Stochastic signal and system analysis H05I9a/H05I7a

Patrick Wambacq

patrick.wambacq@esat.kuleuven.be

slides version 29Apr2015

KU Leuven
Department Electrical Engineering (ESAT)
Kasteelpark Arenberg 10 box 2441
3001 Leuven (Heverlee)
Belgium

Practical matters

- 9 lectures of 2 hours
- 4 exercise sessions of 2.5 hours, assignments on toledo, solutions available a few days later
- Q&A session at the end
- exam in June: oral with written preparation
- open book exam: syllabus, slides, documents on toledo allowed
- exam questions: exercises, examples on toledo
- textbook: "Random Processes for Image and Signal Processing" by E. Dougherty
- these slides: toledo, vtk
- extra information (texts, applets, . . .) on toledo

Table of contents

Part I

Introduction

Complete randomness?

- does it exist? ⇒ theology, quantum physics (uncertainty principle)
  determinism vs. indeterminism
- existence is irrelevant for the validity of a probabilistic approach, eg. prediction of i(t) in a thermally excited resistor R would require tracking the position and interaction of 10^{23} or so electrons ⇒ inconceivable, use probability theory
- probability theory also needed because in the real world not all influences to an effect can be calculated or measured

Kinds of probability

1. intuition: “he probably drove too fast”, “there is maybe an alligator in the pond”, lottery tickets, ... ⇒ theory of ”intuitive probability” (Koopman)
2. classical (non experimental): compute a priori ratio of favourable outcomes/all possible outcomes, eg. probability of obtaining a total of 7 with two dice

2nd die	1st die					
	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

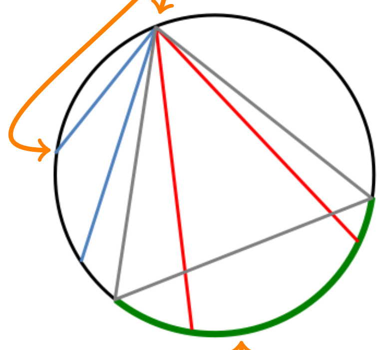
⇒P(total=7)
=6/36=1/6

Classic probability

- problem 1: all outcomes need to be equally probable
- problem 2: ambiguity is possible in case of uncountable number of outcomes, eg. Bertrand paradox
- Bertrand paradox: equilateral triangle inscribed in a circle; what is the probability that a randomly chosen chord in the circle is longer than the side of the triangle?
  Different valid approaches lead to different conclusions

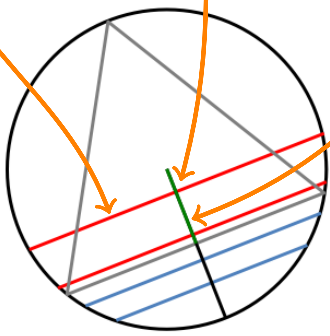
Bertrand paradox

- first approach: choose two endpoints at random $\rightarrow P = \frac{1}{3}$ (arc length $= \frac{1}{3}$ of the circumference)



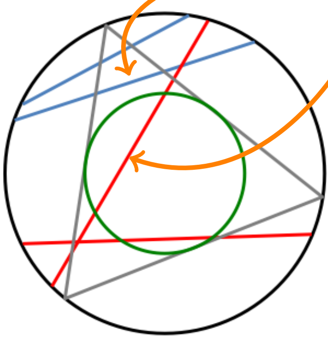
Bertrand paradox

- second approach: choose a point on the radius at random and construct chord perpendicular to this radius $\rightarrow P = \frac{1}{2}$



Bertrand paradox

- third approach: choose at random the mid point of the chord $\rightarrow P = \frac{1}{4}$ = area of the smaller concentric circle

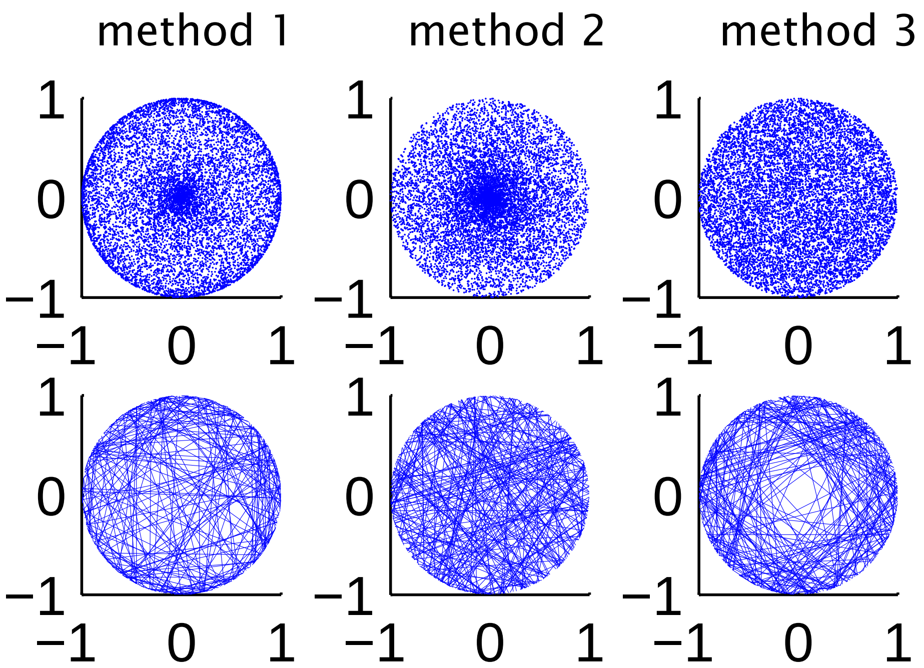


Bertrand paradox

- solution: specify the method of random selection, this leads to unique solution
- principle of “maximum ignorance”: don’t use any information that is not given in the problem statement; here: use scale and translation invariance, solution must be invariant to size and position of the circle → method 2

Bertrand paradox

distribution of the midpoints and of the chords



Kinds of probability

3. experimental: measure frequency of occurrence

$$P(E) = \lim_{n\to\infty} \frac{n_E}{n}$$

   problem: very low P that when flipping a coin 1000 times, we get exactly 500 heads (and P decreases with increasing experiment size)
4. based on axioms: modern approach used in this course

Misuses and paradoxes

- eg. defendant in a murder trial tries to convince the jury of the innocence of his client with the following statement: "probability that someone who beats his wife, also actually murders her, is 0.001"
  But: more significant: "given that a woman was beaten by her husband and that she was murdered: probability that the husband is the murderer is > 0.5"
- women aged 35 to 50: 4/100 chance on breast cancer within the year.
  other group of women aged 45 to 90: chance 11/100
  $\rightarrow$ probability for particular woman "Mrs. Smith" aged 49?
  What if her mother also suffered from breast cancer and she is a smoker?
  What if only two women are very much like Mrs. Smith and one of them gets breast cancer?

APPLICATIONS: very large diversity

- Communication
- Modelling in electronics, electrical engineering
- Instrumentation
- Acoustics
- Digital audio
- Systems reliability
- Geosciences, ecologic systems
- Computer design, physics, chemistry, mechanical engineering
- Radio astronomy
- ...

Applications in electronics

- consumer electronics: microwave ovens, hifi audio, CD/DVD, ...
- research and development:
  - electron emission
  - recombination
  - ionization
  - lifetime of charge carriers
  - physical constants of materials and electronic components
  - IC testing
  - ...

Applications in electronics, continued

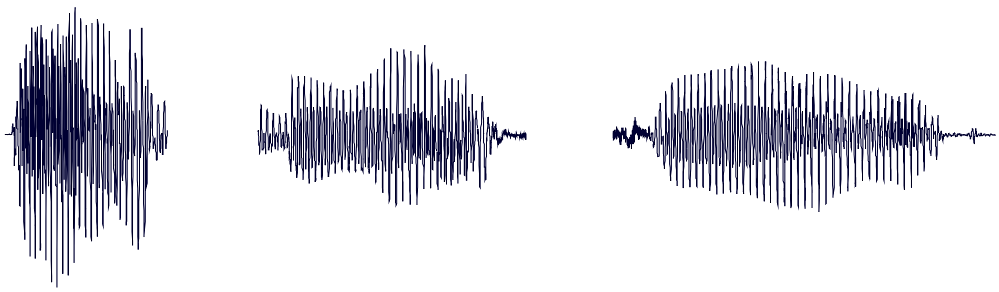
- image and speech processing
- circuits modelling
- biological and medical applications
- generation of test signals for antenna systems, communication, electronic countermeasures, signal detection, . . .

Example: speech processing

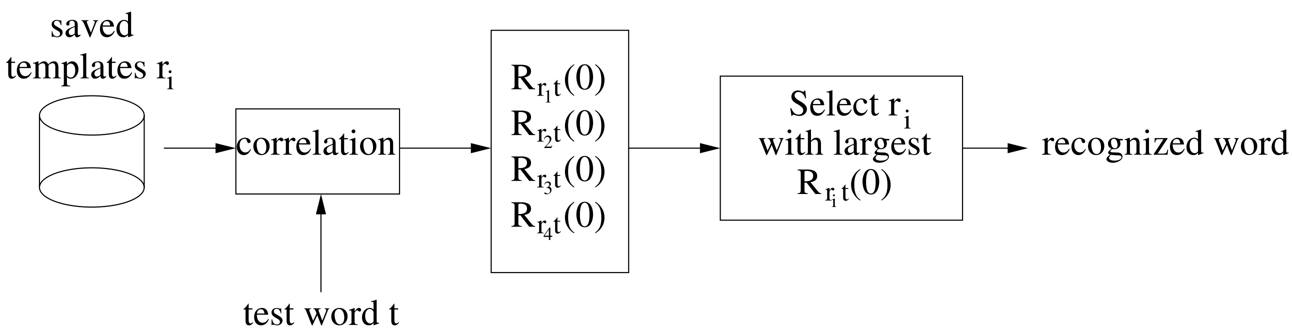
- recognition of spoken commands through template matching
- digitize small vocabulary (waveform, or spectrum, or extracted features)
- correlation between digitized vocabulary and test example (alignment required)
- highest value of $R_{rt}(0)$ gives recognition result
- ! this is a greatly simplified procedure ! (environmental influences, speaker variation, vocabulary size, . . .)

Example: speech processing

vocabulary	to	recognize	speech
template			



Example: speech processing



Variability exists between speakers and between several utterances of the same speaker ⇒ describe in terms of probability and random processes ⇒ models for speech production and speech recognition.

Example: noise suppression

- measure for amount of noise: SNR, S/N ratio
- when models are known, filter can be designed that maximizes S/N
- other criteria are possible (eg perceptive)

Example: error correction

- error detection possible through addition of redundancy
- CD: Reed-Solomon codes: up to 4000 bit errors in a row are repairable (=1/20 sec at 16bits/sample and 44.1kHz)
- design of such codes (eg also Huffman): probability theory, stochastic signals

Why study stochastics?

- digital signal processing: study of signals in time and frequency domain, however not accounting for variability and noise
- probability and stochastic processes can model variability and noise, which is crucial for the performance of many systems
- deterministic signal $\Leftrightarrow$ random phenomenon?
- noise can have deterministic as well as random properties
- describe random properties in terms of properties of an ensemble (collection of realizations) – such as in speech recognition example: many speakers and many examples of every speaker

Part II

Probability Theory

Outline

Events

- sample space S contains all possible outcomes of an experiment
- experiment = observation of all physical quantities associated with a stochastic variable
- sample $s \in S$ is undividable result of an experiment
- event $E \subseteq S$ does/does not happen
- stochastic variable X is function of S, maps S on $\mathbb{R}$
- $x = X(s)$ is a realization of X
- eg: tossing coins, throwing dice, measure voltage

σ-algebra

- σ-algebra = collection $\mathcal{E}$ of subsets of S, satisfying three conditions:
  - $S \in \mathcal{E}$
  - if $E \in \mathcal{E}$, then also $E^c \in \mathcal{E}$ with E^c the complement of E
  - if (possibly infinite but countable) collection $E_1, E_2, \ldots \in \mathcal{E}$ then also $E_1 \cup E_2 \cup \cdots \in \mathcal{E}$
    (countable: mapping to $\mathbb{N}$ is possible)
- elements of $\mathcal{E}$ are *events*
- S is an event $\Rightarrow \emptyset$ is an event
- "normal" algebra: same conditions except no infinite unions: if $A, B \in \mathcal{E}$ then $A \cup B \in \mathcal{E}$
- if $S = \{1, 2, 3\}$, then $\mathcal{E} = \{\emptyset, \{1\}, \{2\}, \{2, 3\}, \{1, 3\}, S\}$ is not an algebra because eg. $\{1\} \cup \{2\} \notin \mathcal{E}$
  $\mathcal{E} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}, S\}$ is an algebra

σ-algebra

- example: $\mathcal{E}$ = set of half open intervals $A_j = (a_j, 1]$ with $0 < a_j < 1$ and $a_j \in \mathbb{Q}$, choose
  $A_j = (1/3, 1], (1/3.1, 1], (1/3.14, 1], (1/3.141, 1], \ldots$
  $\rightarrow \bigcup_{j=1}^{\infty} A_j = (\pi^{-1}, 1] \notin \mathcal{E}$ because $\pi^{-1} \notin \mathbb{Q}$, hence this is not a σ-algebra.
- probability measure could also be defined on an algebra instead of on a σ-algebra, but then some meaningful and useful results would not hold, eg. the laws of large numbers

σ-algebra

- De Morgan: $\bigcap_{i=1}^{\infty} E_i = \left(\bigcup_{i=1}^{\infty} E_i^c \right)^c$
- $\Rightarrow$ intersection of events is an event, in particular set difference $E_2 - E_1 = E_2 \cap E_1^c$ is an event
- *Borel* σ algebra: smallest possible σ algebra over $\mathbb{R}$ consisting of all open intervals (a, b) with $-\infty \leq a < b \leq \infty$ $\Rightarrow$ contains all intervals (open, closed, half-open-half-closed) and all their unions and intersections

Probability measure

- given S and σ algebra $\mathcal{E}$, probability measure P is a real-valued function defined on the events in $\mathcal{E}$, such that
  1. $P(E) \geq 0$ for all $E \in \mathcal{E}$
  2. $P(S) = 1$
  3. if $E_1, E_2, \ldots$ is a disjoint collection of events then

$$P\left(\bigcup_{n=1}^{\infty} E_n \right) = \sum_{n=1}^{\infty} P(E_n)$$

  (if not disjoint then $P\left(\bigcup_{n=1}^{\infty} E_n \right) \leq \sum_{n=1}^{\infty} P(E_n)$)

- triplet $(S, \mathcal{E}, P)$ = probability space

Probability measure: example

- $S = \{a_1, a_2, \ldots, a_n\}$
- for every event $\{a_i\} \in \mathcal{E}$ assign a nonnegative value $P(\{a_i\})$ such that $\sum_{i=1}^{n} P(\{a_i\}) = 1$
- for every event $E = \{e_1, e_2, \ldots, e_m\} \subset S$ define $P(E) = \sum_{i=1}^{m} P(\{e_i\})$ and define $P(\emptyset) = 0$
- then P is a probability measure on $\mathcal{E}$
- for simplicity denote $P(\{a_i\})$ as $P(a_i)$
- special case: equiprobability, $P(a_i) = 1/n$, then $P(E) = m/n$ with m the cardinality of E

Probability measure: properties

- derived from axioms
- $P(E^c) = 1 - P(E)$
- $P(\emptyset) = P(S^c) = 1 - P(S) = 0$
- for $E_1, E_2 \in \mathcal{E}$ and $E_1 \subset E_2$: $P(E_2 - E_1) = P(E_2) - P(E_1)$ and since $P(E_2 - E_1) \geq 0$ it follows that $P(E_1) \leq P(E_2)$
- $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$
- theorem

$$P\left(\bigcup_{k=1}^{n} E_k\right) = \sum_{j=1}^{n} (-1)^{j+1} \sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq n} P\left(\bigcap_{k=1}^{j} E_{i_k}\right)$$

compare to $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

Probability measure: properties

- theorems about continuity:

$$\text{if } E_1 \subset E_2 \subset E_3 \subset \cdots : \quad P\left(\bigcup_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty} P(E_n)$$

(continuity from below)

$$\text{if } E_1 \supset E_2 \supset E_3 \supset \cdots : \quad P\left(\bigcap_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty} P(E_n)$$

(continuity from above)

Conditional probability

- probability of event E given that F has happened
- $P(E|F) = \dfrac{P(E \cap F)}{P(F)}$
- theorem: $P(\cdot|F)$ satisfies the probability axioms: $P(E|F) \geq 0$, $P(S|F) = 1$ and

$$P\left(\bigcup_{n=1}^{\infty} E_n|F\right) = \sum_{n=1}^{\infty} P(E_n|F) \quad \text{(if } E_i \text{ are disjoint)}$$

- $P(E \cap F) = P(F)P(E|F)$, for more events: $P(E_1 \cap E_2 \cap \cdots \cap E_n) =$ $P(E_1)P(E_2|E_1)P(E_3|E_1, E_2) \cdots P(E_n|E_1, E_2, \ldots E_{n-1})$ where $P(E_3|E_1, E_2) = P(E_3|E_1 \cap E_2)$

Bayes' rule

- $P(F|E) = \dfrac{P(F \cap E)}{P(E)} = \dfrac{P(E \cap F)}{P(E)} = \dfrac{P(F)P(E|F)}{P(E)}$
- assume that $F_1, F_2, \ldots, F_n$ form a partition of S
- for event $E \subset S$ it holds that $P(E) = \sum_{k=1}^{n} P(E \cap F_k)$
- Bayes' rule:

$$P(F_k|E) = \frac{P(F_k)P(E|F_k)}{\sum_{i=1}^{n} P(F_i)P(E|F_i)}$$

- *a priori probabilities* $P(F_i)$ and $P(E|F_i)$ are determined empirically and lead to *a posteriori probability* $P(F_k|E)$

Bayes' rule: example

- Speech recognition: what word W (from a vocabulary $\{W_k\}$) has the highest probability of being spoken, given the observed acoustic data with feature vector $\{x_{1\ldots n}\}$?
- $$W = \arg\max_k P(W_k|x_{1\ldots n}) = \arg\max_k \frac{P(x_{1\ldots n}|W_k)P(W_k)}{P(x_{1\ldots n})}$$
  $$= \arg\max_k P(x_{1\ldots n}|W_k)P(W_k)$$
- $P(x_{1\ldots n}|W_k)$ = probability of observations for the given vocabulary (established by prior training)
- $P(W_k)$ = a priori probability of the occurence of a some word (established by prior training)
- search strategy is needed to find W

Independence

- E and F are independent if $P(E \cap F) = P(E)P(F)$, or (if $P(F) > 0$) if and only if $P(E|F) = P(E)$
- in general: $E_1, E_2, \ldots, E_n$ are independent if for any subset $\{E_{i_1}, E_{i_2}, \ldots, E_{i_m}\} \subset \{E_1, E_2, \ldots, E_n\}$ it holds that

$$P\left(\bigcap_{j=1}^{m} E_{i_j}\right) = \prod_{j=1}^{m} P(E_{i_j})$$

- !! pairwise independence of $E_1, E_2, \ldots, E_n$ does not guarantee independence of the entire set!!

Independence: example

- a system composed of m components $C_1, C_2, \ldots, C_m$ that can break independently from each other (events $F_1, F_2, \ldots, F_m$). Event F means failure of the entire system
- *series arrangement*: system fails when any component fails $F = \bigcup_{k=1}^{m} F_k$

$$P(F) = 1 - P(F^c) = 1 - P\left(\bigcap_{k=1}^{m} F_k^c\right) = 1 - \prod_{k=1}^{m}(1 - P(F_k))$$

- *parallel arrangement*: system fails only when all components fail

$$P(F) = P\left(\bigcap_{k=1}^{m} F_k\right) = \prod_{k=1}^{m} P(F_k)$$

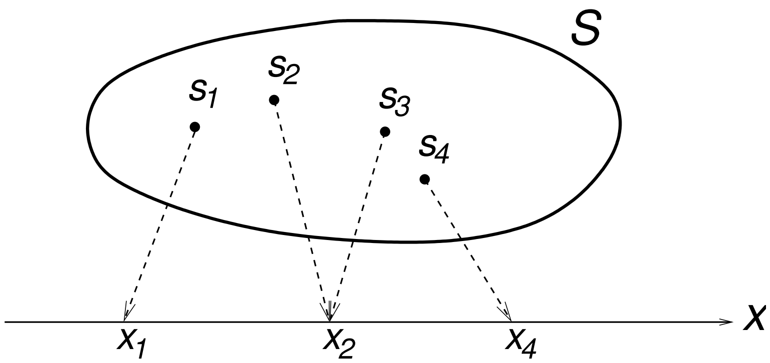
Outline

2 **Probability theory**

- Probability Space
- Random Variables
- Important Probability Distributions
- Multivariate Distributions
- Functions of Several Random Variables
- Laws of Large Numbers
- Parametric Estimation via Random Samples
- Maximum-Likelihood Estimation
- Entropy

Random variables

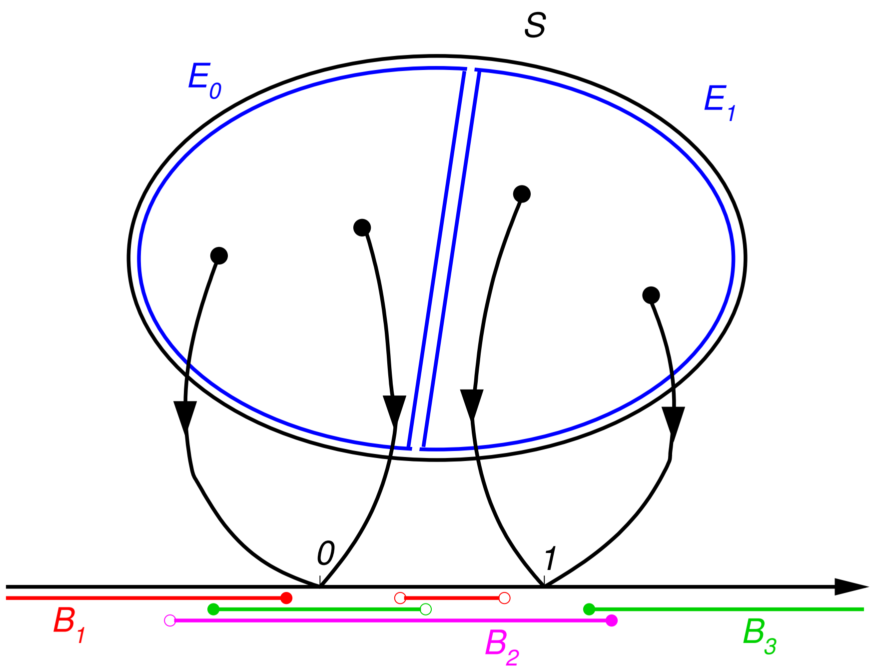
- *random variable* X is a mapping $X : S \to \mathbb{R}$ such that $X^{-1}((-\infty, x]) = \{z \in S : X(z) \leq x\}$ belongs to $\mathcal{E}$ (an event) $\forall x \in \mathbb{R}$
- if X is a random variable, then $X^{-1}(B)$ is an event ($\in \mathcal{E}$) for any Borel set $B \subset \mathbb{R}$, in particular if B is an open set, closed set, intersection of open sets, union of closed sets



Random variables

- theorem: a random variable X on a probability space $(S, \mathcal{E}, P)$ induces a probability measure P_X on the Borel σ-algebra in $\mathbb{R}$ by
$P(X \in B) = P_X(B) = P(X^{-1}(B)) = P(\{z \in S : X(z) \in B\})$
- hence, a random variable X induces a probability space $(\mathbb{R}, \mathcal{B}, P_X)$ on the real axis. If we are only concerned with X and its inclusion probabilities $P(X \in B)$, we only need P_X and need not to worry about the original sample space S. Modeling over the real axis suffices!

Random variables: example



$P_X(B_1) = 0$ because $B_1 \cap \{0, 1\} = \emptyset$
$P_X(B_2) = 1$ because $B_2 \cap \{0, 1\} = \{0, 1\}$
$P_X(B_3) = P(E_0)$ because $B_3 \cap \{0, 1\} = \{0\}$

Probability distribution

- *probability distribution function* $F_X : \mathbb{R} \to [0,1]$, defined as $F_X(x) = P(X \leq x) = P_X((-\infty, x])$
- interval probabilities ($a < b$):

$$
\begin{aligned}
P(a < X \leq b) &= F_X(b) - F_X(a) \\
P(a \leq X \leq b) &= F_X(b) - F_X(a) + P(X = a) \\
P(a < X < b) &= F_X(b) - F_X(a) - P(X = b) \\
P(a \leq X < b) &= F_X(b) - F_X(a) + P(X = a) - P(X = b)
\end{aligned}
$$

Probability distribution

- theorem: if F_X is the probability distribution function for the random variable X, then
  1. F_X is increasing
  2. F_X is continuous from the right
  3. $\lim_{x \to -\infty} F_X(x) = 0$
  4. $\lim_{x \to \infty} F_X(x) = 1$
- conversely, for any function F satisfying these properties, there exists a probability space and a random variable X such that the probability distribution function for X is given by F

Probability density

- non negative function $f(x)$ for which $\int_{-\infty}^{\infty} f(x)dx = 1$
- $f(x)$ is a probability density and yields a distribution:

$$F(x) = \int_{-\infty}^{x} f(t)dt \quad \text{and therefore also}$$

$$F'(x) = \frac{d}{dx}F(x) = f(x) \quad \text{anywhere continuity holds}$$

- $P(X = b) = F(b) - F(a) - P(a < x < b)$, and because of continuity from above $P(a < X < b) \to 0$ if $a \to b$; also $F(b) - F(a) \to 0$ if $a \to b$ from the left. Hence $P(X = b) = 0$. Point probabilities of continuous distributions are zero!!
- $P(a < X \leq b) = \int_{a}^{b} f(t)dt$

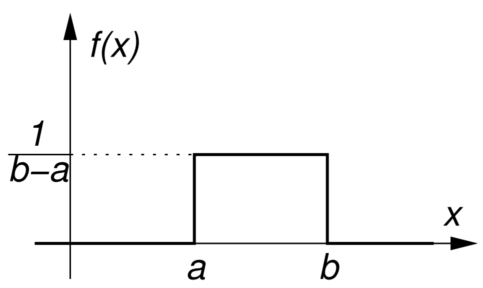
Probability density: example

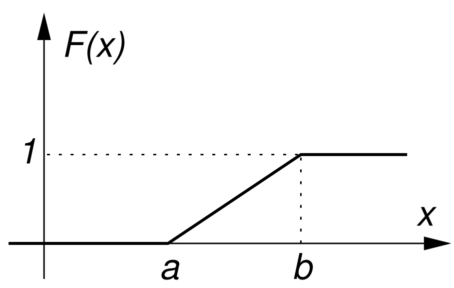
- uniform distribution over interval $[a, b]$ (with $a < b$) characterized by probability density $f(x) = 1/(b - a)$ for $a \leq x \leq b$ and $f(x) = 0$ otherwise.
  Distribution

$$F(X) = 0 \text{ for } x < a$$

$$F(X) = \frac{x-a}{b-a} \text{ for } a \leq x \leq b$$

$$F(X) = 1 \text{ for } x > b$$

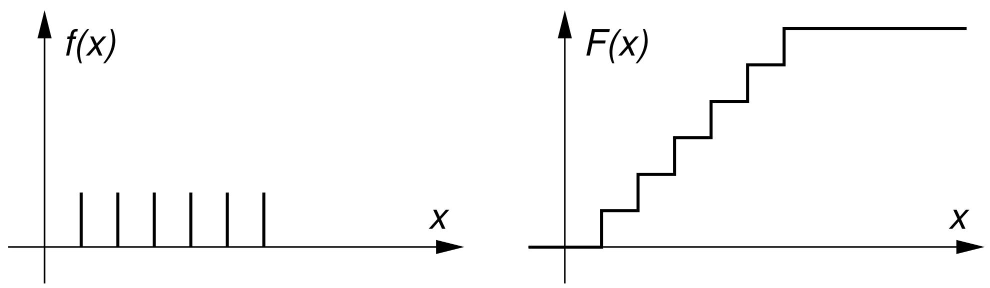




Discrete random variable

- discrete random variable modeled by countable range of points $\Omega_X = \{x_1, x_2, \ldots\}$ and non negative discrete density $f(x)$ with $f(x) > 0$ if and only if $x \in \Omega_X$ and $\sum_{k=1}^{\infty} f(x_k) = 1$
- distribution $F(x) = \sum_{\{k: x_k \leq x\}} f(x_k)$
- $F(x)$ has jumps at x_k and is constant on $[x_k, x_{k+1})$
- point probabilities $P(X = x_k) = f(x_k)$ differ from zero!!
- $P(a < X \leq b) = \sum_{\{k: a < x_k \leq b\}} f(x_k)$

Discrete random variable



- unification of continuous and discrete distributions: use delta functions: $f(x) = \sum_{k=1}^{\infty} f(x_k)\delta(x - x_k)$
- there exist *mixed* distributions: neither continuous nor discrete

Functions of a random variable

- function $Y = g(X)$ of a discrete random variable X
- $f_Y(y) = P(Y = y) = \sum_{\{x:g(x)=y\}} f_X(x)$
- if g is one-to-one then $\{x : g(x) = y\}$ consists of the single element $g^{-1}(y)$ and $f_Y(y) = f_X[g^{-1}(y)]$
- more difficult for continuous random variables; possible approach is via distribution

Functions of a random variable

- example: $Y = aX + b, a \neq 0$
- for $a > 0$

$$F_Y(y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right)$$

- for $a < 0$

$$F_Y(y) = P\left(X \geq \frac{y-b}{a}\right) = 1 - F_X\left(\frac{y-b}{a}\right)$$

- differentiation yields $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$

Functions of a random variable

- in general: if $y = g(x)$ is differentiable for all x and has a strict positive or strict negative derivative then the derivative of $x = g^{-1}(y)$ exists (Jacobian, $J(x; y) = \frac{d}{dy} g^{-1}(y)$)
- theorem:

$$f_Y(y) = \begin{cases} f_X[g^{-1}(y)]\,|J(x;y)| & \text{if } y_1 < y < y_2 \\ 0 & \text{otherwise} \end{cases}$$

with $y_1 < y_2$ such that $\forall y : y_1 < y < y_2$ there exists a single value of x such that $y = g(x)$ (possibly $y_1 = -\infty$ and/or $y_2 = \infty$)

Functions of a random variable

- example: $y = g(x) = e^{tx}, t > 0$
- for $y > 0$:

$$g^{-1}(y) = \frac{\log y}{t}$$

$$J(x; y) = \frac{d}{dy}\left(\frac{\log y}{t}\right) = \frac{1}{ty}$$

$$f_Y(y) = \frac{1}{ty} f_X\left(\frac{\log y}{t}\right)$$

- $f_Y(y) = 0$ for $y < 0$; for $y = 0$, $f_Y(0)$ can be chosen arbitrarily.

