Probability and Random Variables; and Classical Estimation Theory

UDRC Summer School, 26th June 2017

Dr James R. Hopgood

James.Hopgood@ed.ac.uk

Room 2.05

Alexander Graham Bell Building

The King's Buildings

Institute for Digital Communications

School of Engineering

College of Science and Engineering

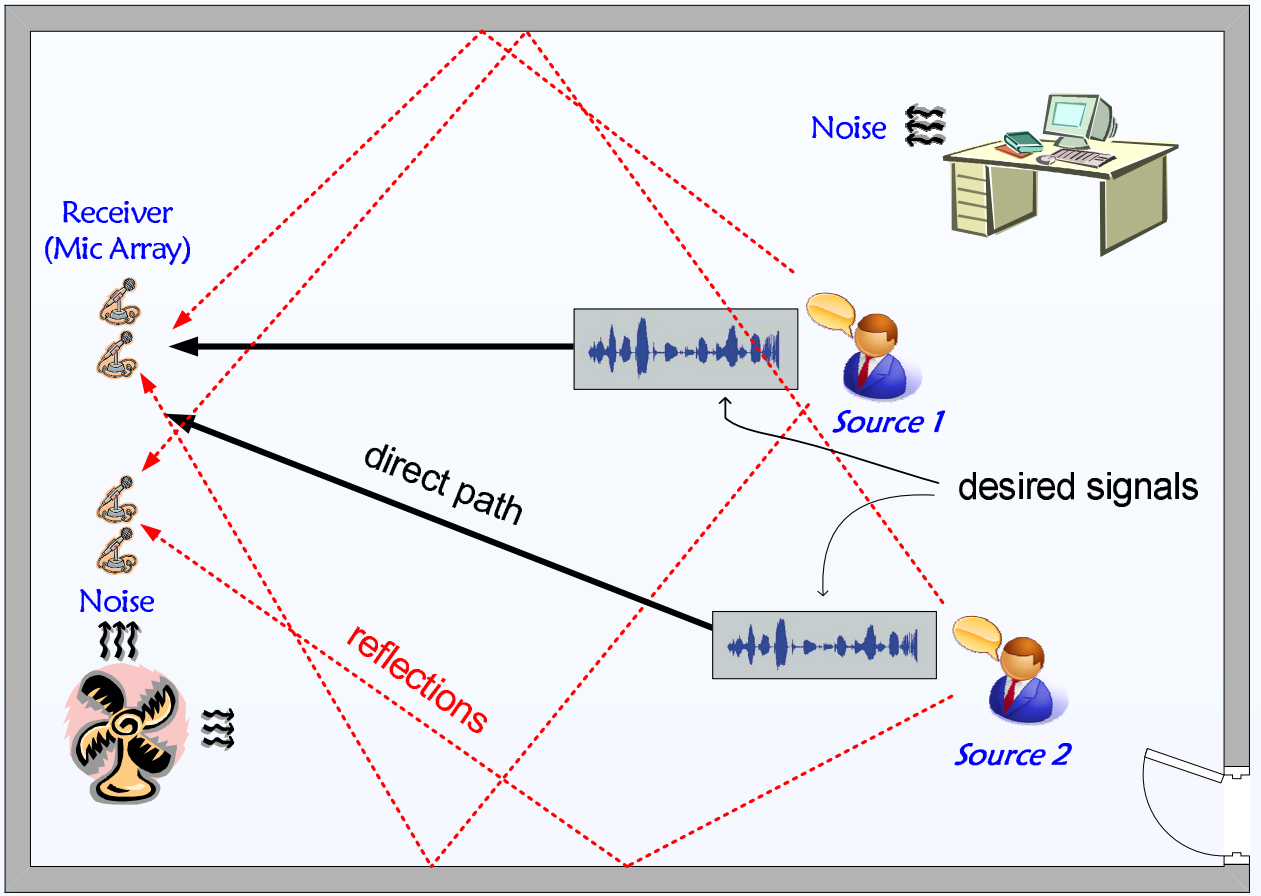
University of Edinburgh

Blank Page

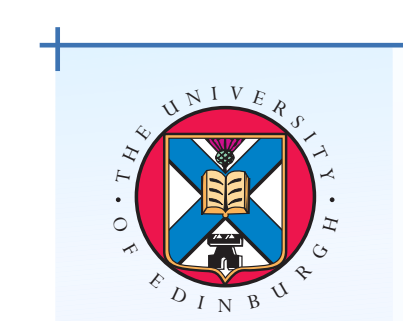
This slide is intentionally left blank.

Handout 1
Aims and Objectives

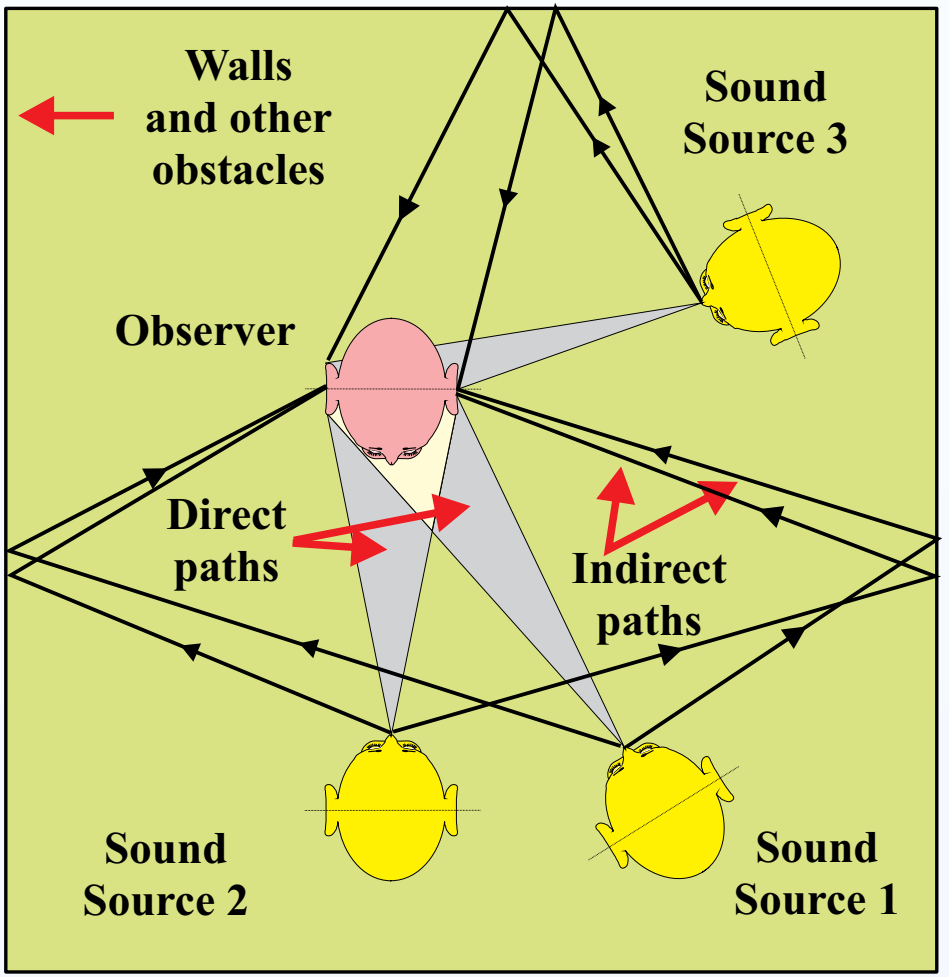
Obtaining the Latest Handouts



Source localisation and blind source separation (BSS). An example of topics using statistical signal processing.



Obtaining the Latest Handouts



Humans turn their head in the direction of interest in order to reduce inteference from other directions; *joint detection, localisation, and enhancement*. An application of probability and estimation theory, and statistical signal processing.

