Statistical Signal Processing & Inference

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 $\not\rightsquigarrow$ **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 $\rightsquigarrow$ **every sample is random too!**

- The sequence is discrete with respect to sample index n, which can be either the standard **discrete time** or some other physical variable, such as the **spatial index** in arrays of sensors

- A random signal $x[n]$ can be real-valued, complex-valued, etc.

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 to amplifier gain or the quality of sensor contact)

Some learning machines require data of a specific mean and variance, yet measured variables are usually of 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 of $\bar{\mu} = 0$, $\bar{\sigma}^2 = 1$. In practice, **raw data** $\{x[n]\}_{n=0:N-1}$ are normalised by subtracting the sample mean, μ, and dividing by the sample standard deviation 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):** $x^{\mathsf{CS}} = \frac{x^{\mathsf{C}}}{\sigma} \quad \looparrowright \quad \mu = 0, \sigma = 1$

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

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

or **bounded** to any lower, l, and upper, u, bound by

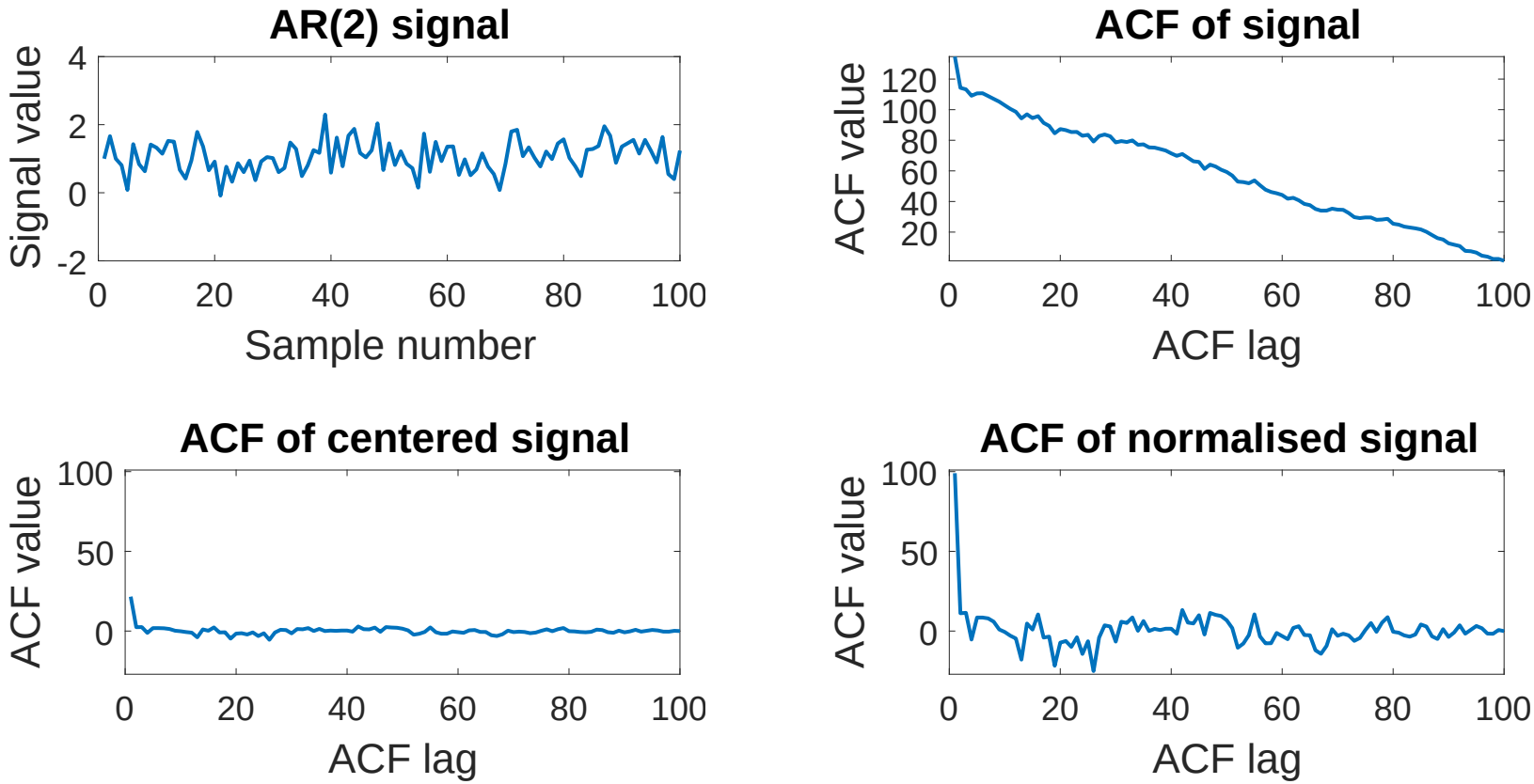
$$x^{\mathsf{ST}} = (u - l)\left(\frac{x(n) - x_{min}}{x_{max} - x_{min}}\right) + l$$

Standardize to zero mean and range $[-1, 1]$ $\looparrowright$ $x(n) = 2\left(\frac{x(n)-x_{min}}{x_{max}-x_{min}}\right) - 1$

Standardisation: Example 1

Autocorrelation under centering and normalisation

Consider an AR(2) signal with the AR parameters: $\mathbf{a} = [0.2, -0.1]^T$

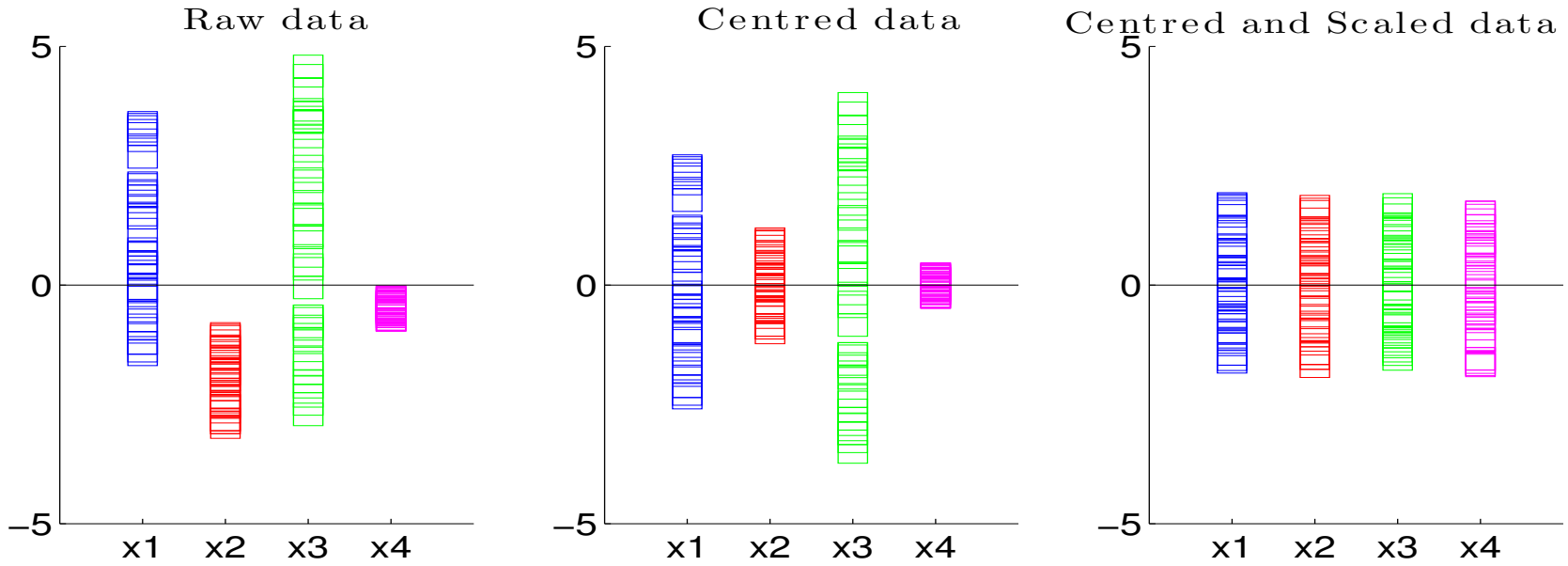


For $\sigma < 1$, normalisation will increase the magnitude the ACF, e.g. for this example $\sigma = 0.5$.

Standardisation: Example 2

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 function $\rightarrow$ convenient and accurate descriptor

- **Cumulative Density Function (CDF)** $\rightarrow$ probability of a random variable falling within a given range, given by

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

$X[n] \rightarrow$ random quantity, $x[n] \rightarrow$ particular fixed value of X.

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

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

☞ **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. $\sin(n) + w(n)$ (w is noise).