Moments

- full description of a random variable requires its probability distribution; often only a partial description is given through its moments.
- *expected value* or *expectation*: $E[X] = \int_{-\infty}^{\infty} xf(x)dx$ if the integral is absolutely convergent
- $E[X]$ = *mean value* = μ_X
- theorem: $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$ with $g(x)$ any piece-wise continuous real-valued function

Moments

- *k-th moment* about the origin $\mu'_k = E[X^k] = \int_{-\infty}^{\infty} x^k f(x)dx$
- $E[X]$ is the first moment of X
- *k-th central moment*

$$\mu_k = E[(X-\mu)^k] = \int_{-\infty}^{\infty} (x-\mu)^k f(x)dx \text{ with } \mu = \mu_X = E[X]$$

- *variance* = 2nd central moment

$$\sigma^2 = \mu_2 = E[(X-\mu)^2] = \mathrm{Var}[X] = \sigma_X^2$$

- standard deviation = $\sqrt{\text{variance}} = \sigma$
- $\sigma^2 = \mu'_2 - \mu^2$
- property: $\mathrm{Var}[aX + b] = a^2\,\mathrm{Var}[X]$

Moments: example

- uniform distribution over interval $[a, b]$
- k-th moment is

$$\mu'_k = E[X^k] = \frac{1}{b-a}\int_a^b x^k dx = \frac{b^{k+1} - a^{k+1}}{(k+1)(b-a)}$$

- for $k = 1, 2$ it follows that

$\mu = (a+b)/2, \mu'_2 = (b^2 + ab + a^2)/3$, furthermore
$\sigma^2 = (b-a)^2/12$

Chebyshev inequalities

- generalized Chebyshev inequality: if X is nonnegative and has mean μ then for any $t > 0$:

$$P(X \geq t) \leq \frac{\mu}{t}$$

- (second) Chebyshev inequality: if X (not necessarily nonnegative) has mean μ and variance σ^2, then for any $t > 0$:

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \text{ or also } P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

or also $P(|X - \mu| < t) \geq 1 - \dfrac{\sigma^2}{t^2}$

- this means that the probability mass in $(\mu - t, \mu + t)$ is bounded from below; if σ^2 is small, then the mass is tightly concentrated about the mean

Moment-generating functions

- $M_X(t) = E[e^{tx}] = \int_{-\infty}^{\infty} e^{tx} f(x) dx$ for all t for which the integral is finite
- $M_{aX+b}(t) = e^{bt} M_X(at)$
- theorem: if $M_X(t) = M_Y(t)$ for all t in some open interval that contains $t = 0$, then X and Y are identically distributed
- $M_X(t)$ can be used to find moments:

$$M_X^{(k)}(0) = \int_{-\infty}^{\infty} x^k f(x) dx = \mu'_k$$

with $M_X^{(k)}(0)$ the k-th derivative of $M_X(t)$, evaluated at $t = 0$

Moment-generating functions: example

- exponential distribution ($b > 0$): $f(x) = be^{-bx}$ for $x \geq 0$ and $f(x) = 0$ for $x < 0$
- moment-generating function:

$$M_X(t) = \int_0^{\infty} e^{tx} be^{-bx} dx = \frac{b}{b-t} \quad \text{for } t < b$$

- taking the derivative: $M_X^{(k)}(t) = \dfrac{k!b}{(b-t)^{k+1}}$
- letting $t = 0$ yields $\mu'_k = k!/b^k$, hence $\mu = 1/b, \mu'_2 = 2/b^2, \sigma^2 = 1/b^2$

Outline

Important distributions/densities

- binomial distribution: repeated independent binary trials
- Poisson distribution: fundamental class of random point processes
- normal distribution: used extensively, model for noise, limiting distribution in many cases
- gamma distribution: family of many useful distributions (eg. exponential distribution); modeling of grain sizes and interarrival times in queues
- beta distribution: takes on many shapes by modifying its parameters and is therefore useful to model various kinds of phenomena

Binomial distribution

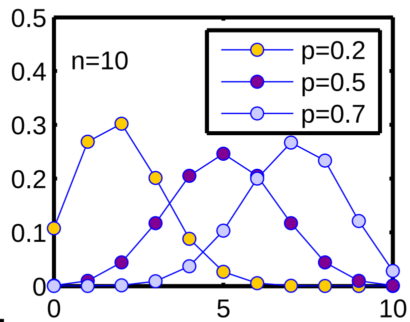
- experiment with $n > 0$ trials
- *Bernoulli trials* if
  1. sample space $S = \{s, f\}$ (success, failure)
  2. $\exists p, 0 < p < 1$ such that for every trial $P(s) = p, P(f) = q = 1 - p$
  3. trials are independent
- can be seen as random selection of n balls with replacement from urn with k black and m white balls. Selecting black ball means success, so $p = k/(k+m)$
- sample space for experiment with n Bernoulli trials $S = \{s, f\}^n$

Binomial distribution

- X counts number of successes in n trials
- density

$$f(x) = P(X = x) = \begin{cases} \binom{n}{x} p^x q^{n-x} & \text{for } x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

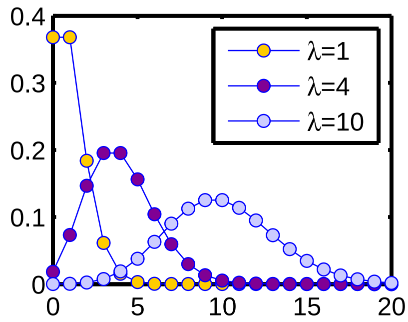
- distribution $F(x) = \sum_{k \le x} \binom{n}{k} p^k q^{n-k}$



- moment-generating function $M_X(t) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x q^{n-x} = (pe^t + q)^n$
- taking derivative and setting $t = 0$ yields $\mu = np, \mu_2' = np(1 + np - p), \sigma^2 = npq$

Poisson distribution

- Poisson distribution results from arrival process in time (see later)
- density $f(x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$ for $x = 0, 1, 2, \ldots$ and with parameter $\lambda > 0$
- distribution $F(x) = \sum_{k \leq x} \dfrac{e^{-\lambda}\lambda^k}{k!}$, hence $F(x) = 0$ for $x < 0$ and $F(x)$ has jumps at $x = 0, 1, 2, \ldots$
- moment-generating function $M_X(t) = \sum_{x=0}^{\infty} \dfrac{e^{tx}e^{-\lambda}\lambda^x}{x!} = \exp[\lambda(e^t - 1)]$
- taking derivative and setting $t = 0$ yields $\mu = \lambda, \mu_2' = \lambda + \lambda^2, \sigma^2 = \lambda$



Relation between binomial and Poisson distributions

- Poisson and binomial distribution are asymptotically related
- $\lim_{n\to\infty} b\left(x; n, \dfrac{\lambda}{n}\right) = \pi(x; \lambda)$ with binomial distribution $b(x; n, p)$ and Poisson distribution $\pi(x; \lambda)$
- hence for large n, $b(x; n, p) \approx \pi(x; np)$
- example: communication channel with error rate of 1 error per 100 (independent) messages. Sending n messages = Bernoulli trial, with $p = 0.01$

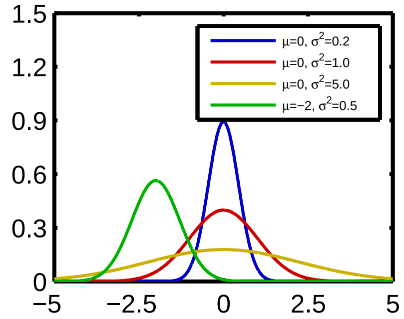
  Poisson approximation: $P(X = x) \approx \dfrac{e^{-np}(np)^x}{x!}$, yields for $n = 200, P(X \geq 3) = 0.3233$

Normal distribution

- normal or *Gaussian* distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

with $-\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$



- distribution

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy$$

- case of $\mu = 0, \sigma = 1$: standard normal distribution

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad \text{en} \quad \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{y^2}{2}} dy$$

Normal distribution

- transformation $Z = (X - \mu)/\sigma$ transforms Gaussian distributed variable X into standard Gaussian distributed variable Z
- moment-generating function

$$\begin{aligned} M_X(t) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\ &= \exp\left[\mu t + \frac{t^2\sigma^2}{2}\right] \end{aligned}$$

- taking derivative and setting $t = 0$ yields mean = μ and variance = σ^2

Gamma distribution

- involves *gamma function*, defined for $x > 0$ as

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

for $x > 0, \Gamma(x+1) = x\Gamma(x)$ and for $x \in \mathbb{N}$, $\Gamma(x+1) = x!$

- *gamma distribution* with parameters $\alpha > 0, \beta > 0$ has density

$$f(x) = \begin{cases} \frac{\beta^{-\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- distribution $F(x) = \frac{\beta^{-\alpha}}{\Gamma(\alpha)} \int_0^x t^{\alpha-1} e^{-t/\beta} dt$ for $x > 0$

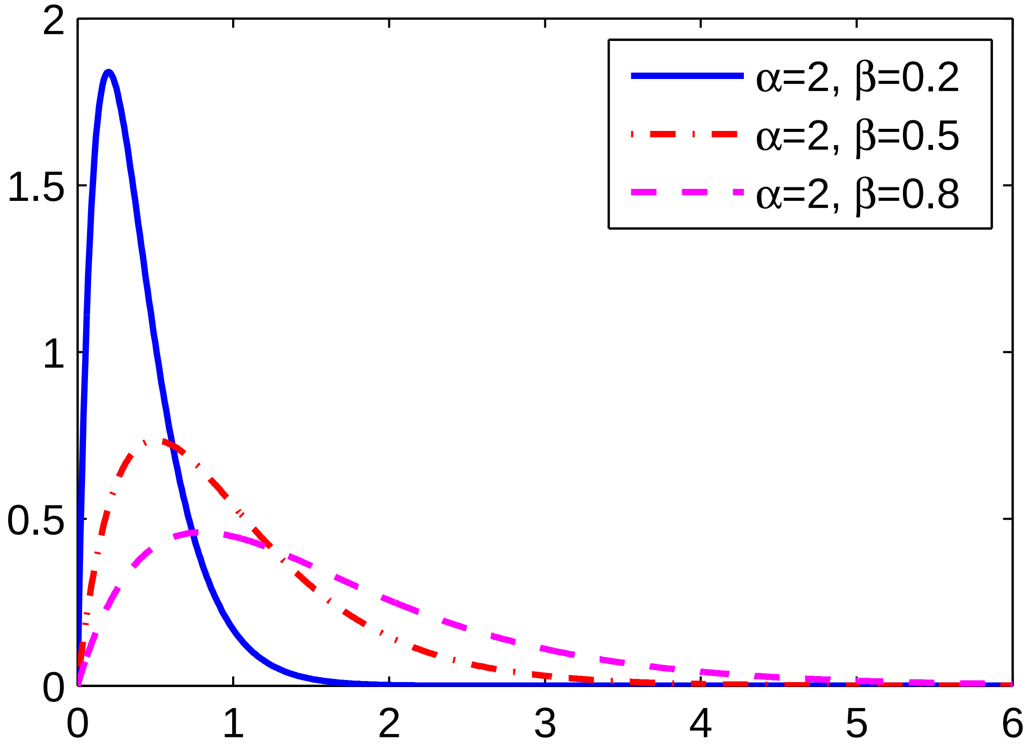
Gamma distribution

- moment-generating function for $t < 1/\beta$

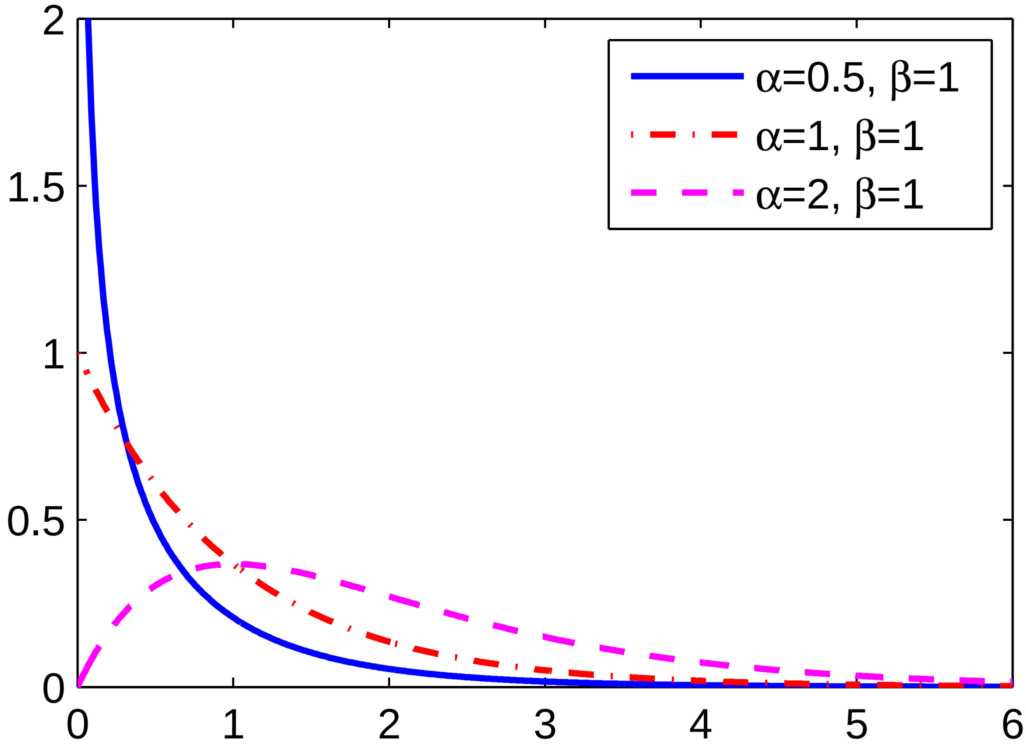
$$M_X(t) = \frac{\beta^{-\alpha}}{\Gamma(\alpha)} \int_{-\infty}^\infty x^{\alpha-1} e^{-[(1/\beta)-t]x} dx = (1 - \beta t)^{-\alpha}$$

- taking derivative and setting $t = 0$ yields $\mu = \alpha\beta, \mu_2' = \beta^2(\alpha+1)\alpha, \sigma^2 = \alpha\beta^2$
- taking $\alpha = 1$ and $\beta = 1/b$ results in the exponential distribution with salient property that it is memoryless: $P(X > x + y | X > y) = P(X > x)$

Gamma distribution



Gamma distribution



Exponential distribution: example

- time-to-failure distribution: distribution of a random variable T measuring the time until system failure.
- *reliability function*

$$R(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)du$$

- $R(t)$ is monotonically decreasing and $R(0) = 1, \lim_{t\to\infty} R(t) = 0$
- *MTTF mean time to failure* $E[T] = \int_0^\infty tf(t)dt$
- *hazard* function $h(t)$ gives the instantaneous failure rate of the system

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{R(t)} = -\frac{R'(t)}{R(t)}$$

Probability theory Important Probability Distributions

Exponential distribution: example

- hence $R(t) = \exp\left[-\int_0^t h(u)du\right]$ and
  $f(t) = h(t)\exp\left[-\int_0^t h(u)du\right]$
- constant $h(t) = q$ is a logical assumption when wear-in period has passed and wear-out stage has not yet been reached
- this gives $f(t) = qe^{-qt}, R(t) = e^{-qt}$, exponential distribution
- memoryless: probability of system working longer than time $t + v$ given that is has worked for time v, is the same as the probability of working for time t from the outset

Beta distribution

- for $\alpha > 0, \beta > 0$ *beta function* is defined as

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt$$

- it can be shown that $B(\alpha, \beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
- *beta distribution*

$$f(x) = \begin{cases} \dfrac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} & 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

- takes on many different shapes (see figures), so it can be used to model many kinds of data distributions

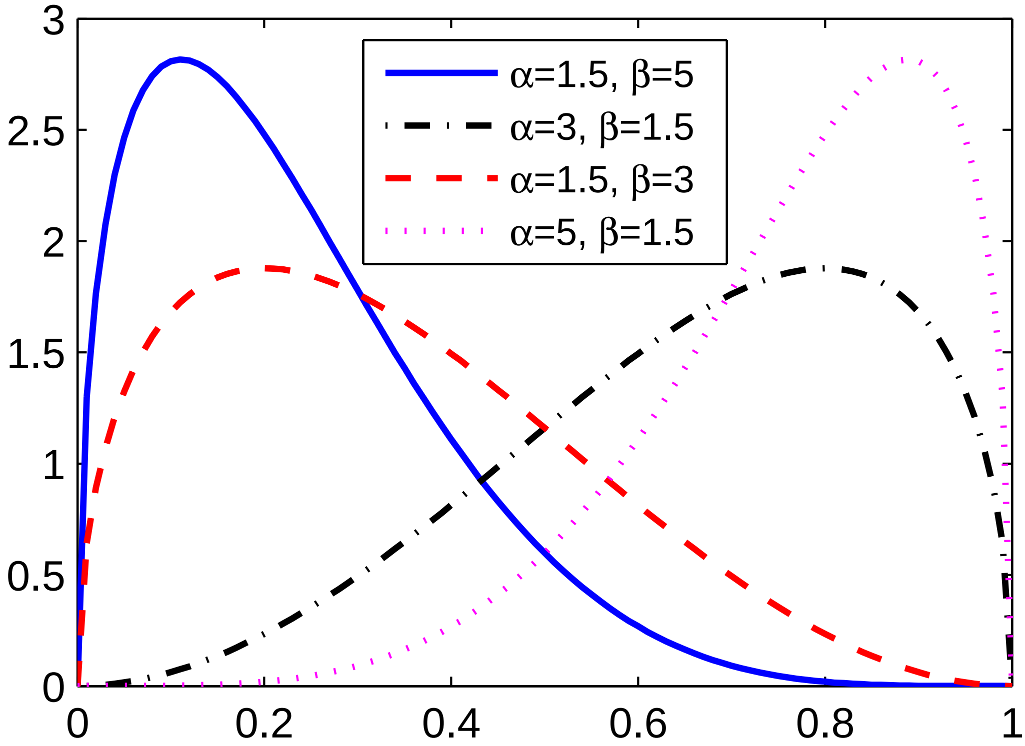
Beta distribution

- $$\mu'_k = \frac{1}{B(\alpha, \beta)} \int_0^1 x^{\alpha+k-1}(1-x)^{\beta-1}\,dx = \frac{B(\alpha+k, \beta)}{B(\alpha, \beta)}$$
  $$= \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+k)}{\Gamma(\alpha)\Gamma(\alpha+\beta+k)}$$
- $\mu = \alpha(\alpha+\beta)^{-1}, \sigma^2 = \alpha\beta(\alpha+\beta)^{-2}(\alpha+\beta+1)^{-1}$
- can be generalized so as to cover interval (a, b): generalized beta distribution has density

$$f(x) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}B(\alpha, \beta)}$$

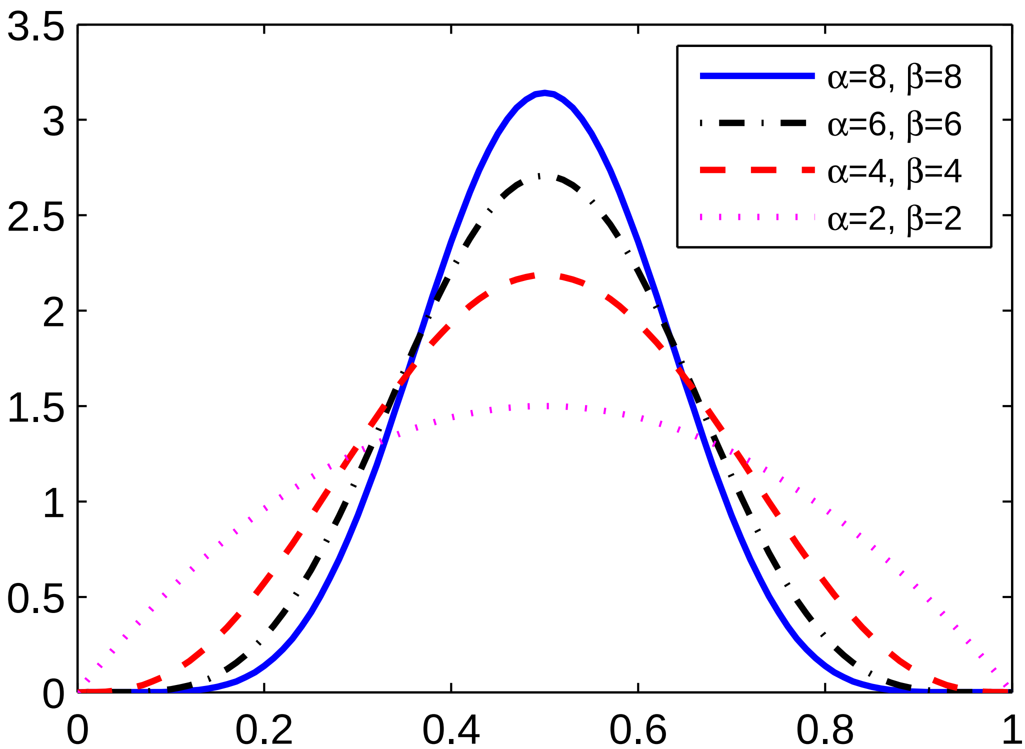
- uniform distribution is a generalized beta distribution with $\alpha = \beta = 1$

Beta distribution

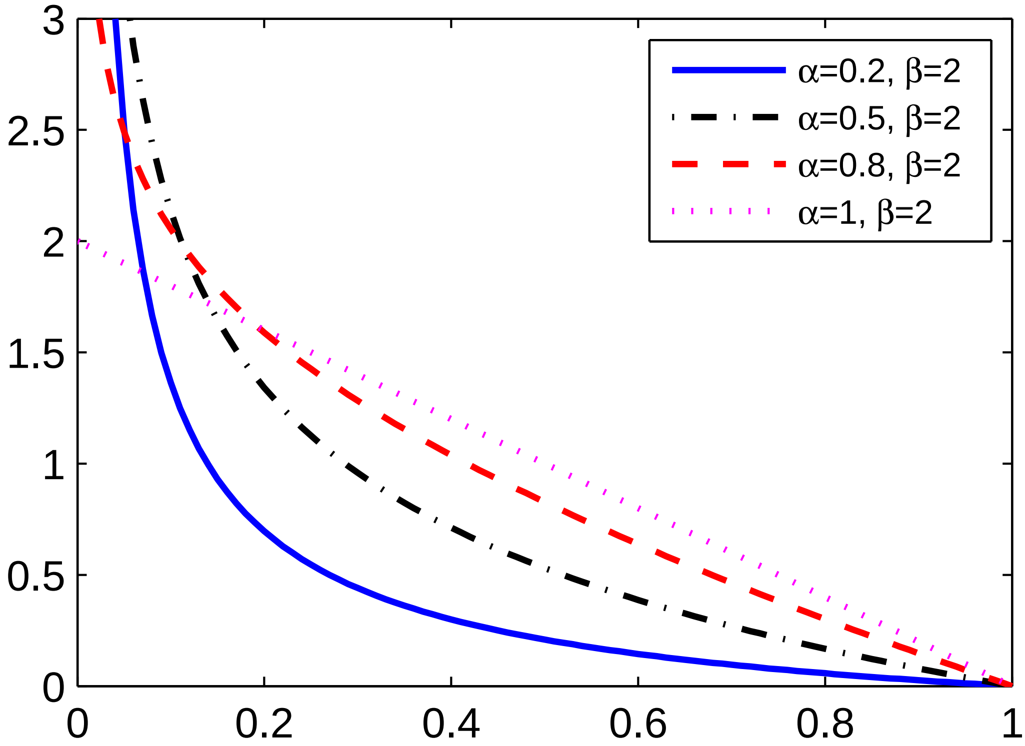


Probability theory | Important Probability Distributions

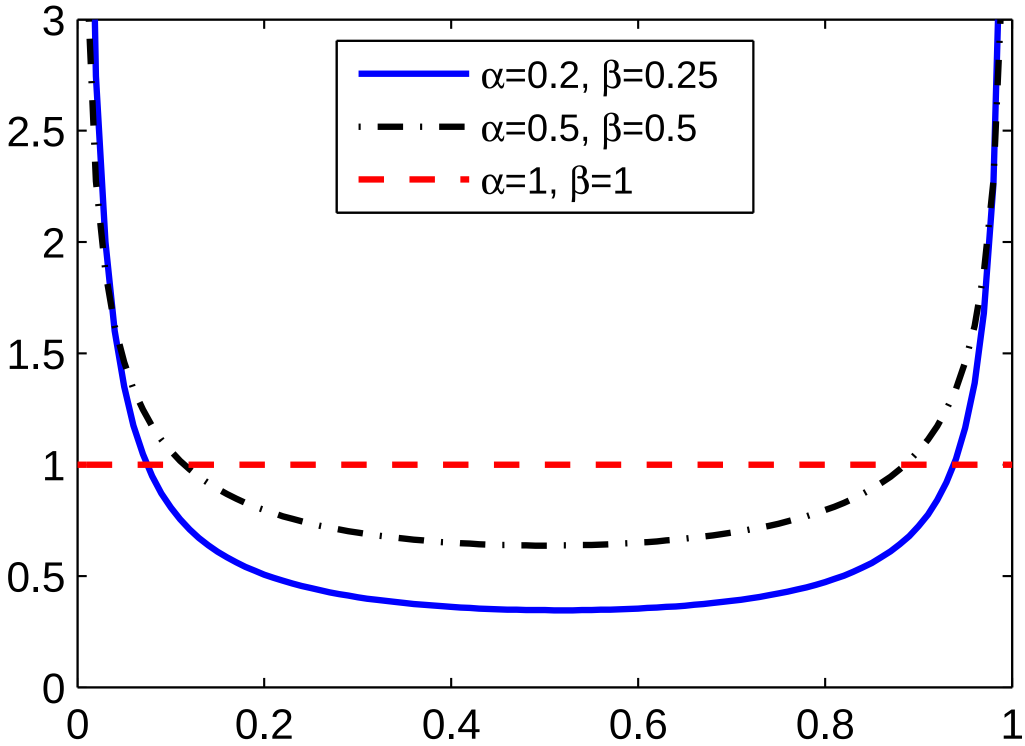
Beta distribution



Beta distribution



Beta distribution



Simulation of distributions

- simulation needed in case of too difficult or impossible analytic description
- generate (input) data according to some distribution and analyze corresponding system output
- based on random number generation: uniformly distributed U over interval $(0, 1)$
- in practice *pseudo random numbers* are generated by *random number generators*
- random number generators for normal distribution also exist

Simulation of distributions

- next step: convert random values generated for uniform distribution into new values with desired distribution
- if F is strictly increasing continuous distribution then $X = F^{-1}(U)$ has probability distribution function F:

$$F_X(x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$$

- example: exponentially distributed X with parameter b:

$$u = F(x) = 1 - e^{-bx} \Rightarrow x = -b^{-1} \log(1 - u)$$

- hence $X = -b^{-1} \log(1 - U)$ has an exponential distribution with mean $1/b$
- can be simplified to $X = -b^{-1} \log U$

Outline

Multivariate distributions

- related phenomena ⇒ related observations, study properties of collections of random variables
- for n random variables $X_1, X_2, \ldots, X_n$ define the *random vector* $\boldsymbol{X}$ as follows:

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

- $\boldsymbol{X}$ is a mapping of the sample space on the n-dimensional Euclidian space $\mathbb{R}^n$

Multivariate distributions

- distributions of $X_1, X_2, \ldots, X_n$ can be determined from the distribution of $\boldsymbol{X}$, but not vice versa (in general)
- simplicity of notation: $\boldsymbol{X}$ is always a column vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)'$
- $\boldsymbol{X}$ induces probability measure on the Borel σ-algebra in $\mathbb{R}^n$, containing all open sets in $\mathbb{R}^n$ by defining the probabilities $P(\boldsymbol{X} \in B)$ for the Borel set $B \subset \mathbb{R}^n$

Jointly distributed random variables

- for n discrete random variables $X_1, X_2, \ldots, X_n$ the *joint (multivariate) distribution* is defined by the *joint probability mass function* $f(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$
- it holds that

$$\sum_{\{(x_1, x_2, \ldots, x_n) : f(x_1, x_2, \ldots, x_n) > 0\}} f(x_1, x_2, \ldots, x_n) = 1$$

and for any Borel set $B \subset \mathbb{R}^n$:

$$P((X_1, X_2, \ldots, X_n)' \in B) = \sum_{\{(x_1, x_2, \ldots, x_n) \in B : f(x_1, x_2, \ldots, x_n) > 0\}} f(x_1, x_2, \ldots, x_n)$$

Jointly distributed random variables

- *continuous* random variables $X_1, X_2, \ldots, X_n$ possess a multivariate distribution defined by the joint density $f(x_1, x_2, \ldots, x_n) \geq 0$ if for any Borel set $B \subset \mathbb{R}^n$:

$$P((X_1, X_2, \ldots, X_n)' \in B) = \int \cdots \int_B f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n$$

- it follows also that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n = 1$$

- if this holds for a function $f(x_1, x_2, \ldots, x_n) \geq 0$ then, conversely, there exist random variables $X_1, X_2, \ldots, X_n$ that have f as their density

Example: multinomial distribution

- experiment that satisfies:
  - n independent trials;
  - every trial has r possible outcomes $w_1, w_2, \ldots, w_r$;
  - values $p_1, p_2, \ldots, p_r$ represent probabilities of the outcomes w_j
- random variable X_j counts number of times that outcome w_j occurs during n trials
- joint density of $X_1, X_2, \ldots, X_r$ is

$$f(x_1, x_2, \ldots, x_r) = \frac{n!}{x_1! x_2! \ldots x_r!} p_1^{x_1} p_2^{x_2} \ldots p_r^{x_r}$$

- multinomial coefficient

$$\binom{n}{x_1, x_2, \ldots, x_r} = \frac{n!}{x_1! x_2! \ldots x_r!}$$

Example: multinomial distribution

$n = 7$ trials
$r = 10$ possible outcomes
$p_j = 0.1$, equiprobable outcomes

Marginal densities

- case of two random variables X and Y with joint density $f(x, y)$: corresponding $f_X(x)$ and $f_Y(y)$ are called the *marginal densities*

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy \text{ , similar for } f_Y(y)$$

- in general

$$f_{X_k}(x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n) dx_n \ldots dx_{k+1} dx_{k-1} \ldots dx_1$$

- joint marginal densities

$$f_{X,Y}(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y,U,V}(x, y, u, v) dvdu$$

- discrete variables: integrals replaced by sums

Joint probability distribution function

- $F(X,Y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(r,s)drds$ with $f(x,y)$ the joint density of X and Y
- it also holds that $\frac{\partial^2}{\partial x \partial y} F(x,y) = f(x,y)$
- theorems:
  1. $\lim_{\min\{x,y\} \to \infty} F(x,y) = 1$
  2. $\lim_{x \to -\infty} F(x,y) = 0, \quad \lim_{y \to -\infty} F(x,y) = 0$
  3. $F(x,y)$ continuous from the right in every variable
  4. if $a < b, c < d$ then $F(b,d) - F(a,d) - F(b,c) + F(a,c) \geq 0$
- in the bivariate continuous case: $P(a < X < b, c < Y < d) = F(b,d) - F(a,d) - F(b,c) + F(a,c)$

Conditioning

- for discrete X and Y it is natural to define

$$P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)} = \frac{f(x,y)}{f_X(x)}$$

- for continuous X and Y the middle expression is undefined
- $\forall x : f_X(x) > 0$ conditional density given by

$$f(y|x) = \frac{f(x,y)}{f_X(x)}$$

- density of the *conditional random variable* $Y|x$
- conditional moments: $E[Y|x] = \int_{-\infty}^{\infty} yf(y|x)dy$ and $\mathrm{Var}[Y|x] = E[(Y|x - \mu_{Y|x})^2]$

Conditioning

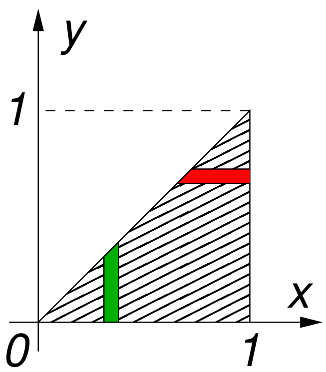
- extend to $n+1$ variables $X_1, X_2, \ldots, X_n, Y$:

$$f(y|x_1, x_2, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n, y)}{f(x_1, x_2, \ldots, x_n)}$$

- also for moments, eg.

$$E[Y|x_1, x_2, \ldots, x_n] = \int_{-\infty}^{\infty} y f(y|x_1, x_2, \ldots, x_n) dy$$

Conditioning: example



- X and Y uniformly distributed over hatched area; $f(x, y) = 2$ in this area and $= 0$ elsewhere
- $f_X(x) = \int_0^x f(x, y) dy = 2x \quad (0 \le x \le 1)$ and

$$f_Y(y) = \int_y^1 f(x, y) dx = 2(1-y) \quad (0 \le y \le 1)$$

Conditioning: example

- hence

$$f(y|x) = \frac{1}{x} \text{ for } 0 \leq y \leq x \text{ and}$$
$$f(x|y) = \frac{1}{1-y} \text{ for } y \leq x \leq 1$$

  so we have uniform distributions for $Y|x$ and $X|y$

- also $E[Y|x] = \int_0^x \frac{y}{x} dy = \frac{x}{2}$ and

$$E[X|y] = \int_y^1 \frac{x}{1-y} dx = \frac{1+y}{2}$$

Independence

- $f(x, y) = f_X(x) f(y|x)$; if it holds that $f(y|x) = f_Y(y)$ then $f(x, y) = f_X(x) f_Y(y)$ = independence
- in general: $X_1, X_2, \ldots, X_n$ are *independent* if $f(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n)$
- if $X_1, X_2, \ldots, X_n$ are independent then so too is any subset of $X_1, X_2, \ldots, X_n$
- marginal densities can be obtained from the joint density. The reverse is not possible generally; if the variables are independent then the joint density can be expressed as the product of the marginal densities

Independence: example

- if $X_1, X_2, \ldots, X_n$ are independent normally distributed random variables with $\mu_1, \mu_2, \ldots, \mu_n$ and $\sigma_1, \sigma_2, \ldots, \sigma_n$, then the multivariate density is expressed as

$$
\begin{aligned}
f(x_1, x_2, \ldots, x_n) &= \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x_k-\mu_k}{\sigma_k}\right)^2} \\
&= \frac{1}{(2\pi)^{n/2}} \left(\prod_{k=1}^{n} \sigma_k^{-1}\right) \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{x_k-\mu_k}{\sigma_k}\right)^2\right)\right] \\
&= \frac{1}{\sqrt{(2\pi)^n \det[\boldsymbol{K}]}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{K}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]
\end{aligned}
$$

Probability theory Multivariate Distributions

Independence: example

- with $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'$ and $\boldsymbol{K}$ a diagonal matrix containing the variances on its diagonal

$$
\boldsymbol{K} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}
$$

Outline

Functions of several random variables

- $Y = g(X_1, X_2, \ldots, X_n)$, and given the joint distribution of the inputs, find the distribution of the output. This is generally very difficult!
- can be worked out analytically for some basic operations by differentiating the distribution functions
- $F_{X+Y}(z) = P(X + Y \le z) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{z} f(x, u - x)\, du$
- differentiation gives $f_{X+Y}(z) = \int_{-\infty}^{\infty} f(x, z - x)\, dx$
- for independent X and Y:

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\, dx$$

= convolution!