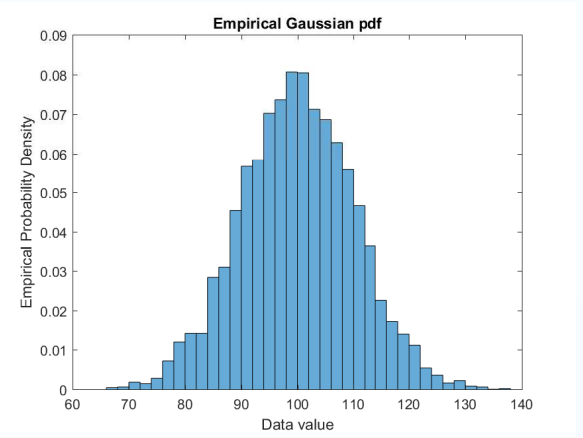
Obtaining the Latest Handouts

- This research tutorial is intended to cover a wide range of aspects which cover the fundamentals of statistical signal processing.
- This tutorial is being continually updated, and feedback is welcomed. The documents published on the USB stick may differ to the slides presented on the day.
- The latest version of this document can be obtained from the author, Dr James R. Hopgood, by emailing him at: at:

  `mailto:james.hopgood@ed.ac.uk`

  (Update: The notes are no longer online due to the desire to maintain copyright control on the document.)
- Extended thanks are given to the many MSc students over the past 12 years who have helped proof-read and improve these documents.

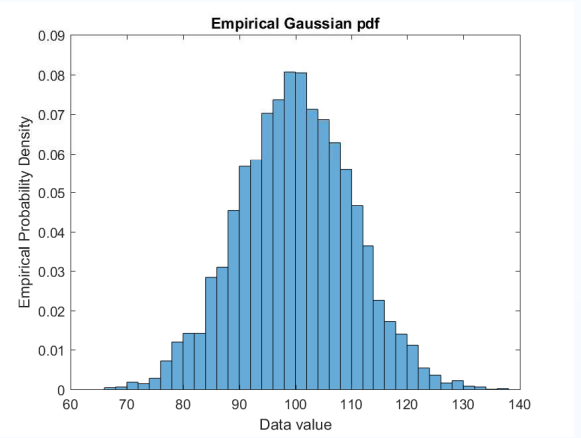
Module Abstract



This topic is covered in two related lecture modules:

1. *Probability, Random Variables, and Estimation Theory*, and
2. *Statistical Signal Processing*,

Module Abstract

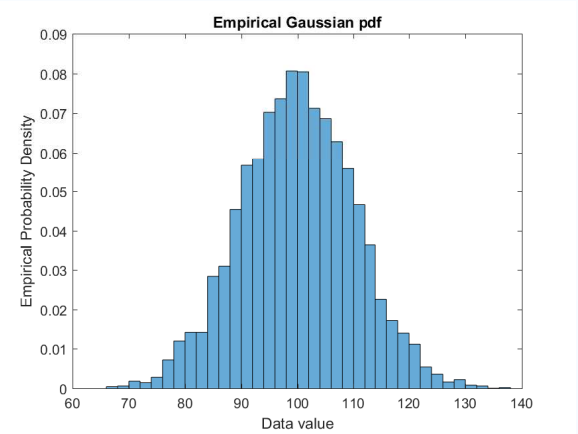


This topic is covered in two related lecture modules:

1. *Probability, Random Variables, and Estimation Theory*, and
2. *Statistical Signal Processing*,

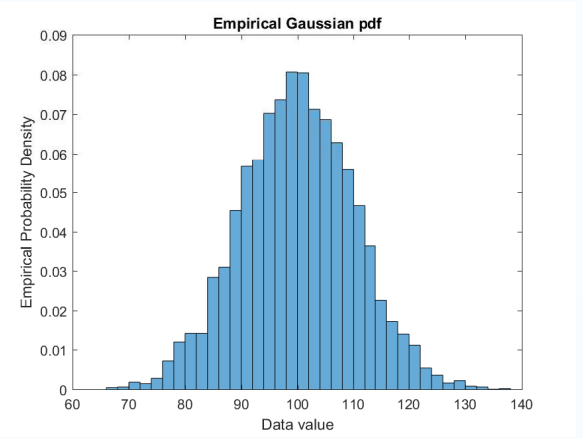
- **Random signals** are extensively used in algorithms, and are:
  - constructively used to model real-world processes;
  - described using *probability and statistics*.

Module Abstract



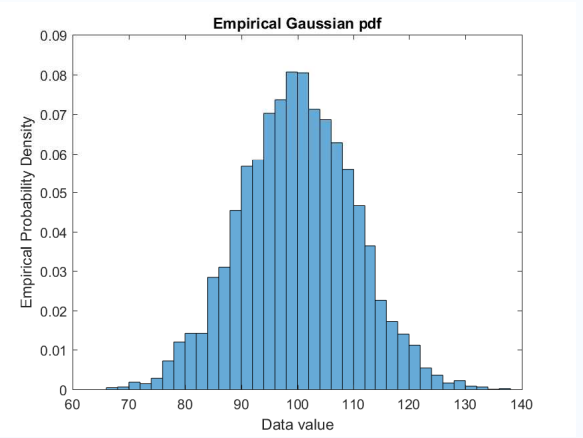
- Their properties are estimated by assumming:
  - an infinite number of observations or data points;
  - time-invariant statistics.

Module Abstract



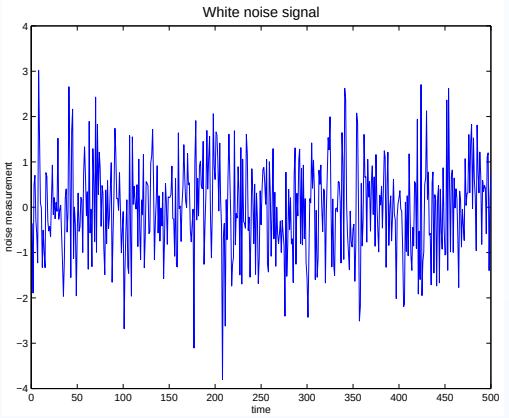
- Their properties are estimated by assumming:
  - an infinite number of observations or data points;
  - time-invariant statistics.
- In practice, these statistics must be estimated from finite-length data signals in noise.

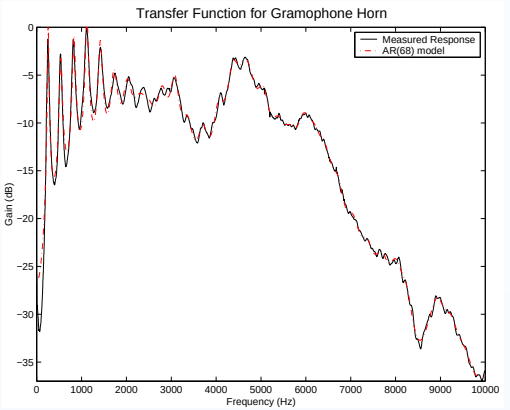
Module Abstract

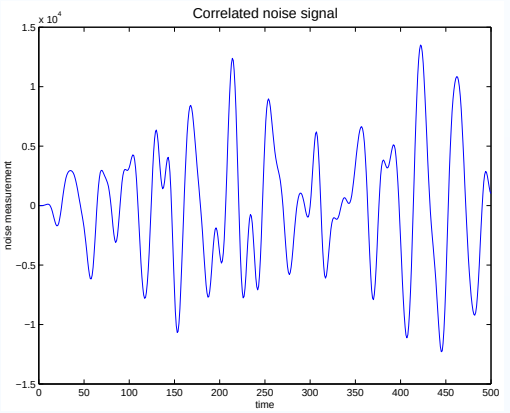


- Their properties are estimated by assumming:
  - an infinite number of observations or data points;
  - time-invariant statistics.
- In practice, these statistics must be estimated from finite-length data signals in noise.
- Module investigates relevant statistical properties, how they are estimated from real signals, and how they are used.

Introduction and Overview







Source Signal
e.g. Clean Speech

→ Channel
e.g. Room Acoustics →

Observed Signal
e.g. Reverberant Speech

Signal processing is concerned with the modification or manipulation of a signal, defined as an information-bearing representation of a real process, to the fulfillment of human needs and aspirations.

Description and Learning Outcomes

Module Aims to provide a unified introduction to the theory, implementation, and applications of statistical signal processing.

Description and Learning Outcomes

Module Aims to provide a unified introduction to the theory, implementation, and applications of statistical signal processing.

