with outliers removed reveals respiration rate [right]





f) Conditional mean estimator for Gaussian random variables

3) If x and y are jointly Gaussian random variables then the optimum estimator for y, given by

$$\hat{y} = g(x)$$

that minimizes the mean square error $\xi = E\{[y = g(x)]^2\}$ is a **linear estimator** in the form

$$\hat{y} = ax + b$$

4) If x is Gaussian with zero mean then

$$E\{x^n\} = \begin{cases} 1 \times 3 \times 5 \times \cdots \times (n-1)\sigma_x^n, & n \quad \text{even} \\ 0, & n \quad \text{odd} \end{cases}$$

e) Ergodic signals

In practice, we often have only one observation of a signal (real–time)

Then, **statistical averages** are replaced by **time averages**.

This is necessary because

- ensemble averages are generally unknown a priori
- only a single realisation of the random signal is often available

Thus, the ensemble average

$$m_x(n) = \frac{1}{L}\sum_{i=1}^{L} x_i(n)$$

is therefore replaced by a time average

$$m_x(N) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)$$

e) Ergodic signals – Example

Consider the random process

$$x(n) = A\cos(n\omega_0)$$

where A is a random variable that is equally likely to assume the value of 1 or 2.

The mean of this process is

$$E\{x(n)\} = E\{A\}\cos(n\omega_0) = 1.5\cos(n\omega_0)$$

However, for a single realisation of this process, for large N, the sample mean is approximately zero

$$m_x \approx 0, \qquad N >> 1$$

$\Rightarrow$ $x(n)$ **is not ergodic and therefore the statistical expectation cannot be computed using time averages on a single realisation.**

e) Ergodicity in the mean

DEF: If the sample mean $\hat{m}_x(N)$ of a WSS process converges to m_x, in the mean–square sense, then the process is said to be **ergodic in the mean**, and we write

$$\lim_{N\to\infty} \hat{m}_x(N) = m_x$$

For the convergence of the sample mean in the mean–square sense

- Asymptotically unbiased

$$\lim_{N\to\infty} E\{\hat{m}_x(N)\} = m_x$$

Consider the variance of the estimate $\to 0$ as $N \to \infty$

$$\lim_{N\to\infty} Var\{\hat{m}_x(N)\} = 0 \qquad \text{(consistent)}$$

e) Ergodicity - Summary

In practice, it is necessary to assume that the single realisation of a discrete time random signal satisfies ergodicity in the mean and autocorrelation.

Mean Ergodic Theorem: Let $x[n]$ be a wide sense stationary (WSS) random process with autocovariance sequence $c_x(k)$, sufficient conditions for $x[n]$ to be ergodic in the mean are that $c_x(k) < \infty$ and

$$\lim_{k \to \infty} c_x(k) = 0$$

Autocorrelation Ergodic Theorem:mean A necessary and sufficient condition for a WSS Gaussian process with covariance $c_x(k)$ to be autocorrelation ergodic is

$$\lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} c_x(k) = 0$$

Taylor series expansion

Most 'smooth' functions can be expanded into their Taylor Series Expansion (TSE)

$$f(x) = f(x_0) + \frac{f'(x_0)}{1}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots = \sum_{n=1}^{\infty} \frac{f^{(n)}(x_0)}{n!}$$

To show this consider the polynomial

$$f(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + a_3(x - x_0)^3 + \cdots$$

1. To get a_0 $\looparrowright$ choose $x = x_0$ $\Rightarrow$ $a_0 = f(x_0)$
2. To get a_1 $\looparrowright$ take derivative of the polynomial above to have

$$\frac{d}{dx}f(x) = a_1 + 2a_2(x - x_0) + 3a_3(x - x_0)^2 + 4a_4(x - x_0)^4 + \cdots$$

choose $x = x_0$ $\Rightarrow$ $a_1 = \frac{df(x)}{dx}|_{x=x_0}$ and so on ... $a_k = \frac{1}{k!}\frac{d^k f(x)}{dx^k}|_{x=x_0}$

Power series - contd.

Consider

$$f(x)=\sum_{n=0}^{\infty}a_nx^n \;\Rightarrow\; f'(x)=\sum_{n=1}^{\infty}na_nx^{n-1} \text{ and } \int_0^x f(t)dt=\sum_{n=0}^{\infty}\frac{a_n}{n+1}x^{n+1}$$

1. Exponential function, cosh, sinh, sin, cos, ...

$$e^x=\sum_{n=0}^{\infty}\frac{x^n}{n!} \quad \text{and} \quad e^{-x}=\sum_{n=0}^{\infty}(-1)^n\frac{x^n}{n!} \quad \Rightarrow \quad \frac{e^x-e^{-x}}{2}=\sum_{n=0}^{\infty}\frac{x^{2n}}{(2n)!}$$

2. other useful formulas

$$\sum_{n=0}^{\infty}x^n=\frac{1}{1-x} \quad \Rightarrow \quad \sum_{n=1}^{\infty}nx^{n-1}=\frac{1}{(1-x)^2} \quad \text{and} \quad \frac{1}{1+x^2}=\sum_{n=0}^{\infty}(-1)^n nx^2$$

Integrate to obtain $atan(x)=\sum_{n=0}^{\infty}(-1)^n\frac{x^{2n+1}}{2n+1}$.