Functions of several random variables

- similarly: $F_{XY}(z) = \int_{-\infty}^{\infty} \frac{dx}{|x|} \int_{-\infty}^{z} f\left(x, \frac{u}{x}\right) du,$

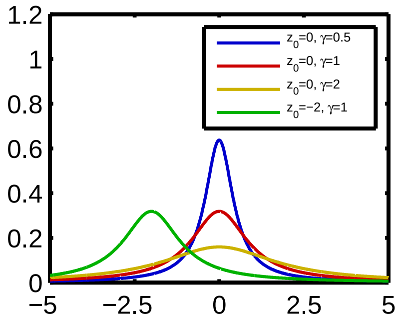
  differentiation leads to $f_{XY}(z) = \int_{-\infty}^{\infty} f\left(x, \frac{z}{x}\right) \frac{dx}{|x|}$
- and also $F_{Y/X}(z) = \int_{-\infty}^{\infty} |x| dx \int_{-\infty}^{z} f(x, ux) du,$

  differentiation leads to $f_{Y/X} = \int_{-\infty}^{\infty} f(x, zx)|x| dx$

Functions of several random variables: example

- if X and Y are independent normally distributed variables and $Z = Y/X$, then it follows that

$$f_Z(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x^2/2} e^{-(zx)^2/2} |x| dx = \frac{1}{\pi(1 + z^2)}$$



- this is the *standard Cauchy density*; has no expectation since the expectation integral does not converge
- in general $f_Z(z) = \frac{1}{\pi}\left[\frac{\gamma}{(z - z_0)^2 + \gamma^2}\right]$

Functions of several random variables

- for the Euclidian norm $Z = \sqrt{X^2 + Y^2}$ it can be found (in polar coordinates):

$$F_Z(z) = \int_0^{2\pi} d\theta \int_0^z f(r\cos\theta, r\sin\theta) r dr$$

for $z \geq 0$ and $= 0$ elsewhere

differentiation yields

$$f_Z(z) = z \int_0^{2\pi} f(z\cos\theta, z\sin\theta) d\theta$$

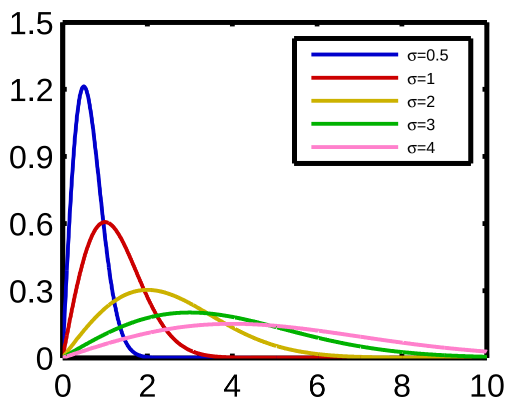
for $z \geq 0$ and $= 0$ elsewhere

Functions of several random variables: example

- Euclidian norm for independent normally distributed variables X and Y with expected value 0 and common variance σ^2:

$$f_Z(z) = \frac{z}{\sigma^2} e^{-\frac{z^2}{2\sigma^2}} \quad \text{for } z \geq 0 \text{ and } = 0 \text{ elsewhere}$$

- this is the *Rayleigh density*



Sum of independent random variables

- important special case of function of several variables $Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$ with constants $a_1, a_2, \ldots, a_n$
- if $X_1, X_2, \ldots, X_n$ are independent, use moment-generating function
- for $Y = X_1 + X_2 + \cdots + X_n$ it is found that

$$M_Y(t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[t \sum_{k=1}^{n} x_k\right] f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{k=1}^{n} e^{tx_k} f(x_k)\, dx_1\, dx_2 \ldots dx_n = \prod_{k=1}^{n} M_{X_k}(t)$$

Sum of independent random variables: example

- for $X_1, X_2, \ldots, X_n$ independent gamma distributed random variables with parameters α_k and β for X_k:

$$M_Y(t) = \prod_{k=1}^{n} (1 - \beta t)^{-\alpha_k} = (1 - \beta t)^{-(\alpha_1 + \alpha_2 + \cdots + \alpha_n)}$$

- hence Y is also gamma distributed with parameters $\alpha_1 + \alpha_2 + \cdots + \alpha_n$ and β
- special case of n independently distributed exponential variables with parameter b: Y is gamma distributed with parameters $\alpha = n$ and $\beta = 1/b$
- hence for $U_1, U_2, \ldots, U_n$ independent uniform random variables over $(0, 1)$, it is found that $X = -b^{-1} \sum_{k=1}^{n} \log U_k$ is gamma distributed with $\alpha = n$ and $\beta = 1/b$, can be used for computer simulation

Sum of independent random variables: example

- for $X_1, X_2, \ldots, X_n$ independent Poisson distributed variables with expected value λ_k for X_k it is found that

$$M_Y(t) = \prod_{k=1}^{n} \exp[\lambda_k(e^t - 1)] = \exp\left[(e^t - 1)\sum_{k=1}^{n} \lambda_k\right]$$

hence Y is Poisson distributed with $\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$

Sum of independent random variables: example

- for $X_1, X_2, \ldots, X_n$ independent normally distributed variables with μ_k and σ_k^2 for X_k it is found that

$$M_Y(t) = \prod_{k=1}^{n} M_{X_k}(a_k t) = \prod_{k=1}^{n} \exp\left[a_k\mu_k t + \frac{a_k^2\sigma_k^2 t^2}{2}\right]$$

$$= \exp\left[\left(\sum_{k=1}^{n} a_k\mu_k\right) t + \left(\sum_{k=1}^{n} a_k^2\sigma_k^2\right)\frac{t^2}{2}\right]$$

hence Y is also normally distributed with

$\mu_Y = \sum_{k=1}^{n} a_k\mu_k$ and $\sigma_Y^2 = \sum_{k=1}^{n} a_k^2\sigma_k^2$

Joint distribution of output random variables

- system with several inputs and several outputs, calculate the joint distribution of the output variables
- case of discrete X and Y at the input, discrete U and V at the output with $U = g(X, Y)$ and $V = h(X, Y)$
- then it follows that

$$\begin{aligned} f_{U,V}(u, v) &= P\left(g(X, Y) = u, h(X, Y) = v\right) \\ &= \sum_{\{(x,y):g(x,y)=u,h(x,y)=v\}} f_{X,Y}(x, y) \end{aligned}$$

Joint distribution of output random variables

- if the mapping $\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} g(x, y) \\ h(x, y) \end{pmatrix}$ is one-to-one with inverse mapping $\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r(u, v) \\ s(u, v) \end{pmatrix}$ then

$$\{(x, y) : g(x, y) = u, h(x, y) = v\} = \{r(u, v), s(u, v)\}$$

and

$$f_{U,V}(u, v) = f_{X,Y}(r(u, v), s(u, v))$$

- can be extended to n variables

Joint distribution of output random variables: example

- X and Y are independent binomially distributed random variables, X with parameters n and p, Y with m and d; then

$$f_{X,Y}(x,y) = \binom{n}{x}\binom{m}{y} p^x d^y (1-p)^{n-x}(1-d)^{m-y}$$

- if $U = X + Y$ and $V = X - Y$ then

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r(u,v) \\ s(u,v) \end{pmatrix} = \begin{pmatrix} (u+v)/2 \\ (u-v)/2 \end{pmatrix}$$

Probability theory | Functions of Several Random Variables

Joint distribution of output random variables: example

- hence

$$f_{U,V}(u,v) = \binom{n}{(u+v)/2}\binom{m}{(u-v)/2} p^{(u+v)/2} d^{(u-v)/2} \times (1-p)^{n-(u+v)/2}(1-d)^{m-(u-v)/2}$$

- constraints on variables u and v can be determined from constraints on x and y

Joint distribution of output random variables

- for continuous random variables: mapping

$$\boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} g_1(x_1, x_2, \ldots, x_n) \\ g_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ g_n(x_1, x_2, \ldots, x_n) \end{pmatrix}$$

and $g_1, g_2, \ldots, g_n$ possess continuous partial derivatives and one-to-one mapping in $A_{\boldsymbol{x}} : f(\boldsymbol{x}) > 0$; also $A_{\boldsymbol{x}} \to A_{\boldsymbol{u}}$

- then

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} r_1(u_1, u_2, \ldots, u_n) \\ r_2(u_1, u_2, \ldots, u_n) \\ \vdots \\ r_n(u_1, u_2, \ldots, u_n) \end{pmatrix}$$

Probability theory Functions of Several Random Variables

Joint distribution of output random variables

- theorem: joint density of $\boldsymbol{U} = (U_1, U_2, \ldots, U_n)'$ is given by

$$f_U(\boldsymbol{u}) = \begin{cases} f_X(r_1(\boldsymbol{u}), r_2(\boldsymbol{u}), \ldots, r_n(\boldsymbol{u}))\,|J(\boldsymbol{x}; \boldsymbol{u})| & \text{for } \boldsymbol{u} \in A_{\boldsymbol{u}} \\ 0 & \text{elsewhere} \end{cases}$$

- $J(\boldsymbol{x}; \boldsymbol{u})$ is the *Jacobian* of the mapping $\boldsymbol{x} \to \boldsymbol{u}$:

$$J(\boldsymbol{x}; \boldsymbol{u}) = \det \begin{pmatrix} \partial x_1/\partial u_1 & \partial x_1/\partial u_2 & \cdots & \partial x_1/\partial u_n \\ \partial x_2/\partial u_1 & \partial x_2/\partial u_2 & \cdots & \partial x_2/\partial u_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial x_n/\partial u_1 & \partial x_n/\partial u_2 & \cdots & \partial x_n/\partial u_n \end{pmatrix}$$

Joint distribution of output random variables: example

- mapping $U = X + Y$ and $V = X - Y$; X and Y possess joint uniform distribution over the unit square $A_x = (0,1)^2$; $A_{\boldsymbol{u}}$ can be found by solving

$$0 < \frac{u+v}{2} < 1, \quad 0 < \frac{u-v}{2} < 1$$

- Jacobian $J(\boldsymbol{x};\boldsymbol{u}) = \det \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix} = -\frac{1}{2}$
- hence $f_{U,V}(u,v) = 1/2$ for $(u,v) \in A_{\boldsymbol{u}}$ and 0 elsewhere

Expectation of function of random variables

- extension of the theorem for one random variable:

$$E[g(X_1, X_2, \ldots, X_n)]$$
$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n$$

with $g(x_1, x_2, \ldots, x_n)$ a piecewise continuous function and $f(x_1, x_2, \ldots, x_n)$ the joint density of $X_1, X_2, \ldots, X_n$
- similar for discrete random variables
- theorem: linearity of expectation

$$E\left[\sum_{k=1}^{n} a_k X_k\right] = \sum_{k=1}^{n} a_k E[X_k]$$

Covariance

- *product moment* of order $(p+q)$ (with $p \geq 0, q \geq 0$):

$$\mu'_{pq} = E[X^p Y^q] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q f(x,y)\,dy\,dx$$

- *central moment* of order $(p+q)$:

$$\mu_{pq} = E[(X-\mu_X)^p (Y-\mu_Y)^q] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_x)^p (y-\mu_y)^q f(x,y)\,dy\,dx$$

- second order central product moment μ_{11} is *covariance*:

$$\mathrm{Cov}[X,Y] = E[(X-\mu_x)(Y-\mu_y)] = \sigma^2_{XY}$$

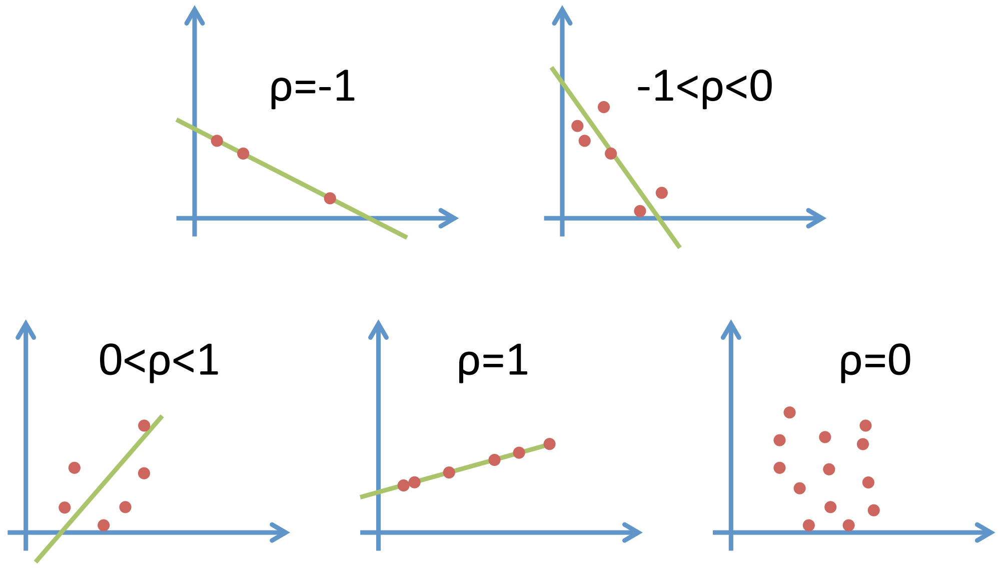
Probability theory | Functions of Several Random Variables

Covariance

- $\mathrm{Cov}[X,Y] = E[XY] - \mu_X \mu_Y$
- if X and Y are independent, then $\mathrm{Cov}[X,Y] = 0$
- covariance = measure of linear relationship between random variables
- *Pearson correlation coefficient* is a normalized measure for linear relationship $\rho_{XY} = \dfrac{\mathrm{Cov}[X,Y]}{\sigma_X \sigma_Y}$
- if $\rho_{XY} = 0$ then X and Y are *uncorrelated*
- independent variables are uncorrelated. The opposite is not valid!
- theorem: $-1 \leq \rho_{XY} \leq 1$; $|\rho_{XY}| = 1$ if and only if there is a linear relationship between X and Y: $P(Y = aX + b) = 1$ with constants a and b ($a \neq 0$)

Correlation coefficient

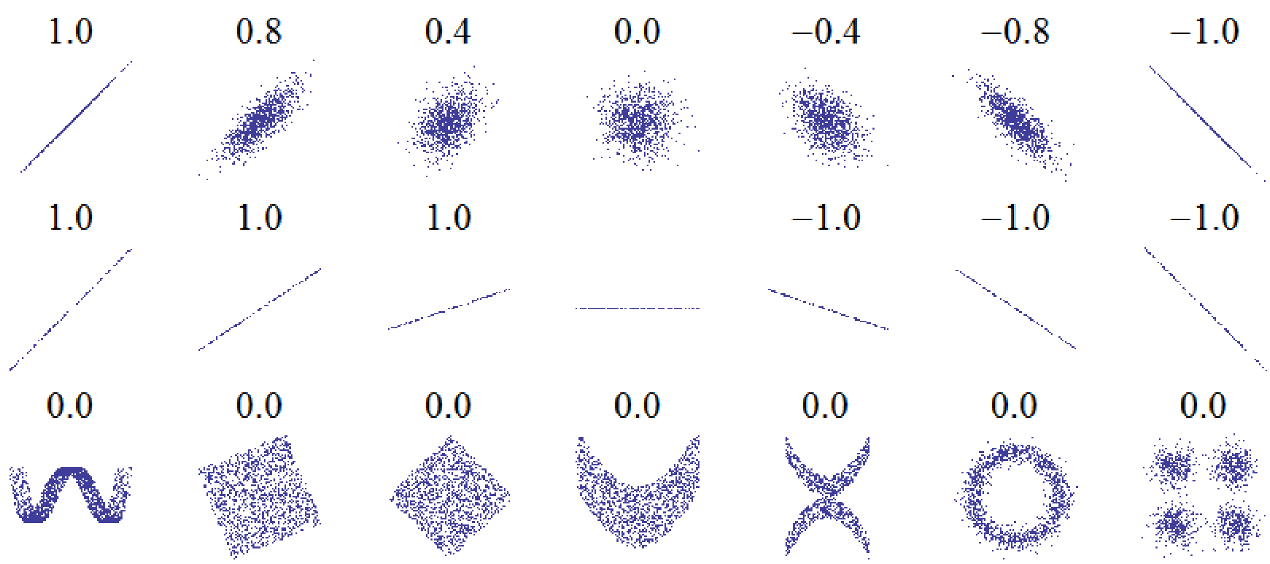
Values for ρ:



Probability theory Functions of Several Random Variables

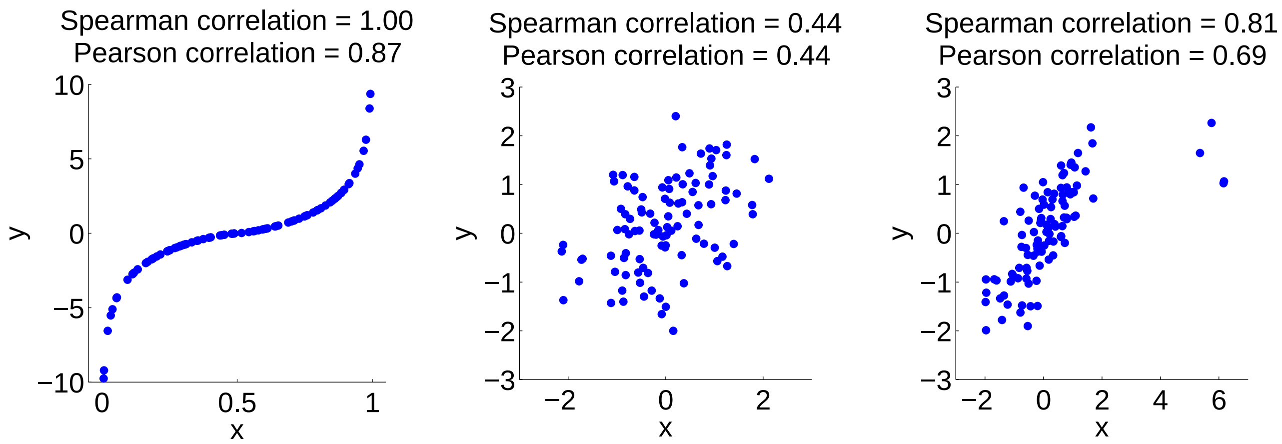
Correlation coefficient

Values for ρ:



Correlation coefficient

Variants exist, e.g. Spearman rank correlation coefficient, captures also relationships between variables based on any monotonic function, and is less sensitive to outliers.



Covariance

- if $Y = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$ then

$$\mathrm{Var}[Y] = \sum_{j=1}^{n} \sum_{k=1}^{n} a_j a_k \, \mathrm{Cov}[X_j, X_k]$$

- if $X_1, X_2, \ldots, X_n$ are uncorrelated then

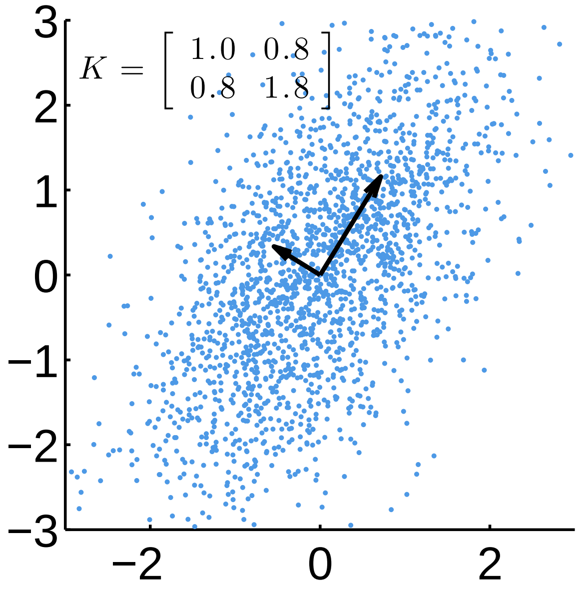
$$\mathrm{Var}[Y] = \sum_{k=1}^{n} a_k^2 \, \mathrm{Var}[X_k]$$

- mean vector and covariance matrix of $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)'$

$$\boldsymbol{\mu} = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{pmatrix} \qquad \boldsymbol{K} = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'] = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \ldots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \ldots & \sigma_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \ldots & \sigma_{nn}^2 \end{pmatrix}$$

- $\boldsymbol{K}$ is real, symmetric and nonnegative definite

Covariance matrix example



X and Y components co-vary, hence σ_X^2 and σ_Y^2 alone do not describe the distribution and σ_{XY}^2 is required → full covariance matrix; directions of arrows correspond to eigenvectors of $\boldsymbol{K}$ and their lengths to square roots of eigenvalues

Multivariate normal distribution

- $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)'$ has a multivariate normal (Gaussian) distribution if

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \det[\boldsymbol{K}]}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{K}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

- properties:
  - mean vector $\boldsymbol{\mu}$
  - covariance matrix $\boldsymbol{K}$ (real, symmetric, nonnegative definite)
  - $X_1, X_2, \ldots, X_n$ independent $\Leftrightarrow$ $\boldsymbol{K}$ diagonal
  - marginal densities are also normally distributed
  - independent if and only if uncorrelated
  - if $\boldsymbol{U} = \boldsymbol{A}\boldsymbol{X}$ with $\boldsymbol{A}$ a non singular $n \times n$ matrix, then $\boldsymbol{U}$ has multivariate normal distribution with mean $\boldsymbol{A}\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{A}\boldsymbol{K}\boldsymbol{A}'$

Outline

Laws of large numbers

- limit properties of sums of random variables
- what is convergence?
- central limit theorem

Weak law of large numbers

- observe X n times, and compute average
- is this average a good approximation of μ_X?
- what is the relationship between the average of random variables and the average of their expectations?
- if $Y_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ then

$$E[Y_n] = \frac{1}{n}\sum_{k=1}^{n} \mu_k \quad \text{and} \quad \mathrm{Var}[Y_n] = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} \mathrm{Cov}[X_j, X_k]$$

- apply Chebyshev inequality to Y_n with $\varepsilon > 0$:

$$P(|Y_n - E[Y_n]| \geq \varepsilon) \leq \frac{1}{\varepsilon^2 n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} \mathrm{Cov}[X_j, X_k]$$

Probability theory | Laws of Large Numbers

Weak law of large numbers

- weak law of large numbers: if

$$\lim_{n\to\infty} \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} \mathrm{Cov}[X_j, X_k] = 0$$

  then, for any $\varepsilon > 0$:

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n}\sum_{k=1}^{n} \mu_k\right| \geq \varepsilon\right) = 0$$

- uses some kind of convergence: *convergence in probability*: sequence $Z_1, Z_2, \ldots$ converges in probability to Z if for any $\varepsilon > 0$:

$$\lim_{n\to\infty} P\left(|Z_n - Z| \geq \varepsilon\right) = 0$$

Weak law of large numbers

Convergence in probability:

- intuitively: probability of “unusual” outcome becomes smaller and smaller (but not zero) as the sequence progresses.
- example 1: estimate by eye the height X of a randomly chosen person, by many observers. Sequence X_n of averages converges in probability to X.
- example 2: archer obtains score X_n in the n-th shot. After years of practice the probability of not hitting the bullseye becomes very small (but not zero). The sequence $\{X_n\}$ will always contain non-perfect scores even if they are becoming increasingly less frequent.

Probability theory Laws of Large Numbers

Weak law of large numbers

- law is mostly applied in specific cases
- if $X_1, X_2, \ldots$ with respective $\sigma_1^2, \sigma_2^2, \ldots$ are uncorrelated and $\exists M : \forall k : \sigma_k^2 \leq M$ then

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} \mathrm{Cov}[X_j, X_k] = \frac{1}{n^2}\sum_{k=1}^{n}\sigma_k^2 \leq \frac{M}{n}$$

  and hence the law holds
- if $X_1, X_2, \ldots$ are independent and identically distributed with equal μ and finite σ^2 then they are uncorrelated and hence

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu\right| \geq \varepsilon\right) = 0$$

  → average converges in probability to the expected value

Weak law of large numbers: example

- example: ∞ number of independent trials; event A occurs with probability p_n at trial n. Set $X_n = 1$ when A occurs and $X_n = 0$ when A does not occur $\Rightarrow$ binomial random variable
- then $E[X_n] = p_n$ and $\text{Var}[X_n] = p_n(1 - p_n)$
- then $Y_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ is the relative frequency of occurrence of A after n trials
- weak law of large numbers holds, hence the relative frequency converges in probability to the average of all p_k.
- when $p_n = p$ = constant, then $\forall \varepsilon > 0$

$$\lim_{n\to\infty} P(|Y_n - p| \geq \varepsilon) = 0$$

Mean square convergence

- the row $Z_1, Z_2, \ldots$ *converges in mean square sense* to Z if

$$\lim_{n\to\infty} E[|Z_n - Z|^2] = 0$$

- it follows from the Chebyshev inequality that mean square convergence implies convergence in probability (but not vice versa!)

Almost-sure convergence

- random variable is regarded as a function and convergence of functions is studied
- regarded as a function, Z_n converges to Z if $\forall w \in S$ it holds that $\lim_{n\to\infty} Z_n(w) = Z(w)$
- Z_n converges to Z *almost surely* if $\exists$ event G such that $P(G) = 0$ and the limit holds $\forall w \in S - G$
- Or also: $\forall w \in S - G, \forall \varepsilon > 0, \exists$ positive integer $N_{w,\varepsilon}$: for $n \geq N_{w,\varepsilon}$: $\quad |Z_n(w) - Z(w)| < \varepsilon$
- written as

$$P\left(\lim_{n\to\infty} Z_n = Z\right) = 1$$

“Almost-sure” vs. “sure”

Subtle difference between something happening with probability 1 and happening always. A sure event always happens. If an event is almost sure, then outcomes not in this event are theoretically possible; however, the probability of such an outcome occurring is smaller than any fixed positive probability, and therefore must be 0. One cannot definitively say that these outcomes will never occur, but can for most purposes assume this to be true.

"Almost-sure" vs. "sure"

- example 1: throw a dart at a unit square wherein the dart will impact exactly one point, and imagine that this square is the only thing in the universe. Event that "the dart hits the diagonal of the unit square exactly" has probability 0. The dart will *almost never* land on the diagonal. Nonetheless the set of points on the diagonal is not empty and a point on the diagonal is no less possible than any other point, therefore theoretically it is possible that the dart actually hits the diagonal.
- example 2: flip a coin; event that every flip results in heads, yielding the sequence $\{H, H, H, \ldots\}$, ad infinitum, is possible in some sense but its probability is zero in an infinite series. There will *almost surely* be at least one tails in an infinite sequence of flips.

Almost-sure convergence

- $P(\lim_{n\to\infty} Z_n = Z) = 1$ means that events for which Z_n does not converge to Z have probability 0.
- example 1: note the exact amount of food that an animal consumes day by day. This sequence can be unpredictable but we are *quite certain* that one day the number will become zero and will stay zero forever after.
- example 2: toss seven coins every morning and give that day a random amount to a certain charity. Stop doing this permanently the first time the result is all tails. We are *almost sure* that one day the amount will be zero and stay zero forever after that.

Almost-uniform convergence

- Z_n converges *almost uniformly* to Z if $\forall \delta > 0, \exists$ event F_δ such that $P(F_\delta) < \delta$ and Z_n converges uniformly to Z over $S - F_\delta$
- uniform convergence: $\forall \varepsilon > 0, \exists$ positive integer $N_{\delta,\varepsilon}$: for $n \geq N_{\delta,\varepsilon}$: $\quad |Z_n(w) - Z(w)| < \varepsilon \quad \forall w \in S - F_\delta$
- fundamental property of random variables: almost sure and almost uniform convergence are equivalent
- almost sure (and almost uniform) convergence implies convergence in probability (not vice versa). Almost sure convergence is a stronger form of convergence.