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

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

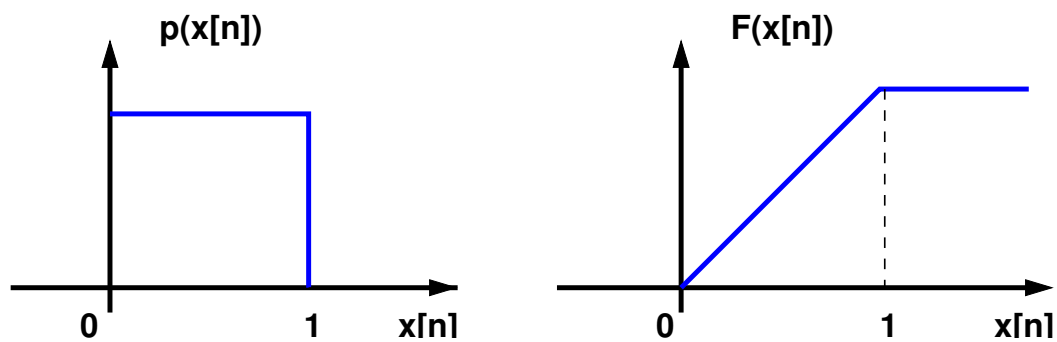
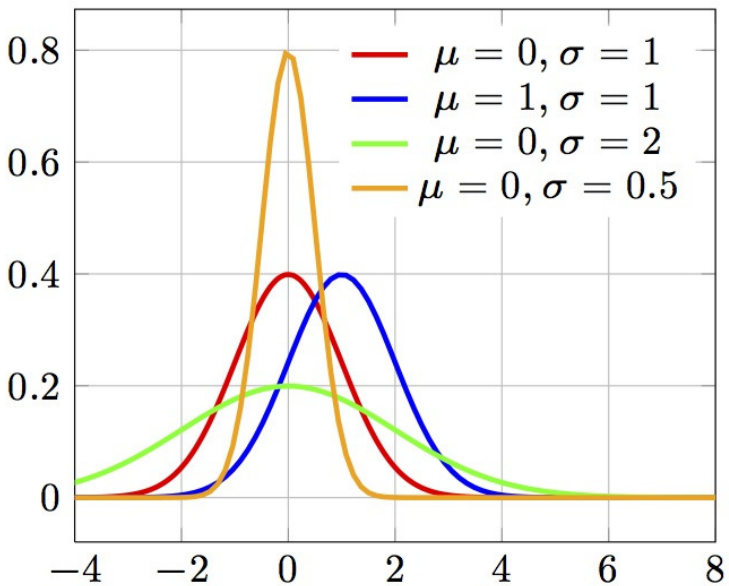


Figure: pdf (left) and CDF (right) for a uniform distribution.

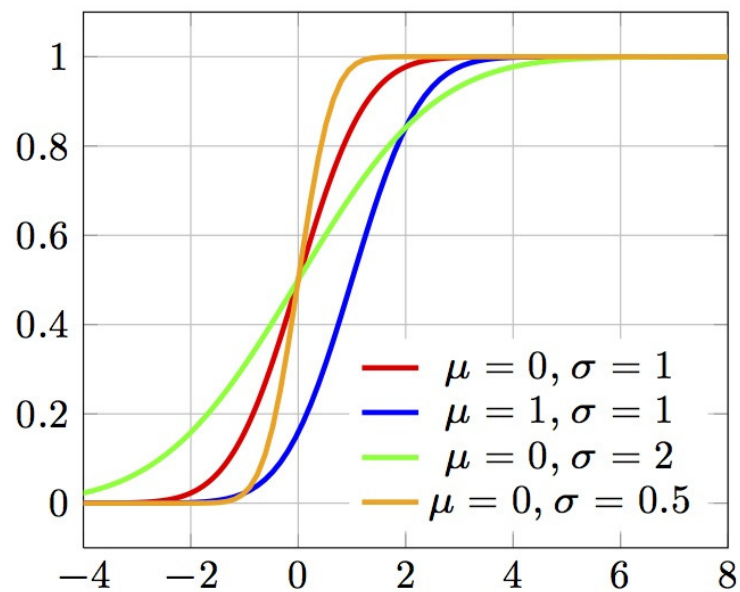
Gaussian probability and cumulative density functions

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

Gaussian pdf



Gaussian CDF



$$p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$P(x; \mu, \sigma) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right]$$

☞ **We read $p(x; \mu, \sigma)$ as 'pdf of x, parametrised by μ and σ'**

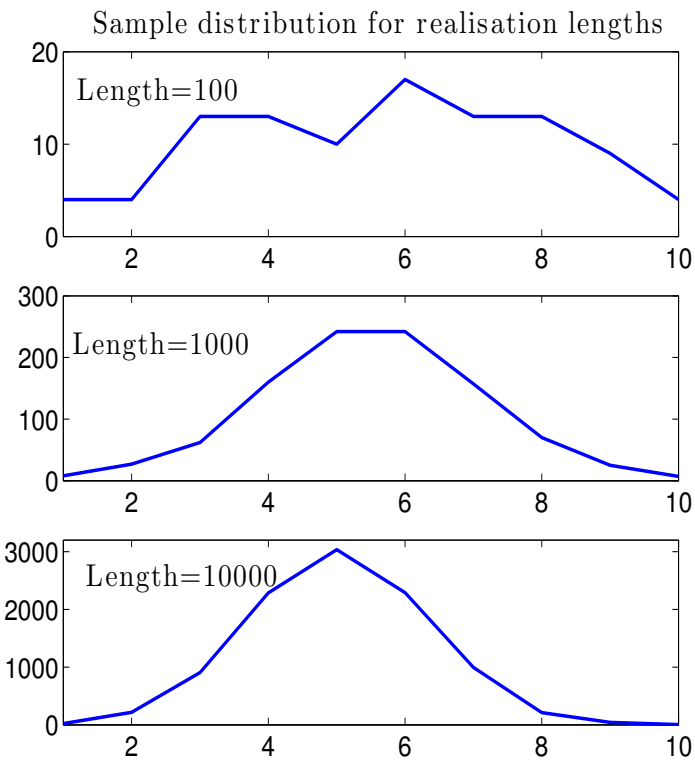
The **standard Gaussian** distribution ($\mu = 0, \sigma = 1$) is given by $p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$

Statistical distributions: Gaussian ↬ randn in Matlab

The accuracy of sample distributions depends on the number of data points

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

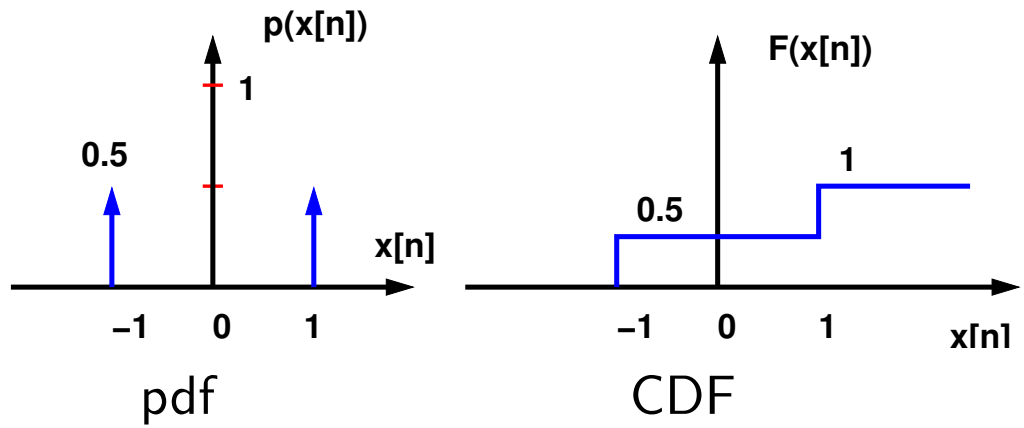
$$p(x[n]) = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{(x[n]-\mu_x)^2}{2\sigma_x^2}} \Rightarrow \log p(x[n]) = -\frac{1}{2}\log\left(2\pi\sigma_x^2\right) - \frac{(x[n]-\mu_x)^2}{2\sigma_x^2}$$



Sample densities for varying N

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

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) ***measure one or more sources with multiple sensors***
- Multidimensional processes (3D inertial bodymotion sensors, radar, vector fields, wind anemometers) – ***intrinsically multidimensional***

Example: Multivariate function with single output (MISO)

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

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

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

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

Joint distributions of delayed samples (temporal)

Joint distribution of $x[n_1]$ and $x[n_2]$ (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 corresponding 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

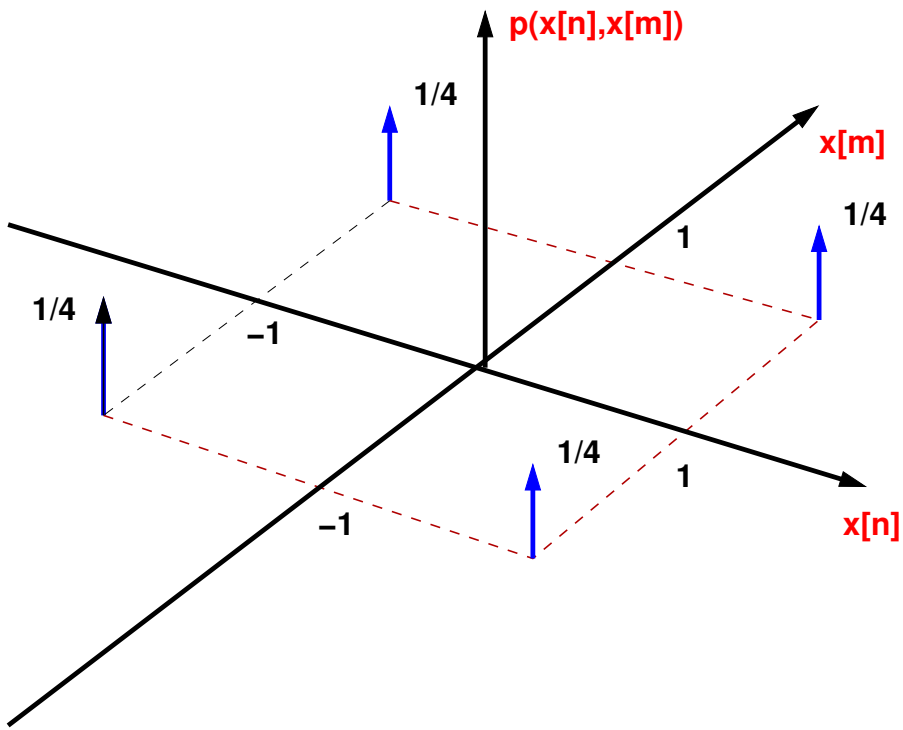
Luckily, real world time series often have "finite memory" (Markov)