Module Objectives At the end of these modules, a student should be able to have:

1. acquired sufficient expertise in this area to understand and implement spectral estimation, signal modelling, parameter estimation, and adaptive filtering techniques;

Description and Learning Outcomes

Module Aims to provide a unified introduction to the theory, implementation, and applications of statistical signal processing.

Module Objectives At the end of these modules, a student should be able to have:

1. acquired sufficient expertise in this area to understand and implement spectral estimation, signal modelling, parameter estimation, and adaptive filtering techniques;

2. developed an understanding of the basic concepts and methodologies in statistical signal processing that provides the foundation for further study, research, and application to new problems.

Structure of the Module

These topics are:

1. review of the fundamentals of **probability theory**;

Structure of the Module

These topics are:

1. review of the fundamentals of **probability theory**;

2. **random variables** and **stochastic processes**;

Structure of the Module

These topics are:

1. review of the fundamentals of **probability theory**;
2. **random variables** and **stochastic processes**;
3. principles of **estimation theory**;

Structure of the Module

These topics are:

1. review of the fundamentals of **probability theory**;
2. **random variables** and **stochastic processes**;
3. principles of **estimation theory**;
4. **Bayesian estimation** theory;

Structure of the Module

These topics are:

1. review of the fundamentals of **probability theory**;
2. **random variables** and **stochastic processes**;
3. principles of **estimation theory**;
4. **Bayesian estimation** theory;
5. review of **Fourier transforms** and **discrete-time systems**;

Structure of the Module

These topics are:

1. review of the fundamentals of **probability theory**;
2. **random variables** and **stochastic processes**;
3. principles of **estimation theory**;
4. **Bayesian estimation** theory;
5. review of **Fourier transforms** and **discrete-time systems**;
6. **linear systems** with stationary random inputs, and **linear system models**;

Structure of the Module

These topics are:

1. review of the fundamentals of **probability theory**;
2. **random variables** and **stochastic processes**;
3. principles of **estimation theory**;
4. **Bayesian estimation** theory;
5. review of **Fourier transforms** and **discrete-time systems**;
6. **linear systems** with stationary random inputs, and **linear system models**;
7. **signal modelling** and **parametric spectral estimation**;

Structure of the Module

These topics are:

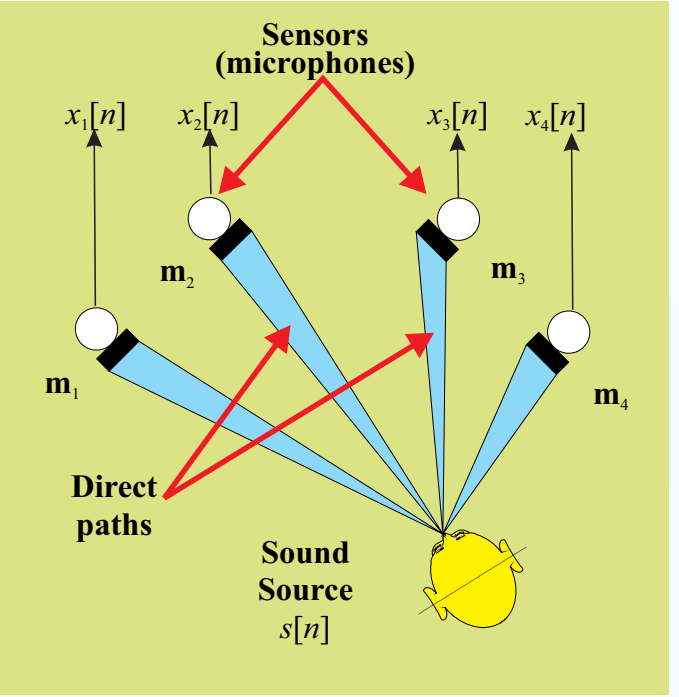
1. review of the fundamentals of **probability theory**;
2. **random variables** and **stochastic processes**;
3. principles of **estimation theory**;
4. **Bayesian estimation** theory;
5. review of **Fourier transforms** and **discrete-time systems**;
6. **linear systems** with stationary random inputs, and **linear system models**;
7. **signal modelling** and **parametric spectral estimation**;
8. an application investigating the estimation of sinusoids in noise, outperforming the Fourier transform.

Passive and Active Target Localisation

A number of signal processing problems rely on knowledge of the desired source position:

1. Tracking methods and target intent inference.
2. Mobile sensor node geometry.
3. Look-direction in beamforming techniques (for example in speech enhancement).
4. Camera steering for audio-visual BSS (including Robot Audition).
5. Speech diarisation.

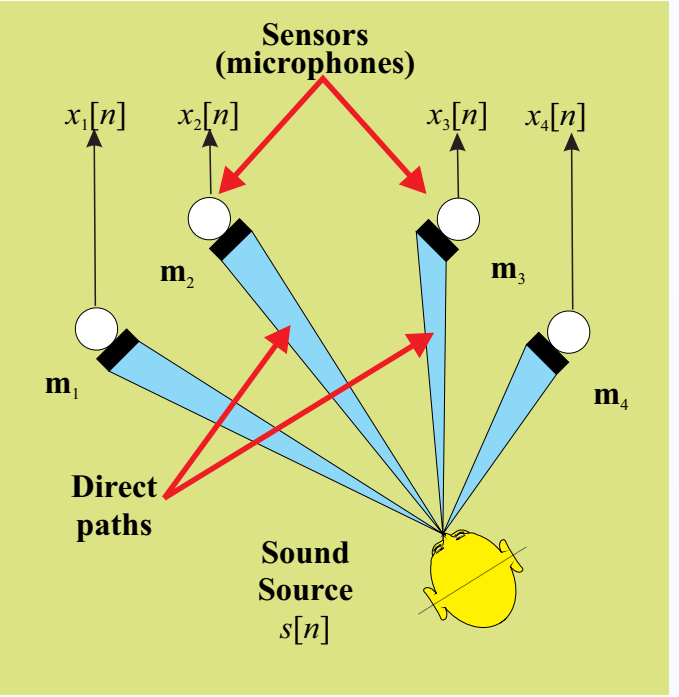
Passive Target Localisation Methodology



Ideal free-field model.

- Most passive target localisation (PTL) techniques rely on the fact that an impinging wavefront reaches one sensor before it reaches another.

Passive Target Localisation Methodology



Ideal free-field model.

- Most PTL techniques rely on the fact that an impinging wavefront reaches one sensor before it reaches another.
- Most PTL algorithms are designed assuming there is no multipath or reverberation present, the *free-field assumption*.

Source Localization Strategies

Existing source localisation methods can loosely be divided into:

1. those based on maximising the steered response power (SRP) of a beamformer:

   - location estimate derived directly from a filtered, weighted, and sum version of the signal data;

Source Localization Strategies

Existing source localisation methods can loosely be divided into:

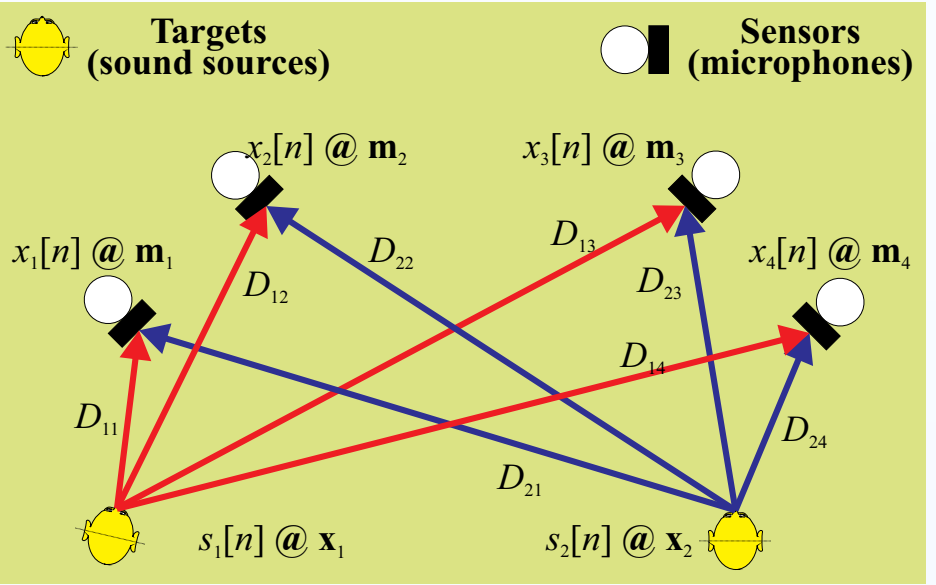
1. those based on maximising the steered response power (SRP) of a beamformer:
   - location estimate derived directly from a filtered, weighted, and sum version of the signal data;
2. techniques adopting high-resolution spectral estimation concepts:
   - any localisation scheme relying upon an application of the signal correlation matrix;

Source Localization Strategies

Existing source localisation methods can loosely be divided into:

1. those based on maximising the steered response power (SRP) of a beamformer:
   - location estimate derived directly from a filtered, weighted, and sum version of the signal data;
2. techniques adopting high-resolution spectral estimation concepts:
   - any localisation scheme relying upon an application of the signal correlation matrix;
3. approaches employing TDOA information:
   - source locations calculated from a set of TDOA estimates measured across various combinations of sensors.

