Statistical Signal Processing

Lecture 1:
Introduction & Estimation

Carlas Smith & Peyman Mohajerin Esfahani

Delft Center for Systems and Control

Delft University of Technology

TODAY

1. **Organizational Details**
2. Stochastic Processes
3. Four Optimal Filtering Problems
4. Course outline/Course Reader
5. Signals/Systems highlights
6. Random variables
7. Estimation

Educational Staff (DCSC)

Lecturer:

- Assistant Professor Peyman Mohajerin Esfahani

  `p.mohajerinesfahani@tudelft.nl`

- Assistant Professor Carlas Smith (???)

  `c.s.smith@tudelft.nl`

Teaching Assistants:

- Dr. Rayyan Sheriff

  `M.R.Sheriff@tudelft.nl`

- Pedro Zattoni Scroccaro

  `P.ZattoniScroccaro@tudelft.nl`

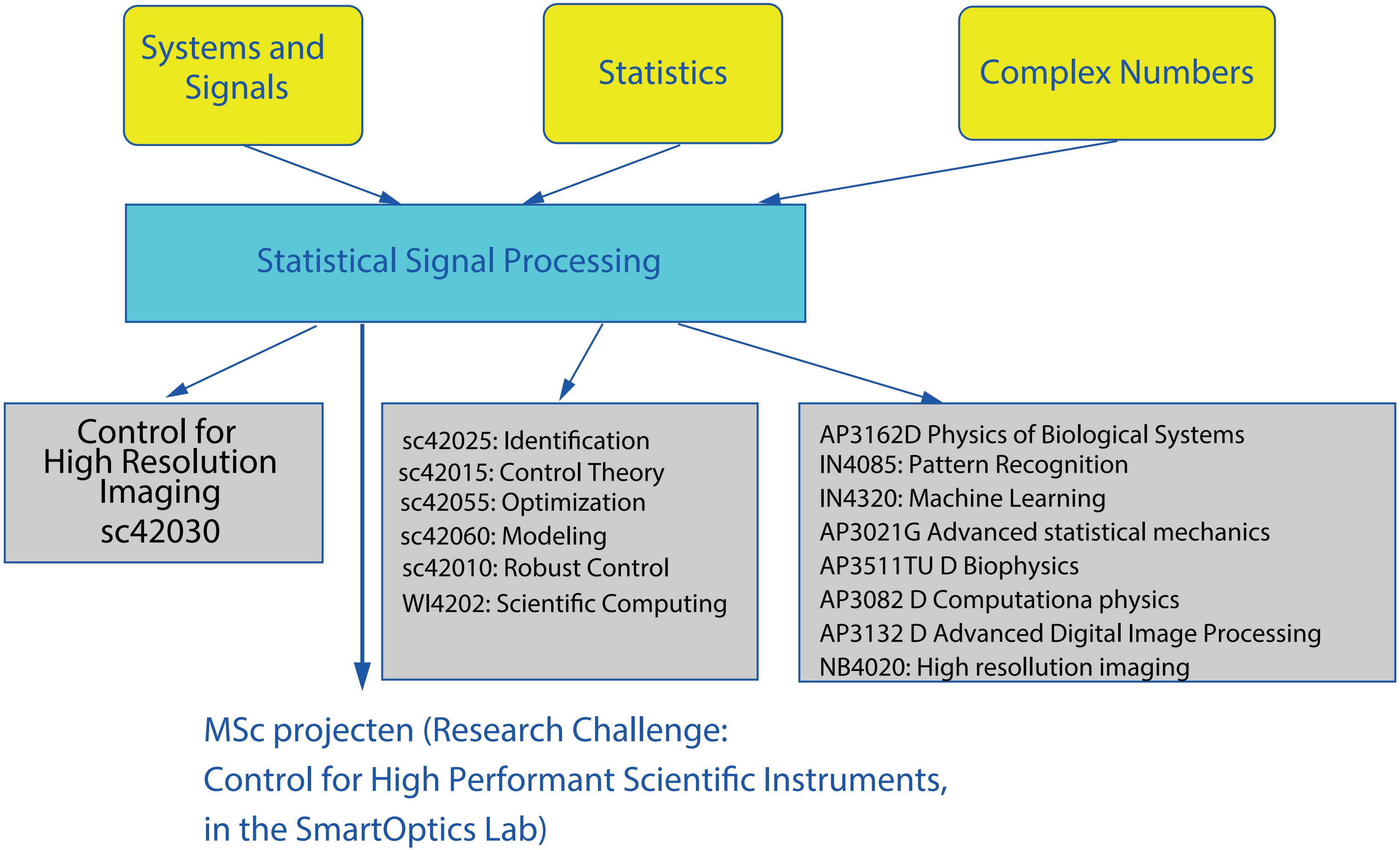
- Mostafa Osman

  `M.E.A.Osman@tudelft.nl`

- Ioannis Dimanidis

  `I.Dimanidis@student.tudelft.nl`

Nested within the DCSC-education program



Course Organization

Announcements/Important Info/downloads via BrightSpace

- Course Schedule: Detailed week planning period September 8th - November 3th
- Course Reader on Slides: Outline of topics and corresponding parts of the book (“diktaat”) to be discussed during each lecture.
- Written Exam/Assignment: November 3th, 2021 — 9.00 - 12.00
- Copies slides of each lecture
- Answers Exercises to be made, etc.

Assistence in linking up . . .

Course Material: Stochastic Processes for Scientists and Engineers with Modern Applications. 2020.

1. “Refresher” (Test Yourself!)
2. **2** Mandatory Python (not Matlab) exercises - see Course Planning.
3. **13** Instruction classes - see course planning
4. Course Reader (planning) on when what is treated
5. Formularium
6. Guideline Preparation (30 % attendance - 40 % preparation Python and instruction classes - 30 % preparation exam).

Final Mark for this course

Marked Homework:

- **2** Python exercises - *Handing in as indicated on the course schedule - "Python Assignment I & II"*

Obligatory! Please contact TAs for questions.

Final Mark:

Formula for your Final Mark $= 0.15H + 0.85E$.

Introduction and Problems

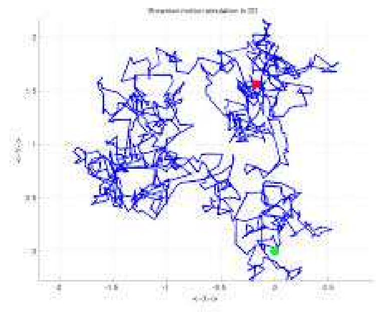
1. Organizational Details
2. **Stochastic Processes in Physics**
3. Four Optimal Filtering Problems
4. Course outline/Course Reader
5. Signals/Systems highlights
6. Random variables
7. Estimation

21st Century Scientific Challenges

Since measurements can be performed with increased accuracy at the nano- (pico-) scale (time, space) a whole new world is to be discovered. At this scale: physical phenomena are “random” rather then deterministic.

Example: Scientific Challenge 1827

Biology: Explain “nature” of apparent random (non-deterministic, non-repeatable) movement of particles (pollen) suspended in fluid.



Albert Einstein (1879-1955)

- In his doctoral dissertation [Zurich, 1905]: Einstein addressed a.o. the question how to describe the evolution over time of the displacement (in x-direction)?

Einstein's Challenge 1827

Refine the question after "experimenting" (in Einstein's days: rigorous mathematical modeling):

`Brownian.m`

Observations

- The time sequences are "non-repeatable" $\Rightarrow$ No deterministic law to "prescribe it"!
- Instead: (partly) describe these "non-repeatable" time sequences via a prescription on how statistical characteristics (such as the mean, variance, etc.) change over time.