Example 1.1. Bivariate pdf

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

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

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



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: **Separability:** $E\{x[m]y[n]\} \neq E\{x[m]\}E\{y[n]\}$

unless $\{x[m]\}$ and $\{y[n]\}$ are independent random processes, that is when $E\{x[m]y[n]\} = E\{x[m]\}E\{y[n]\}$

P3: **Nonlinear transformation of variables:** 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 above)

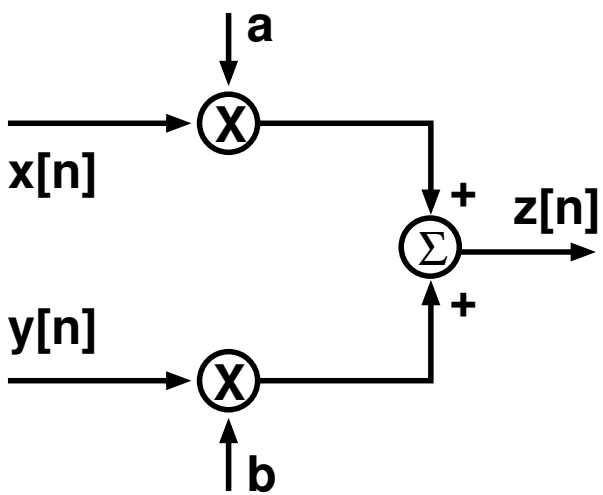
Task: Consider a general linear system given by $z[n] = ax[n] + by[n]$. Find the means, $E\{x[n]\} = \mu_x, E\{y[n]\} = \mu_y$, where $x \perp y$.

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 property is a consequence of the linearity of the $E\{\cdot\}$ operator, and is very useful in the analysis of adaptive learning systems. (see Lecture 7)

Example 1.3. Mean for nonlinear systems (use P3 on Slide 13)

think about e.g. estimating the variance empirically

For a nonlinear system, say the sensor nonlinearity is given by (*cf.* variance)

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

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

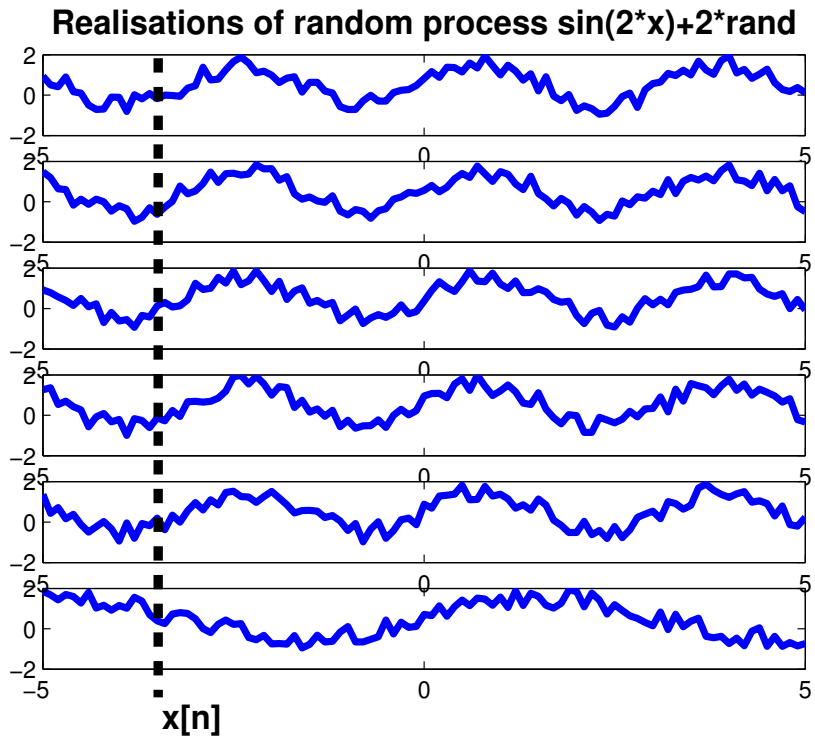
$$\mu_x(n) = \frac{1}{N}\sum_{i=1}^{N} x_i[n]$$

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

For $N \to \infty$ we have

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

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

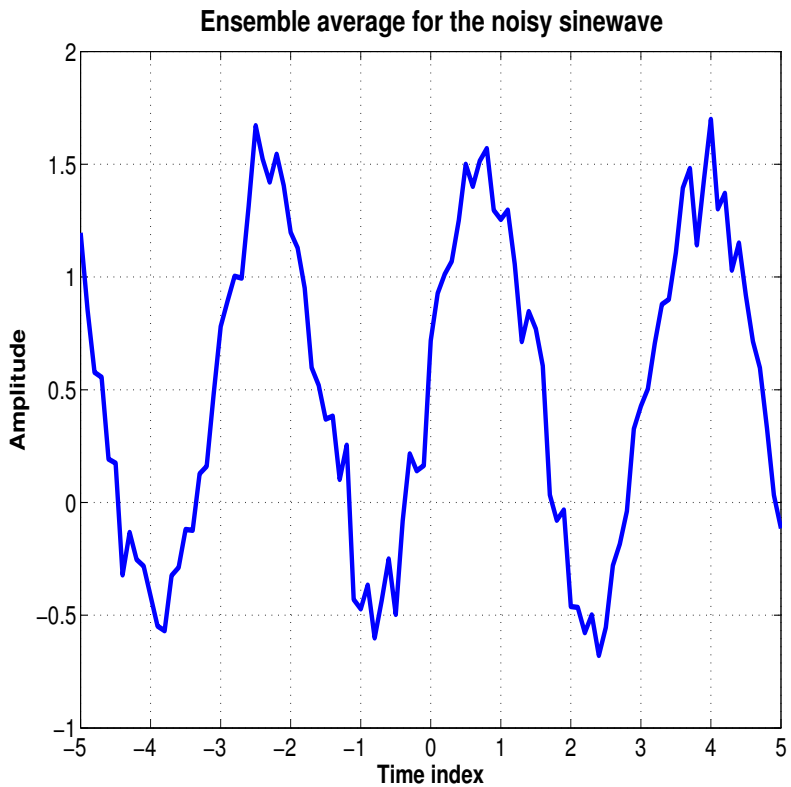


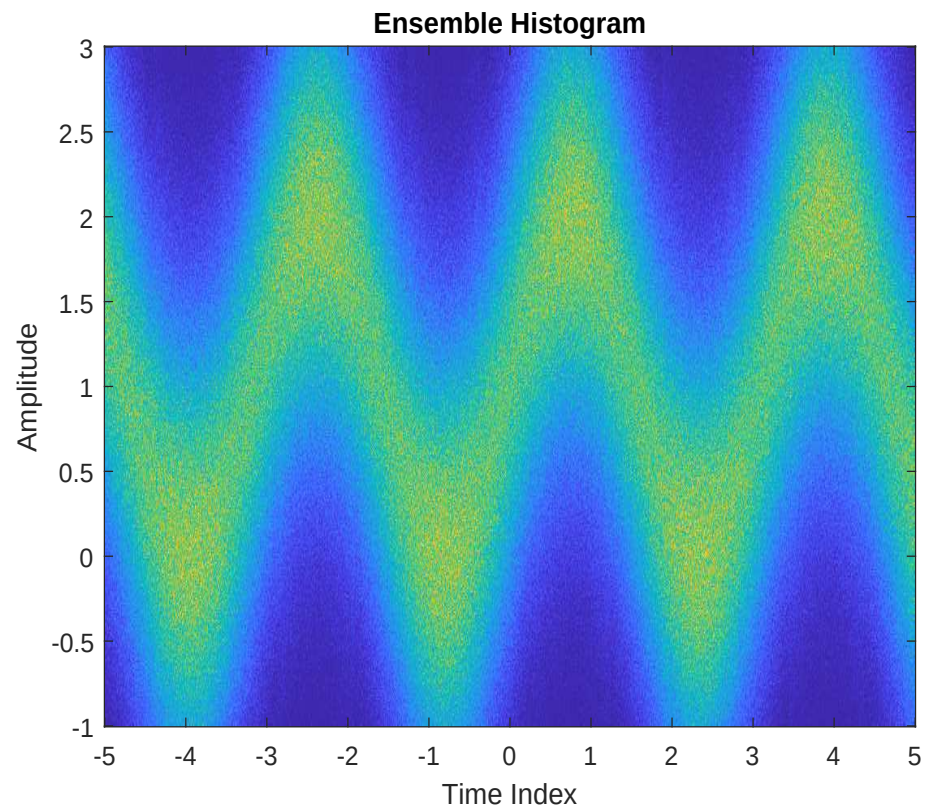
$$\text{Average Statistically} \quad E\{x[n]\} = \mu_x = \int_{-\infty}^{\infty} x[n]p(x[n])dx[n]$$