Geometric Layout

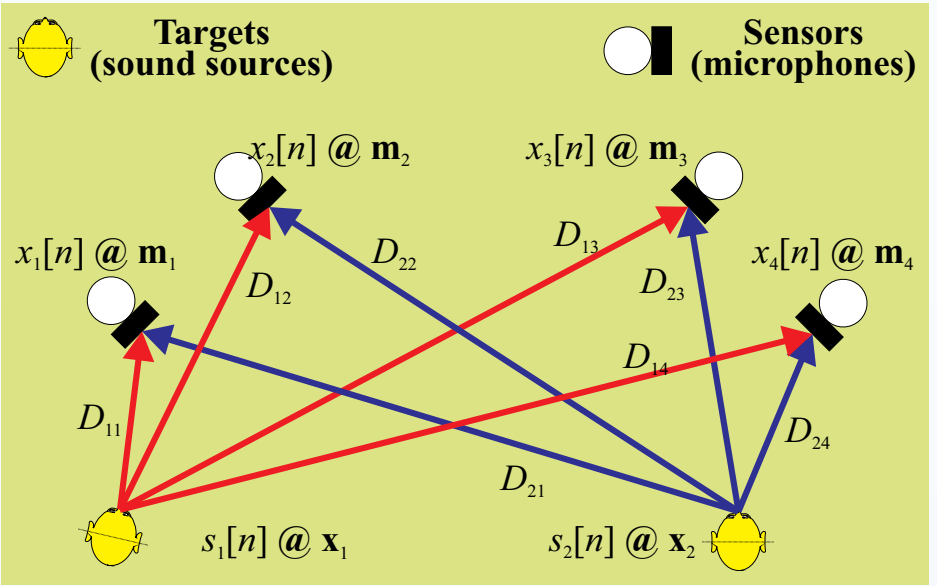


Geometry assuming a free-field model.

Suppose there is a:

- sensor array consisting of N nodes located at positions $\mathbf{m}_i \in \mathbb{R}^3$, for $i \in \{0, \ldots, N-1\}$,
- M talkers (or targets) at positions $\mathbf{x}_k \in \mathbb{R}^3$, for $k \in \{0, \ldots, M-1\}$.

Geometric Layout



Geometry assuming a free-field model.

The TDOA between the sensor node at position $\mathbf{m}_i$ and $\mathbf{m}_j$ due to a source at $\mathbf{x}_k$ can be expressed as:

$$T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right) \triangleq T_{ij}\left(\mathbf{x}_k\right) = \frac{\left|\mathbf{x}_k - \mathbf{m}_i\right| - \left|\mathbf{x}_k - \mathbf{m}_j\right|}{c}$$

where c is the speed of the impinging wavefront.

Ideal Free-field Model

- In an anechoic free-field environment, the signal from source k, denoted $s_k(t)$, propagates to the i-th sensor at time t as:

$$x_{ik}(t) = \alpha_{ik}\, s_k(t - \tau_{ik}) + b_{ik}(t)$$

  where $b_{ik}(t)$ denotes additive noise.

- Note that, in the frequency domain, this expression becomes:

$$X_{ik}(\omega) = \alpha_{ik}\, S_k(\omega)\, e^{-j\omega\,\tau_{ik}} + B_{ik}(\omega)$$

- The additive noise source is assumed to be uncorrelated with the source and noise sources at other sensors.

Ideal Free-field Model

- In an anechoic free-field environment, the signal from source k, denoted $s_k(t)$, propagates to the i-th sensor at time t as:

$$x_{ik}(t) = \alpha_{ik}\, s_k(t - \tau_{ik}) + b_{ik}(t)$$

  where $b_{ik}(t)$ denotes additive noise.

- Note that, in the frequency domain, this expression becomes:

$$X_{ik}(\omega) = \alpha_{ik}\, S_k(\omega)\, e^{-j\omega\,\tau_{ik}} + B_{ik}(\omega)$$

- The additive noise source is assumed to be uncorrelated with the source and noise sources at other sensors.

- The TDOA between the i-th and j-th sensor is given by:

$$\tau_{ijk} = \tau_{ik} - \tau_{jk} = T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k)$$

Indirect TDOA-based Methods

This is typically a two-step procedure in which:

- Typically, TDOAs are extracted using the GCC function, or an adaptive eigenvalue decomposition (AED) algorithm.

Indirect TDOA-based Methods

This is typically a two-step procedure in which:

- Typically, TDOAs are extracted using the GCC function, or an adaptive eigenvalue decomposition (AED) algorithm.
- A hypothesised spatial position of the target can be used to predict the expected TDOAs (or corresponding range) at the sensor.

Indirect TDOA-based Methods

This is typically a two-step procedure in which:

- Typically, TDOAs are extracted using the GCC function, or an adaptive eigenvalue decomposition (AED) algorithm.
- A hypothesised spatial position of the target can be used to predict the expected TDOAs (or corresponding range) at the sensor.
- The error between the measured and hypothesised TDOAs is then minimised.

Indirect TDOA-based Methods

This is typically a two-step procedure in which:

- Typically, TDOAs are extracted using the GCC function, or an adaptive eigenvalue decomposition (AED) algorithm.
- A hypothesised spatial position of the target can be used to predict the expected TDOAs (or corresponding range) at the sensor.
- The error between the measured and hypothesised TDOAs is then minimised.
- Accurate and robust TDOA estimation is the key to the effectiveness of this class of PTL methods.

Indirect TDOA-based Methods

This is typically a two-step procedure in which:

- Typically, TDOAs are extracted using the GCC function, or an adaptive eigenvalue decomposition (AED) algorithm.
- A hypothesised spatial position of the target can be used to predict the expected TDOAs (or corresponding range) at the sensor.
- The error between the measured and hypothesised TDOAs is then minimised.
- Accurate and robust TDOA estimation is the key to the effectiveness of this class of PTL methods.
- An alternative way of viewing these solutions is to consider what **spatial positions** of the target could lead to the estimated TDOA.

Hyperbolic Least Squares Error Function

- If a TDOA is estimated between two sensor nodes i and j, then the error between this and modelled TDOA is

$$\epsilon_{ij}(\mathbf{x}_k) = \tau_{ijk} - T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right)$$

- The total error as a function of target position

$$J(\mathbf{x}_k) = \sum_{i=1}^{N} \sum_{j \neq i=1}^{N} \epsilon_{ij}(\mathbf{x}_k) = \sum_{i=1}^{N} \sum_{j \neq i=1}^{N} \left(\tau_{ijk} - T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right)\right)^2$$

where

$$T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right) \triangleq T_{ij}\left(\mathbf{x}_k\right) = \frac{\left|\mathbf{x}_k - \mathbf{m}_i\right| - \left|\mathbf{x}_k - \mathbf{m}_j\right|}{c}$$

- Unfortunately, since $T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right)$ is a nonlinear function of $\mathbf{x}_k$, the minimum least-squares estimate (LSE) does not possess a closed-form solution.

TDOA estimation methods

Two key methods for TDOA estimation are using the GCC function and the adaptive eigenvalue decomposition (AED) algorithm.