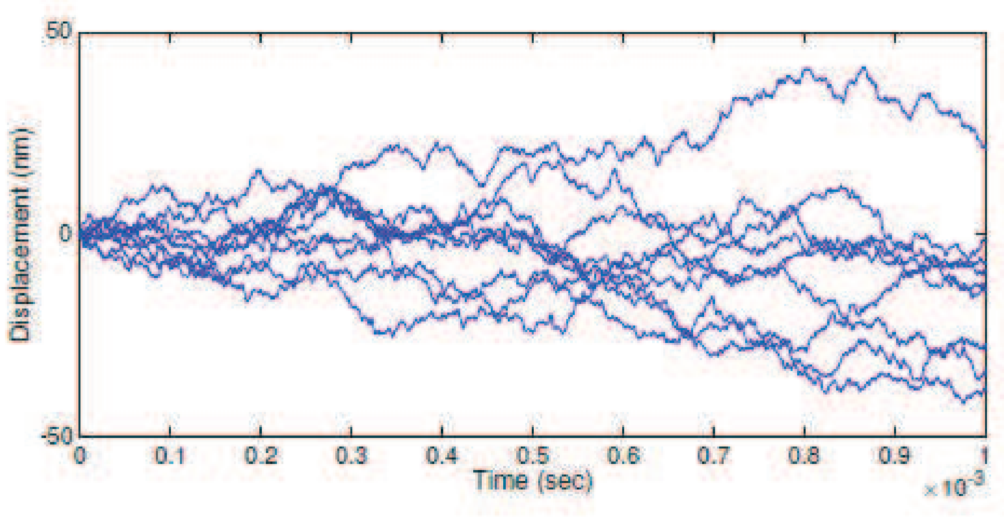
One Result PhD thesis Einstein

The mean-square displacement (variance) in the $x-$direction at time t

$$E[\Delta x(t)^2] = 2Dt$$

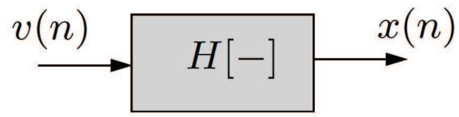
with D is the diffusion constant.

Questions to be addressed in this Course?



(1.a) How to describe "non-repeatable" time sequences — i.e. stochastic processes?

(1.b) *Forward Modeling:* How does the description change when filtering a stochastic process? (1.a+b = TOOLSET)



(2) *Inverse Modeling:* Based on "an" observation of a stochastic process, how can we find a model (filter $H[-]$ and input $v(n)$) such that we can "reproduce" other realizations of $x(n)$

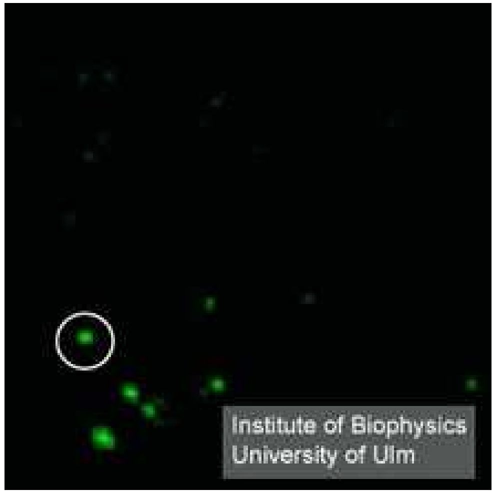
(3) *Optimal filtering:* e.g. how to "remove" noise from an observed time sequence?

Introduction and Problems

1. Organizational Details
2. Stochastic Processes in Physics
3. **Four Optimal Filtering Problems**
   - Estimation
   - Denoising of Signals
   - Deconvolution
   - Active Noise Cancellation
4. Course outline/Course Reader
5. Signals/Systems highlights
6. Random variables
7. Estimation

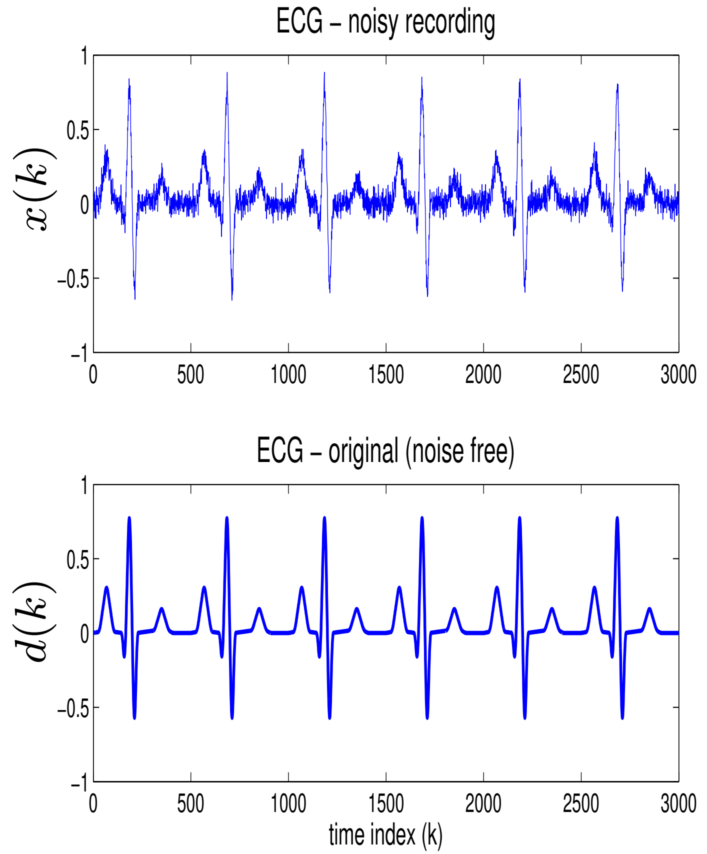
Estimation

Biology: Tracking a particle, i.e. Estimate the series of positions of a particle is crucial in understanding the cellular kinetics of particles (such as proteins (HIV-1))



Denoising of signals

ECG recordings



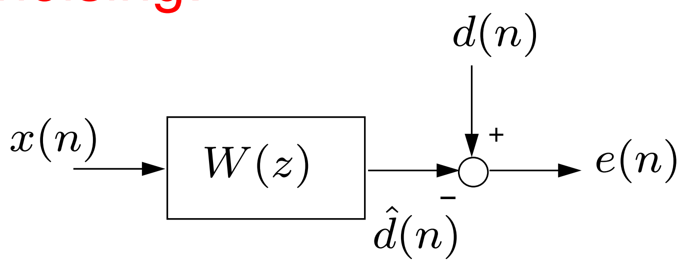
Observation model

$$x(n) = d(n) + v(n)$$

$d(n)$ — "desired" - signal of interest

$v(n)$ — "noise" - disturbance (additive)!

Denoising:



"Design" $W(z)$ to "minimize the error $e(n)$"?

Deblurring images (Deconvolution)

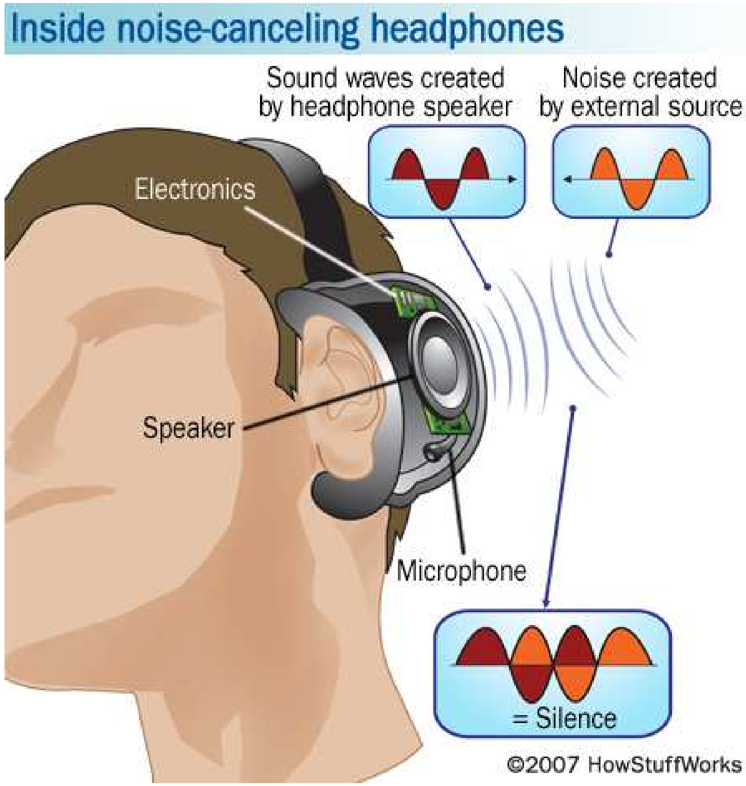
Original Image (Object)