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

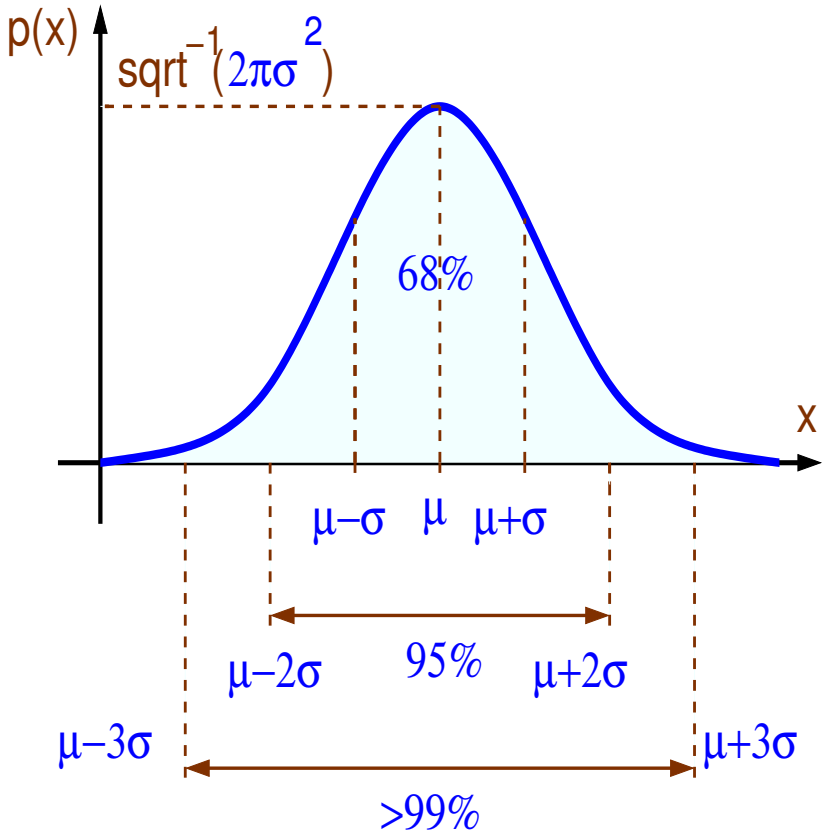
$$\sin(2x) + 2 * randn + 1$$

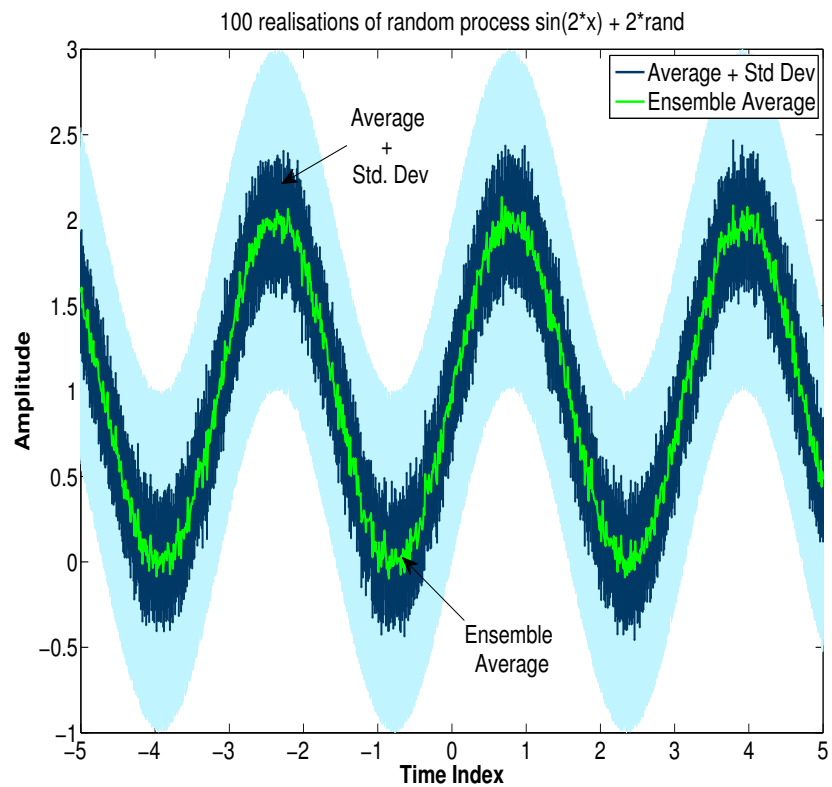
Left: 6 realisations, **Right:** 100 realisations (and the overlay plot)

Ensemble average of a noisy sinewave

A more precise probability distribution for every sample

Every sample in our ensemble average is a random process and has its pdf





Left: Area under the Gaussian vs σ

Right: Histogram for each random sample

Second order statistics: 1) Correlation

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

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

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

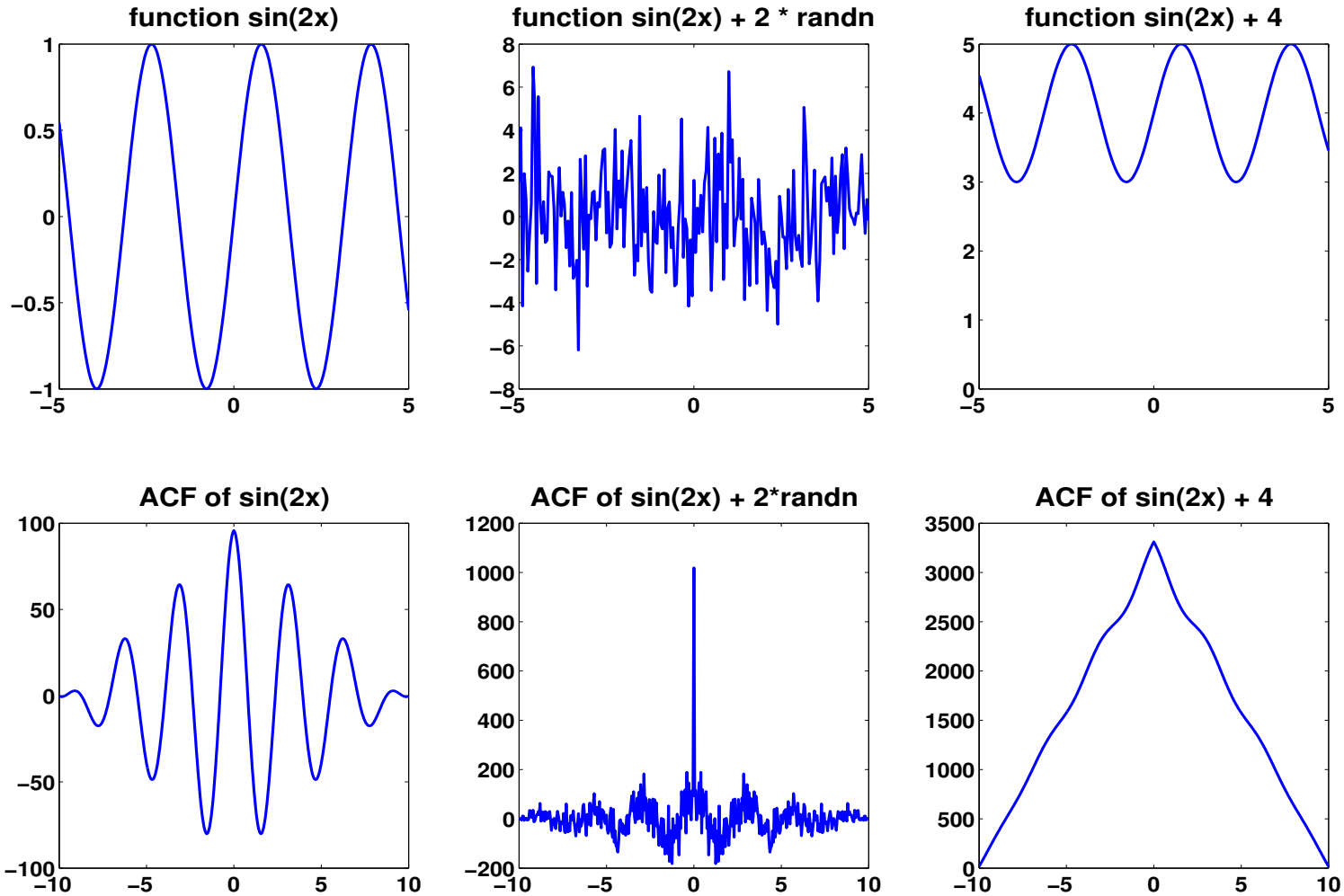
  - In practice, for **ergodic signals** we calculate correlations from the **relative frequency perspective**