GCC algorithm most popular approach assuming an ideal free-field movel

- computationally efficient, and hence short decision delays;
- perform fairly well in moderately noisy and reverberant environments.

TDOA estimation methods

Two key methods for TDOA estimation are using the GCC function and the adaptive eigenvalue decomposition (AED) algorithm.

GCC algorithm most popular approach assuming an ideal free-field movel

- computationally efficient, and hence short decision delays;
- perform fairly well in moderately noisy and reverberant environments.

However, GCC-based methods

- fail when multipath is high;
- focus of current research is on combating the effect of multipath.

TDOA estimation methods

Two key methods for TDOA estimation are using the GCC function and the adaptive eigenvalue decomposition (AED) algorithm.

AED Algorithm Approaches the TDOA estimation approach from a different point of view from the *traditional* GCC method.

- adopts a multipath rather than free-field model;
- computationally more expensive than GCC;
- can fail when there are common-zeros in the channel.

GCC TDOA estimation

The GCC algorithm proposed by *Knapp and Carter* is the most widely used approach to TDOA estimation.

- The TDOA estimate between two microphones i and j

$$\hat{\tau_{ij}} = \arg\max_{\ell} r_{x_i\, x_j}[\ell]$$

- The cross-correlation function is given by

$$r_{x_i\, x_j}[\ell] = \mathcal{F}^{-1}\left(\Phi\left(e^{j\omega T_s}\right) P_{x_1 x_2}\left(e^{j\omega T_s}\right)\right)$$

where the cross-power spectral density (CPSD) is given by

$$P_{x_1 x_2}\left(e^{j\omega T_s}\right) = \mathbb{E}\left[X_1\left(e^{j\omega T_s}\right) X_2\left(e^{j\omega T_s}\right)\right]$$

GCC TDOA estimation

The GCC algorithm proposed by *Knapp and Carter* is the most widely used approach to TDOA estimation.

- The TDOA estimate between two microphones i and j

$$\hat{\tau_{ij}} = \arg\max_{\ell} r_{x_i\, x_j}[\ell]$$

- The cross-correlation function is given by

$$r_{x_i\, x_j}[\ell] = \mathcal{F}^{-1}\left(\Phi\left(e^{j\omega T_s}\right) P_{x_1 x_2}\left(e^{j\omega T_s}\right)\right)$$

where the CPSD is given by

$$P_{x_1 x_2}\left(e^{j\omega T_s}\right) = \mathbb{E}\left[X_1\left(e^{j\omega T_s}\right) X_2\left(e^{j\omega T_s}\right)\right]$$

- For the free-field model, it can be shown that:

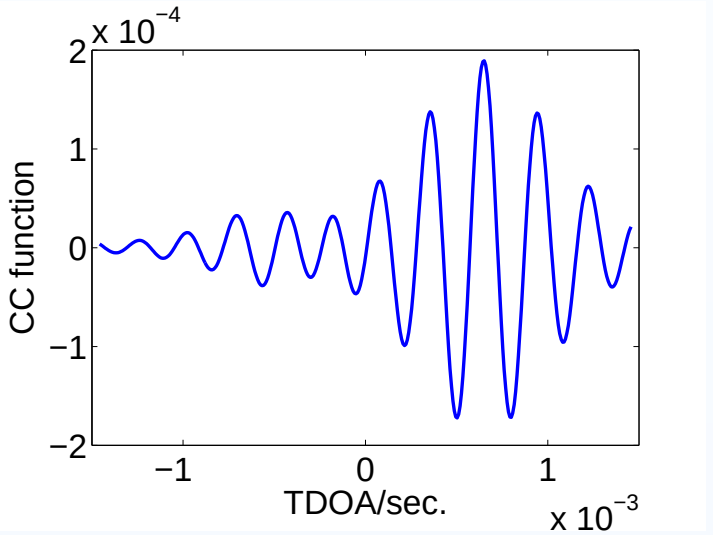
$$\angle P_{x_i x_j}(\omega) = -j\omega\, T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right)$$

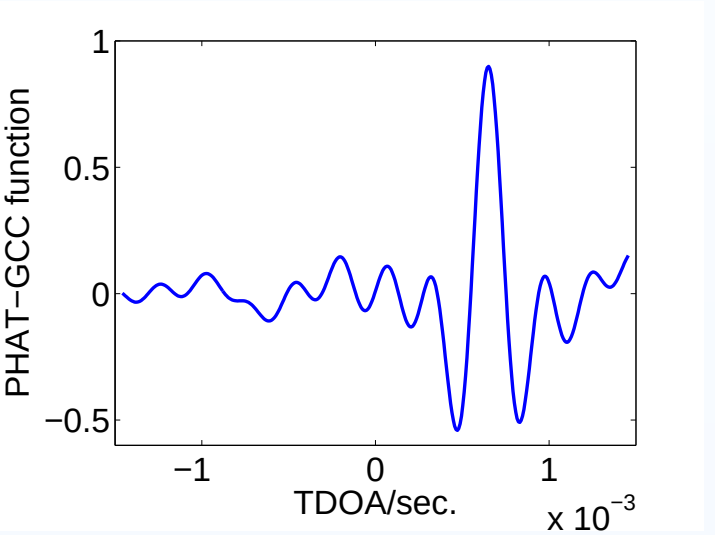
GCC Processors

Processor Name	Frequency Function
Cross Correlation	1
PHAT	$\frac{1}{\left\|P_{x_1x_2}\left(e^{j\omega T_s}\right)\right\|}$
Roth Impulse Response	$\frac{1}{P_{x_1x_1}\left(e^{j\omega T_s}\right)}$ or $\frac{1}{P_{x_2x_2}\left(e^{j\omega T_s}\right)}$
SCOT	$\frac{1}{\sqrt{P_{x_1x_1}\left(e^{j\omega T_s}\right)P_{x_2x_2}\left(e^{j\omega T_s}\right)}}$
Eckart	$\frac{P_{s_1s_1}\left(e^{j\omega T_s}\right)}{P_{n_1n_1}\left(e^{j\omega T_s}\right)P_{n_2n_2}\left(e^{j\omega T_s}\right)}$
Hannon-Thomson or ML	$\frac{\left\|\gamma_{x_1x_2}\left(e^{j\omega T_s}\right)\right\|^2}{\left\|P_{x_1x_2}\left(e^{j\omega T_s}\right)\right\|\left(1-\left\|\gamma_{x_1x_2}\left(e^{j\omega T_s}\right)\right\|^2\right)}$

where $\gamma_{x_1x_2}\left(e^{j\omega T_s}\right)$ is the normalised CPSD or **coherence function**

GCC Processors





Normal cross-correlation and GCC-phase transform (PHAT) (GCC-PHAT) functions for a frame of speech.

Direct Localisation Methods

- Direct localisation methods have the advantage that the relationship between the measurement and the state is linear.
- However, extracting the position measurement requires a multi-dimensional search over the state space and is usually computationally expensive.

Steered Response Power Function

The steered beamformer (SBF) or SRP function is a measure of correlation across *all pairs* of microphone signals for a set of relative delays that arise from a hypothesised source location.

The frequency domain **delay-and-sum beamformer** steered to a spatial position $\hat{\mathbf{x}}_k$ such that $\hat{\tau}_{pk} = |\hat{\mathbf{x}} - \mathbf{m}_p|$:

$$S(\hat{\mathbf{x}}) = \int_{\Omega} \left| \sum_{p=1}^{N} W_p\left(e^{j\omega T_s}\right) X_p\left(e^{j\omega T_s}\right) e^{j\omega \hat{\tau}_{pk}} \right|^2 d\omega$$

Steered Response Power Function

The SBF or SRP function is a measure of correlation across *all pairs* of microphone signals for a set of relative delays that arise from a hypothesised source location.