Recorded (blurred) Image

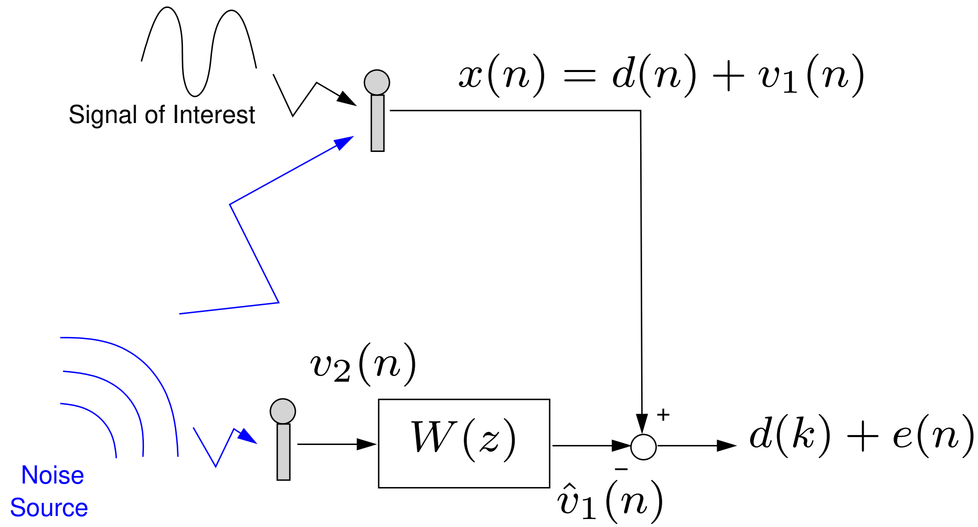
More than just denoising: $x(p) = PSF(p) \star d(p) + v(p)$ (in 2D)

Active Noise Cancellation

Communicating in a "noisy" environment



Challenge: Signal modeling AND cancelling



with $e(n) = v_1(n) - \hat{v}_1(n)$.

Introduction and Problems

1. Organizational Details
2. Stochastic Processes in Physics
3. Four Optimal Filtering Problems
4. **Course outline/Course Reader**
5. Signals/Systems highlights
6. Random variables
7. Estimation

Lecture 1: Introduction

- Motivation, course plan and organization
- Recap - some notions Signals/Systems & Statistics -
- Estimating the parameters of a probability distribution
- Chapter 2, 3 & 4

↓

TOOLSET - Lecture 2: Random processes/Signals

- Characterizing discrete *complex* random processes: Time-domain & Frequency domain
- Ergodicity of discrete *complex* random processes
- Chapter 5

↓

TOOLSET - Lecture 3: Filtering Random Processes/Signals

- Changing characteristics of general RPs by LSI filtering
- Changing characteristics of specific RPs by LSI filtering (ARMA)
- Chapter 6

TOOLSET - Lecture 4: The inverse problem

- From Power Spectra (Frequency domain) to generating a stochastic process
- From Autocorrelation (Time domain) to generating a stochastic process
- Chapter 7 & 8

Optimal filtering - Lecture 5: Optimal filtering of RPs

- The optimal (“minimum variance”) FIR & IIR Wiener Filter
- Mixed causal, anti-causal solution
- Chapter 9

Optimal filtering - Lecture 6: Optimal filtering of RPs

- The optimal (“minimum variance”) IIR Wiener Filter
- causal solution
- Chapter 9

Introduction and Problems

1. Organizational Details
2. Stochastic Processes in Physics
3. Four Optimal Filtering Problems
4. Course outline/Course Reader
5. **Signals/Systems highlights**
6. Random variables
7. Estimation

Discrete-time Signals (Time Domain)

A discrete time sequence: $x(n)$ given as

$\cdots x(-1), \boxed{x(0)}, x(1), \cdots \qquad$ for $x(n) \in \mathbb{C}$

This can mathematically be represented as a summation:

$$x(n) = \sum_{k=-\infty}^{\infty} x(k)\Delta(n-k)$$

with $\Delta(n) = \begin{cases} 1 & \text{for } n = 0 \\ 0 & \text{for } n \neq 0 \end{cases}$

Discrete-time Signals (Frequency Domain)

Fourier Transform

For a signal $x(n\Delta T)$ the DTFT is:

$$\begin{aligned}\mathcal{F}[x](e^{j\omega}) &= X(e^{j\omega}) \\ &= \sum_{n=-\infty}^{\infty} x(n\Delta T)e^{-j\omega n\Delta T}\end{aligned}$$

for $\omega \in \mathbb{R}$

Existence?

z-transform

For a signal $x(n)$ the z-transform is:

$$X(z) = \mathcal{Z}[x](z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} \quad z = re^{j\omega} \in \mathbb{C}$$

Existence? (ROC)

Signal	z-transform	ROC
$\Delta(n)$	1	$\mathbb{C}$
$a^n u(n) \quad a \in \mathbb{R}$	$\frac{1}{1-az^{-1}}$	$\lvert z\rvert > a$
$-a^n u(-n-1) \quad a \in \mathbb{R}$	$\frac{1}{1-az^{-1}}$	$\lvert z\rvert < a$
$a^{\lvert n\rvert}$	$\frac{1-a^2}{(1-az^{-1})(1-az)}$	$a < \lvert z\rvert < \frac{1}{a}$

Discrete-time Fourier Transform (DTFT)

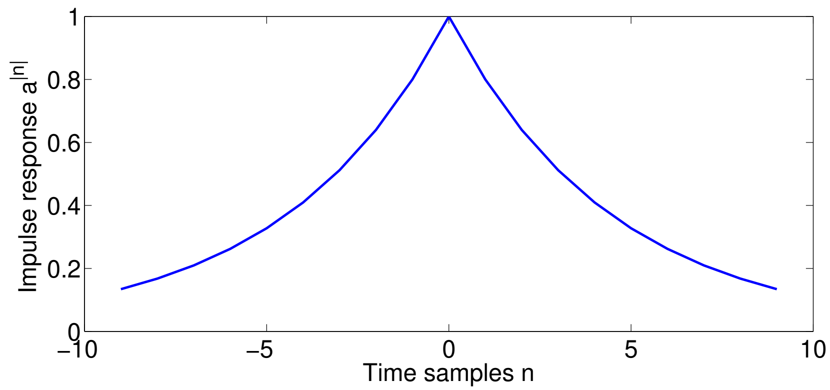
sequence	DTFT	z-transform
$\{x(n)\}$ "existence"	$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-jn\omega}$ $\sum_{n=-\infty}^{\infty} \lvert x(n)\rvert < \infty$	$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n}$ $R_- < \lvert z\rvert < R_+$
$\{x^*(-n)\}$ $x(n-\alpha)$	$X^*(e^{j\omega})$	$X^*(1/z^*)$ $z^{-\alpha}X(z)$
$\delta(n)$ $h(n) = a^{\lvert n\rvert} \quad a \in \mathbb{R}$		1 $\frac{1-a^2}{(1-az^{-1})(1-az)}$
Parseval $\sum_{n=-\infty}^{\infty} x(n)y^*(n)$	$\frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega})Y^*(e^{j\omega})d\omega$	

Discrete-time Systems

$$x(n) \longrightarrow \boxed{T[-]} \longrightarrow y(n) = T[x](n)$$

$T[-]$ will be assumed LTI. It is fully characterized via its impulse response $h(n)$, and its output $y(n)$ is given as,

$$y(n) = \sum_{k=-\infty}^{\infty} h(k)x(n-k)$$



System properties:

- BIBO stability, Causality and anti-causality and the inverse of a system
- minimum phase systems

Introduction and Problems

1. Organizational Details
2. Stochastic Processes in Physics
3. Four Optimal Filtering Problems
4. Course outline/Course Reader
5. Signals/Systems highlights
6. **Random variables**
7. Estimation

Random Variables:

1. **An Example**
2. Discrete and Continuous Random Variables (RVs)
3. Characterization of RVs:
   - (theoretical) via the Probability Distribution/Density Function.
   - (practical) via Ensemble Averages
4. Joint (multiple) RVs

Definitions for Discrete Random Variable

The sample space Ω (“uitkomsten”)

Examples of a discrete sample space:

1. Flipping coins: $\Omega = \{H, T\}$ $\quad H$ is an event $\subset \Omega$
2. Throwing dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$

When the sample space is only linguistic an additional mapping $f(.)$ is invoked to assign real numbers to each event.

$$f : \Omega \to \mathbb{R}$$

Example: Flipping coins: $\omega_1 = \{H\} \Rightarrow x = 1$ and $\omega_2 = \{T\} \Rightarrow x = -1$

Remark: A random variable (RV) may be complex, e.g.

$$z = a + bj \quad j = \sqrt{-1}$$

with a - throwing of a white die and b - throwing a black die.