Strong law of large numbers

- based on almost-sure convergence
- if $X_1, X_2, \ldots$ are independent identically distributed random variables with finite μ then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k = \mu \qquad \text{almost sure}$$

- or, equivalently:

$$P\left(\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k = \mu\right) = 1$$

- conclusion is stronger than in weak law, but so are the assumptions

Convergence in distribution

- convergence in distribution if $\lim_{n\to\infty} F_{X_n}(a) = F_X(a)$ in all points a where F_X is continuous, or also (if continuity also holds in b)

$$\lim_{n\to\infty} P(a < X_n \leq b) = P(a < X \leq b)$$

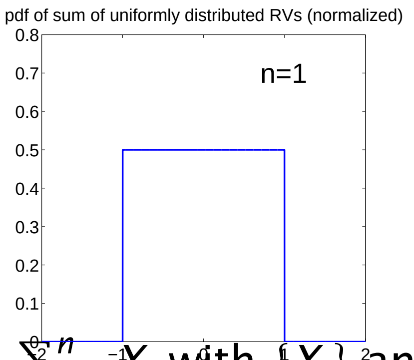
- for continuous X also

$$\lim_{n\to\infty} \int_a^b f_{X_n}(x)\,dx = \int_a^b f_X(x)\,dx$$

- for large n:

$$P(a < X_n \leq b) \approx P(a < X \leq b)$$

Convergence in distribution: example



$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$ with $\{X_i\}$ an identically distributed sequence of uniformly (between -1 and 1) distributed variables. The distribution of Z_n approaches the normal distribution $N(0, \frac{1}{3})$.

Central limit theorem

- for $X_1, X_2, \ldots$ independent identically distributed random variables with μ and σ^2 it holds that, $\forall z$:

$$\lim_{n\to\infty} P\left(\frac{\frac{1}{n}\sum_{k=1}^{n} X_k - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-y^2/2} dy$$

- normalized mean

$$Z_n = \frac{\frac{1}{n}\sum_{k=1}^{n} X_k - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Central limit theorem

- if Z is standardized normally distributed random variable, then

$$\lim_{n\to\infty} Z_n = Z \quad \text{in distribution}$$

or also

$$\lim_{n\to\infty} P(a < Z_n \leq b) = \Phi(b) - \Phi(a)$$

Central limit theorem: example

- individual trials for a binomial distribution are independent and identically distributed. Binomial random variable for n trials is $X^n = \sum_{k=1}^{n} X_k$
- X^n possesses expectation np and variance $np(1-p)$, average value X^n/n possesses expectation p and variance $p(1-p)/n$
- central limit theorem: $\lim_{n\to\infty} \frac{X^n/n - p}{\sqrt{p(1-p)/n}} = Z$
- or also $\lim_{n\to\infty} P\left(a < \frac{X^n - np}{\sqrt{np(1-p)}} \leq b\right) = \Phi(b) - \Phi(a)$

  This is the *De Moivre/Laplace theorem*

Variant: Liapounov theorem

- variant where $X_1, X_2, \ldots$ are not identically distributed: different $\mu_1, \mu_2, \ldots$ and $\sigma_1^2, \sigma_2^2, \ldots$
- notation:

$$S_n = \sum_{k=1}^{n} X_k \qquad m_n = \sum_{k=1}^{n} \mu_k = \mu_{S_n} \qquad s_n^2 = \sum_{k=1}^{n} \sigma_k^2 = \mathrm{Var}[S_n]$$

- theorem: if $\exists \delta > 0 : E[|X_n - \mu_n|^{2+\delta}] < \infty$ for $n = 1, 2, \ldots,$ and $\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{k=1}^{n} E[|X_k - \mu_k|^{2+\delta}] = 0$ then

$$\lim_{n\to\infty} \frac{S_n - m_n}{s_n} = Z \quad \text{in distribution}$$

Variant: Liapounov theorem

- limiting condition is difficult to apprehend, but there is an easily applied corollary: if $\exists C : |X_n| \leq C \quad \forall n$ and $\lim_{n\to\infty} s_n^2 = \sum_{k=1}^{\infty} \sigma_k^2 = \infty$ then the theorem holds
- in practice: real world signals are bounded and the sum of variances diverges. Hence, for real world signals the Liapounov theorem generally holds.
- other sufficient and necessary conditions exist

Outline

Parameter estimation

- given the observations of a random variable, find the parameters of the underlying distribution
- random sample of X: set of random variables $X_1, X_2, \ldots, X_n$ that are independent and have the same distribution as X
- if X has density $f(x)$ then $f(x_1, x_2, \ldots, x_n) = \prod_{k=1}^{n} f(x_k)$
- write density of X as $f(x; \theta)$ with θ an unknown parameter

Parameter estimation

- find a function $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ that provides an estimate of θ
- $\hat{\theta}(X_1, X_2, \ldots, X_n)$ itself is a random variable and is an *estimator* of θ
- $\hat{\theta}(x_1, x_2, \ldots, x_n)$ is an *estimate* of θ for a given observation
- $\hat{\theta}(X_1, X_2, \ldots, X_n)$ is a *statistic* if the estimation rule is free of unknown parameters (but the distribution of $\hat{\theta}$ may still have unknown parameters)

Parameter estimation

- different samples give rise to different estimated values for θ
- which estimate is good? Which one is better than another one? Which one is the best one?
- measure needed for the quality of an estimate/estimator
- an estimator $\hat{\theta}$ is an *unbiased* estimator if $E[\hat{\theta}] = \theta$ (independent of the value of θ)
- often the bias decreases with increasing sample size: $E[\hat{\theta}] \to \theta$ when $n \to \infty$
  $\Rightarrow$ *asymptotically unbiased estimator*

Probability theory Parametric Estimation via Random Samples

Parameter estimation

- precision: can not be guaranteed, but $P(|\hat{\theta} - \theta| < r)$ for $r > 0$ can be determined
- $\hat{\theta}$ is *consistent* estimator if, for all $r > 0$:

$$\lim_{n \to \infty} P(|\hat{\theta} - \theta| < r) = 1$$

  (or also: $\hat{\theta}$ converges in probability to θ)
- Chebyshev for unbiased estimator:

$$P(|\hat{\theta} - \theta| < r) \geq 1 - \frac{\mathrm{Var}[\hat{\theta}]}{r^2}$$

- hence for an unbiased estimator, if $\mathrm{Var}[\hat{\theta}] \to 0$ when $n \to \infty$, $\hat{\theta}$ is a consistent estimator (holds also for asymptotically unbiased estimator)

Sample mean

- sample mean = average value of samples = random variable!
- sample mean $\bar{X}$ is often used as an estimator for expected value μ:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- *empirical mean* $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$
- sample mean is unbiased: $E[\bar{X}] = \mu$; variance of sample mean $\mathrm{Var}[\bar{X}] = \sigma^2/n$
- hence sample mean is consistent estimator of expected value

Sample mean

- precision is given by weak law of large numbers:

$$P(|\bar{X} - \mu| < r) \geq 1 - \frac{\sigma^2}{nr^2}$$

- strong law of large numbers also holds: sample mean converges almost surely to expected value
- central limit theorem: standardized version of sample mean converges in distribution to standardized normally distributed random variable, so for large n:

$$P\left(a < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq b\right) \approx \Phi(b) - \Phi(a)$$

Sample variance

- most often used estimator for the variance: *sample variance*
$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$$
- unbiased estimator: $E[S^2] = \mathrm{Var}[X]$
- theorem from statistics: for $X_1, X_2, \ldots, X_n$ derived from normally distributed X with μ and σ^2, $\bar{X}$ and S^2 are independent and $(n-1)S^2/\sigma^2$ obeys a gamma distribution with $\alpha = (n-1)/2$ and $\beta = 2$. Variance of gamma distribution is $\alpha\beta^2$, hence:

$$\mathrm{Var}[S^2] = \mathrm{Var}\left[\frac{\sigma^2}{n-1}\left(\frac{(n-1)S^2}{\sigma^2}\right)\right] = \frac{2\sigma^4}{n-1}$$

Sample variance

- hence $\mathrm{Var}[S^2] \to 0$ when $n \to \infty$
- hence S^2 is a consistent estimator for a normally distributed random variable
- practical problem: when $\mathrm{Var}[X]$ is large, $2\sigma^4$ is even larger and n has to be very large to obtain a low $\mathrm{Var}[S^2]$
- compare to precision of sample mean:

$$P(|S^2 - \sigma^2| < r) \geq 1 - \frac{2\sigma^4}{(n-1)r^2}$$

Minimum variance unbiased estimators

- *MVUE (minimum variance unbiased estimator)*
- comparison of unbiased estimators: one with a smaller variance gives greater lower bound for $P(|\hat{\theta} - \theta| < r)$ (Chebyshev)
- other criterion for comparison: mean square error $E[|\hat{\theta} - \theta|^2]$
- convergence of estimator $\hat{\theta}$ to θ in mean square sense:

$$\lim_{n \to \infty} E[|\hat{\theta} - \theta|^2] = 0 \quad ?$$

Minimum variance unbiased estimators

- $E[|\hat{\theta} - \theta|^2] = \mathrm{Var}[\hat{\theta}] + (E[\hat{\theta}] - \theta)^2$
- hence for (asymptotically) unbiased estimator

$$\lim_{n \to \infty} E[|\hat{\theta} - \theta|^2] = \lim_{n \to \infty} \mathrm{Var}[\hat{\theta}]$$

- $\hat{\theta} \to \theta$ in mean square sense when $n \to \infty$ if and only if $\mathrm{Var}[\hat{\theta}] \to 0$ when $n \to \infty$
- comparison of unbiased estimators: the one with lowest variance yields lowest mean square error
- $\hat{\theta}$ is MVUE if for any other unbiased estimator $\hat{\theta}_0$ it holds that $\mathrm{Var}[\hat{\theta}] \leq \mathrm{Var}[\hat{\theta}_0]$
- MVUE can also be determined for some class $\mathcal{C}$ of estimators

Minimum variance unbiased estimators: example

- C is class of linear unbiased estimators of μ:
$$\hat{\mu} = \sum_{k=1}^{n} a_k X_k \quad \text{with} \quad E[\hat{\mu}] = \mu$$
- $E[\hat{\mu}] = \left(\sum_{k=1}^{n} a_k\right)\mu$, hence $a_1 + a_2 + \cdots + a_n = 1$
- $\sigma_{\hat{\mu}}^2 = \sum_{k=1}^{n} a_k^2 \sigma^2 = \left(\sum_{k=1}^{n-1} a_k^2 + \left(1 - \sum_{k=1}^{n-1} a_k\right)^2\right)\sigma^2$
- setting derivative to zero yields system with unique solution $a_j = 1/n$ for $j = 1, 2, \ldots, n$
- hence sample mean is best unbiased (MVUE) linear estimator

Cramer-Rao bound

- finding MVUE is generally difficult; checking whether any given unbiased estimator is MVUE, is easier
- if variance of unbiased estimator equals Cramer-Rao bound then it is MVUE
- under certain conditions, with $f(x;\theta)$ the density of X:

$$\mathrm{Var}[\hat{\theta}] \geq \frac{1}{nE\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right]}$$

Cramer-Rao bound: example

- sample mean is MVUE for μ belonging to normally distributed random variable
- $f(x;\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ hence

$$\frac{\partial}{\partial \mu} \log f(x;\mu) = \frac{x-\mu}{\sigma^2} \text{ and}$$

$$E\left[\left(\frac{\partial}{\partial \theta} \log f(X;\theta)\right)^2\right] = \frac{E[(X-\mu)^2]}{\sigma^4} = \frac{1}{\sigma^2}$$

- Cramer-Rao bound σ^2/n = variance(sample mean)
- difference with previous example: sample mean is best possible unbiased estimator for normal distribution, not restricted to any class of estimators $\mathcal{C}$

In search for a good estimator

- best estimator often cannot be found $\Rightarrow$ find “a” good estimator
- different methods yield different estimators with differing properties, depending on the distribution
- two often used methods: *maximum-likelihood* and the *method of moments*
- maximum-likelihood often gives better results but the method of moments can usually be applied when maximum-likelihood is mathematically untractable

Method of moments

- *r-th sample moment* of $X_1, X_2, \ldots, X_n$ is

$$M'_r = \frac{1}{n} \sum_{k=1}^{n} X_k^r$$

- $r = 1 \Rightarrow$ sample mean; $r = 2 \Rightarrow M'_2 = \frac{n-1}{n} S^2 + \bar{X}^2$
- $X_1^r, X_2^r, \ldots, X_n^r$ constitute a random sample of the random variable X^r, hence M'_r is an unbiased and consistent estimator of $E[X^r]$
- for X with density $f(x; \theta_1, \theta_2, \ldots, \theta_p)$ with $\theta_1, \theta_2, \ldots, \theta_p$ parameters to be estimated, $E[X^r] = h_r(\theta_1, \theta_2, \ldots, \theta_p)$

Method of moments

- method of moments: set $E[X^r] = M'_r$, this gives a system

$$\begin{aligned} h_1(\theta_1, \theta_2, \ldots, \theta_p) &= M'_1 \\ h_2(\theta_1, \theta_2, \ldots, \theta_p) &= M'_2 \\ &\vdots \\ h_N(\theta_1, \theta_2, \ldots, \theta_p) &= M'_N \end{aligned}$$

- choose N such that a unique solution exists, yielding $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p$
- then investigate properties of the estimator

Method of moments: example

- gamma distribution with parameters α and β has $E[X] = \alpha\beta$ and $E[X^2] = (\alpha+1)\alpha\beta^2$
- method of moments:

$$\alpha\beta = M_1' = \bar{X}$$
$$(\alpha+1)\alpha\beta^2 = M_2' = \frac{n-1}{n}S^2 + \bar{X}^2$$

- solving this yields

$$\hat{\alpha} = \frac{\bar{X}^2}{\frac{n-1}{n}S^2} \quad \text{and} \quad \hat{\beta} = \frac{\frac{n-1}{n}S^2}{\bar{X}}$$

Order statistics

- order from least to greatest: reorder $X_1, X_2, \ldots, X_n$ to the n *order statistics* $Y_1, Y_2, \ldots, Y_n$ with $Y_1 \leq Y_2 \leq \cdots \leq Y_n$
- every order statistic is a function of the sample:

$$Y_1 = \min\{X_1, X_2, \ldots, X_n\}$$
$$Y_n = \max\{X_1, X_2, \ldots, X_n\}$$
$$\textit{sample median} \quad \tilde{X} = Y_{(n+1)/2} \quad \text{for odd } n$$

- median often used as an estimate for the mean in case of symmetric distributions

Order statistics

- theorem: densities of the n order statistics are given by

$$f_{Y_k}(y) = \frac{n!}{(n-k)!(k-1)!} F_X(y)^{k-1}[1 - F_X(y)]^{n-k} f_X(y)$$

- in particular

$$f_{Y_1}(y) = n[1 - F_X(y)]^{n-1} f_X(y)$$

and

$$f_{Y_n}(y) = nF_X(y)^{n-1} f_X(y)$$

and for odd n

$$f_{\tilde{X}}(y) = \frac{n!}{[((n-1)/2)!]^2} F_X(y)^{(n-1)/2}[1 - F_X(y)]^{(n-1)/2} f_X(y)$$

Outline

Maximum-likelihood estimation

- good estimators in many cases
- joint density of identically distributed independent variables $X_1, X_2, \ldots, X_n$ is the *likelihood function*

$$L(x_1, x_2, \ldots, x_n; \theta) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta) = L(\theta)$$

- *maximum-likelihood* estimation: value of θ that maximizes $L(\theta)$ (for given $x_1, x_2, \ldots, x_n$) $\Rightarrow$ estimator $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$

Maximum-likelihood estimation

- intuitive understanding: for discrete X

$$L(x_1, x_2, \ldots, x_n; \theta) = \prod_{k=1}^{n} P(X = x_k; \theta)$$

- if there exists a θ' for which

$$L(x_1, x_2, \ldots, x_n; \theta') \geq L(x_1, x_2, \ldots, x_n; \theta), \forall \theta$$

then θ' maximizes

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

Maximum-likelihood estimation

- when several parameters need to be estimated: vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)'$ and we obtain a vector estimator:

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(X_1, X_2, \ldots, X_n) = \begin{pmatrix} \hat{\theta}_1(X_1, X_2, \ldots, X_n) \\ \hat{\theta}_2(X_1, X_2, \ldots, X_n) \\ \vdots \\ \hat{\theta}_m(X_1, X_2, \ldots, X_n) \end{pmatrix}$$

- invariance property: if $\hat{\theta}$ is a maximum likelihood estimator of θ and g is a one-to-one function with $\phi = g(\theta)$ then $\hat{\phi} = g(\hat{\theta})$ is a maximum likelihood estimator of ϕ
- favourable properties: often MVUE

Maximum-likelihood estimation: example

- for normally distributed X with unknown μ and known σ^2

$$L(x_1, x_2, \ldots, x_n; \mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2} \sum_{k=1}^{n} \left(\frac{x_k - \mu}{\sigma} \right)^2 \right]$$

- maximized when logarithm is maximized:

$$\log L(\mu) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2} \sum_{k=1}^{n} \left(\frac{x_k - \mu}{\sigma} \right)^2$$

and

$$\frac{d}{d\mu} \log L(\mu) = \frac{1}{\sigma^2} \sum_{k=1}^{n} (x_k - \mu)$$

Maximum-likelihood estimation: example

- setting derivative to zero shows the maximum to be the mean of $x_1, x_2, \ldots, x_n$
- holds for any $x_1, x_2, \ldots, x_n$, hence maximum likelihood estimator $\hat{\mu} = \bar{X}$ = sample mean

Maximum-likelihood estimation: example

- same normally distributed variable, but now also σ^2 unknown $\Rightarrow$ parameter vector $(\mu, \sigma^2)'$
- set partial derivatives w.r.t. every parameter to zero

$$\frac{\partial}{\partial \mu} \log L(\mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{k=1}^{n} (x_k - \mu)$$

$$\frac{\partial}{\partial \sigma^2} \log L(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{k=1}^{n} (x_k - \mu)^2$$

- solutions $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{n} (X_k - \bar{X})^2 = \frac{n-1}{n} S^2$
- unbiased estimator for expected value and asymptotically unbiased estimator for variance

Additive noise

- simple model: constant discrete signal corrupted by independent identically distributed additive noise values with expected value 0

$$X(i) = \theta + N(i)$$

- θ needs to be estimated
- maximum-likelihood estimator = filter with inputs $X_i = X(i)$ and output the estimated value of θ
- in practice: filter with sliding window over the observations, eg sample mean becomes moving average

Additive noise: example

- *Laplace distribution*
  $f(x) = \frac{\alpha}{2} e^{-\alpha|x-\mu|} \quad \forall x \in \mathbb{R}, \alpha > 0$ and $-\infty < \mu < \infty$ with expected value μ and variance $2/\alpha^2$
- $N(i)$ has Laplace density with mean 0 and variance $2/\alpha^2$
- observations $X_1, X_2, \ldots, X_n$ are samples of Laplace distribution with mean θ and variance $2/\alpha^2$
- likelihood function

$$L(x_1, x_2, \ldots, x_n; \theta) = \left(\frac{\alpha}{2}\right)^n \exp\left[-\alpha \sum_{i=1}^{n} |x_i - \theta|\right]$$

- maximize = minimize sum in exponent
  $\Rightarrow \hat{\theta}$ = moving median (with odd n)

Additive noise: example

	Gaussian	Laplacian
Var[sample mean]	$\frac{\sigma^2}{n}$	$\frac{\sigma^2}{n}$
Asympt. Var[sample median]	$\frac{\pi\sigma^2}{2n}$	$\frac{\sigma^2}{2n}$

Additive noise: example

- noise uniformly distributed over $[-\beta, 0]$ with $\beta > 0$
- likelihood function

$$L(x_1, x_2, \dots, x_n; \theta) = \frac{1}{\beta^n} \prod_{k=1}^{n} I_{[\theta-\beta,\theta]}(x_k)$$

  with $I_{[\theta-\beta,\theta]}$ indicator function on interval $[\theta - \beta, \theta]$
- likelihood function becomes:

$$L(x_1, x_2, \dots, x_n; \theta) = \begin{cases} \beta^{-n} \text{ for } \theta - \beta \leq x_k \leq \theta, k = 1, \dots, n \\ 0 \quad \text{elsewhere} \end{cases}$$

Additive noise: example

- $L(\theta)$ maximum when $\theta - \beta \leq x_k \leq \theta$ for all x_k, otherwise $L(\theta) = 0$, hence

$$\max\{x_1, x_2, \ldots, x_n\} \leq \theta \leq \min\{x_1, x_2, \ldots, x_n\} + \beta$$

- hence choose $\hat{\theta}$ such that

$$\max\{X_1, X_2, \ldots, X_n\} \leq \hat{\theta} \leq \min\{X_1, X_2, \ldots, X_n\} + \beta$$

- when β unknown, choose

$$\hat{\theta} = \max\{X_1, X_2, \ldots, X_n\} \quad = \text{moving maximum}$$

- case of uniformly distributed noise over $[0, \beta]$ yields

$$\hat{\theta} = \min\{X_1, X_2, \ldots, X_n\}$$

Additive noise: weighted median

- attribute more weight to some observations: for $x_1, x_2, \ldots, x_n$, integer weights $\gamma_1, \gamma_2, \ldots, \gamma_n$ indicate that x_i needs to be replicated γ_i times when calculating the median
- assume Laplacian distributed noise for X_i with $\mu = 0$ and $\sigma_i^2 = 2/\gamma_i^2$
- likelihood function

$$L(x_1, x_2, \ldots, x_n; \theta) = \frac{\gamma_1 \gamma_2 \cdots \gamma_n}{2^n} \exp\left[-\sum_{k=1}^{n} \gamma_k |x_k - \theta|\right]$$

- maximum when sum in exponent minimum $\Rightarrow$ weighted median
- filter in sliding window is adaptive because the noise is not assumed to be identically distributed

Minimum noise

- signal corrupted by minimum noise: $X(i) = \theta \wedge N(i)$ with $N(i)$ independent identically distributed variables: $X(i)$ less or equal then θ
- distribution $F_{X(i)}(x) = P(X(i) \leq x) = \begin{cases} 1 & \text{for } x \geq \theta \\ F_N(x) & \text{for } x < \theta \end{cases}$
- taking derivative yields $f_{X(i)}(x)$ and likelihood function

$$\begin{aligned} L(x_1, x_2, \ldots, x_n; \theta) \\ = \prod_{k=1}^{n} [f_N(x_k) I_{(-\infty,\theta)}(x_k) + (1 - F_N(\theta))\delta(x_k - \theta)] \end{aligned}$$

- maximize: finally $\hat{\theta} = \max\{X_1, X_2, \ldots, X_n\}$
- dual argument applies to maximum noise

Outline

Entropy: example

- does a piece of information about some person reveal the person's identity?
- identity of an unknown person has an entropy of approximately 33 bits ($2^{33} \approx 8$ billion = world population)
- entropy is a measure of uncertainty, expressed in bits
- knowledge of a piece of information lowers the uncertainty, i.e. the entropy $\Delta H = -\log_2 P(X = x)$
  eg. $\Delta H = -\log_2 P(\text{starsign=fish}) = -\log_2(1/12) = 3.58$ bits of information
  eg. $\Delta H = -\log_2 P(\text{birthday=January 2nd}) = -\log_2(1/365) = 8.51$ bits of information
- conditional entropy! Knowledge of starsign does not reveal extra information if birthday is known, gives partial information if birth month is known

Entropy: example

- also possible for non uniform likelihoods:
  knowing the person's ZIP code is B3000 (Leuven):
  $\Delta H = -\log_2 P(95,984/6,625,000,000) = 16.07$ bits
  if the ZIP code is B3717 (tiny town Herstappe):
  $\Delta H = -\log_2 P(80/6,625,000,000) = 26.30$ bits
  if you live in Moscow:
  $\Delta H = -\log_2 P(10,524,400/6,625,000,000) = 9.30$ bits
  (2007 population figures)
- if you know that the ZIP code is B8957 (town of Mesen, population approx. 950) and you know the birthday then $\Delta H = 31.24$ bits.
  Almost there!!

Entropy

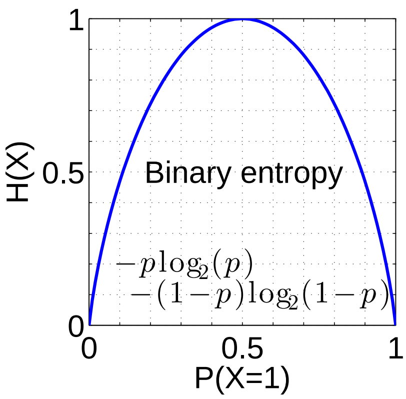
- *entropy* quantifies uncertainty
- criteria for measure of uncertainty (assume discrete X with n possible outcomes with probabilities $p_1, p_2, \ldots, p_n$):
  - non-negative; equal to zero when $\exists i : p_i = 1$
  - maximum when all p_i equal
  - when X and Y have n, resp m equally probable outcomes and $n < m$ then the uncertainty about X is smaller then the uncertainty about Y
  - uncertainty is continuous function of $p_1, p_2, \ldots, p_n$

Entropy

- definition:

$$H[X] = -\sum_{i=1}^{n} p_i \log_2 p_i$$

  assuming that $p_i \log_2 p_i = 0$ when $p_i = 0$



- entropy $H[X]$ measured in bits
- can also be seen as $H[X] = -E[\log_2 f(X)]$
- location on x-axis is not important: only probabilities matter
- $H[X]$ defined here satisfies four mentioned criteria

Conditional entropy

- observation of X can influence uncertainty about Y
- *conditional entropy* of Y given x_i defined as

$$H[Y|x_i] = -\sum_{j=1}^{m} f(y_j|x_i) \log_2 f(y_j|x_i)$$

- in case of varying $x_i \Rightarrow H[Y|X]$ is also random variable
- expected conditional entropy of Y relative to X

$$\bar{H}[Y|X] = E[H[Y|X]] = -E[\log_2 f(Y|X)]$$

- when X and Y are independent then $\bar{H}[Y|X] = H[Y]$

Entropy for vectors

- generalization

$$H[X, Y] = -E[\log_2 f(X, Y)]$$

- can be written as

$$H[X, Y] = H[X] + \bar{H}[Y|X]$$

- for independent X and Y we find $H[X, Y] = H[X] + H[Y]$

Information

- when X and Y are dependent, then observation of X yields information about Y
- *expected amount of information* for Y obtained through observation of X defined as $I_X[Y] = H[Y] - \bar{H}[Y|X]$
- properties:
    - when X and Y are independent then $I_X[Y] = 0$
    - since $H[X|x_i] = 0$, $I_X[X] = H[X]$
    - $I_X[Y] = E\left[\log_2 \dfrac{f(X,Y)}{f_X(X)f_Y(Y)}\right]$
    - symmetry $I_X[Y] = I_Y[X]$
    - $I_X[Y] \geq 0$, hence: $\bar{H}[Y|X] \leq H[Y]$ and $H[X,Y] \leq H[X] + H[Y]$

Entropy of a random vector

- entropy of vector $(X_1, X_2, \dots, X_r)'$ defined as

$$\begin{aligned} H[X_1, X_2, \dots, X_r] &= -E[\log_2 f(X_1, X_2, \dots, X_r)] \\ &= -\sum_{x_1, x_2, \dots, x_r} f(x_1, x_2, \dots, x_r) \log_2 f(x_1, x_2, \dots, x_r) \end{aligned}$$

- properties for 2 variables also hold for $r > 2$ variables:
    - $H[X_1, X_2, \dots, X_r] \leq H[X_1] + H[X_2] + \cdots + H[X_r]$; equality when $X_1, X_2, \dots, X_r$ are independent
    - $\bar{H}[X_r | X_1, X_2, \dots, X_{r-1}] = -E[\log_2 f(X_r | X_1, X_2, \dots, X_{r-1})]$

Entropy of a random vector

- properties:
  - when $\{Z_1, Z_2, \ldots, Z_s\} \subset \{X_1, X_2, \ldots, X_{r-1}\}$ then $\bar{H}[X_r|X_1, X_2, \ldots, X_{r-1}] \leq \bar{H}[X_r|Z_1, Z_2, \ldots, Z_s]$
  - hence $\bar{H}[X_r|X_{r-1}, \ldots, X_{r-k}] \leq \bar{H}[X_r|X_{r-1}, \ldots, X_{r-k+1}]$
  - hence $\bar{H}[X_r|X_{r-1}, \ldots, X_{r-k}] \searrow$ when $k \to \infty$; also bounded below by 0
  - hence $\lim_{k\to\infty} \bar{H}[X_r|X_{r-1}, X_{r-2}, \ldots, X_{r-k}] = \bar{H}_c[X_r]$ exists
  - $\bar{H}_c[X_r]$ is the expected conditional entropy and is measure of the present X_r given the observation of the entire past

Part III

Random Processes

Outline

Random processes

- *Random processes*: example: signal $x(t) = a \cos bt$
- with varying amplitude and frequency and in the presence of noise: $x_1(t) = a_1 \cos b_1 t + n_1(t)$ is a signal generated at some time with a_1, b_1, $n_1(t)$ specific values for **this** signal
- subsequent generated signal has probably other parameter values
- model as random process $X(t)$:

$$X(t) = A \cos Bt + N(t)$$

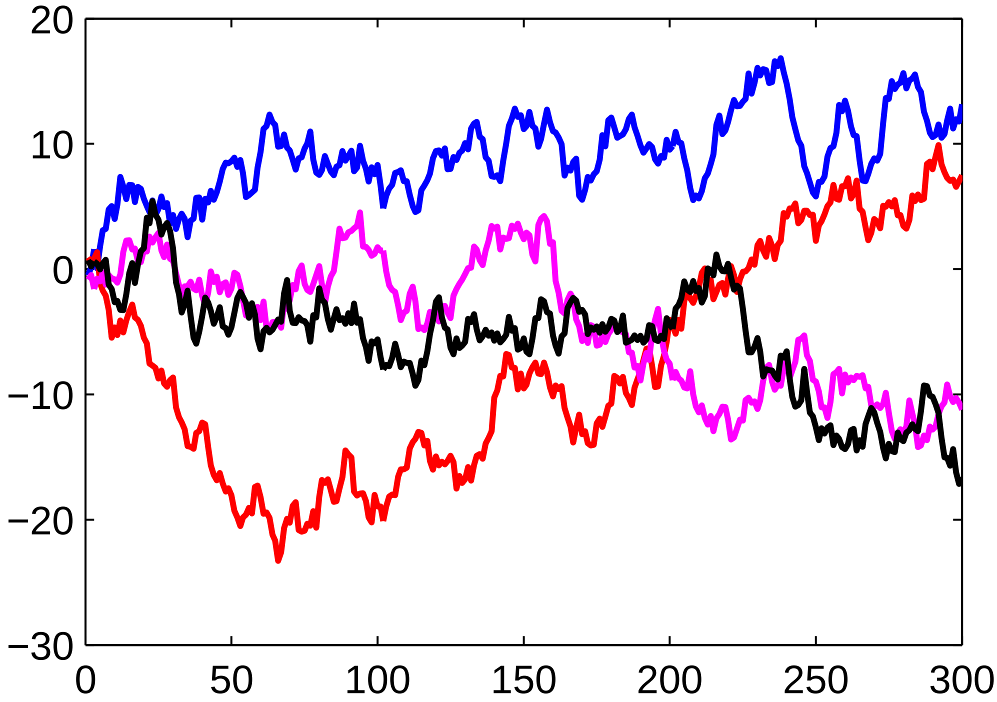
with A, B and $N(t)$ (for every t) random variables