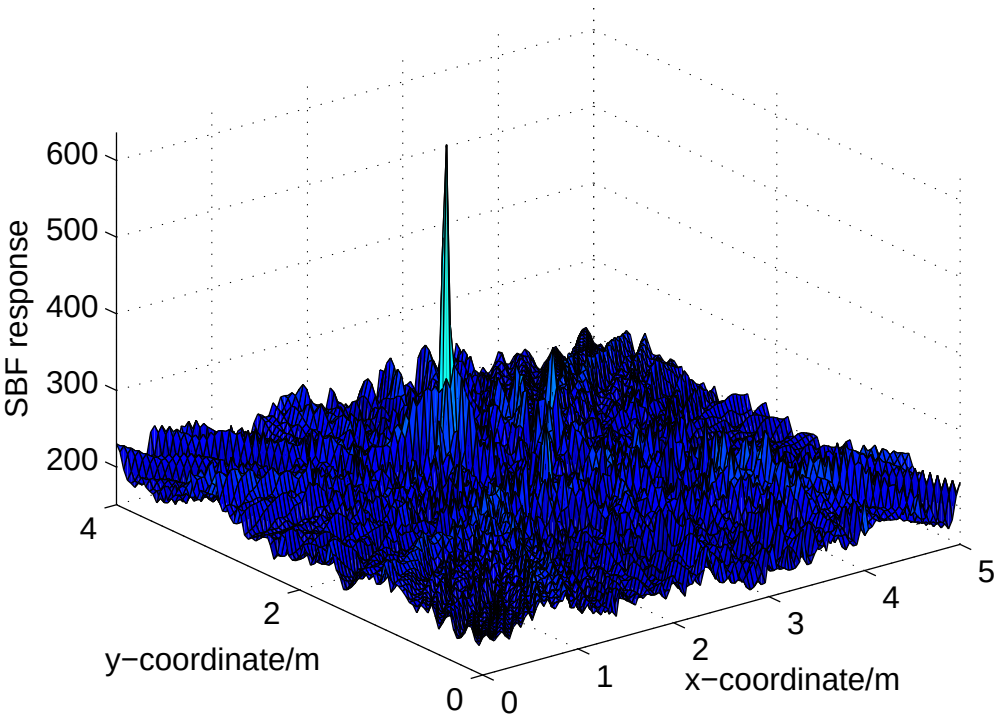
The frequency domain **delay-and-sum beamformer** steered to a spatial position $\hat{\mathbf{x}}_k$ such that $\hat{\tau}_{pk} = |\hat{\mathbf{x}} - \mathbf{m}_p|$:

$$S(\hat{\mathbf{x}}) = \int_{\Omega} \left| \sum_{p=1}^{N} W_p\left(e^{j\omega T_s}\right) X_p\left(e^{j\omega T_s}\right) e^{j\omega \hat{\tau}_{pk}} \right|^2 d\omega$$

$$\mathbb{E}\left[S(\hat{\mathbf{x}})\right] = \sum_{p=1}^{N}\sum_{q=1}^{N} r_{x_i x_j}\left[\hat{\tau}_{pqk}\right]$$

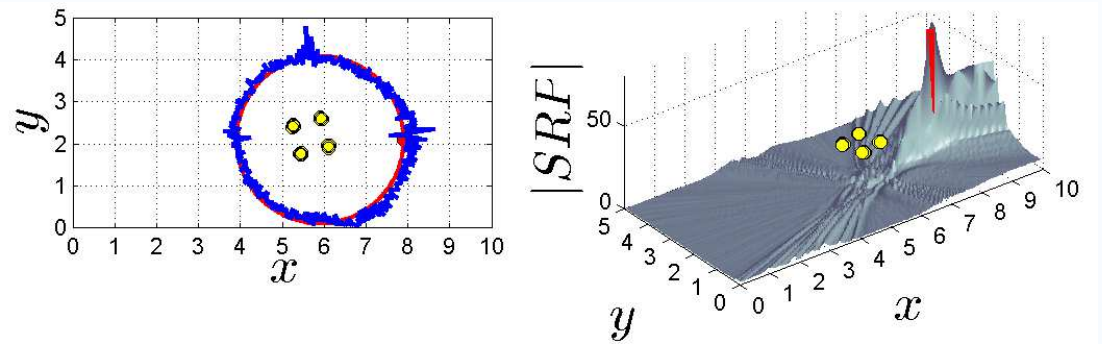
$$\equiv \sum_{p=1}^{N}\sum_{q=1}^{N} r_{x_i x_j}\left[\frac{|\mathbf{x}_k - \mathbf{m}_i| - |\mathbf{x}_k - \mathbf{m}_j|}{c}\right]$$

Steered Response Power Function



SBF response from a frame of speech signal. The integration frequency range is 300 to 3500 Hz. The true source position is at $[2.0, 2.5]m$. The grid density is set to 40 mm.

Steered Response Power Function



An example video showing the SBF changing as the source location moves.

- Show video!

Conclusions

To fully appreciate the algorithms in PTL, we need:

1. Signal analysis in time and frequency domain.
2. Least Squares Estimation Theory.
3. Expectations and frequency-domain statistical analysis.
4. Correlation and power-spectral density theory.
5. And, of course, all the theory to explain the above!

Handout 2
Probability Theory

Introduction

How many water taxis are there in Venice?

Introduction

How many water taxis are there in Venice?

How does your answer change when you see more taxis?

Introduction

- The theory of probability deals with averages of mass phenomena occurring sequentially or simultaneously;
  - this might include radar detection, signal detection, anomaly detection, parameter estimation, ...
- By considering fundamentals such as the probability of individual events, we can develop a probabilistic framework for analysing signals.
- It is *observed* that certain averages approach a constant value as the number of observations increases; and that this value remains the same if the averages are evaluated over any sub-sequence specified before the experiment is performed.

Introduction

If an experiment is performed n times, and the event A occurs n_A times, then with a *high degree of certainty*, the relative frequency n_A/n is *close to* $\Pr(A)$, such that:

$$\Pr(A) \approx \frac{n_A}{n}$$

provided that n is *sufficiently large*.

Note that this interpretation and the language used is all very imprecise.

Classical Definition of Probability

For several centuries, the theory of probability was based on the *classical definition*, which states that the probability $\Pr(A)$ of an event A is determine *a priori* without actual experimentation. It is given by the ratio:

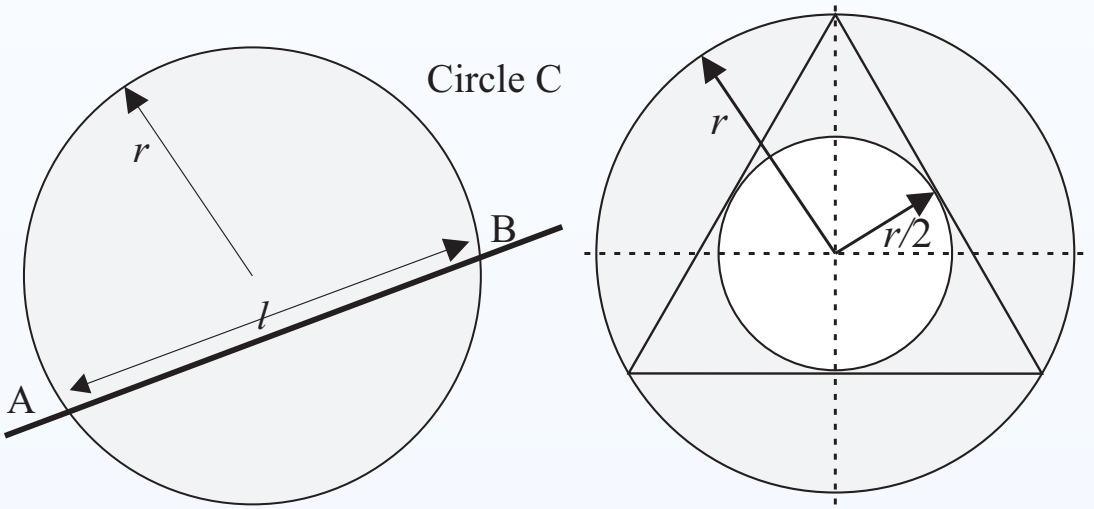
$$\Pr(A) = \frac{N_A}{N}$$

where:

- N is the total number of outcomes,
- and N_A is the total number of outcomes that are favourable to the event A, provided that *all outcomes are equally probable*.