Characterization of Discrete Random Variable

"A RV (Random Variable) is characterized by its frequency of occurrence (probability)"

Example: Flipping a "fair" coin N_T times yields n_H times `head` and n_T times `tail`. If N_T is large enough:

$$\frac{n_H}{N_T} \approx 0.5 \quad \frac{n_T}{N_T} \approx 0.5$$

Definition: A discrete RV with sample space $\Omega = \{\omega_i\}_{i=1}^{N}$ is fully characterized if we assign a probability to each elementary event:

$$Pr\{\omega_i\} = p_i \in [0, 1]$$

The probability that all events can happen is one: $Pr\{\Omega\} = 1$

Example: A Bernoulli RV ($x = \pm 1$): $Pr\{x = 1\} = p \quad Pr\{x = -1\} = 1 - p$

Continuous Random Variables

Example: An ∞ resolution roulette wheel:

$$\Omega = \{\omega : 0 \leq \omega \leq 1\}$$

Probability assignment:

$$Pr\{\alpha_1 < \omega \leq \alpha_2\} = f(\alpha_1, \alpha_2)$$

For a “fair” roulette (all outcomes should be equally plausible),

$$Pr\{\alpha_1 < \omega \leq \alpha_2\} = \alpha_2 - \alpha_1$$

Axioms:

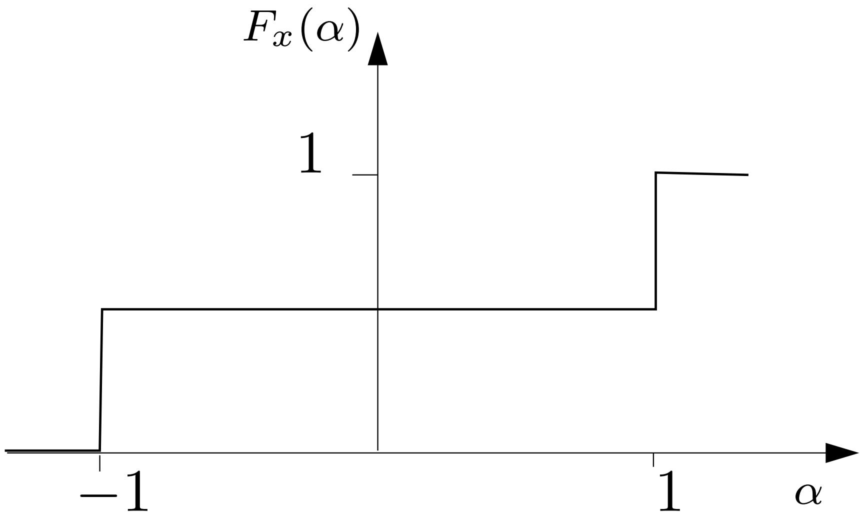
1. $0 \leq Pr(A) \leq 1$ for every event $A \subset \Omega$.
2. $Pr(\Omega) = 1$ for the certain event Ω.
3. For any two mutual exclusive events A_1 and A_2, $Pr(A_1 \cup A_2) = Pr(A_1) + Pr(A_2)$.

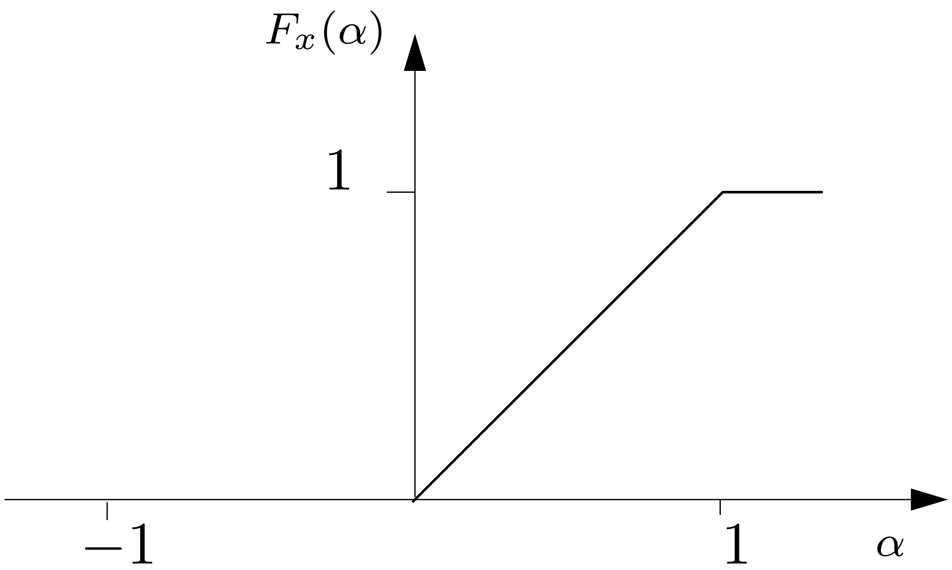
Statistical Characterization of a RV

Definition Probability distribution function (PDF): For a real-valued RV x the PDF $F_x(\alpha)$ is given by,

$$F_x(\alpha) = Pr\{x \leq \alpha\}$$

Examples: Tossing a “fair” coin a “fair” roulette



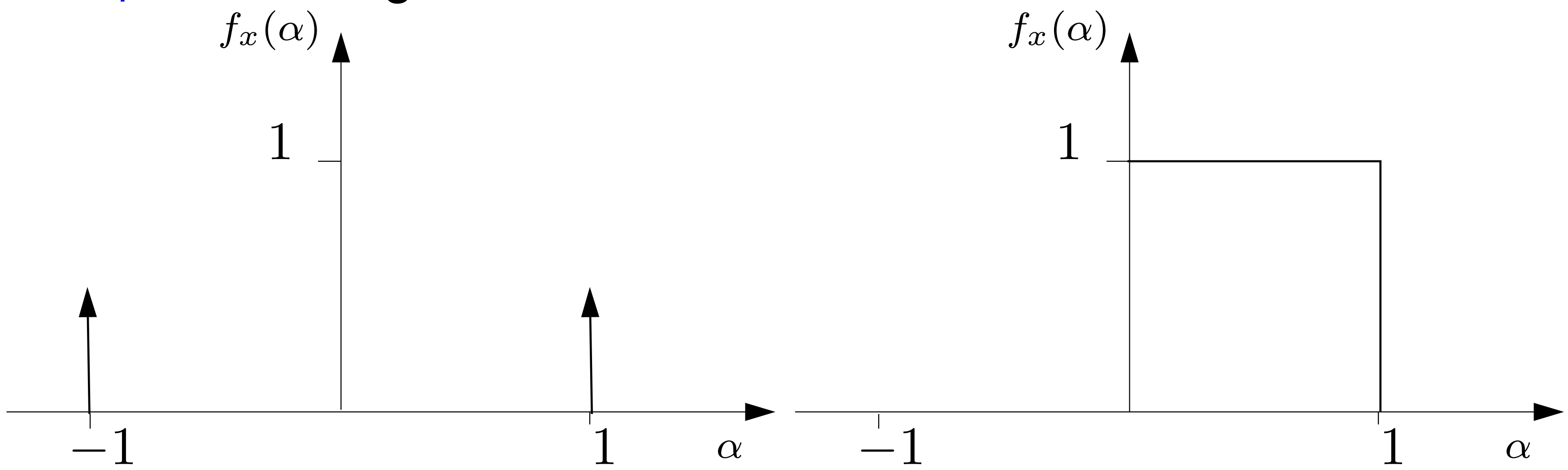


Statistical Characterization of a RV (2)

Definition Probability density function (pdf): For a real-valued RV x the pdf $f_x(\alpha)$ is derived from its PDF as,

$$f_x(\alpha) = \frac{dF_x(\alpha)}{d\alpha}$$

Examples: Tossing a “fair” coin a “fair” roulette



Approximation: An (un-normalized) approximation of a pdf of a RV is given by the Histogram based on empirical trials.