Random processes

- random function or random process is a family of random variables $\{X(\omega;t)\}$ where $t \in T$ and for every fixed t, random variable $X(\omega;t)$ is defined over sample space S $(\omega \in S)$
- if $T \subset \mathbb{R} \Rightarrow$ random signal
- if $T \subset \mathbb{R}^2 \Rightarrow$ random image
- simplicity of notation: $X(t)$
- for every t there is a *1st order distribution* and *1st order density*:

$$F(x;t) = P(X(t) \leq x) \quad \text{and} \quad f(x;t) = \frac{d}{dx}F(x;t)$$

Random processes



Random processes: example

- digital image with sample space $S = \{a, b, c, d, e\}$ with probabilities
  $P(a) = P(b) = 1/8; \quad P(c) = P(d) = P(e) = 1/4$
- realizations of random image $X(\omega; t)$:

$$x_a = \begin{pmatrix} 1 & -1 \\ \mathbf{0} & 1 \end{pmatrix} \quad x_b = \begin{pmatrix} 0 & 2 \\ \mathbf{2} & 0 \end{pmatrix} \quad x_c = \begin{pmatrix} 1 & 2 \\ \mathbf{0} & 1 \end{pmatrix}$$

$$x_d = \begin{pmatrix} 0 & -1 \\ \mathbf{0} & 1 \end{pmatrix} \quad x_e = \begin{pmatrix} 2 & 1 \\ \mathbf{1} & 2 \end{pmatrix}$$

with $T = \{(0, 1), (0, 0), (1, 1), (1, 0)\}$

Random Processes Random Functions

Random processes: example

- hence

$$f(x; 0, 1) = 3/8\delta(x) + 3/8\delta(x-1) + 1/4\delta(x-2)$$
$$f(x; 0, 0) = 5/8\delta(x) + 1/4\delta(x-1) + 1/8\delta(x-2)$$
$$f(x; 1, 1) = 3/8\delta(x+1) + 1/4\delta(x-1) + 3/8\delta(x-2)$$
$$f(x; 1, 0) = 1/8\delta(x) + 5/8\delta(x-1) + 1/4\delta(x-2)$$

Random processes: example

- A and B are random variables with joint density $f_{A,B}(a,b)$; define $X(t) = At + B$
- every realization is a line
- distribution is given by

$$F(x;t) = P(At + B \leq x) = \int_{-\infty}^{\infty} \int_{-\infty}^{x-at} f_{A,B}(a,b)\, db\, da$$

Random processes

- for fixed t, $f(x;t)$ describes the behaviour of $X(t)$
- in general all joint (*n-th order*) densities need to be examined:

$$F(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n) = \\ P(X(t_1) \leq x_1, X(t_2) \leq x_2, \ldots, X(t_n) \leq x_n)$$

- marginal densities can be obtained via integration
- full characterization of $X(t)$ is in general not possible if T contains infinite number of points

Random processes

- we limit ourselves to using joint densities up to some order
- if $\forall\{t_1, t_2, \ldots, t_n\}$ $X(t_1), X(t_2), \ldots, X(t_n)$ are independent then
  $f(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n) = f(x_1; t_1)f(x_2; t_2)\cdots f(x_n; t_n)$
- important class: Gaussian random processes: if $\forall\{t_1, t_2, \ldots, t_n\}$ $X(t_1), X(t_2), \ldots, X(t_n)$ possess a multivariate normal distribution, completely characterized by $\boldsymbol{\mu}$ and $\boldsymbol{K}$

Outline

Moments of random processes

- characterization through distributions is difficult ⇒ use less complete descriptions such as moments
- expected value of $X(t)$:

$$\mu_X(t) = E[X(t)] = \int_{-\infty}^{\infty} xf(x;t)dx$$

- variance

$$\begin{aligned}\mathrm{Var}[X(t)] &= E[(X(t) - \mu_X(t))^2] \\ &= \int_{-\infty}^{\infty} (x - \mu_X(t))^2 f(x;t)dx = \sigma_X^2(t)\end{aligned}$$

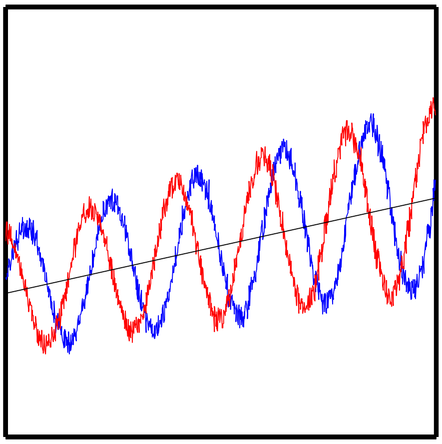
Moments of random processes

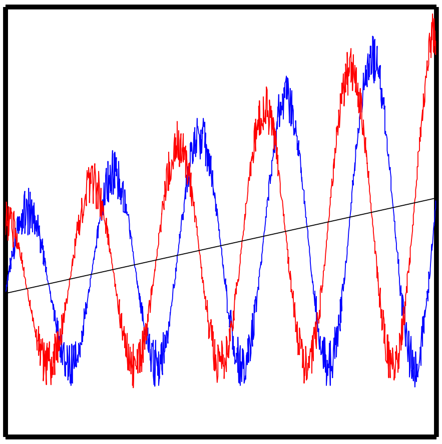
- covariance

$$\begin{aligned}K_X(t,t') &= E[(X(t) - \mu_X(t))(X(t') - \mu_X(t'))] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mu_X(t))(x' - \mu_X(t'))f(x,x';t,t')dx\,dx'\end{aligned}$$

- $K_X(t,t)] = \mathrm{Var}[X(t)]$
- symmetry: $K_X(t,t') = K_X(t',t)$
- correlation-coefficient function $\rho_X(t,t') = \dfrac{K_X(t,t')}{\sigma_X(t)\sigma_X(t')}$
- $|\rho(t,t')| \le 1$;
  $|\rho(t,t')| = 1 \Leftrightarrow P(X(t') = a_{t,t'}X(t) + b_{t,t'}) = 1$

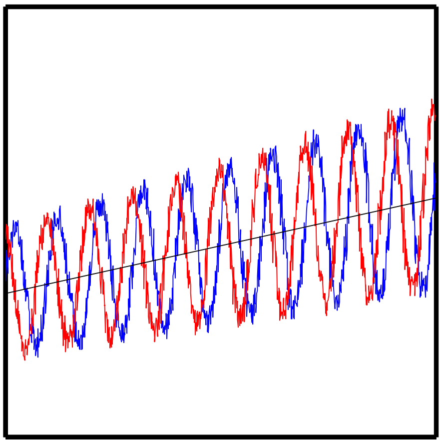
Moments of random processes

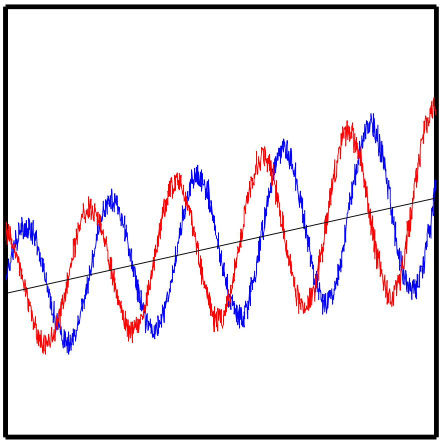




larger variance on the right than on the left

Moments of random processes





equal variance on the right and left

Moments of random processes

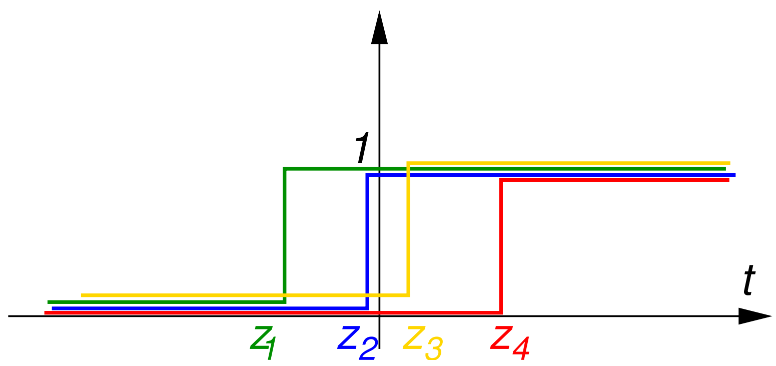
- *autocorrelation function*
  $R_X(t,t') = E[X(t)X(t')] = K_X(t,t') + \mu_X(t)\mu_X(t')$
- *cross-covariance function*
  $K_{XY}(t,s) = E[(X(t) - \mu_X(t))(Y(s) - \mu_Y(s))]$
- if $K_{XY}(t,s) = 0$ then $X(t)$ and $Y(s)$ are uncorrelated
- similarly: $K_{XY}(t,s) = K_{YX}(s,t)$; $\rho_{XY}(t,s) = \frac{K_{XY}(t,s)}{\sigma_X(t)\sigma_Y(s)}$ and $R_{XY}(t,s) = E[X(t)Y(s)]$
- also similar: higher order moments
- mixed moments of order $p_1 + p_2 + \cdots + p_n$:

$$\mu'_{p_1,p_2,\ldots,p_n}(t_1,t_2,\ldots,t_n) = E[X(t_1)^{p_1}X(t_2)^{p_2}\cdots X(t_n)^{p_n}]$$

  and mixed central moments $\mu_{p_1,p_2,\ldots,p_n}(t_1,t_2,\ldots,t_n)$

Moments of random processes: example

- example: $X(Z;t) = I_{[Z,\infty)}(t)$ with Z the standard normal variable



- for every observation z, $X(z;t)$ is a unit step function
- for fixed t, $X(t)$ is a binomial variable

Moments of random processes: example

- probabilities $P(X(t) = 1) = P(Z \leq t) = \Phi(t)$ and $P(X(t) = 0) = 1 - \Phi(t)$
- expected value $\mu_X(t) = \Phi(t)$
- autocorrelation $R_X(t,t')$: $P(X(t)X(t') = 1) = \Phi(\min(t,t'))$ and $P(X(t)X(t') = 0) = 1 - \Phi(\min(t,t'))$ hence $R_X(t,t') = E[X(t)X(t')] = \Phi(\min(t,t'))$
- covariance $K_X(t,t') = \Phi(\min(t,t')) - \Phi(t)\Phi(t')$
- variance $\mathrm{Var}[X(t)] = \Phi(t) - \Phi(t)^2$

Moments of random processes: example

- example: $X(W;t) = I_{[W,\infty)}(t)$ with W uniformly distributed over $[0,2]$
- for fixed t:

$$\mu_X(t) = P(X(t) = 1) = P(W \leq t) = \begin{cases} 0 & t < 0 \\ t/2 & 0 \leq t \leq 2 \\ 1 & t > 2 \end{cases}$$

- autocorrelation $E[X(t)X(t')] = P(W \leq \min(t,t'))$; leads to covariance

Moments of random processes: example

- example: $X(W;t) = I_{[W,\infty)}(t)$ with W a binomial variable with equiprobable outcomes 0 and 1
- then

$$\mu_X(t) = \begin{cases} 0 & t < 0 \\ 1/2 & 0 \leq t < 1 \\ 1 & t \geq 1 \end{cases}$$

- and autocorrelation

$$E[X(t)X(t')] = P(W \leq \min(t,t')) = \begin{cases} 0 & \min(t,t') < 0 \\ 1/2 & 0 \leq \min(t,t') < 1 \\ 1 & 1 \leq \min(t,t') \end{cases}$$

Moments of random processes: example

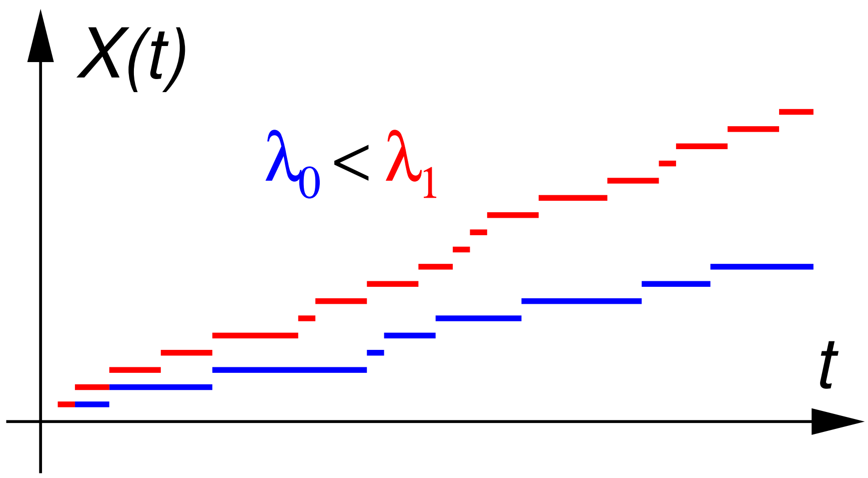
- random function $X(t) = At + B$
- $\mu_X(t) = E[A]t + E[B]$
  $E[X(t)X(t')] = E[A^2]tt' + E[AB](t+t') + E[B^2]$
  $K_X(t,t') = \text{Var}[A]tt' + \text{Cov}[A,B](t+t') + \text{Var}[B]$
  and when A and B are uncorrelated then
  $K_X(t,t') = \text{Var}[A]tt' + \text{Var}[B]$

Moments of random processes: example

- *Poisson process* $X(t)$ counts arrivals in $[0, t]$, eg models arrival of compute jobs for supercomputer:
- assumptions:
    - numbers of arrivals in non overlapping intervals are independent
    - P(1 arrival) in interval of length t is $\lambda t + o(t)$
    - P(more than 1 arrival) in interval of length t is $o(t)$
    - λ constant
- then $P(X(t) = k) = e^{-\lambda t} \dfrac{(\lambda t)^k}{k!}$ for $k = 0, 1, 2, \ldots$

Random Processes | Moments of a Random Function

Moments of random processes: example



two realizations of Poisson process

Moments of random processes: example

- random Euclidian image $X(Z; u, v)$ is indicator function of right half plane $\{(u, v) : u \geq Z\}$ with Z the standard normal variable
- then for a random binary image $X(u, v)$:
  $\mu_X(u, v) = P(X(u, v) = 1) = P(Z \leq u) = \Phi(u)$
  $R_X((u, v), (u', v')) = \Phi(\min(u, u'))$
  $K_X((u, v), (u', v')) = \Phi(\min(u, u')) - \Phi(u)\Phi(u')$

Moments of random processes: example

- random Euclidian image $X(R; u, v)$ is indicator function of random disk with radius R centered at the origin, and $f(r) = be^{-br} I_{[0,\infty)}(r)$ (exponential distribution)
  Let $s = \sqrt{u^2 + v^2}$
- then it follows that
  $\mu_X(u, v) = P(R \geq s) = e^{-bs}$
  $R_X((u, v), (u', v')) = P(R \geq s, R \geq s') = e^{-b \max(s,s')}$
  $K_X((u, v), (u', v')) = e^{-b \max(s,s')} - e^{-b(s+s')}$
  $\text{Var}[X(u, v)] = e^{-bs}(1 - e^{-bs})$
  $\Rightarrow$ has a maximum occuring at $s = b^{-1} \log 2$
- if (u, v) and (u', v') are located on circles with the same radius, then the correlation $\rho_X((u, v), (u', v')) = 1$

Moments of random processess: example

- realizations of discrete random image

$$x_a = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ \mathbf{0} & 1 & 1 \end{pmatrix} \; x_b = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ \mathbf{0} & 0 & 1 \end{pmatrix} \; x_c = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ \mathbf{0} & 0 & 0 \end{pmatrix} \; x_d = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ \mathbf{0} & 0 & 0 \end{pmatrix}$$

- sample space $S = \{a, b, c, d\}$ with $P(a) = P(d) = 1/6$ and $P(b) = P(c) = 1/3$
- this gives

$$\mu_x = \begin{pmatrix} 1/2 & 5/6 & 1 \\ 1/6 & 1/2 & 5/6 \\ \mathbf{0} & 1/6 & 1/2 \end{pmatrix}$$

- covariance: $E[X(0,1)X(1,1)] = 1/6$, hence $K_X((0,1),(1,1)) = 1/12$, also $K_X((2,0),(1,1)) = 1/4$; $\sigma_X((2,0)) = \sigma_X((1,1)) = 1/2 \Rightarrow \rho_X((2,0),(1,1)) = 1$

Mean and covariance of a sum

- linearity yields $\mu_{aX+bY}(t) = a\mu_X(t) + b\mu_Y(t)$
- for the covariance: $K_{X+Y}(t,t') = K_X(t,t') + K_Y(t,t') + K_{XY}(t,t') + K_{YX}(t,t')$ in case of uncorrelated X and Y: $K_{X+Y}(t,t') = K_X(t,t') + K_Y(t,t')$
- generalization to n functions: for $W(t) = \sum_{j=1}^{n} X_j(t)$ it is found that

$$\mu_W(t) = \sum_{j=1}^{n} \mu_{X_j}(t) \text{ and } K_W(t,t') = \sum_{i=1}^{n}\sum_{j=1}^{n} K_{X_iX_j}(t,t')$$

for mutually uncorrelated X_j:

$$K_W(t,t') = \sum_{j=1}^{n} K_{X_j}(t,t')$$

Outline

Differentiation of random functions

- problem: limit of difference quotient interacts with randomness
- suggestion: derivative of random process can be found as the process that consists of the derivatives of the realizations

$$\frac{d}{dt}\{X(\omega;t)\} \stackrel{?}{=} \{\frac{d}{dt}X(\omega;t)\}$$

- let $\Delta_X(\omega;t,h) = \dfrac{X(\omega;t+h) - X(\omega;t)}{h}$
- then for fixed ω the deterministic derivative equals

$$\frac{d}{dt}X(\omega;t) = \lim_{h\to 0}\Delta_X(\omega;t,h)$$

Differentiation of random functions

- linearity leads to

$$\mu_{X'}(t) = E[X'(t)] = E\left[\lim_{h\to 0} \frac{X(\omega; t+h) - X(\omega; t)}{h}\right]$$
$$= \frac{d}{dt}E[X(t)] = \frac{d}{dt}\mu_X(t)$$

- not rigorous: $\Delta_X(\omega; t, h)$ is a random variable and interchanging lim and $E[\cdot]$ is not justified
- similar for covariance:

$$K_{X'}(u, v) = \frac{\partial^2}{\partial u \partial v} K_X(u, v)$$

Mean-square differentiability

- define derivative through mean-square convergence: $X_h(t) \to X(t)$ in mean-square sense if

$$\lim_{h\to 0} E[|X_h(t) - X(t)|^2] = 0$$

- notation: $\underset{h\to 0}{\mathrm{l.i.m.}}\, X_h(t) = X(t)$
- $X(t)$ is differentiable in mean-square sense and $X'(t)$ is the mean-square derivative in t if

$$X'(\omega; t) = \underset{h\to 0}{\mathrm{l.i.m.}}\, \Delta_X(\omega; t, h) \qquad \text{or}$$
$$\lim_{h\to 0} E\left[\left|\frac{X(\omega; t+h) - X(\omega, t)}{h} - X'(\omega; t)\right|^2\right] = 0$$

Mean-square differentiability

- theorem: $X(t)$ is mean-square differentiable on interval T $\Leftrightarrow$ expected value is differentiable on T and covariance has second order partial derivatives with respect to u and v on T. In that case

$$\mu_{X'}(t) = \frac{d}{dt}\mu_X(t)$$

and

$$K_{X'}(u,v) = \frac{\partial^2}{\partial u \partial v} K_X(u,v)$$

- this can be extended to higher dimensions and partial derivatives

Mean-square differentiability: example

- parabola $X(Z;t) = (t-Z)^2$, translated by random distance Z; fixing Z and taking derivative yields $Y(Z;t) = 2(t-Z)$
- $Y(Z;t)$ is the MS (mean square) derivative:

$$\begin{aligned}
&E[|\Delta_X(Z;t,h) - Y(Z;t)|^2] \\
&= E\left[\left|\frac{(t+h-Z)^2-(t-Z)^2}{h} - 2(t-Z)\right|^2\right] \\
&= E[h^2] = h^2
\end{aligned}$$

- letting $h \to 0$ shows that $X'(Z;t) = 2(t-Z)$

Mean-square differentiability: example

- random sine wave $X(t) = Z\sin(t - W)$ with Z and W independent
- differentiation of realizations yields $Y(t) = Z\cos(t - W)$
- this is the MS derivative!

$$
\begin{aligned}
&\Delta_X(Z; t, h) - Y(Z; t) \\
&= \frac{Z\sin(t + h - W) - Z\sin(t - W)}{h} - Y(Z; t) = \cdots \\
&= Z[\cos W(\xi \cos(t + h/2) - \cos t) \\
&\quad + \sin W(\xi \sin(t + h/2) - \sin t)] \\
&\qquad\qquad\qquad\qquad \text{with } \xi = \sin(h/2)/(h/2)
\end{aligned}
$$

- hence $\lim_{h\to 0} E[|\Delta_X(Z; t, h) - Y(Z; t)|^2] = 0$

Random Processes | Differentiation

Mean-square differentiability: example

- according to theorem:

$$
\begin{aligned}
\mu_X(t) &= E[Z\sin(t - W)] \\
&= E[Z](E[\cos W]\sin t - E[\sin W]\cos t)
\end{aligned}
$$

and

$$
\begin{aligned}
\mu_{X'}(t) &= E[Z\cos(t - W)] \\
&= E[Z](E[\cos W]\cos t + E[\sin W]\sin t)
\end{aligned}
$$

- 2nd condition: $K_X(u, v) = \cdots$ and $K_{X'}(u, v) = \cdots$ and it is found that $K_{X'}(u, v) = \dfrac{\partial^2}{\partial u \partial v} K_X(u, v)$
- power of theorem is in the $\Leftarrow$ direction

Mean-square differentiability: example

- counterexample: $X(Z;t) = I_{[Z,\infty)}(t)$ with Z the standard normal variable
- all realizations $x_z(t)$ are differentiable except single one where $t = z$
- removal of that realization ($P(Z = z) = 0!!$) yields differentiable process. But there is no differentiability over an interval!
- use the theorem:

$$(1) \quad \frac{d}{dt}\mu_X(t) = \frac{d}{dt}\Phi(t) = f_Z(t)$$

Random Processes Differentiation

Mean-square differentiability: example

- (2) $\dfrac{\partial^2}{\partial u \partial v} K_X(u,v)$: first calculate

$$\frac{\partial}{\partial u} K_X(u,v) = \frac{\partial}{\partial u}[\Phi(\min(u,v)) - \Phi(u)\Phi(v)]$$

$$= \begin{cases} f_Z(u) - f_Z(u)\Phi(v) & u < v \\ -f_Z(u)\Phi(v) & v < u \end{cases}$$

- subsequent calculation of $\frac{\partial}{\partial v}$ is not possible except in generalized sense:

$$\frac{\partial^2}{\partial u \partial v} K_X(u,v) = f_Z(u)\delta(v-u) - f_Z(u) f_Z(v)$$

- hence $X(t)$ is not MS differentiable

Outline

Integration of random functions

- similar problem as with differentiation: convergence of limit needs careful definition

$$\int_a^b x(t)\,dt = \lim_{\|\Delta t_k\| \to 0} \sum_{k=1}^{n} x(t_k')\Delta t_k$$

  with $a = t_0 < t_1 < t_2 < \cdots < t_n = b, \quad \Delta t_k = t_k - t_{k-1}$,
  $\|\Delta t_k\|$ is the maximum of all Δt_k, $t_k' \in [t_{k-1}, t_k]$
  = limit of Riemann sums
- derivation for two dimensions $t = (u, v) \in \mathbb{R}^2$ and $T \subset \mathbb{R}^2$
- partition $\Xi = \{R_k\}$ of T consisting of disjoint collection of rectangles R_k such that $T = \bigcup_k R_k$

Integration of random functions

- Riemann sum for realization of $X(\omega; u, v)$ and partition Ξ:

$$\Sigma_X(\omega; \Xi) = \sum_{k=1}^{n} X(\omega; u'_k, v'_k) A(R_k)$$

with $A(R_k)$ the area of R_k and $(u'_k, v'_k) \in R_k$

- integral: with $\|R_k\|$ the maximum of the rectangle dimensions, take limit over all partitions for which $\|R_k\| \to 0$

$$\iint_T X(\omega; u, v)\, du\, dv = \lim_{\Xi, \|R_k\| \to 0} \Sigma_X(\omega; \Xi)$$

Integration of random functions

- problem: lim and $E[\cdot]$ cannot be interchanged $\Rightarrow$ mean-square convergence
- $X(\omega; u, v)$ is mean-square integrable with integral I

$$I = \iint_T X(u, v)\, du\, dv \qquad \text{(= random variable)}$$

if and only if

$$\lim_{\Xi, \|R_k\| \to 0} E[|I - \Sigma_X(\omega; \Xi)|^2] = 0$$

Integration of random functions

- theorem: if $Y(t)$ exists in MS sense

$$Y(t) = \int_T g(t,s)X(s)\,ds$$

(with $g(t,s)$ a deterministic function), then

$$(1) \quad \mu_Y(t) = \int_T g(t,s)\mu_X(s)\,ds$$

$$(2) \quad K_Y(t,t') = \int_T \int_T g(t,s)g(t',s')K_X(s,s')\,ds\,ds'$$

- conversely: if (1) and (2) exist, then $Y(t)$ exists in MS sense

Integration of random functions

- consequence: covariance is *nonnegative definite*:

$$\int_T \int_T K_X(t,t')g(t)g(t')\,dt\,dt' \geq 0$$

- $\Rightarrow$ not every symmetric function can be a covariance function!

Integration of random functions

- theorem boils down to interchange of lim and $E[\cdot]$:

$$\begin{aligned}\mu_Y(t) &= E\left[\lim_{\|\Delta s_k\|\to 0}\sum_{k=1}^{n} g(t,s'_k)X(\omega;s'_k)\Delta s_k\right]\\ &= \lim_{\|\Delta s_k\|\to 0}\sum_{k=1}^{n} g(t,s'_k)E[X(\omega;s'_k)]\Delta s_k\\ &= \int_T g(t,s)E[X(\omega;s)]ds\\ &= \int_T g(t,s)\mu_X(s)ds\end{aligned}$$

Integration of random functions: example

- example: random sine wave $X(t) = Z\sin(t - W)$ with $g(t,s) = 1$
- $\mu_X(t)$ and $K_X(t,t')$ are integrable, hence $X(t)$ is integrable in MS sense
- for $T = [0, 2\pi[$ it is found that

$$\begin{aligned}\mu_Y(t) &= \int_0^{2\pi} \mu_X(s)\,ds\\ &= E[Z]\left(E[\cos W]\int_0^{2\pi}\sin t\,dt - E[\sin W]\int_0^{2\pi}\cos t\,dt\right) = 0\end{aligned}$$

Integration of random functions: example

- form of the theorem lends itself to integral transforms such as the Fourier series
- coefficients of Fourier series are themselves random variables:

$$A_k = \frac{1}{\pi}\int_0^{2\pi} X(t)\cos kt\,dt \quad \text{and} \quad B_k = \frac{1}{\pi}\int_0^{2\pi} X(t)\sin kt\,dt$$

- if all A_k and B_k exist (cf. theorem) then the Fourier series of $X(t)$ exists
- partial sum

$$X_n(t) = \frac{A_0}{2} + \sum_{k=1}^{n} A_k \cos kt + B_k \sin kt$$

Outline

Mean ergodicity

- estimate needed of parameters of random processes, eg. expected value
- for discrete process $X(k)$ with $k = 1, 2, \ldots$ with constant mean $E[X(t)] = \mu$, set

$$Y(n) = \frac{1}{n} \sum_{k=1}^{n} X(k)$$

- then $\mathrm{Var}[Y(n)] = E[|Y(n) - \mu|^2] = \cdots = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_X(i,j)$
- hence $Y(n)$ converges to μ in MS sense $\Leftrightarrow$

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_X(i,j) = 0$$

Mean ergodicity

- use Chebyshev inequality in present context:

$$P(|Y(n) - \mu| \geq \varepsilon) \leq \frac{1}{\varepsilon^2 n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_X(i,j)$$

  and hence if lim $= 0$ then

$$P(|Y(n) - \mu| \geq \varepsilon) = 0$$

- generalization required to one or more continuous random variables

Mean ergodicity

- two-dimensional continuous random process $X(\omega; u, v)$ is averaged over square $T = [-r, r] \times [-r, r]$ with area $4r^2$

$$Y = \frac{1}{(2r)^2} \int_{-r}^{r} \int_{-r}^{r} X(u, v)\, du\, dv$$

- Y is random variable, not a function, and

$$\mu_Y = \frac{1}{(2r)^2} \int_{-r}^{r} \int_{-r}^{r} \mu_X(u, v)\, du\, dv$$

$$\mathrm{Var}[Y] = \frac{1}{(2r)^4} \int_{-r}^{r} \int_{-r}^{r} \int_{-r}^{r} \int_{-r}^{r} K_X((u, v), (u', v'))\, du\, dv\, du'\, dv'$$

Random Processes Mean Ergodicity

Mean ergodicity

- let $X_0(u, v) = X(u, v) - \mu_X(u, v)$, then

$$\mathrm{Var}[Y] = E[|Y - \mu_Y|^2] = E\left[\left|\frac{1}{(2r)^2} \int_{-r}^{r} \int_{-r}^{r} X_0(u, v)\, du\, dv\right|^2\right]$$

$$= \frac{1}{(2r)^4} \int_{-r}^{r} \int_{-r}^{r} \int_{-r}^{r} \int_{-r}^{r} K_X((u, v), (u', v'))\, du\, dv\, du'\, dv'$$

- hence $\underset{r\to\infty}{\mathrm{l.i.m.}} \frac{1}{(2r)^2} \int_{-r}^{r} \int_{-r}^{r} X_0(u, v)\, du\, dv = 0$ if and only if fourfold integral = 0 for $r \to \infty$

Mean ergodicity

- if $X(u,v)$ has constant μ_X then the limit becomes

$$\underset{r\to\infty}{\text{l.i.m.}} \frac{1}{(2r)^2} \int_{-r}^{r}\int_{-r}^{r} X(u,v)\,du\,dv = \mu_X$$

- hence MS limit of the average of $X(u,v)$ over T equals the expected value of $X(u,v)$ if *ergodicity* holds

Mean ergodicity

- theorem: if n-dimensional $X(t)$ has constant μ_X, and with $A[X;r]$ the average of $X(t)$ over the n-dimensional square T with side $2r$ around the origin, then it holds that

$$\underset{r\to\infty}{\text{l.i.m.}}\, A[X;r] = \mu_X$$

if and only if

$$\lim_{r\to\infty} \frac{1}{(2r)^{2n}} \int_T \int_T K_X(u,v)\,du\,dv = 0$$

(2n fold integral)

Mean ergodicity

- importance of ergodicity: estimate of μ can be found using only one realization by averaging over sufficiently large area
- without ergodicity: estimate requires averaging over many observations for every t
- necessary and sufficient condition is difficult to work with; will be simplified later on for class of random processes (stationary processes)

Outline

Poisson process

- one-dimensional Poisson process describes points arriving randomly in time, $X(t)$ counts the number of points arriving in $[0, t]$
- assumptions:
  - numbers of arrivals in non overlapping intervals are independent
  - P of exactly 1 arrival in interval of length t is $\lambda t + o(t)$
  - P of two or more arrivals in interval of length t is $o(t)$
- parameter λ is constant and $o(t)$ is any function $g(t)$ for which $\lim_{t\to 0} g(t)/t = 0$
- random arrival times are *Poisson points*

Poisson process

- density of Poisson process $P(X(t) = k) = e^{-\lambda t}\dfrac{(\lambda t)^k}{k!}$ for $k = 0, 1, 2, \ldots$
  with expected value = variance = λt
- Poisson process has *independent increments*: for $t < t' < u < u'$, $X(u') - X(u)$ and $X(t') - X(t)$ are independent
- using this we can find that

$$K_X(t, t') = \lambda \min(t, t') \text{ and } \rho(t, t') = \frac{\min(t, t')}{\sqrt{tt'}}$$

and hence $\rho(t, t') \to 0$ if $|t - t'| \to \infty$

Derivative of the Poisson process

- Poisson process is not differentiable in MS sense
- generalized derivative of covariance:

$$\frac{\partial^2 K_X(t,t')}{\partial t\,\partial t'} = \lambda\delta(t-t')$$

$\Rightarrow$ theorem applied in generalized sense

- derivative of expected value = λ
- *generalized derivative of Poisson process*: random process with expected value λ and covariance $\lambda\delta(t-t')$