For $x=1$ we have $\frac{\pi}{4}=1=1/3+1/5-1/7+\cdots$

Numerical derivatives - examples

- Two-point approximation:
  - Forward: $f'(0) = \frac{f(1)-f(0)}{h}$
  - Backward: $f'(0) = \frac{f(-1)-f(0)}{h}$

- Three-point approximation:
  - $f'(0) = \frac{f(1)-2f(0)+f(-1)}{2h}$
  - $f''(0) = \frac{f(1)-2f(0)+f(-1)}{h^2}$

- Five-point approximation (also look up for stencil):
  - $f'(0) = \frac{f(-2)-8f(-1)+8f(1)-f(2)}{12h}$
  - $f''(0) = \frac{-f(-2)+16f(-1)-30f(0)+16f(1)-f(2)}{12h^2}$

Constrained optimisation using Lagrange multipliers: Basic principles

Consider a two-dimensional problem:

$$\text{maximize} \qquad \underbrace{f(x,y)}_{function\ to\ max/min}$$

$$\text{subject to} \qquad \underbrace{g(x,y)=c}_{constraint}$$

$\looparrowright$ **we look for point(s) where curves f & g touch (but do not cross).**

In those points, the tangent lines for f and g are parallel $\Rightarrow$ so too are the gradients $\boldsymbol{\nabla_{x,y} f} \parallel \boldsymbol{\lambda \nabla_{x,y} g}$, where λ is a scaling constant.

Although the two gradient vectors are parallel they can have different magnitudes

Therefore, we are looking for $\max$ or $\min$ points (x,y) of $f(x,y)$ for which

$$\nabla_{x,y} f(x,y) = -\lambda \nabla_{x,y} g(x,y) \quad \text{where } \nabla_{x,y} f = \Big(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\Big) \text{ and } \nabla_{x,y} g = \Big(\frac{\partial g}{\partial x}, \frac{\partial g}{\partial y}\Big)$$

We can now combine these conditions into one equation as:

$$F(x,y,\lambda) = f(x,y) - \lambda\big(g(x,y) - c\big) \quad \text{and solve} \quad \nabla_{x,y,z} F(x,y,\lambda) = \mathbf{0}$$

Obviously, $\nabla_\lambda F(x,y,\lambda) = 0 \quad \Leftrightarrow \quad g(x,y) = c$

The method of Lagrange multipliers in a nutshell

max/min of a function $f(x, y, z)$ where x, y, z are coupled

Since x, y, z **are not independent** there exists a constant $g(x, y, z) = c$

Solution: Form a new function

$F(x, y, z, \lambda) = f(x, y, z) - \lambda\big(g(x, y, z) - c\big)$ and calculate $F'_x, F'_y, F'_z, F'_\lambda$

Set $F'_x, F'_y, F'_z, F'_\lambda = 0$ and solve for the unknown x, y, z, λ.

Example 1: Economics

Two factories, A and B make TVs, at a cost $\boldsymbol{f(x, y) = 6x^2 + 12y^2}$ $(x, y = \#TV \in A, B)$. Minimise the cost of producing 90 TVs, by finding optimal numbers $\#x$ and $\#y$ at factories A and B.

Solution: The constraint g: (x+y=90), so

$\boldsymbol{F(x, y, \lambda) = 6x^2 + 12y^2 - \lambda(x + y - 90)}$

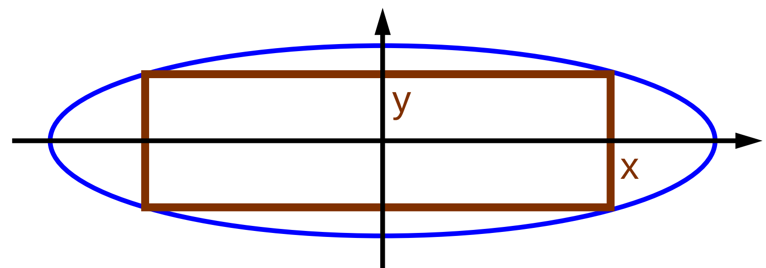
Then: $F'_x = 12x - \lambda$, $F'_y = 24y - \lambda$, $F'_\lambda = -x - y - 90$, and for $\min$ / $\max$ $\nabla F = \mathbf{0}$

Set to zero $\Rightarrow x = 60, y = 30, \lambda = 720$

Example 2: Geometry

Find the rectangle of maximal perimeter, inscribed in the ellipse $x^2 + 4y^2 = 4$.

Solution: Constraint $(x^2 + 4y^2 = 4)$



The perimeter $P(x, y) = 4x + 4y$ so

$F(x, y, \lambda) = 4x + 4y - \lambda(x^2 + 4y^2 - 4)$

$\Rightarrow P'_x = \lambda g'_x, P'_y = \lambda g'_y \looparrowright x = 4y$

Solve to give: $x = 4/\sqrt{5}, P = 4\sqrt{5}$.

Notes:

○

Notes:

○