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

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

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

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



☞ Useful information becomes obscured in additive noise or under a DC offset

Second order statistics: 2) Covariance

- **Covariance** is defined as

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

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

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

  - 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

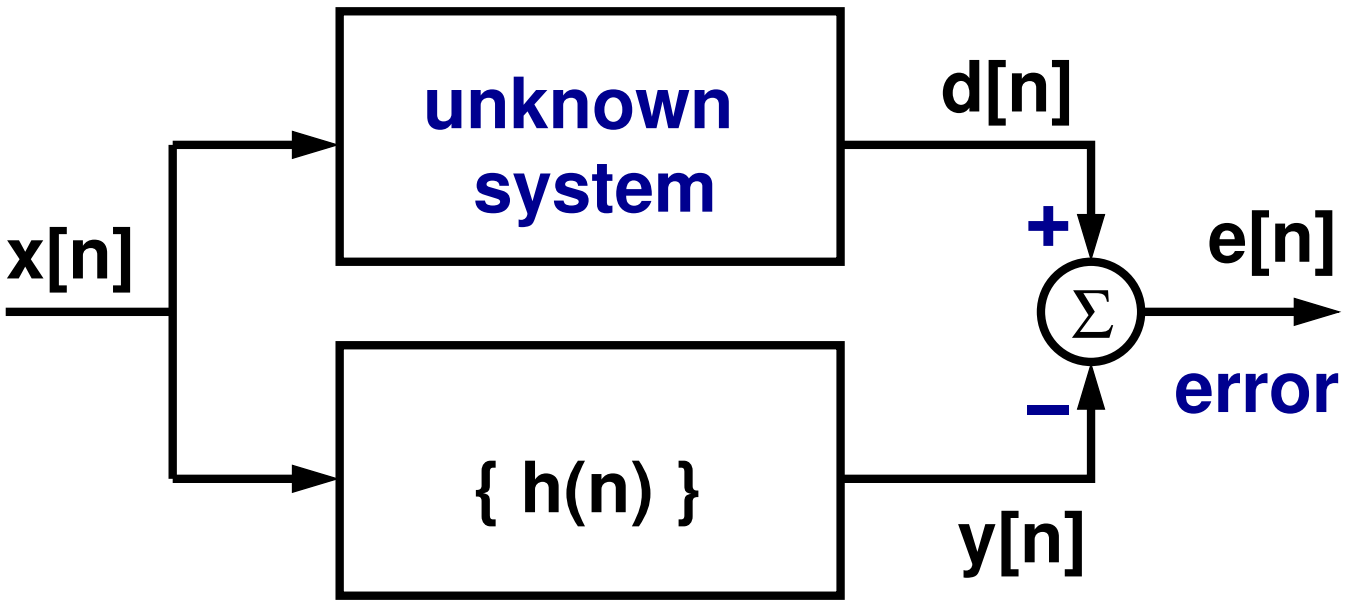
$$R_N(l_1, l_2, \ldots, l_n) = E\{x[l_1]x[l_2]\cdots x[l_n]\}$$

Higher order moments can be used to form Gaussian 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 output error should be zero mean, white, and uncorrelated with the output signal ↬ See Lecture 7

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

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

☞ The optimum coefficients are **inversely proportional** to the autocorrelation matrix and **directly proportional** to the estimate of the crosscorrelation between the teaching signal, $d(n)$ and the input, $x(n)$.

For more detail, see Lecture 7, e.g. the Wiener filter or acoustic echo cancellation in concert venues.

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 **"nonlinearly" dependent but "linearly" 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, they are:
  - **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)

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

As SSS is too strict for practical applications, we consider the more 'relaxed' stationarity condition $\looparrowright$ wide sense stationarity.

- **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 statistical moments (mean and covariance/correlation) 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) $r(0) \geq |r(m)|$ with a 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 (constant sub-diagonals)**.

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

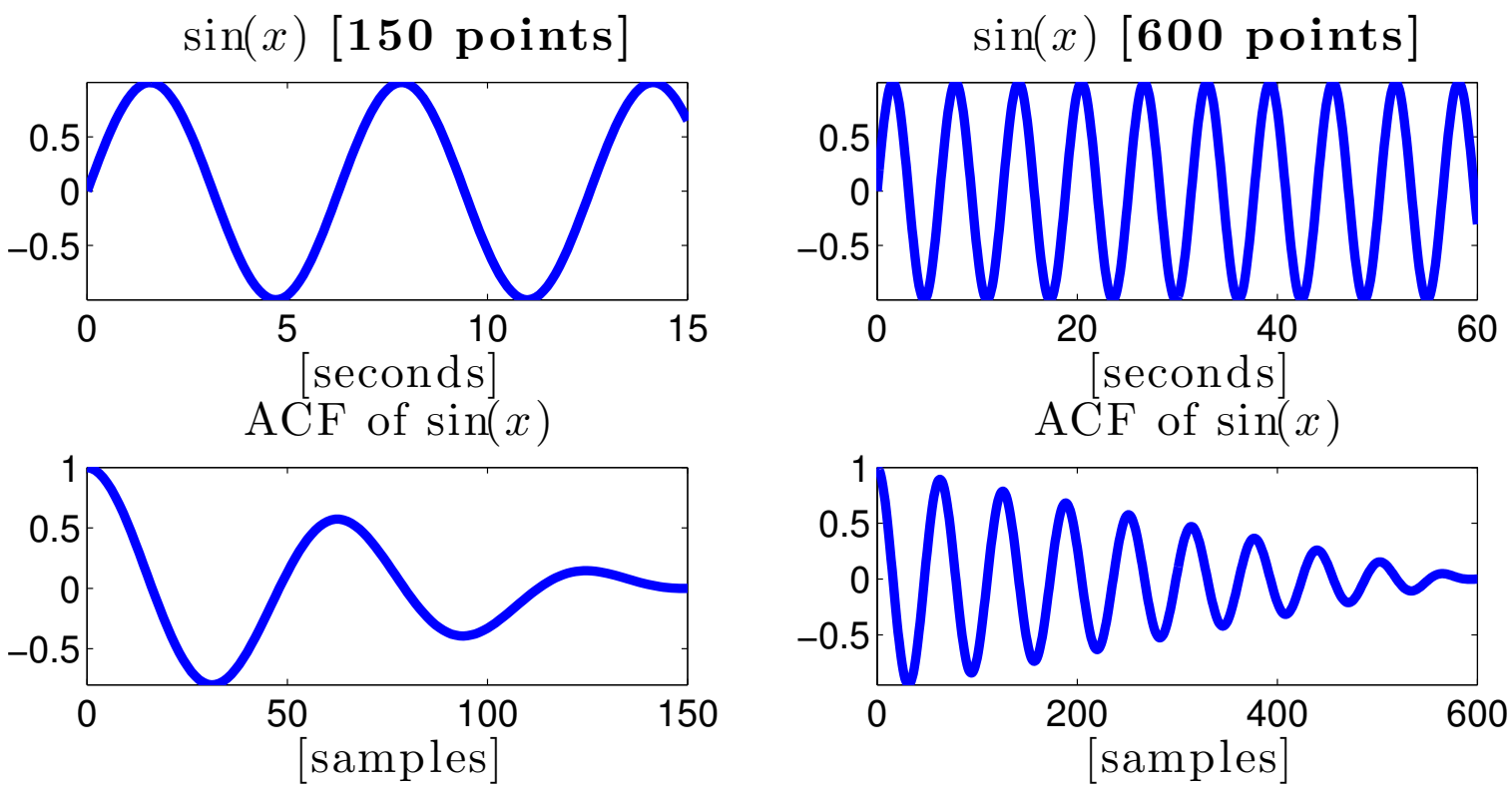
$$\mathbf{a}^T \mathbf{R} \mathbf{a} \geq 0 \qquad \forall \mathbf{a} \neq 0$$

This follows from $y = \mathbf{a}^T \mathbf{x}$ and $y^T = \mathbf{x}^T \mathbf{a}$, so (e.g. output of a linear system)

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

Properties of $r(m)$ – contd II

vi) Autocorrelation function **reflects the basic shape of a signal**, e.g. if the signal is periodic, then its autocorrelation function will be periodic and with the same period. (you also see here the effects of rectangular window)



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 $\rightsquigarrow$ 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) see Slide 26

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 $x = A\cos(\omega n + \theta)$

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

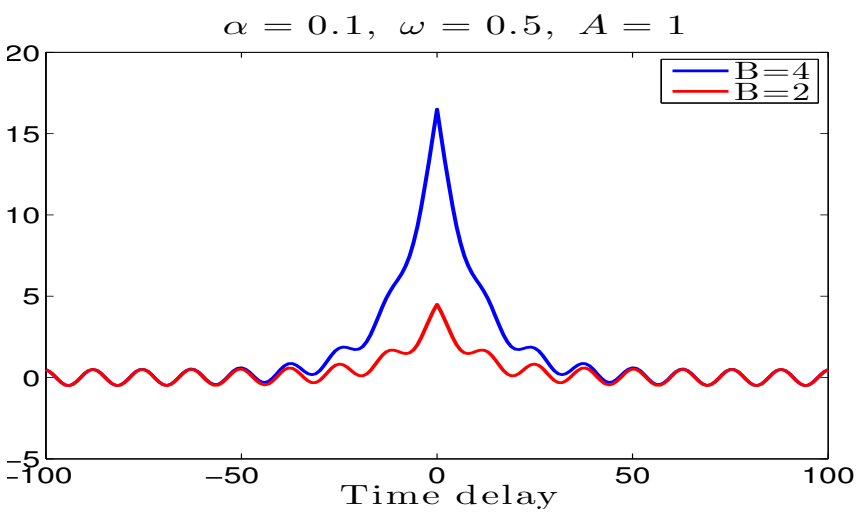
$$\textbf{ACF}: \; R(m) = E\big[y[n]y[n+m]\big] =$$