Derivative of the Poisson process

- process can be written as

$$X'(t) = \sum_{k=1}^{\infty} \delta(t-Z_k)$$

with Z_k a sequence of Poisson points
$\Rightarrow$ *Poisson impulse process* with $\mu_{X'}(t) = \lambda$ and $K_{X'}(t,t') = \lambda\delta(t-t')$

Poisson points

- time distribution between Poisson points: distribution of k-th Poisson point following given point
- let Y = time at which k-th Poisson point occurs, then

$$F_Y(t) = P(Y \leq t) = P(X(t) \geq k) = \sum_{x=k}^{\infty} \frac{e^{-\lambda t}(\lambda t)^x}{x!}$$

  taking derivative yields

$$f_Y(t) = \frac{d}{dt} F_Y(t) = \frac{\lambda e^{-\lambda t}(\lambda t)^{k-1}}{(k-1)!}$$

  $\Rightarrow$ gamma distribution with $\alpha = k$ and $\beta = 1/\lambda$
- for $k = 1$: exponential distribution with expected value $1/\lambda$

Poisson points

- Poisson points model complete randomness: uniform distribution of points in $[0, \infty)$ $\Rightarrow$ physical phenomena
- in finite interval $[a, b]$: n Poisson points in this interval determine n random variables that are independent and uniformly distributed over the interval
- if points in the interval are ordered: $\tau_1 < \tau_2 < \cdots < \tau_n$ in $[0, T]$; it can be shown that these have the same distribution as the order statistics of a random sample of size n taken from a uniform distribution over $[0, T]$, with density

$$f_{\tau_1, \tau_2, \ldots, \tau_n}(t_1, t_2, \ldots, t_n) = \begin{cases} n!/T^n & 0 \leq t_1 \leq t_2 \leq \cdots \leq t_n \leq T \\ 0 & \text{elsewhere} \end{cases}$$

Poisson points

- functions of Poisson points: apply g to Poisson points and take sum
- expected value of sum $E\left[\sum_{i=1}^{\infty} g(t_i)\right] = \lambda \int_{-\infty}^{\infty} g(t)\,dt$
- follows from uniform distribution and chain rule:

$$E\left[\sum_{i=1}^{n} g(s_i)\right] = \sum_{i=1}^{n} E[g(s_i)] = \frac{n}{b-a}\int_a^b g(t)\,dt$$

for n points s_i in $[a, b]$

Poisson points

- in fact, number of points in $[a, b]$ is a random variable N and

$$\frac{n}{b-a}\int_a^b g(t)\,dt = E[W|N=n] \quad \text{with} \quad W = \sum_{a \le t_i \le b} g(t_i)$$

- then

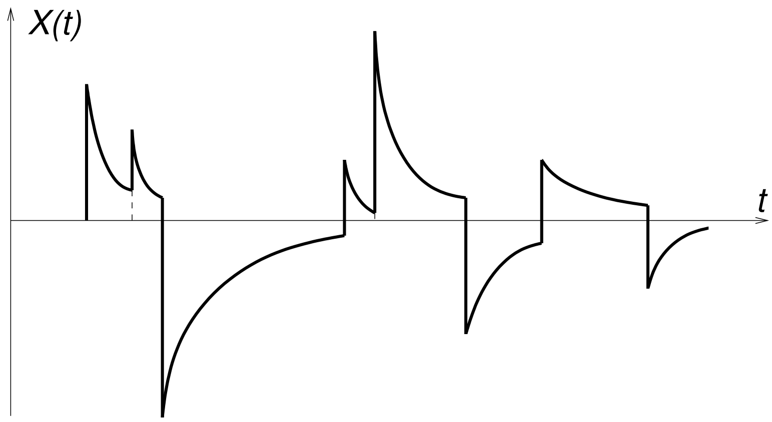
$$\begin{aligned} E[W] &= \sum_{n=0}^{\infty} E[W|N=n]P(N=n) \\ &= \frac{1}{b-a}\int_a^b g(t)\,dt \sum_{n=0}^{\infty} nP(N=n) \\ &= \lambda \int_a^b g(t)\,dt \end{aligned}$$

Poisson points

- it also holds that

$$E\left[\prod_{i=1}^{\infty}(1+g(t_i))\right] = \exp\left(\lambda \int_{-\infty}^{\infty} g(t)\,dt\right)$$

- filtered Poisson impulse process



Poisson process: axiomatic formulation

- $X(t)$ with $t \geq 0$ has *independent increments* if $X(0) = 0$ and the random variables $X(t_2) - X(t_1)$, $X(t_3) - X(t_2)$, ..., $X(t_n) - X(t_{n-1})$ are independent for all $t_1 < t_2 < \cdots < t_n$
- *stationary* independent increments: if $X(t+r) - X(t'+r)$ is identically distributed as $X(t) - X(t')$, $\forall t, t', r$
- axiomatic definition of Poisson process with parameter λ:
  - $X(t)$ has values in $\{0, 1, 2, \dots\}$
  - $X(t)$ has stationary independent increments
  - for $s < t$, $X(t) - X(s)$ has a Poisson distribution with expected value $\lambda(t-s)$

Poisson process: axiomatic formulation

- extension to higher dimensions: Poisson points in space $\Rightarrow$ modeling of grain, texture, ...
- points are randomly distributed over $D \subset \mathbb{R}^n$ according to a spatial Poisson process if
  - for disjoint domains $D_1, D_2, \ldots, D_r$ the counts $N(D_1), N(D_2), \ldots, N(D_r)$ are independent random variables
  - for every D, $N(D)$ has a Poisson distribution with expected value $\lambda v(D)$ with $v(D)$ = volume(D)
- theorem: if $\{t_i\}$ is Poisson point process in $\mathbb{R}^n$ with intensity λ, $\{s_i\}$ is sequence of independent identically distributed random variables, independent from $\{t_i\}$, then $\{t_i + s_i\}$ is Poisson point process with intensity λ

Outline

White noise

- *discrete white noise* $X(k)$: random function with $\mu_X = 0$ and $X(k)$ and $X(j)$ uncorrelated for $k \neq j$
- covariance of white noise:

$$K_X(k,j) = E[X(k)X(j)] = \begin{cases} \mathrm{Var}[X(k)] & k = j \\ 0 & k \neq j \end{cases}$$

- hence it follows for a function g that

$$\sum_{i=1}^{\infty} K_X(k,i)g(i) = \mathrm{Var}[X(k)]g(k)$$

White noise

- suppose a similar process exists for the continuous setting:

$$\int_{-\infty}^{\infty} K_X(t,t')g(t')dt' = I(t)g(t)$$

  where $I(t)$ plays the role played by $\mathrm{Var}[X(k)]$ for the discrete case
- this is possible if we set $K_X(t,t') = I(t)\delta(t-t')$
- $\Rightarrow$ *continuous white noise* represented by any random process with $\mu = 0$ and covariance of the form $I(t)\delta(t-t')$

White noise

- $\Rightarrow$ infinite variance and uncorrelated variables
- $I(t)$ is the *intensity* of white noise
- approximation in 1 dimension: process with

$$K_X(t, t') = be^{-b|t-t'|} \quad (b > 0)$$

  exists and if $b \nearrow\nearrow$, $K_X(t, t')$ behaves like covariance of white noise

Random walk

- discrete process in which at every time step a particle moves over a unit distance either to the left or to the right, starting from 0
- steps are independent; P for step to the right = p; to the left $q = 1 - p$
- onedimensional *random walk* is discrete process $X(n)$ that gives the distance to the right after n steps; $X(n)$ can take values from $\{-n, -n+2, \ldots, n-2, n\}$

Random walk

- for $X(n) = x$ there must be $(n+x)/2$ steps to the right and $(n-x)/2$ to the left, hence, with Y binomial random variable for n trials

$$f_{X(n)}(x) = P\left(Y = \frac{n+x}{2}\right) = \binom{n}{\frac{n+x}{2}} p^{(n+x)/2} q^{(n-x)/2}$$

- moments of $X(n)$ follow from moments of Y since $X(n) = 2Y - n$:

$$E[X(n)^m] = \sum_{y=0}^{n} (2y-n)^m P(Y=y)$$

Random walk

- and therefore

$$E[X(n)] = 2np - n$$
$$E[X(n)^2] = 4(npq + n^2p^2) + (1-4p)n^2$$

- important case: $p = q = 1/2$, then $E[X(n)] = 0$ and $\mathrm{Var}[X(n)] = n$
- process has stationary independent increments; use this to find that

$$K_X(n,n') = \min(n,n')$$
$$\rho_X(n,n') = \frac{\min(n,n')}{\sqrt{nn'}}$$
$$\rho_X(n,n') \to 0 \quad \text{if} \quad |n-n'| \to \infty$$

Wiener process

- *Brownian motion* of particles: take a random walk process in several dimensions and decrease step size
- assumption of stationary independent increments is plausible
- assumption that displacement $X(t)$ is normally distributed with $\mu = 0$ is empirically verified
- *Wiener process* $X(t)$ with $t \geq 0$ satifies the conditions
  - $X(0) = 0$
  - $X(t)$ has stationary independent increments
  - $E[X(t)] = 0$
  - for every t, $X(t)$ is normally distributed

Wiener process

- from these assumptions it can be shown that:
  - increment $X(t) - X(t')$ has $\mu = 0$ and variance $\sigma^2|t - t'|$ with σ^2 a parameter to be determined empirically
  - $\mathrm{Var}[X(t)] = \sigma^2 t \quad (t \geq 0)$
  - $K_X(t, t') = \sigma^2 \min(t, t')$
  - Wiener and Poisson process have the same covariance up to a multiplicative factor
  - derivative of Wiener process is white noise with

$$\frac{\partial^2 K_X(t, t')}{\partial t \partial t'} = \sigma^2 \delta(t - t')$$

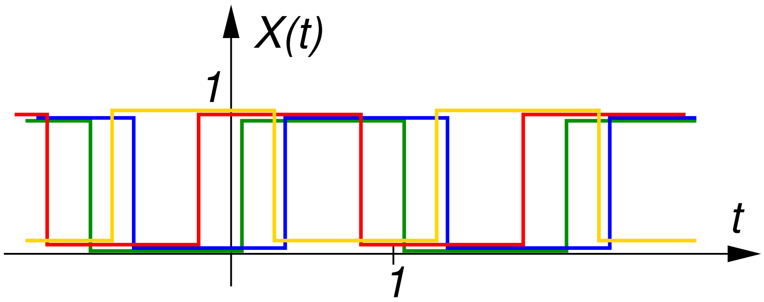
Outline

Stationarity

- relation between distributions at different sets of time points
- *WSS: wide sense stationarity* when $X(t)$ has constant μ_X and when $K_X(t, t') = k_X(\tau)$ with $\tau = t - t'$
- hence $\mathrm{Var}[X(t)] = K_X(t, t) = k_X(0) = \text{constant}$
- also
  - $k_X(-\tau) = k_X(\tau)$
  - $\rho_X(t - t') = \rho_X(\tau) = \dfrac{k_X(\tau)}{k_X(0)}$, hence also $|k_X(\tau)| \leq k_X(0)$
  - $R_X(t, t') = k_X(\tau) + \mu_X^2 = r_X(\tau)$
- WS stationarity also implies translation invariance: $\forall h : K_X(t + h, t' + h) = K_X(t, t')$

Stationarity: example

- waveform $g(t) = \begin{cases} 1 & 0 \le t < 1 \\ 0 & 1 \le t < 2 \end{cases}$ and defined over $\mathbb{R}$ by periodic extension, and Z uniformly distributed random variable over $[0, 2]$, define new random process $X(t) = g(t - Z)$



- from uniformity of Z it follows that for every t, $X(t)$ is a binary random variable with equiprobable outcomes 0 and 1 and with $\mu_X = 1/2$

Random Processes Stationarity

Stationarity: example

- covariance: first suppose $|t - t'| \le 1$: $R_X(t, t') = P(X(t) = X(t') = 1)$ and to have $X(t) = X(t') = 1$ both t and t' must be located under a block of the waveform. Because Z is distributed uniformly,

$$P(X(t) = X(t') = 1) = \frac{1 - |t - t'|}{2}$$

- for $1 < |t - t'| \le 2$: $P(X(t) = X(t') = 1) = \dfrac{|t - t'| - 1}{2}$
- hence

$$R_X(t, t') = \frac{|1 - |t - t'||}{2} \quad \text{and} \quad k_X(\tau) = \frac{2|1 - |\tau|| - 1}{4}$$

(for all τ by periodic extension)
- $\mathrm{Var}[X(t)] = k_X(0) = 1/4$
- hence $X(t)$ is WS stationary

Stationarity: example

- let $Y(t)$ be the Poisson process with $\mu_Y(t) = \lambda t$ and $r > 0$ a constant, then $X(t) = Y(t+r) - Y(t)$ is a Poisson increment process; counts # points in $(t, t+r]$
- $\mu_X(t) = E[Y(t+r)] - E[Y(t)] = \lambda r$
- covariance: if $|t - t'| \geq r$ then $k_X(\tau) = 0$ because the intervals defined by t and t' don't overlap
- for $t < t'$, $E[Y(t)Y(t')] = \lambda t + \lambda^2 tt'$, hence for $|t - t'| < r$: $K_X(t, t') = \ldots = \lambda(r - (t' - t))$
because of symmetry $K_X(t, t') = \lambda(r - |t - t'|)$
- hence $X(t)$ is WS stationary with

$$k_X(\tau) = \begin{cases} \lambda(r - |\tau|) & |\tau| \leq r \\ 0 & |\tau| > r \end{cases}$$

Stationarity: example

- $X(t)$ with values -1 and 1, changing at Poisson points and remaining constant in between; also $X(0) = 1$,

$$X(t) = \begin{cases} 1 & k \quad \text{even} \\ -1 & k \quad \text{odd} \end{cases} \quad k = \#\ \text{Poisson points in } (0, t]$$

- hence
$$P(X(t) = 1) = P(k \text{ even}) = \sum_{k=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^{2k}}{(2k)!} = e^{-\lambda t} \cosh \lambda t$$
and also
$P(X(t) = -1) = e^{-\lambda t} \sinh \lambda t$
- from this $\mu_X(t) = e^{-\lambda t} \cosh \lambda t - e^{-\lambda t} \sinh \lambda t = e^{-2\lambda t}$
- finally $R_X(t, t') = E[X(t)X(t')] = \cdots = e^{-2\lambda |t - t'|}$
- not stationary because $\mu_X(t)$ is not a constant!

Stationarity: example

- slight change: allow value 1 as well as -1 at origin, equiprobable
- set $Y(0)$ a binomial random variable with values -1 or 1, with probability 1/2 each, and independent from the Poisson process that determines $X(t)$, define

$$Y(t) = Y(0)X(t) = \begin{cases} X(t) & Y(0) = 1 \\ -X(t) & Y(0) = -1 \end{cases}$$

- $Y(t)$ is the random telegraph signal
- $\mu_Y(t) = E[Y(0)]E[X(t)] = 0$ and $K_Y(t,t') = E[Y(t)Y(t')] = E[Y(0)^2]E[X(t)X(t')] = e^{-2\lambda|t-t'|}$
- hence $Y(t)$ is WS stationary!

Stationarity: example

- the random signal $X(t) = Z\cos bt + W\sin bt$ with b a constant and Z and W uncorrelated variables with expected value 0 and variance σ^2
- then $\mu_X(t) = E[Z]\cos bt + E[W]\sin bt = 0$
- covariance:

$$\begin{aligned} K_X(t,t') &= E[X(t)X(t')] \\ &= E[Z^2]\cos bt\cos bt' + E[W^2]\sin bt\sin bt' \\ &\quad + E[ZW](\cos bt\sin bt' + \sin bt\cos bt') \\ &= \sigma^2\cos b(t-t') \end{aligned}$$

- hence $X(t)$ is WS stationary

Ergodicity for WS stationary processes

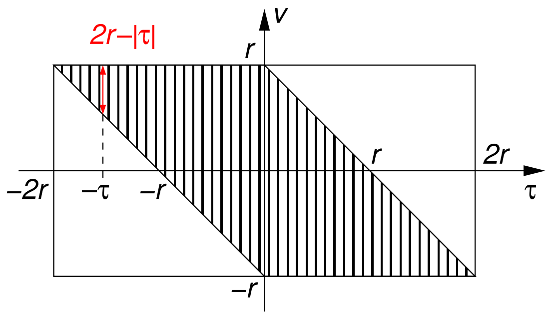
- theorem about ergodicity of expected value contains a complex necessary and sufficient condition for onedimensional WSS process:

$$\lim_{r\to\infty}\frac{1}{(2r)^2}\int_{-r}^{r}\int_{-r}^{r}k_X(u-v)\,du\,dv$$

$$=\lim_{r\to\infty}\frac{1}{(2r)^2}\int_{-r}^{r}\int_{-r-v}^{r-v}k_X(\tau)\,d\tau\,dv=0$$

Ergodicity for WS stationary processes

- define $G(\tau,v)=k_X(\tau)$ in shaded area and $=0$ elsewhere



- then

$$\frac{1}{(2r)^2}\int_{-r}^{r}\int_{-r-v}^{r-v}k_X(\tau)\,d\tau\,dv=\frac{1}{(2r)^2}\int_{-r}^{r}\int_{-2r}^{2r}G(\tau,v)\,d\tau\,dv$$

$$=\frac{1}{(2r)^2}\int_{-2r}^{2r}\int_{-r}^{r}G(\tau,v)\,dv\,d\tau=\frac{1}{(2r)^2}\int_{-2r}^{2r}(2r-|\tau|)k_X(\tau)\,d\tau$$

Ergodicity for WS stationary processes

- hence the theorem: WS stationary process is mean ergodic

$$\Leftrightarrow \quad \lim_{r\to\infty} \frac{1}{2r} \int_{-2r}^{2r} \left(1 - \frac{|\tau|}{2r}\right) k_X(\tau)\, d\tau = 0$$

- readily extendible to n dimensions
- simpler sufficient conditions:
  - with integrable covariance: $\left| \left(1 - \frac{|\tau|}{2r}\right) k_X(\tau) \right| \leq |k_X(\tau)|$,

    hence if $\int_{-\infty}^{\infty} |k_X(\tau)|\, d\tau$ exists $\Rightarrow$ OK
  - $\lim_{|\tau|\to\infty} k_X(\tau) = 0$ (eg: random telegraph signal)

Covariance ergodicity for WSS processes

- WS stationary process $X(t)$ is covariance ergodic if (with $X_0(t) = X(t) - \mu_X$)

$$k_X(\tau) = \underset{r\to\infty}{\mathrm{l.i.m.}} \frac{1}{2r} \int_{-r}^{r} X_0(t+\tau) X_0(t)\, dt$$

- then covariance can be estimated by averaging realization of $X_0(t+\tau)X_0(t)$ over sufficiently large interval
- random process $Y(t) = X_0(t+\tau)X_0(t)$ is WS stationary if $X(t)$ is Gaussian with $\mu_Y(t) = k_X(\tau)$ and $K_Y(t,t') = k_X^2(t-t') + k_X(t-t'+\tau)k_X(t-t'-\tau)$

Covariance ergodicity for WSS processes

- apply mean ergodicity theorem to $Y(t)$
- this gives the theorem: WSS Gaussian process is covariance ergodic if and only if

$$\lim_{r\to\infty} \frac{1}{2r} \int_{-2r}^{2r} \left(1 - \frac{|t|}{2r}\right) [k_X^2(t) + k_X(t+\tau)k_X(t-\tau)]\, dt = 0$$

- $\lim_{|\tau|\to\infty} k_X(\tau) = 0$ is sufficient condition

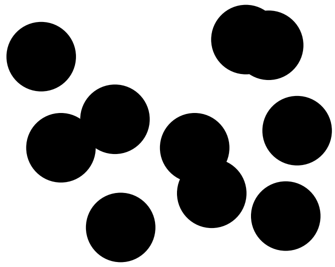
Strict Sense Stationarity

- stronger form of stationarity: $X(t)$ is strict sense stationary if $\forall\, t_1, t_2, \ldots, t_n,\ h$ the n-th order distribution obeys

$$F(x_1, x_2, \ldots, x_n; t_1 + h, t_2 + h, \ldots, t_n + h) \\ = F(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n)$$

  $\Rightarrow$ similarly for densities
- translation over h yields new process with random variables exhibiting the same multivariate distribution for all t
- SS stationarity implies WS stationarity
- for Gaussian process:
  SS stationarity = WS stationarity

Strict Sense Stationarity: example



- Poisson point process in $\mathbb{R}^2$
- $N(D)$ is # points in domain D,
  $P(N(D) = k) = e^{-\lambda v(D)} \frac{(\lambda v(D))^k}{k!}$
  with $v(D)$ the volume of D
- at every Poisson point there is a disc with radius r; S is union of all discs
- process $X(t) = 1$ when $t \in S$, and $= 0$ otherwise
- every realization of S is a partial coverage of $\mathbb{R}^2$, and coverage $\nearrow$ when $\lambda \nearrow$
- $X(t)$ is SS stationary because $v(D(t_j + h)) = v(D(t_j))\ \forall t_j, h$

Outline

Parameter estimation

- for mean and covariance ergodic WSS process: estimate μ_X and k_X by averaging a single sufficiently large observation $\Rightarrow$ ergodicity is an appealing property for WSS processes
- increase precision by enlarging observation interval
- estimates:

$$\hat{\mu}_X = \frac{1}{r}\int_0^r X(t)\,dt \quad \text{and} \quad \hat{k}_X(\tau) = \frac{1}{r-\tau}\int_0^{r-\tau} X_0(t+\tau)X_0(t)\,dt$$

with $X_0(t) = X(t) - \mu_X$

- for discrete case: $\int$ becomes $\sum$, comparable to sample mean and sample covariance

Parameter estimation

- estimators are unbiased:

$$E[\hat{\mu}_X] = \mu_X \quad \text{and} \quad E[\hat{k}_X(\tau)] = k_X(\tau)$$

- for unbiased estimator, MS error equals variance of estimator, and MS convergence is equivalent to convergence of variance to 0
- theorem:

(1) $\hat{\mu}_X$ unbiased, $\mathrm{Var}[\hat{\mu}_X] = \frac{2}{r}\int_0^r \left(1 - \frac{\tau}{r}\right) k_X(\tau)\,d\tau$

(2) $\hat{k}_X(\tau)$ unbiased and in case of Gaussian process,

$$\mathrm{Var}[\hat{k}_X(\tau)] = \frac{2}{r-\tau}\int_0^{r-\tau}\left(1 - \frac{t}{r-\tau}\right)[k_X^2(t) + k_X(t+\tau)k_X(t-\tau)]\,dt$$

Parameter estimation

- hence ergodicity of expected value means

$$\lim_{r\to\infty} \mathrm{Var}[\hat{\mu}_X] = 0$$

- ergodicity of covariance for Gaussian process means

$$\lim_{r\to\infty} \mathrm{Var}[\hat{k}_X(\tau)] = 0$$

- convergence in MS sense
- precision (in MS sense) increases when interval length $\to \infty$

Parameter estimation: example

- random telegraph signal $X(t)$ has covariance function $k_X(\tau) = e^{-2|\tau|}$
- then

$$\mathrm{Var}[\hat{\mu}_X] = \frac{2}{r}\int_0^r \left(1 - \frac{\tau}{r}\right) e^{-2\tau}\, d\tau \le \frac{2}{r}\int_0^r e^{-2\tau}\, d\tau = \frac{1 - e^{-2r}}{r}$$

can be made arbitrarily small by choosing r large enough

Parameter estimation

- problem: formulae for variance of estimators contain covariance which is probably unknown ⇒ use conservative estimates
- for digital computation: discrete approximation of integrals, partitioning of $[0, r]$ in n equal subintervals of length r/n and midpoints $t_1, t_2, \ldots, t_n$:

$$\hat{m}_X = \frac{1}{r}\sum_{i=1}^{n}\frac{r}{n}X(t_i) = \frac{1}{n}\sum_{i=1}^{n}X(t_i) \quad \text{and}$$

$$\hat{k}_X\left(\frac{mr}{n}\right) = \frac{1}{n-m}\sum_{i=1}^{n-m}\left(X(t_{m+i}) - \hat{m}_X\right)\left(X(t_i) - \hat{m}_X\right)$$

Parameter estimation

- here also variances can be calculated:

$$\mathrm{Var}[\hat{m}_X] = \frac{1}{n}\left[k_X(0) + 2\sum_{i=1}^{n-1}\left(1 - \frac{i}{n}\right)k_X\left(\frac{ir}{n}\right)\right]$$

$$\mathrm{Var}\left[\hat{k}_X\left(\frac{mr}{n}\right)\right] = \frac{1}{n-m}\left[k_X^2(0) + k_X^2\left(\frac{mr}{n}\right)\right.$$
$$\left. +2\sum_{i=1}^{n-m-1}\left(1 - \frac{i}{n-m}\right)\left(k_X^2\left(\frac{ir}{n}\right) + k_X\left(\frac{i+m}{n}r\right)k_X\left(\frac{i-m}{n}r\right)\right)\right]$$

- here also Var → 0 when $r \nearrow$ for the mean in case of mean-ergodicity and for the variance in case of covariance-ergodicity + normality

Outline

Linear systems

- problem formulation: Ψ is linear system, and given n-th order distributions of $X(t)$, find the distributions of $\Psi(X)$, eg. find relationship between input covariance $K_X(t, t')$ and output covariance $K_{\Psi(X)}(t, t')$ $\Rightarrow$ *analysis*
- 2nd problem: *synthesis*: find some optimal system w.r.t. some specific goal $\Rightarrow$ constrain the system: linearity $\Rightarrow$ find optimal linear filter

Linearity

- linear operator Ψ: $\Psi(a_1 X_1 + a_2 X_2) = a_1 \Psi(X_1) + a_2 \Psi(X_2)$

$$\begin{tikzcd} X(t) \arrow[r, "\Psi"] \arrow[d, "E"'] & Y(s) \arrow[d, "E"] \\ \mu_X(t) \arrow[r, "?"] & \mu_Y(s) \end{tikzcd}$$

$$\begin{tikzcd} X(t) \arrow[r, "\Psi"] \arrow[d, "\mathrm{Cov}"'] & Y(s) \arrow[d, "\mathrm{Cov}"] \\ K_X(t,t') \arrow[r, "?"] & K_Y(s,s') \end{tikzcd}$$

- is this allowed: $E[\Psi(X)] \stackrel{?}{=} \Psi(E[X]) \Rightarrow$ under some conditions; OK in all practical situations

Linear operators for deterministic functions

- integral operator $y(s) = \int_T g(s,t)x(t)\,dt = \Psi(x)(s)$

  (more generally: $y(s) = \sum_{k=0}^{\infty} \int_T g_k(s,t)x^{(k)}(t)\,dt$)

  (discrete case: use δ functions)

- $x(t)$ itself can be a linear combination: $x(t) = \sum_{k=1}^{n} a_k x_k(t)$

  or more generally $x(t) = \int_U a(u)Q(t,u)\,du$

Linear operators for deterministic functions

- can order of Ψ and $\int$ be interchanged?
- yields $y(s) = \int_U a(u)[\Psi_t(Q(t,u))]\, du$
  with $\Psi_t Q(t,u)(s) = \int_T g(s,t) Q(t,u)\, dt$
  (Ψ_t: apply Ψ to t, for fixed u)
- interchange is allowed under some conditions, which are most often satisfied, certainly for discrete finite cases

Linear operators for deterministic functions: example

- apply to the case $x(t) = \int_{-\infty}^{\infty} x(u)\delta(t-u)\, du$
- yields $y(s) = \int_{-\infty}^{\infty} x(u)\Psi_t\delta(t-u)(s)\, du$ with $\Psi_t\delta(t-u)(s)$ the impulse response
- example: $y(t) = x'(t), s = t$ yields

$$y(t) = -\int_{-\infty}^{\infty} \delta'(t-u)x(u)\, du$$

hence the impulse response of the derivative operator equals $-\delta'(t-u)$; more generally: n-th order derivative operator has impulse response $(-1)^n\delta^{(n)}(t-u)$

Output covariance

- with $Y_0(s) = Y(s) - \mu_Y(s) = \Psi[X(t) - \mu_X(t)](s) = \Psi[X_0(t)](s)$ it can be obtained that $K_Y(s,s') = E[Y_0(s)Y_0(s')] = \cdots = \Psi_t\Psi_{t'}K_X(t,t')$
- theorem: when it holds for $X(t)$ that $\Psi EX = E\Psi X$ then
  1. $\mu_{\Psi X}(s) = \Psi(\mu_X(t))$
  2. $K_{\Psi X}(s,s') = \Psi_t\Psi_{t'}K_X(t,t') = \Psi_{t'}\Psi_t K_X(t,t')$
- if $g(t,u)$ = impulse response of Ψ then

  $Y(t) = \int_{-\infty}^{\infty} g(t,u)X(u)\,du$ and from the theorem:
  1. $\mu_Y(t) = \int_{-\infty}^{\infty} g(t,u)\mu_X(u)\,du$
  2. $K_Y(t,t') = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(t,u)g(t',u')K_X(u,u')\,du\,du'$

Output covariance: example

- if Ψ is the derivative operator, then its impulse response equals $g(t,u) = -\delta'(t-u)$
- this yields

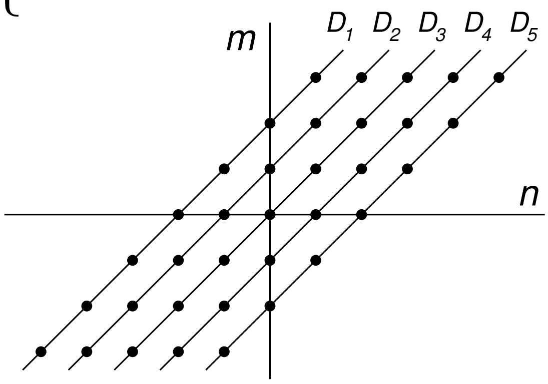
$$
\begin{aligned}
K_Y(t,t') &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta'(t-u)\delta'(t'-u')K_X(u,u')\,du\,du' \\
&= \int_{-\infty}^{\infty} \delta'(t'-u')\left(\int_{-\infty}^{\infty} \delta'(t-u)K_X(u,u')\,du\right)du' \\
&= -\int_{-\infty}^{\infty} \delta'(t'-u')\frac{\partial}{\partial t}K_X(t,u')\,du' \\
&= \frac{\partial}{\partial t'}\frac{\partial}{\partial t}K_X(t,t')
\end{aligned}
$$

Output covariance: example

- moving average filter with coefficients (a_1, a_2, a_3) defines linear operator $Y(n) = a_1X(n) + a_2X(n+1) + a_3X(n+2)$
- assume input $X(n)$ is white noise with variance = 1, then

  input covariance $K_X(n,m) = \begin{cases} 1 & n = m \\ 0 & n \neq m \end{cases}$

- first apply filter to $K_X(n,m)$ w.r.t. n: nonzero results on D_1, D_2, D_3, intermediate result $K_Z(n,m)$ (not a true covariance)



Output covariance: example

- $K_Z(n,m) = a_1K_X(n,m) + a_2K_X(n+1,m) + a_3K_X(n+2,m)$
- on $D_1\,(m = n+2)$:
  $K_Z(n,n+2) = a_1K_X(n,n+2) + a_2K_X(n+1,n+2) + a_3K_X(n+2,n+2) = a_1 0 + a_2 0 + a_3 1 = a_3$
- on $D_2\,(m = n+1)$:
  $K_Z(n,n+1) = a_1K_X(n,n+1) + a_2K_X(n+1,n+1) + a_3K_X(n+2,n+1) = a_1 0 + a_2 1 + a_3 0 = a_2$
- on $D_3\,(m = n)$:
  $K_Z(n,n) = a_1K_X(n,n) + a_2K_X(n+1,n) + a_3K_X(n+2,n) = a_1 1 + a_2 0 + a_3 0 = a_1$
- everywhere else $K_Z(n,m) = 0$

Output covariance: example

- next step: filtering w.r.t. m :
  $K_Y(n,m) = a_1 K_Z(n,m) + a_2 K_Z(n,m+1) + a_3 K_Z(n,m+2)$
- on $D_5\,(n = m+2)$:
  $K_Y(n,m) = a_1 K_Z(m+2,m) + a_2 K_Z(m+2,m+1) +$
  $a_3 K_Z(m+2,m+2) = a_1 0 + a_2 0 + a_3 a_1$
- on $D_4\,(n = m+1)$: $K_Y(n,m) = a_1 0 + a_2 a_1 + a_3 a_2$
- on $D_3\,(n = m)$: $K_Y(n,m) = a_1 a_1 + a_2 a_2 + a_3 a_3$
- on $D_2\,(n = m-1)$: $K_Y(n,m) = a_1 a_2 + a_2 a_3 + a_3 0$
- on $D_1\,(n = m-2)$: $K_Y(n,m) = a_1 a_3 + a_2 0 + a_3 0$
- everywhere else $K_Y(n,m) = 0$
- $\Rightarrow$ WSS process

Output covariance: example

- again via theorem (discrete version) with

$$Y(n) = \sum_{k=-\infty}^{\infty} g(n,k) X(k) \quad \text{and}$$

$$g(n,k) = \begin{cases} a_{k-n+1} & k = n, n+1, n+2 \\ 0 & \text{elsewhere} \end{cases}$$

- $\mu_Y(n) = a_1 \mu_X(n) + a_2 \mu_X(n+1) + a_3 \mu_X(n+2)$
- and $K_Y(n,m) = \sum_{k=n}^{n+2} \sum_{l=m}^{m+2} a_{k-n+1} a_{l-m+1} K_X(k,l) = \ldots$
  gives same result as before (assuming white noise)

Part IV

Power spectral density

Power spectral density

- *power spectral density* of WSS process $X(t)$ is Fourier transform of $r_X(\tau)$:

$$S_X(\omega) = \int_{-\infty}^{\infty} r_X(\tau) e^{-j\omega\tau} d\tau$$

  (on the condition that $r_X(\tau)$ is integrable)
- since $r_X(\tau) = \overline{r}_X(-\tau)$, $S_X(\omega)$ is real valued
- for real valued $X(t)$, $S_X(\omega)$ is an even real function
- inverse transform

$$r_X(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_X(\omega) e^{j\omega\tau} d\omega$$

Power spectral density

- further assume that $\mu_X = 0$, then also

$$\mathrm{Var}[X(t)] = k_X(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_X(\omega)d\omega$$

- theorem: with $\hat{X}_T(\omega) = \int_{-T}^{T} X(t)e^{-j\omega t}dt$ it holds that

$$\lim_{T\to\infty}\frac{1}{2T}\mathrm{Var}[\hat{X}_T(\omega)] = S_X(\omega)$$

- consequence: $S_X(\omega) \geq 0$

Power spectral density

Power spectral density

- this follows from

$$\begin{aligned}
\mathrm{Var}[\hat{X}_T(\omega)] &= E[|\hat{X}_T(\omega)|^2] \\
&= E\left[\int_{-T}^{T} X(t)e^{-j\omega t}dt \int_{-T}^{T}\overline{X(s)}e^{-j\omega s}ds\right] \\
&= \int_{-T}^{T}\int_{-T}^{T} k_X(t-s)e^{-j\omega(t-s)}dt\,ds \\
&= 2T\int_{-2T}^{2T}\left(1-\frac{|\tau|}{2T}\right)k_X(\tau)e^{-j\omega\tau}d\tau
\end{aligned}$$

Power spectral density

- it can be shown that with $\omega_0 < \omega_1$ it holds that

$$\lim_{\omega_1 \to \omega_0} \underset{T \to \infty}{\mathrm{l.i.m.}} \frac{1}{2T(\omega_1 - \omega_0)} \int_{\omega_0}^{\omega_1} |\hat{X}_T(\omega))|^2 d\omega = S_X(\omega_0)$$

(lim and l.i.m. are not interchangeable!)