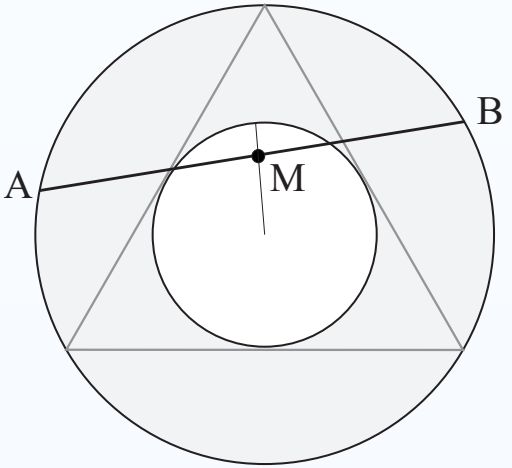
Bertrand's Paradox

Consider a circle C of radius r; what is the probability p that the length ℓ of a *randomly selected* cord AB is greater than the length, $r\sqrt{3}$, of the inscribed equilateral triangle?



Bertrand's paradox, problem definition.

Bertrand's Paradox

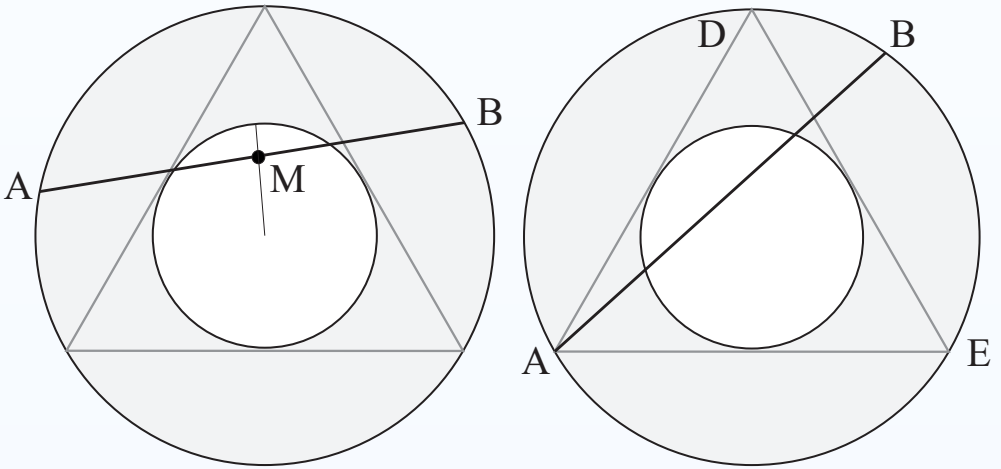


Different selection methods.

1. In the **random midpoints** method, a cord is selected by choosing a point M anywhere in the full circle, and two end-points A and B on the circumference of the circle, such that the resulting chord AB through these chosen points has M as its midpoint.

$$p = \frac{\pi\left(\frac{r}{2}\right)^2}{\pi r^2} = \frac{1}{4}$$

Bertrand's Paradox

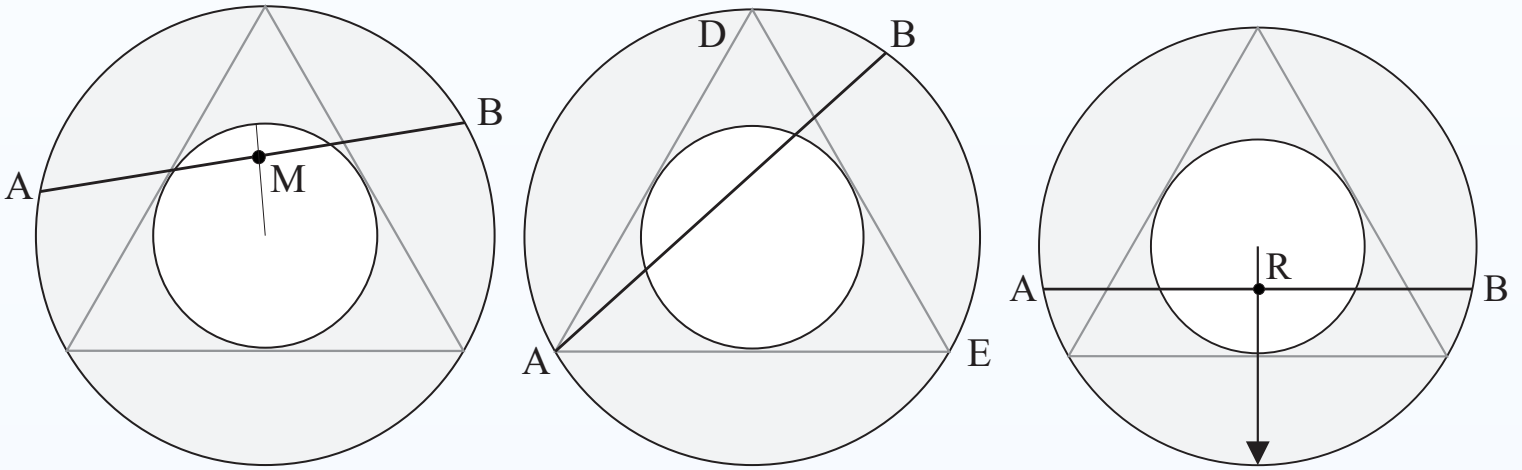


Different selection methods.

1. In the **random endpoints** method, consider selecting two random points on the circumference of the (outer) circle, A and B, and drawing a chord between them.

$$p = \frac{\frac{2\pi r}{3}}{2\pi r} = \frac{1}{3}$$

Bertrand's Paradox



Different selection methods.

1. Finally, in the **random radius method**, a radius of the circle is chosen at random, and a point on the radius is chosen at random. The chord AB is constructed as a line perpendicular to the chosen radius through the chosen point.

$$p = \frac{r}{2r} = \frac{1}{2}$$

There are thus three different but reasonable solutions to the same problem. Which one is valid?

Difficulties with the Classical Definition

1. The term **equally probable** in the definition of probability is making use of a concept still to be defined!

Difficulties with the Classical Definition

1. The term **equally probable** in the definition of probability is making use of a concept still to be defined!

2. The definition can only be applied to a limited class of problems.

   In the die experiment, for example, it is applicable only if the six faces have the same probability. If the die is loaded and the probability of a "4" equals 0.2, say, then this cannot be determined from the classical ratio.

Difficulties with the Classical Definition

1. The term **equally probable** in the definition of probability is making use of a concept still to be defined!

2. The definition can only be applied to a limited class of problems.

   In the die experiment, for example, it is applicable only if the six faces have the same probability. If the die is loaded and the probability of a "4" equals 0.2, say, then this cannot be determined from the classical ratio.

3. If the number of possible outcomes is infinite, then some other measure of infinity for determining the classical probability ratio is needed, such as length, or area. This leads to difficulties, as discussed in Bertrand's paradox.

Axiomatic Definition

The axiomatic approach to probability is based on the following three postulates and *on nothing else*:

1. The probability $\Pr(A)$ of an event A is a non-negative number assigned to this event:

$$\Pr(A) \geq 0$$

2. Defining the **certain event**, S, as the event that occurs in every trial, then the probability of the certain event equals 1, such that:

$$\Pr(S) = 1$$

3. If the events A and B are **mutually exclusive**, then the probability of one event or the other occurring separately is:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B)$$

Set Theory

Unions and Intersections Unions and intersections are commutative, associative, and distributive, such that:

$$A \cup B = B \cup A, \quad (A \cup B) \cup C = A \cup (B \cup C)$$

$$AB = BA, \quad (AB)C = A(BC), \quad A(B \cup C) = AB \cup AC$$

Set Theory

Unions and Intersections Unions and intersections are commutative, associative, and distributive, such that:

$$A \cup B = B \cup A, \quad (A \cup B) \cup C = A \cup (B \cup C)$$

$$AB = BA, \quad (AB)C = A(BC), \quad A(B \cup C) = AB \cup AC$$

Complements The complement $\overline{A}$ of a set $A \subset S$ is the set consisting of all elements of S that are not in A. Note that:

$$A \cup \overline{A} = S \quad \text{and} \quad A \cap \overline{A} \equiv A\overline{A} = \{\emptyset\}$$