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

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:

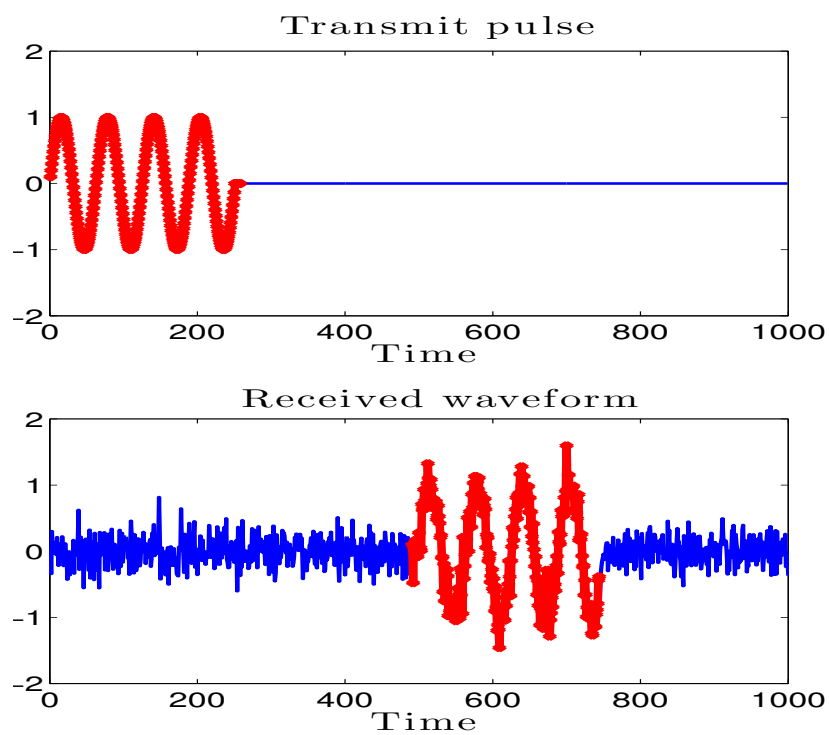
The received signal (see previous slide)

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

$$\begin{aligned} R_{xy}(\tau) &= E\{x(n)y(n+\tau)\} \\ &= aR_x(\tau - T_0) + R_{xw}(\tau) \end{aligned}$$

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 units 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

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

- ACF tells us about the correlation/power within a signal $\rightsquigarrow$ **Average**
- PSD tell us about the distribution of power across frequencies $\rightsquigarrow$ **Density**

PSD properties

i) $S(f)$ is a **non-negative and real (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

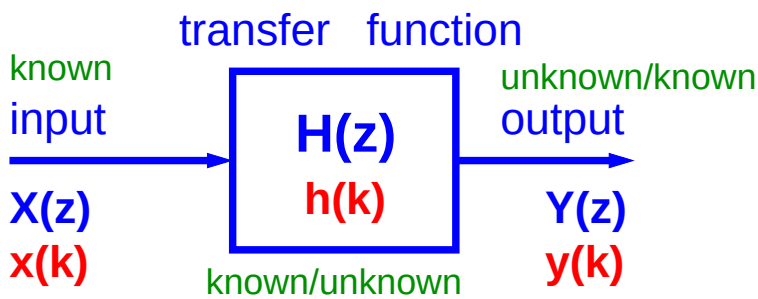
$$S(f) = \sum_{m=-\infty}^{\infty} r_{xx}(m)\cos(2\pi m f) = r_{xx}(0) + 2\sum_{m=1}^{\infty} r_{xx}(m)\cos(2\pi m f)$$

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

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

☞ 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

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

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

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 ↬ 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 × 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, then

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

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

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

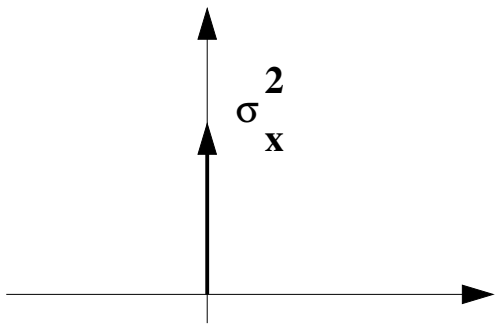
where the variance (signal power) $\sigma^2 = E\{x^2[n]\}$ and

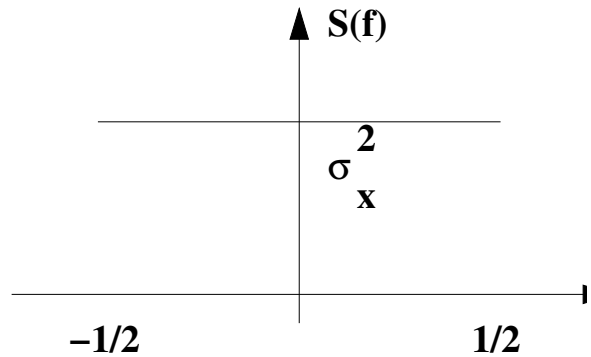
$$\delta(n-m) = \begin{cases} 1, & n = m \\ 0, & \text{elsewhere} \end{cases}$$

with $\delta(n)$ as 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 (see also Lecture 2)

(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 Statistical Signal Processing and Inference paradigms, learning algorithms often rely on the minimum phase property of a signal for stability (of e.g. the inverse system in channel equalisation – Lecture 7) and to be able to admit real-time implementation (causality).

d) Gaussian random signals

Consider a signal whereby each of the L samples is Gaussian distributed

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

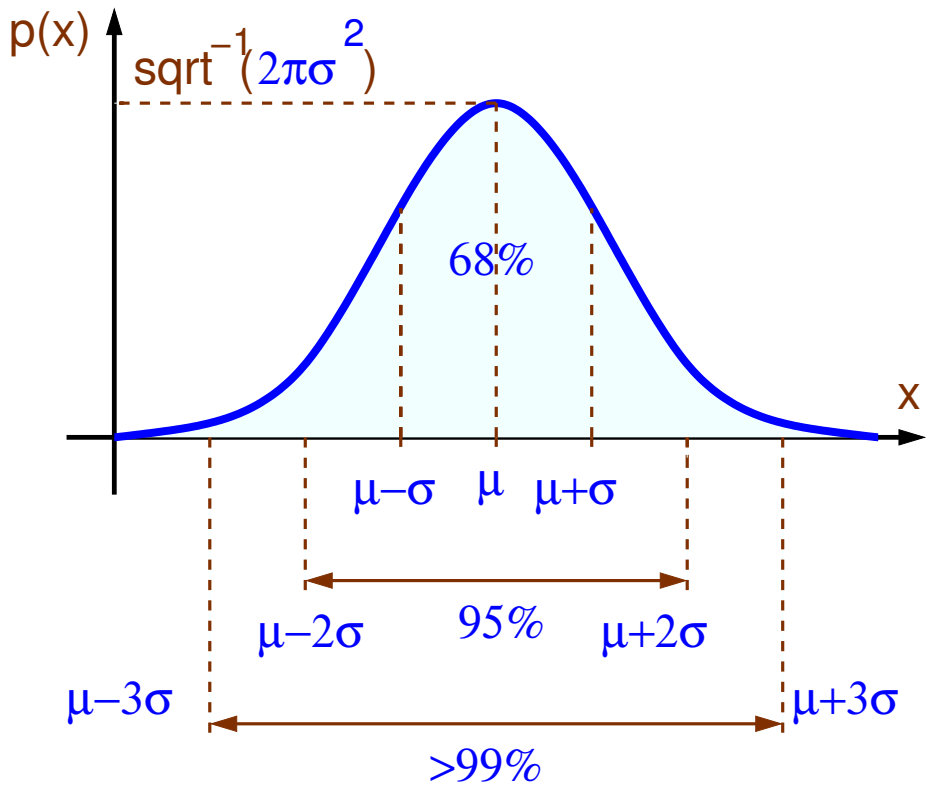
This situation is denoted by $\mathcal{N}(\mu(i), \sigma_i^2)$.