- everything can be generalized to more dimensions

Power spectral density and linear operators

- with $h(t)$ a deterministic function, define linear operator on WSS input $X(t)$:

$$Y(t) = \int_{-\infty}^{\infty} h(\tau)X(t-\tau)\,d\tau$$

- then $\mu_Y = \mu_X H(0)$ with $H(\omega)$ Fourier transform of $h(t)$
- in the same way $r_Y(\tau) = (r_{YX} \otimes h_0)(\tau)$ with $h_0(v) = \overline{h(-v)}$
- this gives

$$S_Y(\omega) = \overline{H(\omega)} S_{YX}(\omega)$$

Power spectral density and linear operators

- also $r_{YX}(\tau) = (h \otimes r_X)(\tau)$
- and hence $S_{YX}(\omega) = H(\omega)S_X(\omega)$ leads to

$$S_Y(\omega) = |H(\omega)|^2 S_X(\omega)$$

- $\Rightarrow$ interpretation of power spectral density: with $\omega_0 < \omega_1$ set $H(\omega) = 1$ for $\omega \in (\omega_0, \omega_1]$ and $= 0$ elsewhere
- then $S_Y(\omega) = S_X(\omega)$ for $\omega \in (\omega_0, \omega_1]$ and $= 0$ elsewhere
- average power in output $Y(t)$ equals

$$E[|Y(t)|^2] = r_Y(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_Y(\omega)d\omega = \frac{1}{2\pi}\int_{\omega_0}^{\omega_1} S_X(\omega)d\omega$$

- hence integration of $S_X(\omega)$ over frequency band gives power in that band

Part V

Optimal Filtering

Outline

5 Optimal filtering

- Optimal Mean-Square-Error Filters
- Optimal Finite-Observation Linear Filters
- Optimal Infinite-Observation Linear Filters

Optimal MSE filters

- problem: estimate outcome of unobserved random variable based on outcomes of a set of observed variables: estimate values of $Y(s)$ based on observed $X(t)$
- filtering approach: find system that given an input $X(t)$, produces output $\hat{Y}(s)$ that best estimates $Y(s)$
- find function Ψ that minimizes mean square error (MSE): $MSE\langle\Psi\rangle = E[|Y - \Psi(X)|^2]$
- $\Psi(X)$ is the optimal MSE estimator
- often estimator is restricted to some class of functions $\Rightarrow$ not always **the** best estimator

Conditional expectation

- defined earlier:

$$E[Y|x] = \int_{-\infty}^{\infty} y f(y|x)\, dy$$

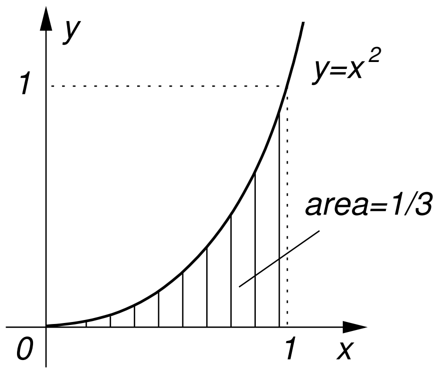
with

$$f(y|x) = \frac{f(x,y)}{f_X(x)}$$

- for given observation x, $E[Y|x]$ is a parameter, but $E[Y|X]$ is a random variable

Conditional expectation: example

- X and Y uniformly distributed over hatched area



$$f_X(x) = 3\int_0^{x^2} dy = 3x^2 \ (0 \le x \le 1)$$

$$f_Y(y) = 3\int_{\sqrt{y}}^{1} dx = 3(1-\sqrt{y}) \ (0 \le y \le 1)$$

- hence $f(y|x) = x^{-2}$ for $0 < x \le 1$ and $0 \le y \le x^2 \Rightarrow$ uniformity of conditional random variable over $[0, x^2] \Rightarrow$ $E[Y|x] = x^2/2$ and $E[Y|X] = X^2/2$
- calculate distribution and density of $E[Y|X]$: $F_{E[Y|X]}(z) = (2z)^{3/2}$ and $f_{E[Y|X]}(z) = 3\sqrt{2z}$ $(0 \le z \le 1/2)$
- this gives $E[E[Y|X]] = 3/10 = E[Y]$!! $\Rightarrow$ $E[Y|X]$ unbiased estimator of $E[Y]$

Conditional expectation

- theorem (chain rule):

$$E[E[Y|X]] = E[Y]$$

  or also

$$E[Y] = \int_{-\infty}^{\infty} E[Y|x] f_X(x)\, dx$$

$$E[Y] = \sum_x E[Y|x] P(X = x)$$

- when X and Y are independent this reduces to $E[Y] = E[Y]$

Chain rule: example

- binary image with random number of disjoint rectangles with random sizes, uniformly distributed rotations
- N rectangles with known μ_N, independent height H and width B, gamma distributed with α_1, β_1 and α_2, β_2
- total area A ?
- for fixed $N = n > 0$:

$$E[A|n] = E\left[\sum_{k=1}^{n} H_k B_k\right] = \sum_{k=1}^{n} E[H_k]E[B_k] = n\alpha_1\beta_1\alpha_2\beta_2$$

- hence also $E[A|N] = N\alpha_1\beta_1\alpha_2\beta_2$
- $E[A] = E[E[A|N]] = \sum_{n=0}^{\infty} n\alpha_1\beta_1\alpha_2\beta_2 P(N = n)$
  $= \mu_N\alpha_1\beta_1\alpha_2\beta_2$

Conditional expectation: variance

- conditional variance $\mathrm{Var}[Y|x] = E[(Y|x - \mu_{Y|x})^2]$; generalize to random variable

$$\mathrm{Var}[Y|X] = E[Y^2|X] - E[Y|X]^2$$

- simple math yields

$$\mathrm{Var}[E[Y|X]] = \mathrm{Var}[Y] - E[\mathrm{Var}[Y|X]]$$

- and hence $\mathrm{Var}[E[Y|X]] \leq \mathrm{Var}[Y]$

Optimal nonlinear filter

- begin with special case: best constant estimate c of Y
- MSE $E[|Y - c|^2]$ minimal if $c = \mu_Y$
- in general: estimate $\hat{Y} = \Psi(X)$; when Ψ is chosen from all possible functions, optimal nonlinear MSE filter is obtained
- without observation of X, $E[Y]$ is best estimator
- with observation of X, $E[Y|x]$ is best estimate and $\Psi(X) = E[Y|X]$ is best estimator
- theorem: $E[Y|X]$ is the optimal MSE estimator for Y based on X:

$$\forall \Psi : E[|Y - E[Y|X]|^2] \leq E[|Y - \Psi(X)|^2]$$

Optimal filter for Gaussian case

- important case: X and Y jointly normal with marginal μ_X, μ_Y, σ_X^2 and σ_Y^2, and with ρ
- to find $E[Y|X]$ we need $f(y|x)$:

$$f(y|x) = \frac{1}{\sqrt{2\pi\sigma_Y^2(1-\rho^2)}} \exp\left\{ -\frac{1}{2}\left[\frac{y - \left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right)}{\sigma_Y\sqrt{1-\rho^2}}\right]^2 \right\}$$

- this is a normal distribution, hence for fixed $X = x$

$$\mu_{Y|x} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X) \qquad \Rightarrow \text{straight line}$$

- therefore best MSE estimator is linear in this case

Optimal filter: example

- normally distributed signal Y with μ_Y and σ_Y^2 is transmitted; received signal $X|y$ is normally distributed with $\mu_{X|y} = y$ and $\sigma_{X|y}^2 = 1$
- estimate transmitted signal based on corrupt received signal and a priori knowledge
- $E[Y|x]$ required;

$$f(y|x) = \cdots = \frac{\exp[A(x)]}{2\pi\sigma_Y f_X(x)} \exp\left[-\frac{1}{2}\frac{\left(y - \frac{\mu_Y + x\sigma_Y^2}{1+\sigma_Y^2}\right)^2}{\frac{\sigma_Y^2}{1+\sigma_Y^2}}\right]$$

- hence $Y|x$ is normally distributed

Optimal filter: example

- density yields

$$E[Y|x] = \frac{1}{1+\sigma_Y^2}\mu_Y + \frac{\sigma_Y^2}{1+\sigma_Y^2}x$$
$$\mathrm{Var}[Y|x] = \frac{\sigma_Y^2}{1+\sigma_Y^2}$$

- best estimator is weighted average of expectation μ_Y and received signal x
- with large σ_Y^2, rely more on x
- when $\sigma_Y^2 \to 0$, take μ_Y

Optimal filter

- in general $E[Y|x]$ is difficult to obtain: density must be known and analytical expression for conditional expectation must be computed
- $\Rightarrow$ confine estimators to restricted class $\mathcal{C}$ of functions: find $\Psi \in \mathcal{C}$ such that

$$E[|Y - \Psi(X)|^2] \leq E[|Y - \xi(X)|^2] \quad \forall \xi \in \mathcal{C}$$

- important class $\mathcal{C}$: linear estimators

Multiple observation variables

- extension: estimate Y based on observation of $X_1, X_2, \ldots, X_n$
- find $\Psi(X_1, X_2, \ldots, X_n)$ such that $\text{MSE}\langle\Psi\rangle = E[|Y - \Psi(X_1, X_2, \ldots, X_n)|^2]$ is minimal
- conditional density $f(y|x_1, x_2, \ldots, x_n)$ needs to be known, expectation $E[Y|X_1, X_2, \ldots, X_n]$ is the optimal estimator

$$E[Y|X_1, X_2, \ldots, X_n] = \int_{-\infty}^{\infty} y f(y|x_1, x_2, \ldots, x_n)\, dy$$

$$E[Y|X_1, X_2, \ldots, X_n] = \sum_{k=0}^{m-1} kP(Y = k|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

Bayesian parametric estimation

- earlier approach: unknown parameter estimated from observed data; Maximum Likelihood Estimation (MLE): find $\hat{\theta}$ that maximizes $f(x_{1\ldots n}|\theta)$
- other approach: treat parameter as random variable with a priori known distribution
- $\Rightarrow$ find the best estimator, given the a priori knowledge and observations
- *Bayesian estimation*: parameters in $f(x; \theta_1, \theta_2, \ldots, \theta_m)$ are random variables $\Theta_1, \Theta_2, \ldots, \Theta_m$ with a priori known distribution $\pi(\theta_1, \theta_2, \ldots, \theta_m)$
- we limit ourselves to 1 parameter

Bayesian parametric estimation

- define *risk function* $E_{\hat{\theta}}[|\Theta - \hat{\theta}|^2]$ (=random variable!)
- compare estimators: define *Bayes risk* of an estimator $\hat{\theta}$:

$$B(\hat{\theta}) = E_{\Theta}[E_{\hat{\theta}}[|\Theta - \hat{\theta}|^2]]$$

- $\hat{\theta}_1$ is better than $\hat{\theta}_2$ if $B(\hat{\theta}_1) < B(\hat{\theta}_2)$
- if appropriate, use restricted class $\mathcal{C}$ of estimators

Bayesian parametric estimation

- minimize Bayes risk

$$\begin{aligned} B(\hat{\theta}) &= \int_{-\infty}^{\infty} E_{\hat{\theta}}[|\theta - \hat{\theta}|^2]\pi(\theta)\, d\theta \\ &= E[|\Theta - \hat{\theta}(X_1, X_2, \ldots, X_n)|^2] \end{aligned}$$

- $\Rightarrow$ Bayes risk is minimized by conditional expectation
- Bayes estimator is $\hat{\theta} = E[\Theta | X_1, X_2, \ldots, X_n]$

Bayesian parametric estimation

- obtaining estimator of Θ based on $X_1, X_2, \ldots, X_n$ requires *a posteriori density*

$$f(\theta|x_1, x_2, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n, \theta)}{f(x_1, x_2, \ldots, x_n)} = \frac{\pi(\theta) \prod_{k=1}^{n} f(x_k|\theta)}{\int_{-\infty}^{\infty} \left(\pi(\theta) \prod_{k=1}^{n} f(x_k|\theta) \right) d\theta}$$

- $\Rightarrow$ a posteriori density expressed as a function of a priori density

Bayesian parametric estimation: example

- X: success or failure with probabilities p and $1 - p$
- p might fluctuate $\Rightarrow$ model as a random variable P with Beta distribution

$$\pi(p) = \begin{cases} \dfrac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)} & 0 \le p \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

- density of X: $f(x|p) = p^x(1-p)^{1-x}$ with $x = 0$ or 1
- this yields $\pi(p) \prod_{k=1}^{n} f(x_k|p) = \dfrac{p^{n\bar{x}+\alpha-1}(1-p)^{n-n\bar{x}+\beta-1}}{B(\alpha, \beta)}$ and

$$\int_{-\infty}^{\infty} \pi(p) \prod_{k=1}^{n} f(x_k|p) dp = \frac{B(n\bar{x}+\alpha, n-n\bar{x}+\beta)}{B(\alpha, \beta)}$$

Bayesian parametric estimation: example

- this leads to the a posteriori density

$$f(p|x_1, x_2, \ldots, x_n) = \frac{p^{n\bar{x}+\alpha-1}(1-p)^{n-n\bar{x}+\beta-1}}{B(n\bar{x}+\alpha, n-n\bar{x}+\beta)}$$

  for $0 \leq p \leq 1$
- this is a Beta distribution with parameters $n\bar{x}+\alpha$ and $n-n\bar{x}+\beta$
- $\Rightarrow$ Bayes estimator is expectation

$$\hat{p} = \frac{n\bar{X}+\alpha}{n+\alpha+\beta}$$

  with $\bar{X}$ the sample mean; if $n \to \infty$ then $\hat{p} \to \bar{X}$

Bayesian parametric estimation

- in the example: convergence of Bayes estimator to maximum likelihood estimator if $n \to \infty$ = typical behaviour of Bayes estimators
- it can be shown that difference between both is small when compared to $n^{-1/2}$
- for small n the difference can be small if the samples are compatible with the a priori distribution, if not, differences can be large

Bayesian parametric estimation: example

- in absence of reliable a priori information: assume uniform a priori distribution
- in previous example this gives $\pi(p) = 1$ for $0 \le p \le 1$ and $\pi(p) = 0$ elsewhere; $f(x|p)$ does not change and a posteriori density becomes

$$f(p|x_1, x_2, \ldots, x_n) = \frac{p^{n\bar{x}}(1-p)^{n-n\bar{x}}}{\int_0^1 p^{n\bar{x}}(1-p)^{n-n\bar{x}}\,dp}$$

- this is a beta distribution with parameters $n\bar{x}+1$ and $n - n\bar{x} + 1$
- Bayes estimator is $\hat{p} = \dfrac{n\bar{X}+1}{n+2}$; here also $\hat{p} \to \bar{X}$ when $n \to \infty$

Conjugate priors

- calculating a posteriori density $f(\theta|x_1, x_2, \ldots, x_n)$ could be impossible analytically (let alone finding its expectation)
- therefore use such $\pi(\theta)$ that product $\pi(\theta)\prod_{k=1}^{n} f(x_k|\theta)$ is a distribution of the same family as $\pi(\theta)$, see previous examples, then the distribution $\pi(\theta)$ is called the *conjugate prior* of the distribution $f(x_1, x_2, \ldots, x_n|\theta)$

likelihood $f(x_{1\ldots n}\|\theta)$	conjugate prior $\pi(\theta)$
binomial	beta
poisson	gamma
normal (σ^2 known)	normal
gamma (α known)	gamma
...	...

Conjugate prior example: gamma likelihood

- assume α known: $f_{X_k}\left(x_k | \frac{1}{\beta}\right) = \dfrac{\beta^{-\alpha}}{\Gamma(\alpha)} x_k^{\alpha-1} e^{-x_k/\beta}$
- conjugate prior for $1/\beta$, with *hyperparameters* α_0, β_0:
  $f_{1/\beta}(1/\beta) = \dfrac{\beta_0^{-\alpha_0}}{\Gamma(\alpha_0)} \left(\dfrac{1}{\beta}\right)^{\alpha_0 - 1} e^{-(1/\beta)/\beta_0}$ (gamma distribution)
- then, with C a constant:
  $\prod_{k=1}^{n} f_{X_k}\left(x_k | \frac{1}{\beta}\right) f_{1/\beta}\left(\frac{1}{\beta}\right) = C\left(\frac{1}{\beta}\right)^{\alpha n + \alpha_0 - 1} e^{\left(-\frac{1}{\beta}\left(\sum_{k=1}^{n} x_k + \frac{1}{\beta_0}\right)\right)}$
- resulting $f\left(\frac{1}{\beta} | x_1 \ldots x_k\right)$ is also gamma distribution with parameters $\alpha' = \alpha n + \alpha_0$ and $1/\beta' = \sum_{k=1}^{n} x_k + 1/\beta_0$

$$E[1/\beta] = \alpha'\beta' = \frac{\alpha n + \alpha_0}{n\bar{x} + \frac{1}{\beta_0}} \text{ and } \hat{\beta} = \frac{n\bar{x} + \frac{1}{\beta_0}}{\alpha n + \alpha_0}$$

Bayesian vs. maximum likelihood estimation

- assume that a set of probability distribution parameters θ best explains dataset $\boldsymbol{x} = x_1 \ldots x_n$
  Bayes' rule: $f_{\theta|\boldsymbol{X}}(\theta|\boldsymbol{x}) = \dfrac{f_{\boldsymbol{X}|\theta}(\boldsymbol{x}|\theta) f_\theta(\theta)}{f_{\boldsymbol{X}}(\boldsymbol{x})}$ or
  posterior $\propto$ *likelihood* $\times$ *prior*
- MLE finds $\hat{\theta}$ that maximizes *likelihood* $f_{\boldsymbol{X}|\theta}(\boldsymbol{x}|\theta)$, treats term $\frac{f_\theta(\theta)}{f_{\boldsymbol{X}}(\boldsymbol{x})}$ as a constant and does not allow to inject prior beliefs about θ
- Bayesian estimation: treat θ as random variable and fully calculate *posterior* $f_{\theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$, then select best estimate $\hat{\theta}$, e.g. expected value; variance can also be calculated.
  But: needs $f_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{-\infty}^{\infty} \left(\pi(\theta) \prod_{k=1}^{n} f(x_k|\theta)\right) d\theta \rightarrow$ difficult, therefore use conjugate priors

Outline

Optimal finite-observation linear filters

- find $a_1, a_2, \ldots, a_n$ and b such that MSE is minimized:

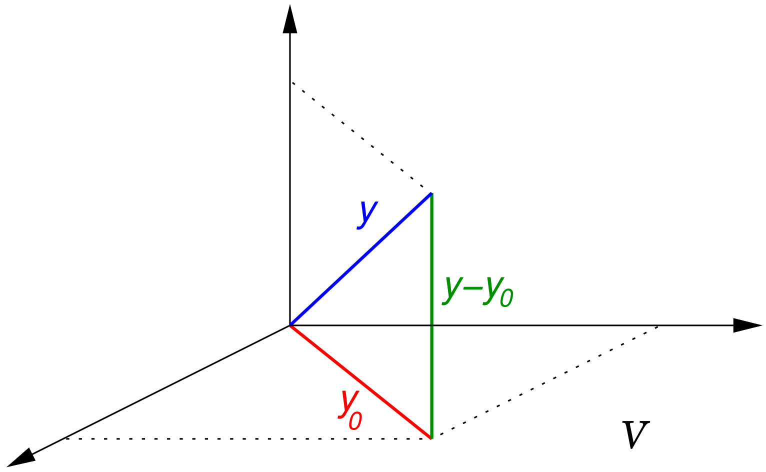
$$\mathrm{MSE}\langle \Psi_A \rangle = E\left[\left| Y - \left(\sum_{k=1}^{n} a_k X_k + b \right) \right|^2 \right]$$

with $\Psi_A(X_1, X_2, \ldots, X_n) = \hat{Y} = \sum_{k=1}^{n} a_k X_k + b$

- if $b = 0 \Rightarrow$ homogeneous filter; nonhomogeneous filter can be treated as special case of homogeneous filter by introducing additional constant variable $X_0 = 1 \Rightarrow$ only homogeneous case is considered here

Optimal finite-observation linear filters

- minimize $E[|Y - \hat{Y}|^2] \Rightarrow |Y - \hat{Y}|$ is smallest when $\hat{Y}$ is projection of Y on subspace S_X spanned by Ψ_A
- y_0 is projection of y on subspace $\mathcal{V}$ if and only if $(y - y_0) \perp v$ or $\langle (y - y_0), v \rangle = 0 \quad \forall v \in \mathcal{V}$



Optimal filtering | Optimal Finite-Observation Linear Filters

Optimal finite-observation linear filters

- inner product of U and V is $E[UV]$
- hence $\hat{Y}$ is the projection of Y on subspace S_X if and only if $\forall V \in S_X$ it holds that $E[(Y - \hat{Y})V] = 0$
- theorem: there exists a set of constants $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_n$ that minimizes MSE$\langle \Psi_A \rangle$ and for any $a_1, a_2, \ldots, a_n$ it holds that

$$E\left[\left(Y - \sum_{k=1}^{n} \hat{a}_k X_k\right)\left(\sum_{j=1}^{n} a_j X_j\right)\right] = 0$$

- $Y - \hat{\mathbf{A}}'X$ orthogonal to $\mathbf{A}'X$ with $\mathbf{A} = (a_1, a_2, \ldots, a_n, b)'$
- if $X_1, X_2, \ldots, X_n$ are linearly independent, then the set $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_n$ is unique

Design of the optimal linear filter

- $\hat{Y}$ is the optimal MSE estimator if and only if

$$\begin{aligned} 0 &= E[(Y - \hat{Y})(a_1X_1 + a_2X_2 + \cdots + a_nX_n)] \\ &= \sum_{k=1}^{n} a_k E[(Y - \hat{Y})X_k] \qquad \forall a_1, a_2, \ldots, a_n \end{aligned}$$

- hence $E[(Y - \hat{Y})X_k] = 0 \quad \forall k$
- solve for $k = 1, 2, \ldots, n$ the equations

$$E\left[\left(Y - \sum_{j=1}^{n} \hat{a}_j X_j\right) X_k\right] = E[YX_k] - \sum_{j=1}^{n} \hat{a}_j E[X_jX_k] = 0$$

Design of the optimal linear filter

- let $R_{kj} = E[X_kX_j] = E[X_jX_k] = R_{jk}$ and $R_k = E[YX_k]$
- solve system of equations

$$\begin{array}{ccccccccc} R_{11}\hat{a}_1 &+& R_{12}\hat{a}_2 &+& \cdots &+& R_{1n}\hat{a}_n &=& R_1 \\ R_{21}\hat{a}_1 &+& R_{22}\hat{a}_2 &+& \cdots &+& R_{2n}\hat{a}_n &=& R_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ R_{n1}\hat{a}_1 &+& R_{n2}\hat{a}_2 &+& \cdots &+& R_{nn}\hat{a}_n &=& R_n \end{array}$$

- in matrix notation $\mathbf{R}\hat{\mathbf{A}} = \mathbf{C}$ with $\mathbf{C} = (R_1, R_2, \ldots, R_n)'$ and $\mathbf{R}$ the matrix composed of R_{jk}
- if $X_1, X_2, \ldots, X_n$ are linearly independent then $\det[\mathbf{R}] \neq 0$ and the solution is found by $\hat{\mathbf{A}} = \mathbf{R}^{-1}\mathbf{C}$

Design of the optimal linear filter

- MSE of the filter:

$$E[|Y - \hat{Y}|^2] = \cdots = E[|Y|^2] - \sum_{k=1}^{n} \hat{a}_k R_k$$

- depends only on second order moments of $X_1, X_2, \ldots, X_n$ and Y
- optimal linear filter is *second order filter*
- nonhomogeneous case: introduce $X_0 = 1$ and use $n + 1$ variables
- in that case solution becomes $\hat{Y} = \hat{a}_0 + \sum_{k=1}^{n} \hat{a}_k X_k$

Design of the optimal linear filter: example

- optimal homogeneous linear estimator of Y in terms of single random variable X_1:
  $\mathbf{R}\hat{\mathbf{A}} = \mathbf{C}$ is reduced to $R_{11}\hat{a}_1 = R_1$ and hence

$$\hat{a}_1 = \frac{R_1}{R_{11}} = \frac{E[X_1 Y]}{E[X_1^2]}$$

- nonhomogeneous case: system of equations

$$\begin{pmatrix} R_{00} & R_{01} \\ R_{10} & R_{11} \end{pmatrix} \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = \begin{pmatrix} R_0 \\ R_1 \end{pmatrix}$$

  with $R_{00} = 1, R_{01} = R_{10} = E[X_1], R_0 = E[Y]$

Design of the optimal linear filter: example

- solution is

$$\begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = \frac{1}{\mathrm{Var}[X_1]} \begin{pmatrix} E[X_1^2] & -E[X_1] \\ -E[X_1] & 1 \end{pmatrix} \begin{pmatrix} E[Y] \\ E[X_1 Y] \end{pmatrix}$$

Design of the optimal linear filter: example

- two observed variables X_1 and X_2, with X_1, X_2, Y independent and uniformly distributed over $[0, 1]$
- then $R_{11} = E[X_1^2] = E[X_2^2] = R_{22} = \frac{1}{3}$,

  $R_{12} = R_{21} = E[X_1 X_2] = \frac{1}{4}$,

  $R_1 = E[X_1 Y] = E[X_2 Y] = R_2 = \frac{1}{4}$
- solving of the system gives $\hat{a}_1 = \hat{a}_2 = \frac{3}{7}$

  hence $\hat{Y} = \frac{3}{7} X_1 + \frac{3}{7} X_2$
- MSE of the estimator

  $E[(Y - \hat{Y})^2] = E[Y^2] - (\hat{a}_1 R_1 + \hat{a}_2 R_2) = \frac{5}{42}$
- biased estimator: $E[\hat{Y}] = 3/7$ while $E[Y] = 1/2$

Design of the optimal linear filter: example

- nonhomogeneous case with $X_0 = 1$ gives

$$\mathbf{R} = \begin{pmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1/3 & 1/4 \\ 1/2 & 1/4 & 1/3 \end{pmatrix} \quad \text{and}$$

$$\begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} = \begin{pmatrix} 7 & -6 & -6 \\ -6 & 12 & 0 \\ -6 & 0 & 12 \end{pmatrix} \begin{pmatrix} 1/2 \\ 1/4 \\ 1/4 \end{pmatrix} = \begin{pmatrix} 1/2 \\ 0 \\ 0 \end{pmatrix}$$

- hence $\hat{Y} = \hat{a}_0 X_0 = 1/2$; could be predicted because Y and X_1, X_2 are independent; hence $E[Y]$ is the best estimator
- homogeneous case has restriction: no constant term
- MSE = 1/12, lower than for homogeneous solution