Summary Definition RVs

A random variable x is fully characterized by

- The definition of its sample space Ω
- The definition of its Probability density function (pdf) $f_x(\alpha)$

Random Variables:

1. An Example
2. Discrete and Continuous Random Variables (RVs)
3. Characterization of RVs:
   - (theoretical) via the Probability Distribution/Density Function.
   - **(practical) via Ensemble Averages**
4. Joint (multiple) RVs

The expectation operator $E[.]$

Example: Let x be the RV representing the number of eyes on a die. Assume that we thrown a fair die N_T times and that the number k appears n_k times, Then the average value that is thrown is given by the (ensemble) sample mean:

$$< x >_{N_T} = \frac{n_1 + 2n_2 + 3n_3 + 4n_4 + 5n_5 + 6n_6}{N_T}$$

Definition mean or expected value: For a discrete (continuous) RV x that assumes values α_k with probability $Pr\{x = \alpha_k\}$ is defined as:

$$E[x] = \sum_{k \in \Omega} \alpha_k Pr\{x = \alpha_k\} (= \int_{-\infty}^{\infty} \alpha f_x(\alpha) d\alpha)$$

Three important ensemble averages ($x \in \mathbb{R}$)

1. *Mean square value:* $E[x^2] = \sum_k \alpha_k^2 Pr\{x = \alpha_k\}$ (for discrete RV) and in general $E[x^2] = \int_{-\infty}^{\infty} \alpha^2 f_x(\alpha) d\alpha$.

2. *Mean square error (MSE) of estimate:* Let x be an RV and let $\hat{x}$ be an estimate of x then the MSE is:

$$E[(x - \hat{x})^2]$$

3. *Variance*:
   Var$(x) = E[(x - E[x])^2] = \int_{-\infty}^{\infty} (\alpha - E[x])^2 f_x(\alpha) d\alpha$.
   It can be shown that,

$$\mathrm{Var}(x) = E[x^2] - \Big(E[x]\Big)^2$$

Random Variables:

1. An Example
2. Discrete and Continuous Random Variables (RVs)
3. Characterization of RVs:
   - (theoretical) via the Probability Distribution/Density Function.
   - (practical) via Ensemble Averages
4. **Joint (multiple) RVs**

Jointly Distributed Random Variables

Rational: When two RVs are "related" it becomes possible to "predict" one from an "observation" of the other.

Example: Flipping two "fair" coins produces the pair of random variables $\Omega = \{(-1,-1),(1,-1),(-1,1),(1,1)\}$. With the probability of each outcome $\frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2}$.

The statistical description ("relationship") of the pair of random variables $(x(1), x(2))$ is provided by the joint distribution function:

$$F_{x(1),x(2)}(\alpha_1, \alpha_2) = Pr(x(1) \leq \alpha_1 \ \underline{\text{and}} \ x(2) \leq \alpha_2)$$

or provided by the joint density function:

$$f_{x(1),x(2)}(\alpha_1, \alpha_2) = \frac{\partial^2}{\partial \alpha_1 \partial \alpha_2} F_{x(1),x(2)}(\alpha_1, \alpha_2)$$

Joint ensemble averages (Joint Moments)

Consider two random variables $x \in \mathbb{C}$ and $y \in \mathbb{C}$ then,

Definition of the correlation r_{xy}: This is the second-order joint moment,

$$r_{xy} = E[xy^*]$$

Definition of the covariance c_{xy}: Let $m_x = E[x], m_y = E[y]$, then,

$$c_{xy} = \text{Cov}(x, y) = E[(x - m_x)(y - m_y)^*] = r_{xy} - m_x m_y^*$$

Definition of the correlation coefficient ρ_{xy}: Let $\sigma_x^2 = E[|x - m_x|^2], \sigma_y^2 = E[|y - m_y|^2]$, then,

$$\rho_{xy} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

Exercise: Show that $|\rho_{xy}| \leq 1$.

Independent, Uncorrelated, Orthogonal rv's

Consider two random variables $x \in \mathbb{C}$ and $y \in \mathbb{C}$ then,

1 x, y Independent $\Leftrightarrow$ $f_{xy}(\alpha, \beta) = f_x(\alpha) f_y(\beta)$

2 x, y Uncorrelated $\Leftrightarrow$ $E[xy^*] = E[x]E[y^*]$

Therefore,

$$c_{xy} = r_{xy} - m_x m_y^* = 0$$

and independent RVs are **always** uncorrelated. The reverse is not necessarily true. Exercise: If x and y are uncorrelated then,

$$\mathrm{Var}(x + y) = \mathrm{Var}(x) + \mathrm{Var}(y)$$

3 x, y Orthogonal $\Leftrightarrow$ $E[xy^*] = 0$

Introduction and Problems

1. Organizational Details
2. Stochastic Processes in Physics
3. Four Optimal Filtering Problems
4. Course outline/Course Reader
5. Signals/Systems highlights
6. Random variables
7. **Estimation**

Estimation:

1. **Least Mean-Square estimation of one RV from another.**
2. Properties of unbiasedness and consistency of an estimator.
3. Gaussian RVs

Linear Mean Square Estimation

Consider two random variables $x \in \mathbb{R}$ and $y \in \mathbb{R}$ then,

We seek to estimate y, denoted by $\hat{y}$ from the random variable x via the linear relationship:

$$\hat{y} = ax + b \quad a, b \in \mathbb{R}$$

The Linear Mean Square Estimate minimizes the mean square error criterium,

$$\xi = E[(y - \hat{y})^2]$$

Exercise: Determine the optimal estimate and optimal value of the criterion.

Estimation:

1. Quality of estimators
2. Example: linear regression
3. Maximum Likelihood Principle
4. The Cramer-Rao lower bound
5. Example: mean of Poisson observations

The Estimator and the Estimate

Let $x_1, x_2, \cdots x_N$ measurable RV with the p.d.f. $f_x(x, \theta_0)$.
Then the function $\hat{\theta}_N = g(x_1, \cdots x_N)$ is a estimator of parameter θ_0.

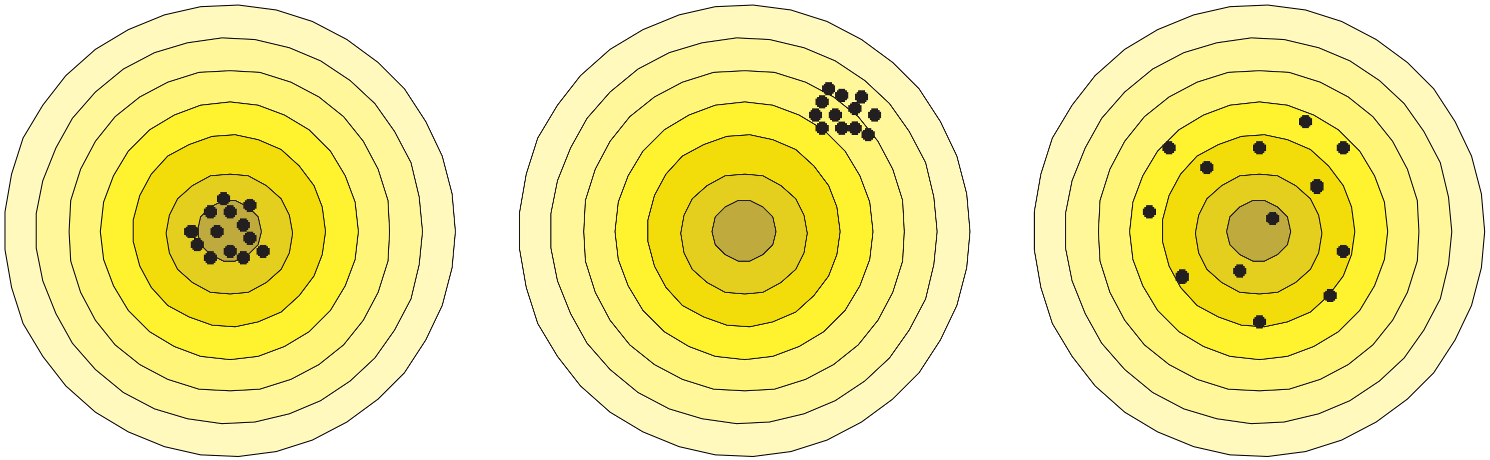
Application of this function to a set of outcomes is called an estimate .

Example estimator:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{or} \quad \tilde{x} = \frac{1}{2}[\max_i x_i + \min_i x_i]$$

as estimator of the expectation value of x.

Quality of estimators $\hat{\theta}_N$ of θ_0



Bull's eye represents θ_0;
left: unbiased estimator with small variance
middle: biased estimator with small variance
right: unbiased estimator with large variance

Parameter Estimation: Bias

More general let $\hat{\theta}_N$ be an estimate of a parameter θ_0 based on a sequence of N random variables (e.g. measurements).