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

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

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.
Hint: A product $\prod$ of individual Gaussian distributions becomes a sum of arguments of the exponentials.

e) Properties of a Gaussian distribution



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

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

$$z = ax + by$$

is also Gaussian with the mean

$$m_z = am_x + bm_y$$

and variance

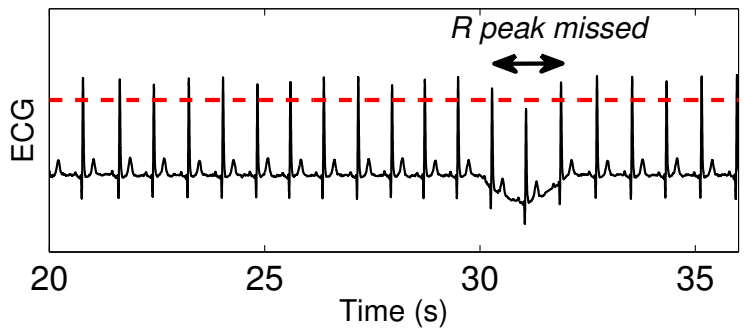
$$\sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\sigma_x\sigma_y\rho_{xy}$$

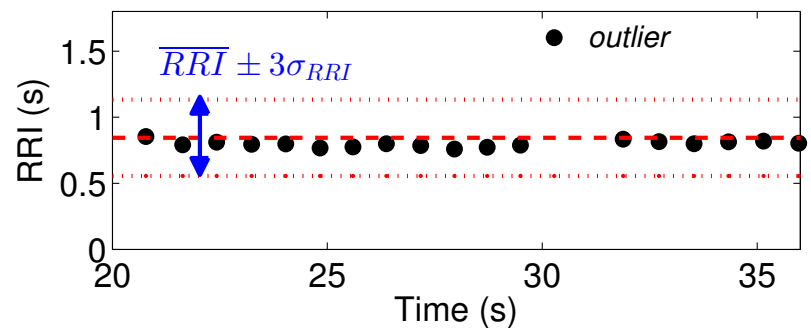
2) If two jointly Gaussian random variables are *uncorrelated* ($\rho_{xy} = 0$) then they are statistically independent,

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

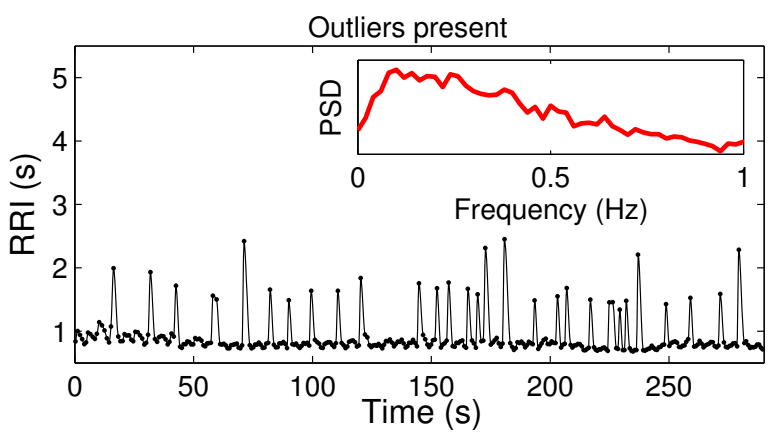
Example 1.11. Rejecting outliers from cardiac data

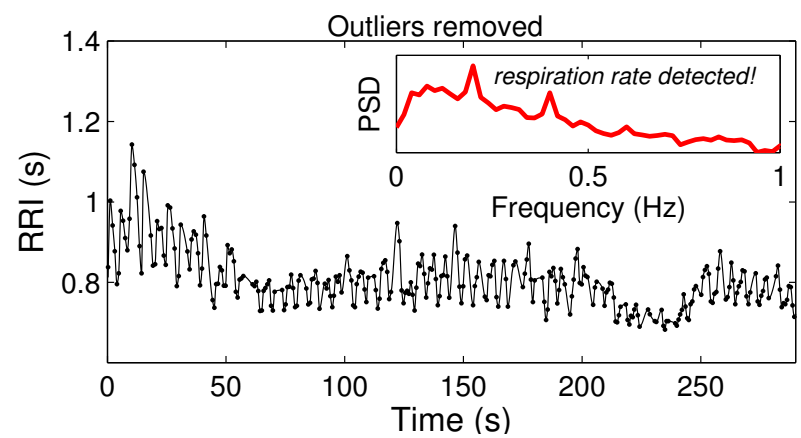
○ Failed detection of R peaks in ECG [left] causes outliers in R-R interval (RRI, time difference between consecutive R peaks) [right]





○ 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, **for ergodic processes, statistical averages** may be 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

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

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

e) Ergodicity in the mean

Definition: 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, it needs to be:

- Asymptotically unbiased

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

Consider the variance of the estimate $\to 0$ as $N \to \infty$ (see Lecture 3)

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

e) Ergodicity - Summary

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

Mean Ergodic Theorem: Let $x[n]$ be a wide sense stationary (WSS) random process with the autocovariance sequence $c_x(k)$. Then, 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: A necessary and sufficient condition for a WSS Gaussian process with covariance $c_x(k)$ to be autocorrelation ergodic is

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

Useful results: Concentration inequalities (Markov and Chebyshev inequalities)

Often in practice we do not know a full pdf, but just its mean (for Markov inequality) or mean and variance (for Chebyshev inequality).

The Markov and Chebyshev inequalities are very useful for putting bounds on probabilities!

We would like to e.g. know the bound on how likely it is for a random variable X, for example:

- To be far from its mean, $E\{X\} = \mu$, that is $P(|X - \mu| > \epsilon$, or
- To be very large, that is, $P(X \geq \epsilon)$

where ϵ is some threshold.

Example 1.12: The marks for an exam range from 0 to 110, with 10 additional marks for solving an optional question. The average mark for the exam is 50. What is the upper bound on the probability of students scoring more than 100 marks?

Example 1.13: The average height of a child in a kindergarden is $100\,cm$. What is the probability of a child being taller than $220\,cm$?

Useful results: Markov inequality (valid only for non-negative random variables, $X \geq 0$)

Q: In the two Examples from the previous slide, we do not know the pdf of X, can we still put some guesses of what our probabilities are going to be?

A: If we know only the mean of a **non–negative** random variable, μ, then the **Markov inequality** states that

$$Prob(X \geq \epsilon) \leq \frac{E\{X\}}{\epsilon} = \frac{\mu}{\epsilon} \qquad \equiv \qquad Prob(X \geq \epsilon\,\mu) \leq \frac{1}{\epsilon}$$

Proof: The random variable X is non-negative with mean μ. Then, $\forall \epsilon \geq 0$

$$\mu = E\{X\} = \int_0^\infty x\,p(x)dx = \int_0^\epsilon x\,p(x)dx + \int_\epsilon^\infty x\,p(x)dx$$

$$\geq \int_\epsilon^\infty x\,p(x)dx \geq \int_\epsilon^\infty \epsilon\,p(x)dx = \epsilon \int_\epsilon^\infty p(x)dx = \epsilon\,Prob(X \geq \epsilon)$$

Hence

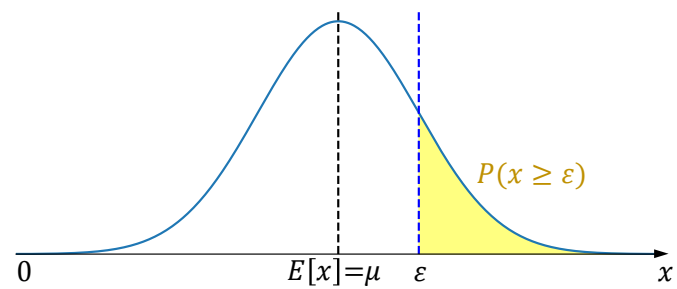
$$Prob(X \geq \epsilon) \leq \frac{E\{X\}}{\epsilon} = \frac{\mu}{\epsilon} \qquad \textbf{Markov inequality}$$

Useful results: Markov inequality, contd.

☞ In Examples 1.12 and 1.13 we are asking ourselves "what is the bound on the probability that a random variable is in the tails of the distribution (outlier)". However, we only know that a random variable is non-negative, i.e. $X \geq 0$, and we know its mean, μ, but not its probability distribution.

Example 1.12: Solution

The mean mark is $\mu = 50$, and we seek the bound on the probability of a student score being greater than 100, that is $Prob(X > 100)$, with the maximum score being 110. Using the Markov inequality, $Prob(X \geq 100) \leq 1/2$.



Here, the threshold ϵ=100. Thus

$$Prob(X \geq 100) \leq \frac{E\{X\}}{\epsilon} = \frac{50}{100} = \frac{1}{2}$$

Example 1.13: Solution. We seek $Prob(child_height \geq 220\,cm)$. The Markov inequality puts a bound on this $P(X \geq 2.2\,m) \leq 1/2.2 = 0.45$.

This bound is generous ↬ we need more information for a tighter bound.

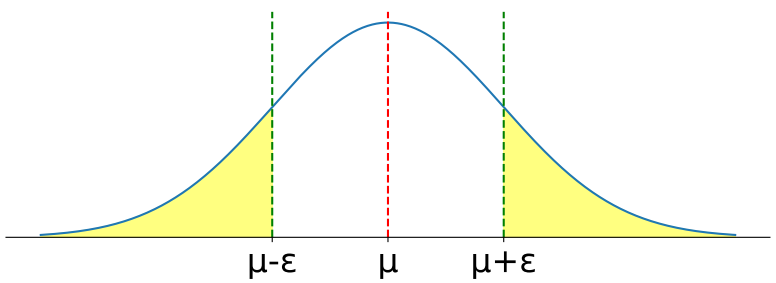
Useful results: Chebyshev inequality

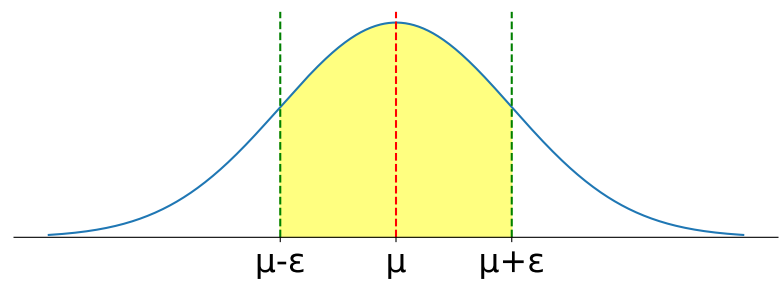
Makes sense, $Prob(|X-\mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$ is proportional to σ^2 and inversely proportional to ϵ^2, and can be used if the pdf is not given, but only μ and σ^2

Intuitively, we can have a tighter bound on the probability of an extreme event, if we know both the mean, μ, and variance, σ^2, of a RV X.