Optimal filtering | Optimal Finite-Observation Linear Filters

Design of the optimal linear filter: example

- more interesting: correlated variables: X_1, X_2 and Y with $\sigma^2 = 1$, $\mu_{X_1} = \mu_{X_2} = 0$, $\mu_Y = \mu$; $\rho, \rho_{1Y}, \rho_{2Y}$ correlation coefficients between X_1 and X_2, X_1 and Y, X_2 and Y
- then $\mathbf{R} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix}$, nonsingular if $|\rho| \neq 1$
- solving the system gives (for $|\rho| \neq 1$)

$$\hat{a}_0 = \mu$$

$$\hat{a}_1 = \frac{\rho_{1Y} - \rho\rho_{2Y}}{1 - \rho^2}$$

$$\hat{a}_2 = \frac{\rho_{2Y} - \rho\rho_{1Y}}{1 - \rho^2}$$

Design of the optimal linear filter: example

- some special cases
- if $\rho_{1Y} = \rho_{2Y} = 0$, then $\hat{a}_1 = \hat{a}_2 = 0$ and $\hat{Y} = \mu$
- if $\rho_{1Y} = 0$ and $\rho_{2Y} = 1$, then $\exists c > 0, d : Y = cX_2 + d$
  - but $\mathrm{Var}[Y] = c^2\,\mathrm{Var}[X_2]$, hence $c = 1$
  - also $E[Y] = E[X_2] + d$ hence $d = \mu$
  - hence $Y = X_2 + \mu$
  - this also means $\rho = 0$
  - solving leads to estimator $\hat{Y} = X_2 + \mu$
- in general: MSE=$1 - \dfrac{\rho_{1Y}^2 + \rho_{2Y}^2 - 2\rho\rho_{1Y}\rho_{2Y}}{1-\rho^2}$
  and for $\rho_{1Y} = \rho_{2Y} = 0$, MSE=1=Var$[Y]$
  for $\rho_{1Y} = 0, \rho_{2Y} = 1$ MSE=0

Optimal filter for the Gaussian case

- it was shown earlier that for X and Y jointly normal, optimal filter is linear
- extend to $X_1, X_2, \ldots, X_n$: if $X_1, X_2, \ldots, X_n, Y$ are jointly normal, then the optimal linear filter is also the optimal MSE filter:

$$E[Y|\mathbf{X}] = \sum_{k=1}^{n} \hat{a}_k E[X_k]$$

  with $\mathbf{X}=(X_1, X_2, \ldots, X_n)'$
- this is one of the reasons why often a joint Gaussian distribution is assumed

Role of wide sense stationarity

- optimal linear estimator is applied to sliding window
- filter is generally dependent on the position z of the window: **R** and **C** depend on z:
  $R_{kj}(z) = E[X_k X_j] = E[S(z + w_k)S(z + w_j)]$ and
  $R_k(z) = E[YX_k] = E[YS(z + w_k)]$
  with $W = \{w_1, w_2, \ldots, w_n\}$ the window and
  $S(z)$ the observation at location z
- hence optimal filter Ψ_{op} is translation variant
- in the case of WS stationarity $\Rightarrow$ translation invariance and optimal filter is identical for all z

Role of wide sense stationarity

- in previous examples **R** and **C** were obtained from process model
- in practice **R** and **C** are usually estimated: with $X_{z,-m}, X_{z,-m+1}, \ldots, X_{z,0}, \ldots, X_{z,m-1}, X_{z,m}$ the values of $X(j)$ translated to z:

$$\hat{\mathbf{R}}_z = \begin{pmatrix} X_{z,-m}X_{z,-m} & X_{z,-m}X_{z,-m+1} & \cdots & X_{z,-m}X_{z,m} \\ X_{z,-m+1}X_{z,-m} & X_{z,-m+1}X_{z,-m+1} & \cdots & X_{z,-m+1}X_{z,m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{z,m}X_{z,-m} & X_{z,m}X_{z,-m+1} & \cdots & X_{z,m}X_{z,m} \end{pmatrix}$$

Role of wide sense stationarity

$$\hat{\mathbf{C}}_z = \begin{pmatrix} X_{z,-m} Y_z \\ X_{z,-m+1} Y_z \\ \vdots \\ X_{z,m} Y_z \end{pmatrix}$$

- and then **R** and **C** are estimated as follows

$$\hat{\mathbf{R}} = \frac{1}{N} \sum_{z \in A} \hat{\mathbf{R}}_z \qquad \text{and} \qquad \hat{\mathbf{C}} = \frac{1}{N} \sum_{z \in A} \hat{\mathbf{C}}_z$$

  with A the domain of interest containing N points

Noise filtering

- *signal plus noise model:* observed $X_1, X_2, \ldots, X_n$ modelled as random variables $U_1, U_2, \ldots, U_n$ that are additively corrupted by noise, represented as random variables $N_1, N_2, \ldots, N_n$:

$$X_k = U_k + N_k \qquad k = 1, \ldots, n$$

- often it is assumed that the noise is uncorrelated with the data and that it is white

Noise filtering: example

- simple case: homogeneous linear estimate of Y given corrupted observation of Y itself and of other variable X
- $\mu_Y = \mu_X = 0$, $\sigma_X^2 = \sigma_Y^2 = 1$, also ρ given
- observations X_1, X_2:

$$\begin{aligned} X_1 &= Y + N_1 \\ X_2 &= X + N_2 \end{aligned}$$

with $\mu_{N_1} = \mu_{N_2} = 0$, $\mathrm{Var}[N_1] = \mathrm{Var}[N_2] = \sigma^2$, N_1, N_2 uncorrelated with X and Y and also mutually uncorrelated

Optimal filtering | Optimal Finite-Observation Linear Filters

Noise filtering: example

- simple math leads to

$$\mathbf{R} = \begin{pmatrix} 1+\sigma^2 & \rho \\ \rho & 1+\sigma^2 \end{pmatrix} \quad \text{and} \quad \mathbf{C} = \begin{pmatrix} 1 \\ \rho \end{pmatrix}$$

- leads to solution

$$\hat{Y} = \frac{1+\sigma^2-\rho^2}{(1+\sigma^2)^2-\rho^2} X_1 + \frac{\rho\sigma^2}{(1+\sigma^2)^2-\rho^2} X_2$$

- without noise ($\sigma^2 = 0$) this becomes $\hat{Y} = X_1$
- with extreme noise ($\sigma^2 \to \infty$) this becomes $\hat{Y} = 0 = E[Y]$

Noise filtering: example

- if $\rho = 0$ then $\hat{Y} = \frac{1}{1+\sigma^2} X_1$
- compute MSE:

$$E[(Y-\hat{Y})^2] = E[Y^2] - (\hat{a}_1 R_1 + \hat{a}_2 R_2)$$
$$= 1 - \frac{1+\sigma^2-\rho^2+\rho^2\sigma^2}{(1+\sigma^2)^2-\rho^2}$$

- if $\sigma^2 = 0$, then MSE=0
- for fixed ρ: $\lim_{\sigma^2 \to \infty} E[(Y-\hat{Y})^2] = 1$

Noise filtering: example

- sampling of random telegraph signal gives discrete process
- this process is corrupted by noise; assume WS stationarity, hence filtering with sliding window is possible
- windows contain $Y_{-m}, Y_{-m+1}, \ldots, Y_{-1}, Y_0, Y_1, \ldots, Y_{m-1}, Y_m$, similarly for N and X
- observations $X_j = Y_j + N_j$ for $j = -m, \ldots, m$
- Y_0 needs to be estimated
- $\mu_Y = \mu_N = 0$, also $\sigma_Y^2 = 1$ and $\sigma_N^2 = \sigma^2$

Noise filtering: example

- this yields $R_{jj} = E[X_j^2] = E[Y_j^2] + E[N_j^2] = 1 + \sigma^2$ and also $R_{ij} = E[X_i X_j] = E[Y_i Y_j] = e^{-2\lambda|i-j|}$ (for $i \neq j$)
- also $R_0 = 1$ and $R_j = E[X_j Y_0] = E[Y_j Y_0] = e^{-2\lambda|j|}$ $(j \neq 0)$
- e.g. in 5 points window:

$$\mathbf{R} = \begin{pmatrix} 1+\sigma^2 & e^{-2\lambda} & e^{-4\lambda} & e^{-6\lambda} & e^{-8\lambda} \\ e^{-2\lambda} & 1+\sigma^2 & e^{-2\lambda} & e^{-4\lambda} & e^{-6\lambda} \\ e^{-4\lambda} & e^{-2\lambda} & 1+\sigma^2 & e^{-2\lambda} & e^{-4\lambda} \\ e^{-6\lambda} & e^{-4\lambda} & e^{-2\lambda} & 1+\sigma^2 & e^{-2\lambda} \\ e^{-8\lambda} & e^{-6\lambda} & e^{-4\lambda} & e^{-2\lambda} & 1+\sigma^2 \end{pmatrix} \quad \mathbf{C} = \begin{pmatrix} e^{-4\lambda} \\ e^{-2\lambda} \\ 1 \\ e^{-2\lambda} \\ e^{-4\lambda} \end{pmatrix}$$

Optimal filtering | Optimal Finite-Observation Linear Filters

Noise filtering: example

- numerical example: $\lambda = 0.02145$ and $\sigma^2 = 0.2$ give

$$\hat{\mathbf{A}} = \mathbf{R}^{-1}\mathbf{C} = \begin{pmatrix} 0.1290 \\ 0.1899 \\ 0.3328 \\ 0.1899 \\ 0.1290 \end{pmatrix}$$

- optimal linear filter is weighted average, suppresses additive noise but blurs jumps (edges) in the signal

Edge detection

- many approaches possible, e.g.:
  - linear filter: gradient, e.g. in 2 dimensions with impulse responses

$$\begin{pmatrix} -1 & 0 & 1 \\ -\lambda & 0 & \lambda \\ -1 & 0 & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & \lambda & 1 \\ 0 & 0 & 0 \\ -1 & -\lambda & -1 \end{pmatrix}$$

  - *morphological* gradient: (max – min) within some window
  - ...
- here: stochastic approach, find best estimator given some model

Edge detection: example

- edge signal $Y(k)$, to be estimated from observed signal $X(k)$
- observed signal contains uncorrelated additive noise $N(k)$ (variance σ^2): $X(k) = U(k) + N(k)$ with $U(k)$ the original non-corrupted signal where edges need to be found
- define $Y(k)$ as a binomial random variable with $P(Y(k) = 1) = p$ and $P(Y(k) = 0) = q$ with $p, q > 0; p + q = 1$
- define $U(k)$ with $P(U(0) = 1) = 1/2$ and $P(U(0) = -1) = 1/2$, as follows: $U(k) = U(k-1)$ if $Y(k) = 0$ and $U(k) = -U(k-1)$ when $Y(k) = 1$
- $U(k)$ behaves more or less like the random telegraph signal, except that it is discrete and based on the binomial distribution

Edge detection: example

- use sliding window with three points X_{-1}, X_0, X_1
- then $E[X_{-1}^2] = E[X_0^2] = E[X_1^2] = 1 + \sigma^2$; also
  $E[X_0 X_1] = E[U_0 U_1] = \cdots = q - p$ and
  $E[X_{-1} X_1] = \cdots = (q-p)^2$
- finally leads to (with $a = q - p$):

$$\mathbf{R} = \begin{pmatrix} 1+\sigma^2 & a & a^2 \\ a & 1+\sigma^2 & a \\ a^2 & a & 1+\sigma^2 \end{pmatrix}$$

- $E[Y_0 X_j] = 0$, hence $\mathbf{C} = (0,0,0)'$ and $\hat{\mathbf{A}} = \mathbf{R}^{-1}\mathbf{C} = \mathbf{0}$
- explanation: linear filter should give edge indication at positive as well as negative jumps

Edge detection: example

- solution: find only positive jumps in $U(k)$, define process $Z(k)$ (similar to $Y(k)$)
- then $E[Z_0 U_{-1}] = -p/2, \ldots$ and finally

$$\mathbf{C} = \frac{p}{2} \begin{pmatrix} -1 \\ 1 \\ a \end{pmatrix}$$

  and solution is given by $\mathbf{R}^{-1}\mathbf{C}$

Edge detection: example

- special case: no noise: leads to

$$\hat{\mathbf{A}} = \frac{1}{1-a^2}\begin{pmatrix} -pq \\ pq \\ 0 \end{pmatrix} \quad \text{and} \quad \psi_{\text{opt}}(\mathbf{x}) = \frac{x_0 - x_{-1}}{4}$$

- for signal $s = (\dots, 1, 1, 1, 1, -1, -1, -1, -1, 1, 1, 1, \dots)$, $\psi_{\text{opt}}(s) = (\dots, 0, 0, 0, 0, -1/2, 0, 0, 0, 1/2, 0, 0, 0, \dots)$
- hence output -1/2 and 1/2 at negative, resp. positive jumps $\Rightarrow$ use threshold on these output values
- construct similar filter for optimal detection of negative jumps, or use absolute value to detect both at the same time

Outline

Optimal infinite-observation linear filters

- estimators until now based on finite observation; solution was found by projecting random variable on subspace spanned by observations
- sometimes filter needed based on observation of entire discrete grid or of continuous random signal $\Rightarrow$ infinite number of observations
- orthogonal projection can also be applied to subspaces with infinite dimension; solution is only guaranteed when subspace is closed

Optimal infinite-observation linear filters

- theorem: if $\mathcal{S}$ is subspace of finite second moment random variables and Y has finite second moment, then $\hat{Y}$ is the optimal MSE estimator of Y, lying in $\mathcal{S}$, if and only if $E[(Y-\hat{Y})U] = 0 \; \forall U \in \mathcal{S}$. If the estimator exists, then it is unique.
- practical problems:
  - procedure to find solution for finite observation is not applicable here because of infinite dimensionality
  - solution may not exist

Optimal infinite-observation linear filters

- we limit ourselves to linear integral operators: for fixed s, estimate the value $Y(s)$ by observation of $X(t)$ over some portion T of its domain
- find estimator of the form

$$W(s) = \int_T g(s,t)X(t)\,dt$$

- minimize MSE:

$$MSE\langle W(s)\rangle = E\left[\left|Y(s) - \int_T g(s,t)X(t)\,dt\right|^2\right]$$

Optimal filtering | Optimal Infinite-Observation Linear Filters

Optimal infinite-observation linear filters

- the optimal estimator $\hat{Y}(s) = \int_T \hat{g}(s,t)X(t)\,dt$ satisfies the relation

$$E[(Y(s) - \hat{Y}(s))W(s)] = 0 \qquad \text{for all } W(s)$$

- are conditions satisfied?
  - do $W(s)$ form a linearly closed subspace? $\Rightarrow$ if $g(s,t)$ belongs to linearly closed class $\mathcal{G}$:
    $\forall g_1(s,t), g_2(s,t) \in \mathcal{G}, c_1, c_2$:
    $g(s,t) = c_1 g_1(s,t) + c_2 g_2(s,t) \in \mathcal{G}$
  - do all $W(s)$ have finite second moment? $\Rightarrow$ assuming $K_X(t,t')$ is square-integrable: yes, if $g(s,t)$ are square-integrable

Optimal infinite-observation linear filters

- expanding $E[(Y(s) - \hat{Y}(s))W(s)] = 0$ yields: $\hat{g}(s,t)$ is optimal MSE estimator if and only if

$$R_{YX}(s,t) = \int_T \hat{g}(s,u) R_X(u,t)\, du \quad \forall t \in T$$

  = *Wiener-Hopf* equation, compare to $\mathbf{R}\hat{\mathbf{A}} = \mathbf{C}$
- theorem does not assert existence of solution
- can be applied to discrete signals $\Rightarrow$ integral replaced by sum
- can be extended to higher dimensions
- also **MSE** $= \mathrm{Var}[Y(s)] - \int_T \hat{g}(s,u) R_{YX}(s,u)\, du$ (if $\mu_Y = 0$)

Wiener filter

- if $X(t)$ and $Y(s)$ are WS stationary and $X(t)$ is observed over all time:

$$r_{YX}(\xi) = \int_{-\infty}^{\infty} \hat{g}(\xi - \tau) r_X(\tau)\, d\tau$$

- Fourier transform yields

$$\hat{G}(\omega) = \frac{S_{YX}(\omega)}{S_X(\omega)}$$

  this is the *Wiener filter*

Wiener filter: example

- in 2 dimensions for digital image: Wiener-Hopf equation becomes

$$r_{YX}(m,n) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \hat{g}(m-k, n-l) r_X(k,l)$$

- suppose $Y(n,m)$ is corrupted by linear operation (e.g. motion blur) and additive noise:

$$X(m,n) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} b(m-k, n-l) Y(k,l) + N(n,m)$$

- then $S_{YX}(\omega_1,\omega_2) = \bar{B}(\omega_1,\omega_2) S_Y(\omega_1,\omega_2)$ and $S_X(\omega_1,\omega_2) = |B(\omega_1,\omega_2)|^2 S_Y(\omega_1,\omega_2) + S_N(\omega_1,\omega_2)$

Wiener filter: example

- this yields the filter

$$\hat{G}(\omega_1,\omega_2) = \frac{\bar{B}(\omega_1,\omega_2) S_Y(\omega_1,\omega_2)}{|B(\omega_1,\omega_2)|^2 S_Y(\omega_1,\omega_2) + S_N(\omega_1,\omega_2)}$$

- in absence of noise, this becomes the inverse filter:

$$\hat{G}(\omega_1,\omega_2) = B(\omega_1,\omega_2)^{-1}$$

- if only noise is present:

$$\hat{G}(\omega_1,\omega_2) = \frac{S_Y(\omega_1,\omega_2)}{S_Y(\omega_1,\omega_2) + S_N(\omega_1,\omega_2)}$$

Part VI

Kalman Filter

Kalman filter: context

Optimal filters discussed earlier:

- linear finite-observation filters:

$$E\left[\left(Y - \sum_{k=1}^{n} \hat{a}_k X_k\right)\left(\sum_{j=1}^{n} a_j X_j\right)\right] = 0 \quad \forall\, a_j$$

$\Rightarrow$ find $\hat{a}_k \Rightarrow \hat{\mathbf{A}} = \mathbf{R}^{-1}\mathbf{C}$

- Wiener filter: infinite observation:

$$E[(Y(s) - \int_T \hat{g}(s,t)X(t)\,dt)\int_T g(s,t)X(t)\,dt] = 0 \quad \forall\, g(s,t)$$

$$\Rightarrow R_{YX}(s,t) = \int_T \hat{g}(s,u)R_X(u,t)\,du$$

WSS $\Rightarrow$ convolution $\Rightarrow$ solve in frequency domain

$$\hat{G}(\omega) = \frac{S_{YX}(\omega)}{S_X(\omega)}$$

Kalman filter: context

Kalman:

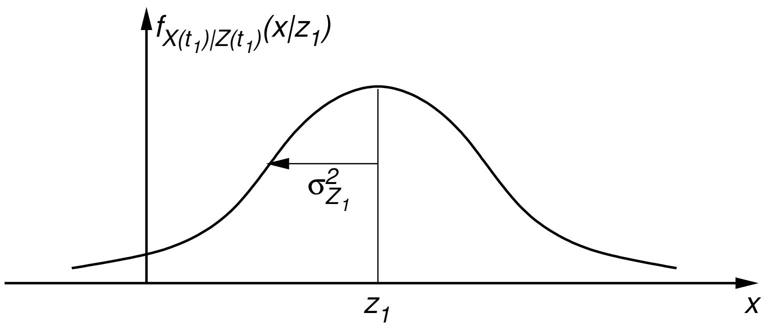
- observations corrupted by white Gaussian noise
- linear (in its basic form)
- optimal filter
- WSS not required
- no matrix inversion required
- recursive solution: compare to sample mean:

$$\hat{Y}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

recursive implementation: $\hat{Y}_{n+1} = \frac{n}{n+1}\hat{Y}_n + \frac{1}{n+1}X_{n+1}$

Kalman filter: simple example

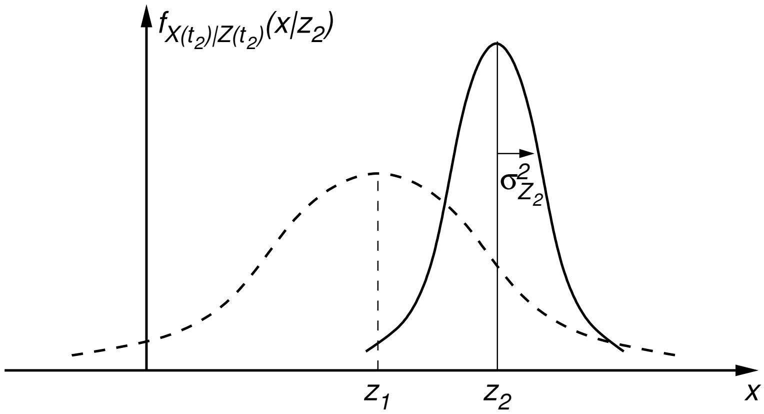
- lost at sea: measure position using stars: at t_1 we are at location z_1 (1D simplification).
- measurement error $\Rightarrow \sigma^2_{Z_1}$



- best estimate $\hat{x}(t_1) = z_1$ with variance $\sigma^2_X(t_1) = \sigma^2_{Z_1}$

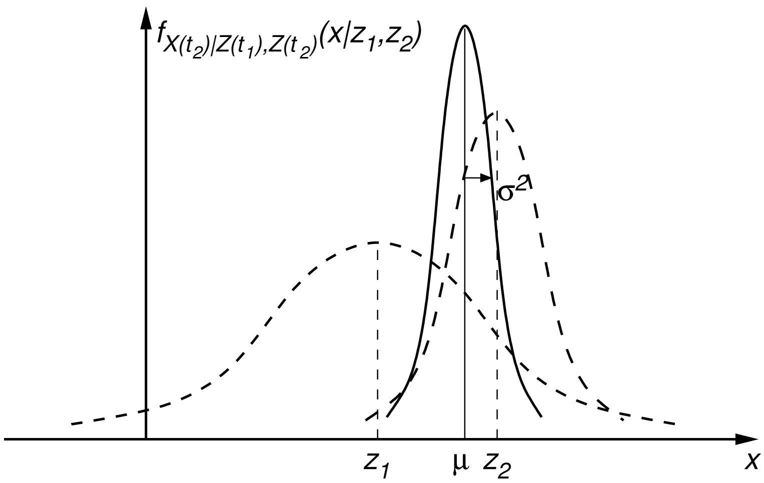
Kalman filter: simple example

- immediately after that at t_2 (real position unchanged): second measurement by more experienced navigator $\Rightarrow z_2, \sigma_{Z_2}^2$



- combination of both measurements?

Kalman filter: simple example



- $\mu = [\sigma_{Z_2}^2/(\sigma_{Z_1}^2 + \sigma_{Z_2}^2)]z_1 + [\sigma_{Z_1}^2/(\sigma_{Z_1}^2 + \sigma_{Z_2}^2)]z_2$
- $1/\sigma^2 = (1/\sigma_{Z_1}^2) + (1/\sigma_{Z_2}^2)$
- best estimate $\hat{x}(t_2) = \mu$

Kalman filter: simple example

- rewrite equation as follows

$$\hat{x}(t_2) = \mu = \hat{x}(t_1) + K(t_2)[z_2 - \hat{x}(t_1)]$$

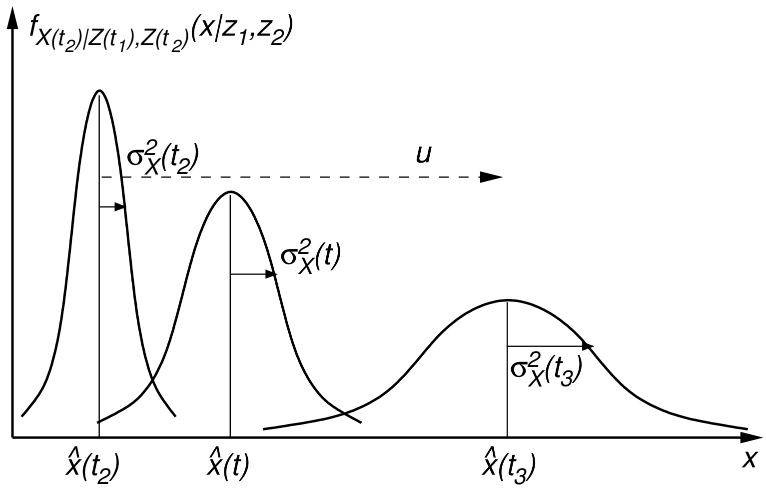
with

$$K(t_2) = \sigma_{Z_1}^2 / (\sigma_{Z_1}^2 + \sigma_{Z_2}^2)$$

- significance: optimal estimate at t_2 equal to best estimate just before t_2, plus correction term based on difference between new measurement and old estimate, using optimal weighting coefficient $K(t_2)$
- also: $\sigma_X^2(t_2) = \sigma_X^2(t_1) - K(t_2)\sigma_X^2(t_1)$

Kalman filter: simple example

- introduce dynamics: sail with some speed until next measurement is done at t_3
- speed $dX/dt = u + W$ with u the nominal speed and W white Gaussian noise with $\mu_W = 0$ and known variance σ_W^2



Kalman filter: simple example

- just before new measurement z_3 at t_3:
  $\hat{x}(t_3^-) = \hat{x}(t_2) + u[t_3 - t_2]$ and
  $\sigma_X^2(t_3^-) = \sigma_X^2(t_2) + \sigma_W^2[t_3 - t_2]$
- immediately after new measurement:
  $\hat{x}(t_3) = \hat{x}(t_3^-) + K(t_3)[z_3 - \hat{x}(t_3^-)]$ and
  $\sigma_X^2(t_3) = \sigma_X^2(t_3^-) - K(t_3)\sigma_X^2(t_3^-)$
  with $K(t_3) = \sigma_X^2(t_3^-)/[\sigma_X^2(t_3^-) + \sigma_{Z_3}^2]$
- interpretation of $K(t_3)$: if measurement accuracy is low ($\sigma_{Z_3}^2$ large) then $K(t_3)$ is small and new measurement has only small influence on result, and vice-versa

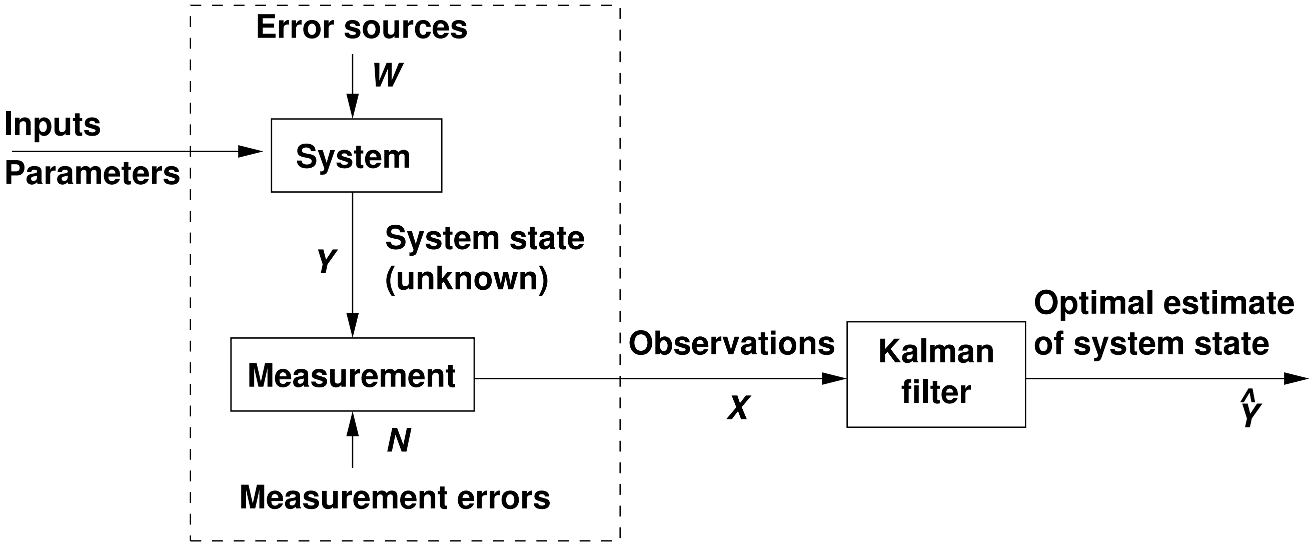
Kalman filter: basic version

- model for data: $Y_k = a_{k-1}Y_{k-1} + W_{k-1}$ with
  $E[Y_0] = 0, \sigma_0^2, E[W_k] = 0, E[W_k W_l] = Q_k \delta_{kl}$
- model for observation: $X_k = c_k Y_k + N_k$ with
  $E[N_k] = 0, E[N_k N_l] = R_k \delta_{kl}$
- find linear estimator as follows:

$$\hat{Y}_k = \sum_{j=1}^{k} h_j(k) X_{k-j} \quad \text{and also} \quad \hat{Y}_{k+1} = \sum_{j=1}^{k+1} h_j(k+1) X_{k+1-j}$$

- transform this equation to recursive form where all estimates and observations from the past are combined into $\hat{Y}_k$

Kalman filter: basic version



Kalman filter: basic version

- MMSE $\Rightarrow$ projection, as before, yields

$$R_{YX}(k,l) = \sum_{j=1}^{k} h_j(k) R_X(k-j,l) \quad l = 0,1,\ldots,k-1$$

$$R_{YX}(k+1,l) = \sum_{j=1}^{k+1} h_j(k+1) R_X(k+1-j,l) \quad l = 0,1,\ldots,k$$

- also

$$R_{YX}(k+1,l) = a_k R_{YX}(k,l) \quad \text{and} \quad R_{YX}(k,l) = R_X(k,l)/c_k$$

- from this, recursive form can be obtained:

$$\hat{Y}_{k+1} = a_k \hat{Y}_k + h_1(k+1)(X_k - c_k \hat{Y}_k)$$

Kalman filter: basic version

- $\hat{Y}_{k+1} = a_k \hat{Y}_k + h_1(k+1)(X_k - c_k \hat{Y}_k) = a_k \hat{Y}_k + K(k) I_k$
- with $K(k)$ the *Kalman gain* (to be determined) and $I_k = X_k - c_k \hat{Y}_k$ the *innovations*
- determine $K(k)$ by minimising $E[e_{k+1}^2]$ with $e_k = Y_k - \hat{Y}_k, k = 1, 2, \ldots, e_0 = 0$, this yields $E[e_{k+1}^2] = [a_k - K(k)c_k]^2 E[e_k^2] + Q_k + K^2(k) R_k$ and
$$K(k) = \frac{a_k c_k E[e_k^2]}{R_k + c_k^2 E[e_k^2]}$$
- if $R_k = 0$ (noise-free measurement), then $K(k) = a_k / c_k$ and $\hat{Y}_{k+1} = a_k X_k$
- if $E[e_k^2] = 0$ (a priori error estimate = 0) then $K(k) = 0$ and $\hat{Y}_{k+1} = a_k \hat{Y}_k$

Kalman filter: algorithm

Kalman Filter

Initialization: $\hat{Y}_0 = 0; E[e_0^2] = \sigma_0$
For $k = 0, 1, 2, \ldots$ **do**

 Set $K(k) = \dfrac{a_k c_k E[e_k^2]}{R_k + c_k^2 E[e_k^2]}$
 Set $E[e_{k+1}^2] = [a_k - K(k)c_k]^2 E[e_k^2] + Q_k + K^2(k) R_k$
 Output $\hat{Y}_{k+1} = a_k \hat{Y}_k + K(k)[X_k - c_k \hat{Y}_k]$

end

Kalman filter: extensions

- higher dimensions: all equations become vector/matrix equations
- nonlinear: *Extended Kalman Filter*: linearize around current mean and covariance, compare to Taylor series
- demo: `www.cs.unc.edu/~welch/kalman/kftool/index.html`