Definition Bias: The bias of the estimate $\hat{\theta}_N$ is,

$$E[\theta - \hat{\theta}_N] \overset{\theta \text{ deterministic}}{=} \theta - E[\hat{\theta}_N]$$

Desired properties,

1. *Unbiased:* $E[\hat{\theta}_N] = \theta$
2. *Asymptotically unbiased:* $\lim_{N\to\infty} E[\hat{\theta}_N] = \theta$

Parameter Estimation: Consistency

Definition Consistency: An estimate $\hat{\theta}_N$ is consistent if,

1. The estimate is unbiased,

$$E[\hat{\theta}_N] = \theta$$

2. Its variance goes to zero as $N \to \infty$,

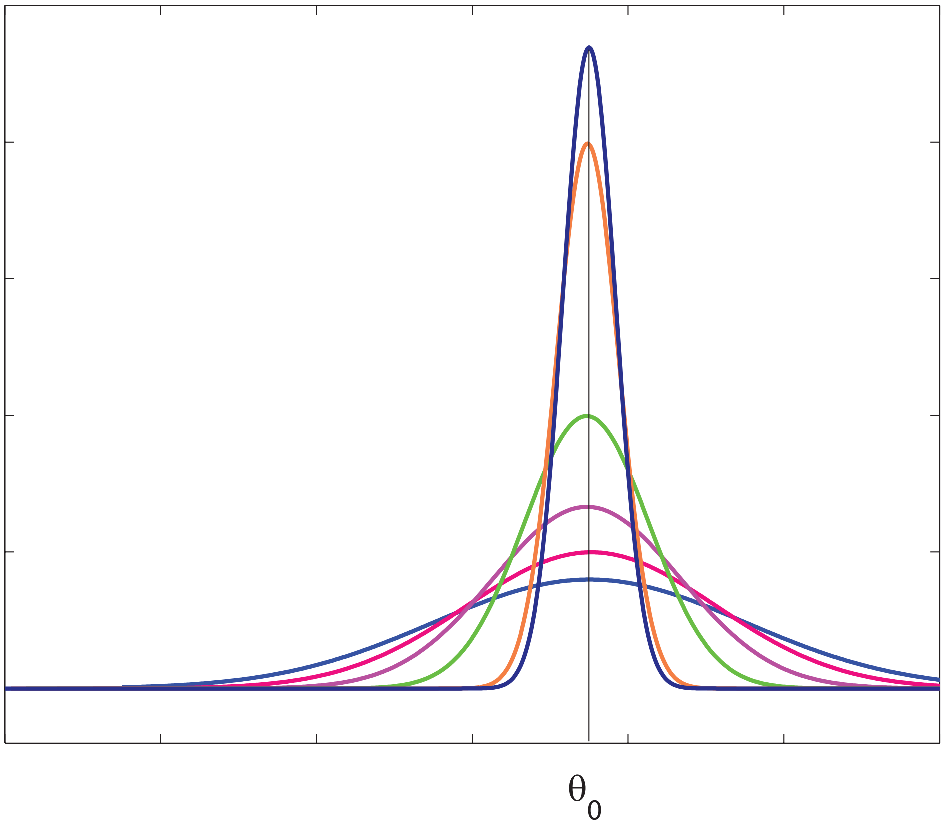
$$\lim_{N \to \infty} E[|\hat{\theta}_N - \theta|^2] = 0$$

This is a form of probabilistic convergence.

Probability density function (pdf) of a consistent estimator

Observation: An estimate is also a RV e.g. the linear mean-square estimate $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

Illustration: $f_{\hat{\boldsymbol{\theta}}_N}(\theta)$ for increasing values of N:



Estimation:

1. Quality of estimators
2. Example: linear regression
3. Maximum Likelihood Principle
4. The Cramer-Rao lower bound
5. Example: mean of Poisson distribution

Lineair regression

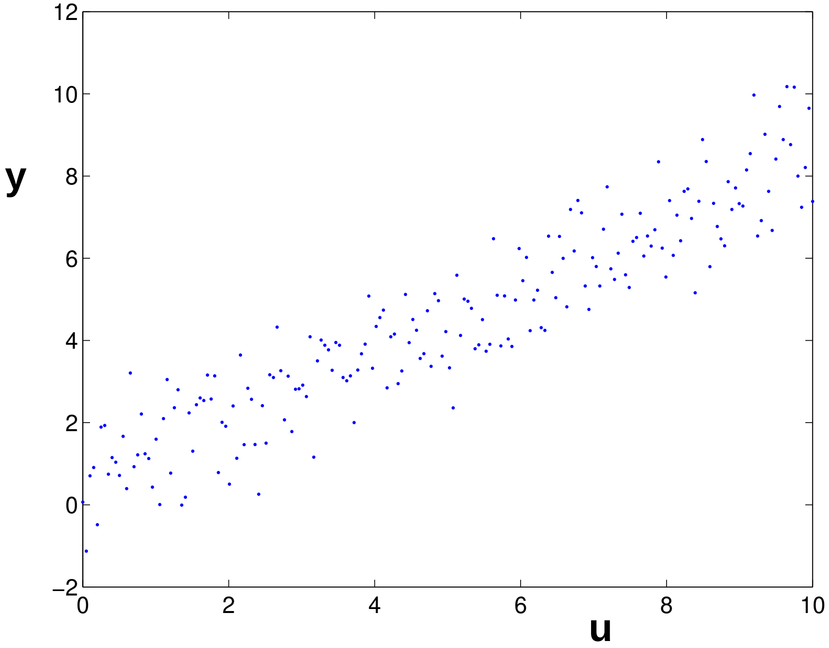
We look for a linear relation between 2 series u_i, y_i, $i = 1, \cdots N$.

Model: $y_i = b_0 + b_1 u_i$ will not match due to “disturbances”

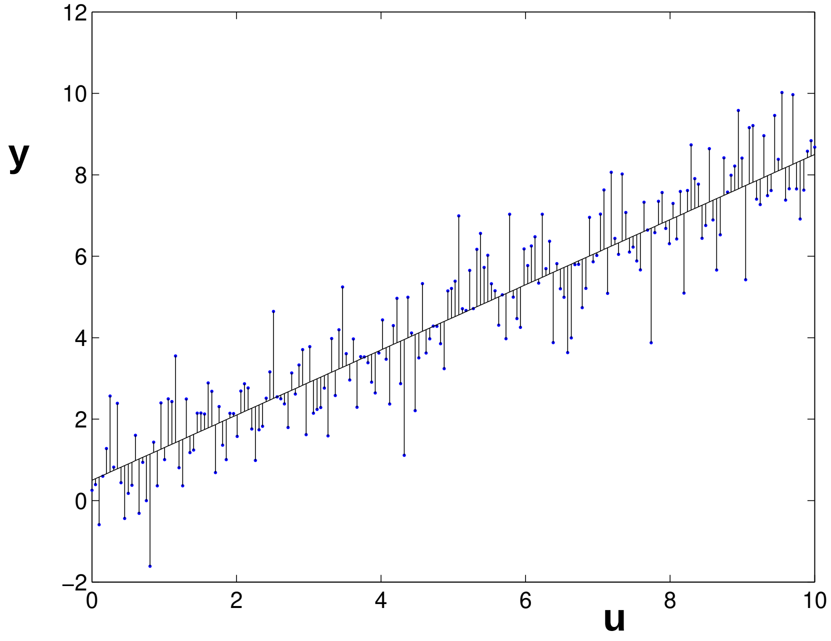
Therefore we will incorporate an error / measurement noise term into our model.

Model: $y_i = b_0 + b_1 u_i + e_i$ to describe our observations. The underlying assumption is that: u_i is noise free and y_i is disturbed.