Chebyshev inequality: For any random variable, X, which can be either negative, 0, or positive, the bound on the probability that X is further away from the mean than some threshold, ϵ, is given by

$$Prob(|X-\mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2} \qquad \equiv \qquad Prob(|X-\mu| < \epsilon) \geq 1 - \frac{\sigma^2}{\epsilon^2}$$





Proof: Since $\sigma^2 = \sigma_X^2 = E\{(X-\mu)^2\}$, we have

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx \geq \int_{|x-\mu|\geq\epsilon} (x-\mu)^2 p(x)dx$$

$$\geq \epsilon^2 \int_{|x-\mu|\geq\epsilon} p(x)dx \geq \epsilon^2 Prob(|X-\mu| \geq \epsilon) \quad \Rightarrow \quad Prob(|X-\mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

Useful results: Chebyshev inequality ↬ Examples

Ex. 1.13: Solution, revisited. We seek $Prob(child_height \geq 220\,cm)$, with the mean height $\mu = 100\,cm$ and the std. $\sigma = 50\,cm$. The Markov inequality puts a bound on this as $P(X \geq 2.2\,m) \leq 0.45$.

Now, using the Chebyshev inequality, we have

$$P(|X-1| \geq 1.2) \leq \frac{0.5^2}{1.2^2} = 0.17$$

probability to be more than 5.5 away from the mean ↗

↖ a much tighter bound than Markov inequality

☞ Although the Chebyshev inequality is an application of Markov inequality, it has different consequences. For example, if the variance is small, the the RV is unlikely to fall too far from the mean.

Application of Chebyshev inequality: The probability that the distance from the mean is at least ϵ standard deviations is (independent of the pdf)

$$Prob(|X-\mu| \geq \epsilon\,\sigma) \leq \frac{\sigma^2}{\epsilon^2\sigma^2} = \frac{1}{\epsilon^2}$$

Useful results: Markov and Chebyshev inequalites, another example and a useful trick

Example (Markov inequality: A router crashes if it receives more than 1000 packets per second. We know that the average load for this network is $\mu = 50\,packets/sec$. What is the bound on the probability of a crash?

Answer: Probab. of a crash is bounded by $Prob(crash) \leq 50/1000 = 0.05$

Trick: Markov inequality produces crude bounds, as it assumes only non-negative RVs. **Remedy:** For symmetric pdf's consider all RVs.

Example (Markov inequality): Consider a variable X which is uniform on $[-4, 4]$, with $\mu = 0$. What is the prob. $Prob(X \geq 3)$?

Answer: Markov's inequality allows us to use only non-negative (absolute) values of X, so that the mean of $|X|$ becomes $\mu = 2$. Then, $Prob(|X| \geq 3) \leq \mu/3 = 2/3$. This is a very crude bound. Now,

$$Prob(X \geq 3) = \tfrac{1}{2} Prob(|X| \geq 3) \leq 1/3 \qquad \leftarrow \quad \text{better bound}$$

non-negative RVs $\nearrow$ $\nwarrow$ two-sided distribution

For sums of independent RVs, much better bounds are Chernoff bounds.

Useful results: Taylor series expansion

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

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

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 obtain a_0 $\hookrightarrow$ choose $x = x_0$ $\Rightarrow$ $a_0 = f(x_0)$
2. To obtain a_1 $\hookrightarrow$ 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 $\hookrightarrow$ $a_k = \frac{1}{k!}\frac{d^k f(x)}{dx^k}_{|x=x_0}$

Useful results: Power series - contd.

Consider

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

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

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

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

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

Useful results: 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):
  - $f'(0) = \frac{f(-2)-8f(-1)+8f(1)-f(2)}{12h}$
  - $f''(0) = \frac{-f(-2)+16f(-1)-30f(0)+16f(1)-f(2)}{12h^2}$

Useful results: Constrained optimisation using Lagrange multipliers: Basic principles

Consider a two-dimensional problem:

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

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

↬ **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

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

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 constraint $g(x, y, z) = c$

Solution: Form a new function

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

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

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

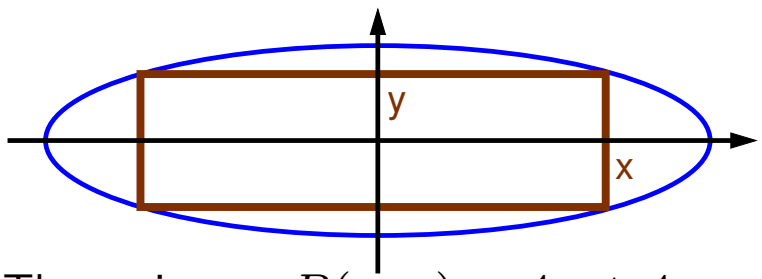
Then: $F'_x = 12x - \lambda$, $F'_y = 24y - \lambda$, $F'_\lambda = -x - y - 90$, and for $\min / \max \nabla F = \mathbf{0}$

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

Appendix: Probability vs. Statistics

(for discrete RVs, $E\{X\} = \sum_{i=1}^{I} x_i P_X(x_i)$, where P_X is the probability function)

Probability: A data modelling view, describes how data **will likely behave**

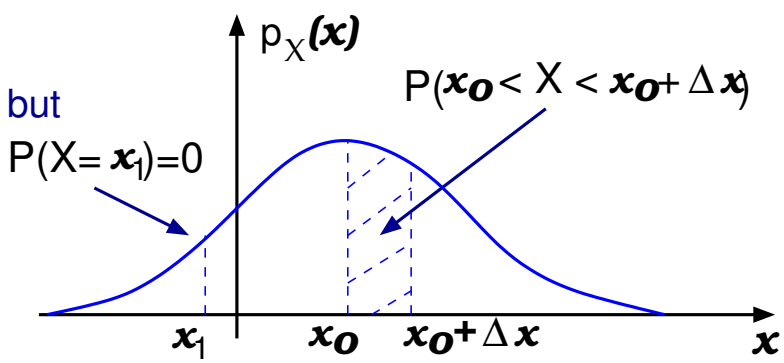
for example: $$average = E\{X\} = \int_{-\infty}^{\infty} x\, p_X(x)\, dx$$ no data here

Notice that there is no explicit mention of data here $\looparrowright$ x is a dummy variable and p_X is the pdf of a random variable X.

Statistics: A data analysis view, determines how data **did behave**

for example: $$average = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$ no pdf here

Vagaries of probability: $P(x_0 < X < x_0 + \Delta x) = \int_{x_0}^{x_0+\Delta x} p_X(x)dx$



Notice that

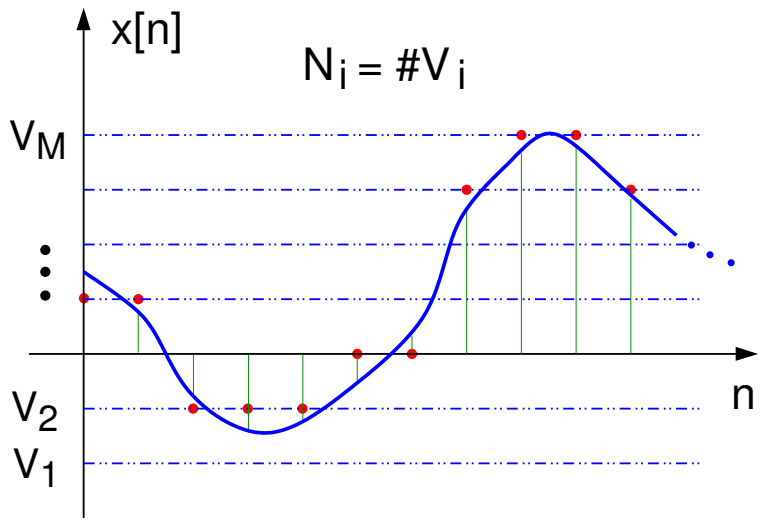
$$P(X = x_1) = 0$$

This appears odd, but otherwise the probabilities sum up to ∞

Appendix: Statistics vs. Probability

Statistical inference ↬ based on the observed data and supported by prob. theory

Vagaries of statistics: Consider N coarse-quantised data points, $x[0], \ldots, x[N-1]$. The quantised signal has $M \ll N$ possible amplitude values, $V_1, \ldots, V_M$, for which the corresponding relative frequencies are, $N_1 = \#V_1, \ldots, N_M = \#V_M$. Calculate the mean, $\bar{x}$.



Solution:

$$\bar{x} = \frac{1}{N}\sum_{n=0}^{N-1} x[n] \quad = \quad \frac{1}{N}\sum_{m=1}^{M} V_m N_m \quad = \quad \sum_{m=1}^{M} V_m \underbrace{\frac{N_m}{N}}_{\approx\ P(x=V_m)}$$

☞ Clearly, the factor $1/N$ does not imply "uniform distribution"

Notes:

○

Notes:

○

Notes:

○

Notes:

○