Suggestion: find the estimator $\hat{b}_0$, $\hat{b}_1$ by minimization $\sum_i e_i^2$



y
u



y
u

Linear regression-estimator

$$y_i = b_0 + b_1 u_i + e_i, \quad i = 1, \cdots, N$$

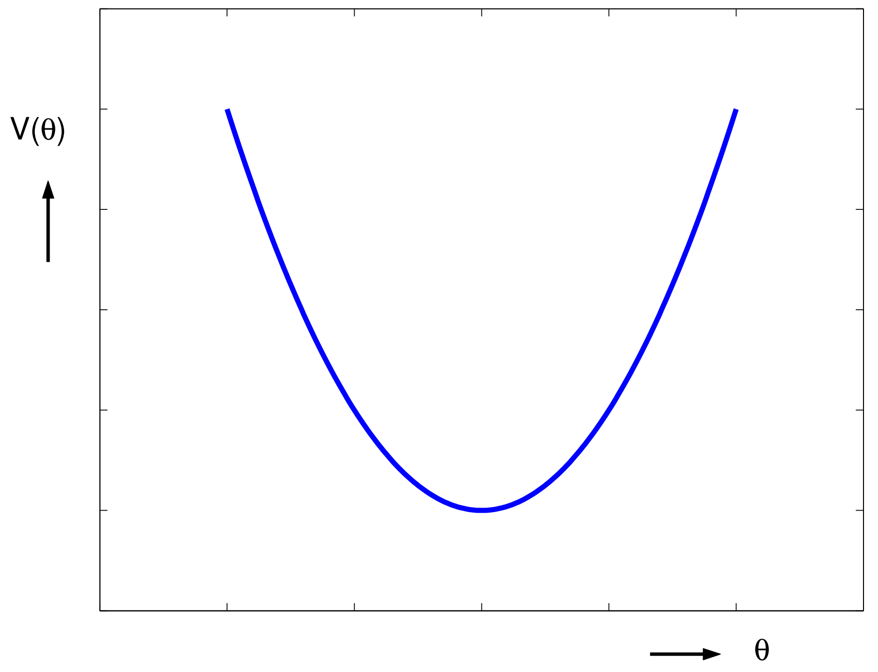
can be rewritten as

$$y_i = \phi_i^T \theta + e_i, \quad \text{where } \phi_i = \begin{bmatrix} 1 \\ u_i \end{bmatrix} \quad \text{and } \theta = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

$$V(\theta) := \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \phi_i^T \theta)^2$$

Linear regression-estimator

$$V(\theta) := \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \phi_i^T \theta)^2$$



The formal “least squares” (LS) estimator

Let u_i be deterministic and y_i a realisation of a RV, then the LS-estimator is:

$$\hat{\theta}_N = (X^T X)^{-1} X^T Y$$

with $Y_N = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$, and $X_N = \begin{bmatrix} \phi_1^T \\ \vdots \\ \phi_N^T \end{bmatrix}$

and the estimate $\hat{\theta}_N$ is a random variable.

When is θ_N unbiased?

$$Y_N = X_N\theta_0 + E_N, \qquad E_N = \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix}$$

$$\begin{aligned} \hat{\theta}_N &= (X_N^T X_N)^{-1} X_N^T Y_N \\ &= (X_N^T X_N)^{-1} X_N^T (X\theta_0 + E) \\ &= \theta_0 + (X_N^T X_N)^{-1} X_N^T E_N. \end{aligned}$$

Unbiased if $E[\hat{\theta}_N] = \theta_0$.

This is the case when $E[E_N] = 0$, i.e. $E[e_i] = 0, \forall i$.

Is also θ_N consistent?

Using

$$\theta_N \quad = \quad \theta_0 + (X_N^T X_N)^{-1} X_N^T E_N.$$

the coveriance of $\hat{\theta}_N$ is

$$E\left[(\hat{\theta}_N - \theta_0)(\hat{\theta}_N - \theta_0)^T\right] = E\left[(X_N^T X_N)^{-1} X_N^T E_N E_N^T X_N (X_N^T X_N)^{-1}\right]$$

When e is white noise with variance σ^2 then

$$E[E_N E_N^T] = \sigma^2 \cdot I$$

and therefore

$$cov(\hat{\theta}_N) = \sigma^2 \cdot (X_N^T X_N)^{-1}, \lim_{N\to\infty} cov(\hat{\theta}_N) = 0$$

Estimation:

1. Quality of estimators
2. Example: linear regression
3. Maximum Likelihood Principle
4. The Cramer-Rao lower bound
5. Example: mean of Poisson distribution

Maximum Likelihood Principle

A general principle for constructing an estimator when you know the probability density function of your observations.

Goal: estimate the unknown parameter θ in the pdf of a rv y on the basis of a set of observations (trekking) y

Example: rv y has a normal distrubtion with unit variance and unknown mean m_y

$$f_y(y;\theta) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(y-\theta)^2}{2}}$$

- For a given θ this is a pdf
- For a given y and unknown θ this is a deterministic function of $\theta \rightarrow$ likelihood function $L(\theta; y)$

Maximum Likelihood Principle

For an obervation y, determine θ so that $L(\theta; y)$ is maximum

(find the pdf that - with hindsight - is the most probable)

For 1 observation y:

$$L(\theta; y) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(y-\theta)^2}{2}}$$

maximalization of $L(\theta)$ leads to: $\hat{\theta} = y$

For n independent observations y_i:

$$L(\theta; y_1, \ldots, y_n) = f_y(y_1, \cdots, y_n; \theta) = \prod_{i=1}^{n} f_{y_i}(y_i; \theta)$$

Maximum likelihood principle

Observations: $y_1, ..., y_n$. Unknown parameter(s): θ.

1. Establish dependence of joint probability density function (pdf) of the observations on the unknown parameters:

$$f_y(y_1, ..., y_n; \theta)$$

2. Substitute available observations $y_1, ..., y_n$ (numbers) for corresponding variables in the joint pdf and consider the parameters θ as variables::

$$L(\theta; y_1, \ldots, y_n) := f_y(y_1, ..., y_n; \theta) \quad \text{Likelihood function}$$

4. Maximum Likelihood estimator:

$$\boxed{\hat{\theta}_{ML} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \log L(\theta)}$$

Example of MLE

For a fixed u we peform measurements y, and our underlying model is

$$y = \theta \cdot u + e$$

where e is a rv with pdf f_e, and θ is an unknown constant.

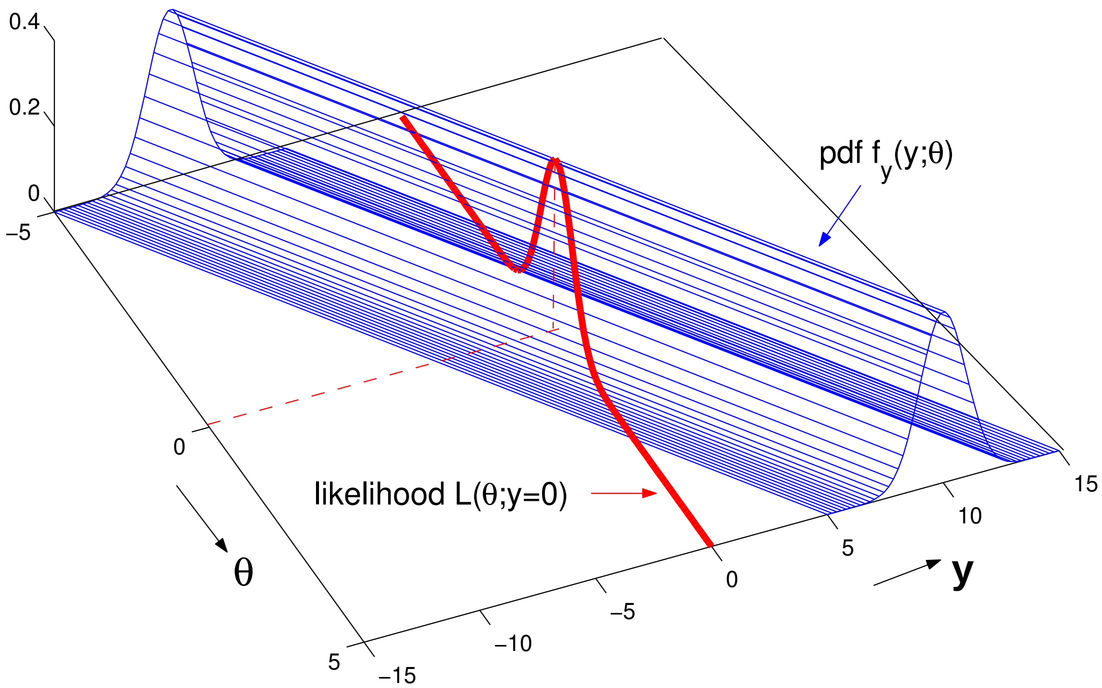
For a given θ and u the pdf of observation y is:

$$f_y(y) = f_e(\underbrace{y - \theta u}_{e})$$

or equivalently: $f_y(y;\theta) = f_e(y - \theta u)$.

Example of MLE - Linear regression

For 1 observation with model $y = \theta \cdot u + e$, at $u = 2$.



If we observe $y = 0$ then $\hat{\theta} = \arg\max_\theta L(\theta; y = 0)$.

MLE - Linear regression

Let $y_i = \phi_i^T \theta + e_i$;   with $\theta = [b_0 \; b_1]^T$, and e_i are independent rv's pdf f_e, than:

$$f_{\mathbf{y}}(y_1, y_2, \ldots, y_n; \theta) = \prod_{i=1}^{n} f_e(y_i - \phi_i^T \theta)$$

If f_e Gaussian:

$$L(\theta; Y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \phi_i^T \theta)^2}{2\sigma^2}}$$

$$-logL(\theta; Y) = \frac{n}{2} log 2\pi + n log \sigma + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \phi_i^T \theta)^2$$

Example of MLE - Linear regression

$$-logL(\theta;Y) = \frac{n}{2}log2\pi + nlog\sigma + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \phi_i^T\theta)^2$$

ML-estimator:

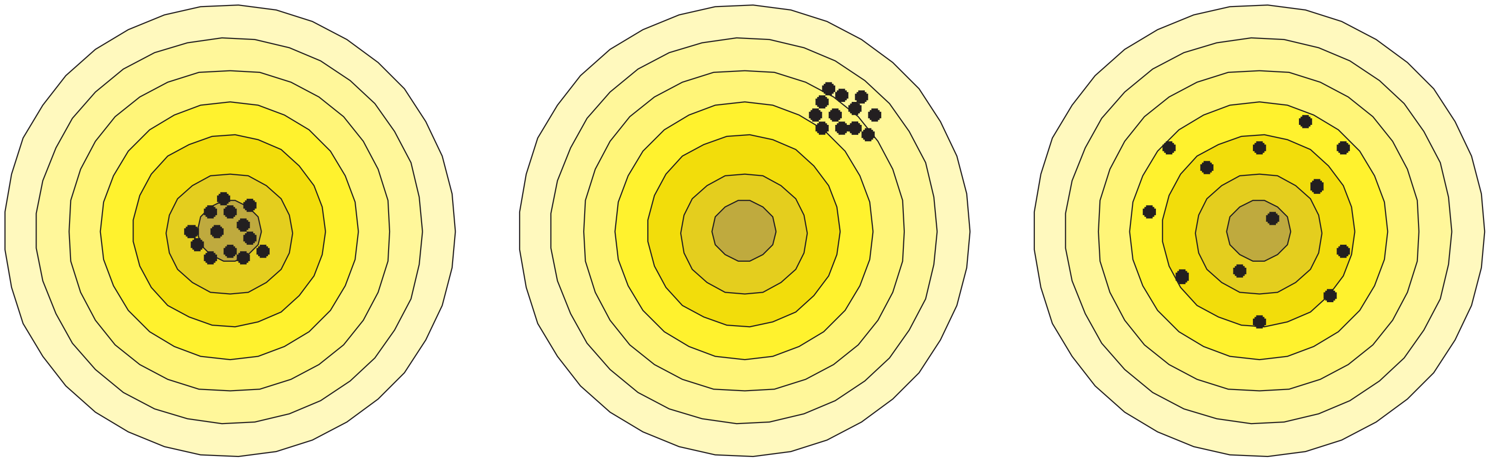
$$\hat{\theta}_{ML} = \arg\min_{\theta}\sum_{i=1}^{n}(y_i - \phi_i^T\theta)^2 = \arg\min_{\theta}\sum_{i=1}^{n}e_i^2 \quad = \text{LS}$$

For n independent observations from a Gaussian distribution with equal variance for all observations, the ML estimator is given by the simple least squares (LS) estimator.

Estimation:

1. Quality of estimators
2. Example: lineair regression
3. Maximum Likelihood Principle
4. The Cramer-Rao lower bound
5. Example: mean of an poisson distribution

Quality of estimators $\hat{\theta}_N$ of θ_0



Bull's eye represents θ_0;
left: unbiased estimator with small variance
middle: biased estimator with small variance
right: unbiased estimator with large variance

Quality of estimators (part II)

- An unbiased estimator $\hat{\theta}$ is called an efficient estimator when

$$cov(\hat{\theta}) \leq cov(\bar{\theta})$$

for all unbiased estimators $\bar{\theta}$.

- The Efficiency of a scalar unbiased estimator $\hat{\theta}_N$ is defined as

$$\frac{var(\hat{\theta}_N^{opt})}{var(\hat{\theta}_N)}$$

with $\hat{\theta}_{opt}$ as the minimum-vaiance estimator out of all unbiased estimators (assuming it exists)

The Cramer-Rao lower bound

Consider observations of a random variable y with pdf $f_y(y, \theta)$, with θ an unknown parameter.

Then for *any* unbiased estimator $\hat{\theta}_N$ of the parameter θ, it's covariance matrix satisfies

$$cov(\hat{\theta}_N) \geq J^{-1}$$

with the Fisher Information Matrix:

$$J = E\left[-\frac{\partial^2}{\partial\theta^2}\log f_y(y, \theta)\right]$$

The proof can be found in the lecture notes

Properties of the ML estimator

The ML-Estimator has the property that for $N \to \infty$

$$\hat{\theta}_N \to \mathcal{N}(\theta_0, J^{-1})$$

with J the Fisher information matrix (and J^{-1} the Cramér-Rao lower bound).
This means that the ML-estimator

- is asymptotically unbiased
- is consistent
- is asymptotically efficient (i.e., it approaches the minimal possible variance (CRLB) of all unbiased estimators)

Estimation:

1. Quality of estimators
2. Example: linear regression
3. Maximum Likelihood Principle
4. The Cramer-Rao lower bound
5. Example: mean of Poisson distribution

MLE for mean of a Poisson distribution

$$f_{y_i}(y_i;\lambda) = \frac{(\lambda)^{y_i}}{(y_i)!}\, e^{-\lambda}, \qquad E[y_i] = var(y_i) = \lambda, \forall i$$

$$L(\lambda; y_1, \ldots, y_N) = \prod_{i=1}^{n} f_{y_i}(y_i) = \prod_{i=1}^{n} \frac{(\lambda)^{y_i}}{(y_i)!}\, e^{-\lambda}$$

$$\hat{\lambda}_{ML} = \arg\max_{\lambda} L(\lambda) = \arg\max_{\lambda} logL(\lambda)$$

$$\log L(\lambda) = \sum_{i=1}^{n} \{-\lambda + y_i \log(\lambda) - \log(y_i!)\}$$

$$\left.\frac{\partial \log L}{\partial \lambda}\right|_{\lambda=\hat{\lambda}_{ML}} = 0 \rightarrow \sum_{n} \left(-1 + \frac{y_i}{\hat{\lambda}_{ML}}\right) = 0 \rightarrow \widehat{\lambda}_{ML} = \frac{1}{n}\sum_{n} y_i$$

$$E[\hat{\lambda}_{ML}] = \lambda; \qquad var(\hat{\lambda}_{ML}) = \frac{n\lambda}{n^2} = \frac{\lambda}{n}$$ = CRLB?

CRLB of mean of a Poisson distribution

$$\log L(\lambda) = \sum_{i=1}^{n} \{-\lambda + y_i \log(\lambda) - \log(y_i!)\}\,, \quad E[y_i] = var(y_i) = \lambda, \forall i$$

$$\frac{\partial^2 \log L}{\partial \lambda^2} = \sum_n -\frac{y_i}{\lambda^2}$$

$$J = E\left[\sum_n \frac{y_i}{\lambda^2}\right] = n\frac{\lambda}{\lambda^2}$$

$$CRLB(\hat{\lambda}) = J^{-1} = \frac{\lambda}{n}$$

$$E[\hat{\lambda}_{ML}] = \lambda; \quad var(\hat{\lambda}_{ML}) = \frac{n\lambda}{n^2} = \frac{\lambda}{n}$$ = CRLB

Next steps forward to improve your chances to succeed ...

Instruction session for explanation of the abstract notions and getting hands-on-exerpience!

Preparation:

Study Chapter (2 & 3 &) 4

Next Instruction/lecture see Course Overview