Set Theory

Unions and Intersections Unions and intersections are commutative, associative, and distributive, such that:

$$A \cup B = B \cup A, \quad (A \cup B) \cup C = A \cup (B \cup C)$$
$$AB = BA, \quad (AB)C = A(BC), \quad A(B \cup C) = AB \cup AC$$

Complements The complement $\overline{A}$ of a set $A \subset S$ is the set consisting of all elements of S that are not in A. Note that:

$$A \cup \overline{A} = S \quad \text{and} \quad A \cap \overline{A} \equiv A\overline{A} = \{\emptyset\}$$

Partitions A partition U of a set S is a collection of mutually exclusive subsets A_i of S whose union equations S:

$$\bigcup_{i=1}^{\infty} A_i = S, \quad A_i \cap A_j = \{\emptyset\}, \quad i \neq j \quad \Rightarrow \quad U = [A_1, \ldots, A_n]$$

Set Theory

De Morgan's Law Using Venn diagrams, it is relatively straightforward to show

$$\overline{A \cup B} = \overline{A} \cap \overline{B} \equiv \overline{A}\,\overline{B} \quad \text{and} \quad \overline{A \cap B} \equiv \overline{AB} = \overline{A} \cup \overline{B}$$

Set Theory

De Morgan's Law Using Venn diagrams, it is relatively straightforward to show

$$\overline{A \cup B} = \overline{A} \cap \overline{B} \equiv \overline{A}\,\overline{B} \quad \text{and} \quad \overline{A \cap B} \equiv \overline{AB} = \overline{A} \cup \overline{B}$$

As an application of this, note that:

$$\begin{aligned} \overline{A \cup BC} &= \overline{A}\,\overline{BC} = \overline{A}\left(\overline{B} \cup \overline{C}\right) \\ &= \left(\overline{A}\,\overline{B}\right) \cup \left(\overline{A}\,\overline{C}\right) \\ &= \overline{A \cup B} \cup \overline{A \cup C} \\ \Rightarrow \quad A \cup BC &= (A \cup B)\,(A \cup C) \end{aligned}$$

Properties of Axiomatic Probability

Impossible Event The probability of the impossible event is 0, and therefore:

$$\Pr(\emptyset) = 0$$

Properties of Axiomatic Probability

Impossible Event The probability of the impossible event is 0, and therefore:

$$\Pr(\emptyset) = 0$$

Complements Since $A \cup \overline{A} = S$ and $A\overline{A} = \{\emptyset\}$, then $\Pr\left(A \cup \overline{A}\right) = \Pr(A) + \Pr\left(\overline{A}\right) = \Pr(S) = 1$, such that:

$$\Pr\left(\overline{A}\right) = 1 - \Pr(A)$$

Properties of Axiomatic Probability

Impossible Event The probability of the impossible event is 0, and therefore:

$$\Pr(\emptyset) = 0$$

Complements Since $A \cup \overline{A} = S$ and $A\overline{A} = \{\emptyset\}$, then $\Pr\left(A \cup \overline{A}\right) = \Pr(A) + \Pr\left(\overline{A}\right) = \Pr(S) = 1$, such that:

$$\Pr\left(\overline{A}\right) = 1 - \Pr(A)$$

Sum Rule The **addition law of probability** or the **sum rule** for any two events A and B is given by:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

Properties of Axiomatic Probability

Example (Proof of the Sum Rule). SOLUTION. To prove this, separately write $A \cup B$ and B as the union of two mutually exclusive events.

- First, note that

$$A \cup B = \left(A \cup \overline{A}\right)(A \cup B) = A \cup \left(\overline{A}\, B\right)$$

and that since $A\left(\overline{A}\, B\right) = \left(A\, \overline{A}\right) B = \{\emptyset\}B = \{\emptyset\}$, then A and $\overline{A}\, B$ are mutually exclusive events.

- Second, note that:

$$B = \left(A \cup \overline{A}\right) B = (A\, B) \cup \left(\overline{A}\, B\right) \qquad \square$$

and that $(A\, B) \cap \left(\overline{A}\, B\right) = A\, \overline{A}\, B = \{\emptyset\}\, B = \{\emptyset\}$ and are therefore mutually exclusive events.

Properties of Axiomatic Probability

Example (Proof of the Sum Rule). SOLUTION. Using these two disjoint unions, then:

$$\Pr(A \cup B) = \Pr\left(A \cup \left(\overline{A}\, B\right)\right) = \Pr(A) + \Pr\left(\overline{A}\, B\right)$$

$$\Pr(B) = \Pr\left((A\, B) \cup \left(\overline{A}\, B\right)\right) = \Pr(A\, B) + \Pr\left(\overline{A}\, B\right)$$

Eliminating $\Pr\left(\overline{A}\, B\right)$ by subtracting these equations gives the desired result:

$$\Pr(A \cup B) - \Pr(B) = \Pr\left(A \cup \left(\overline{A}\, B\right)\right) = \Pr(A) - \Pr(A\, B) \quad \square$$

Properties of Axiomatic Probability

Example (Sum Rule). Let A and B be events with probabilities $\Pr(A) = 3/4$ and $\Pr(B) = 1/3$. Show that $1/12 \leq \Pr(A\,B) \leq 1/3$.

Properties of Axiomatic Probability

Example (Sum Rule). Let A and B be events with probabilities $\Pr(A) = 3/4$ and $\Pr(B) = 1/3$. Show that $1/12 \leq \Pr(A\,B) \leq 1/3$.

SOLUTION. Using the sum rule, that:

$$\Pr(A\,B) = \Pr(A)+\Pr(B)-\Pr(A \cup B) \geq \Pr(A)+\Pr(B)-1 = \frac{1}{12} \quad \square$$

which is the case when the whole **sample space** is covered by the two events. The second bound occurs since $A \cap B \subset B$ and similarly $A \cap B \subset A$, where $\subset$ denotes subset. Therefore, it can be deduced $\Pr(A\,B) \leq \min\{\Pr(A), \Pr(B)\} = 1/3$.

The Real Line

If the **certain event**, S, consists of a non-countable infinity of elements, then its probabilities cannot be determined in terms of the probabilities of elementary events.

The Real Line

If the **certain event**, S, consists of a non-countable infinity of elements, then its probabilities cannot be determined in terms of the probabilities of elementary events.

Suppose that S is the set of all real numbers. To construct a probability space on the real line, consider events as intervals $x_1 < x \leq x_2$, and their countable unions and intersections.

The Real Line

If the **certain event**, S, consists of a non-countable infinity of elements, then its probabilities cannot be determined in terms of the probabilities of elementary events.

Suppose that S is the set of all real numbers. To construct a probability space on the real line, consider events as intervals $x_1 < x \leq x_2$, and their countable unions and intersections.

To complete the specification of probabilities for this set, it suffices to assign probabilities to the events $\{x \leq x_i\}$.

The Real Line

If the **certain event**, S, consists of a non-countable infinity of elements, then its probabilities cannot be determined in terms of the probabilities of elementary events.

Suppose that S is the set of all real numbers. To construct a probability space on the real line, consider events as intervals $x_1 < x \leq x_2$, and their countable unions and intersections.

To complete the specification of probabilities for this set, it suffices to assign probabilities to the events $\{x \leq x_i\}$.

This notion leads to **cumulative distribution functions (cdfs)** and **probability density functions (pdfs)** in the next handout.

Conditional Probability

If an experiment is repeated n times, and on each occasion the occurrences or non-occurrences of two events A and B are observed. Suppose that only those outcomes for which B occurs are considered, and all other experiments are disregarded.

Conditional Probability

If an experiment is repeated n times, and on each occasion the occurrences or non-occurrences of two events A and B are observed. Suppose that only those outcomes for which B occurs are considered, and all other experiments are disregarded.

In this smaller collection of trials, the proportion of times that A occurs, given that B has occurred, is:

$$\Pr\left(A \,\middle|\, B\right) \approx \frac{n_{AB}}{n_B} = \frac{n_{AB}/n}{n_B/n} = \frac{\Pr\left(AB\right)}{\Pr\left(B\right)}$$

provided that n is sufficiently large.

It can be shown that this definition satisfies the **Kolmogorov Axioms**.

Conditional Probability

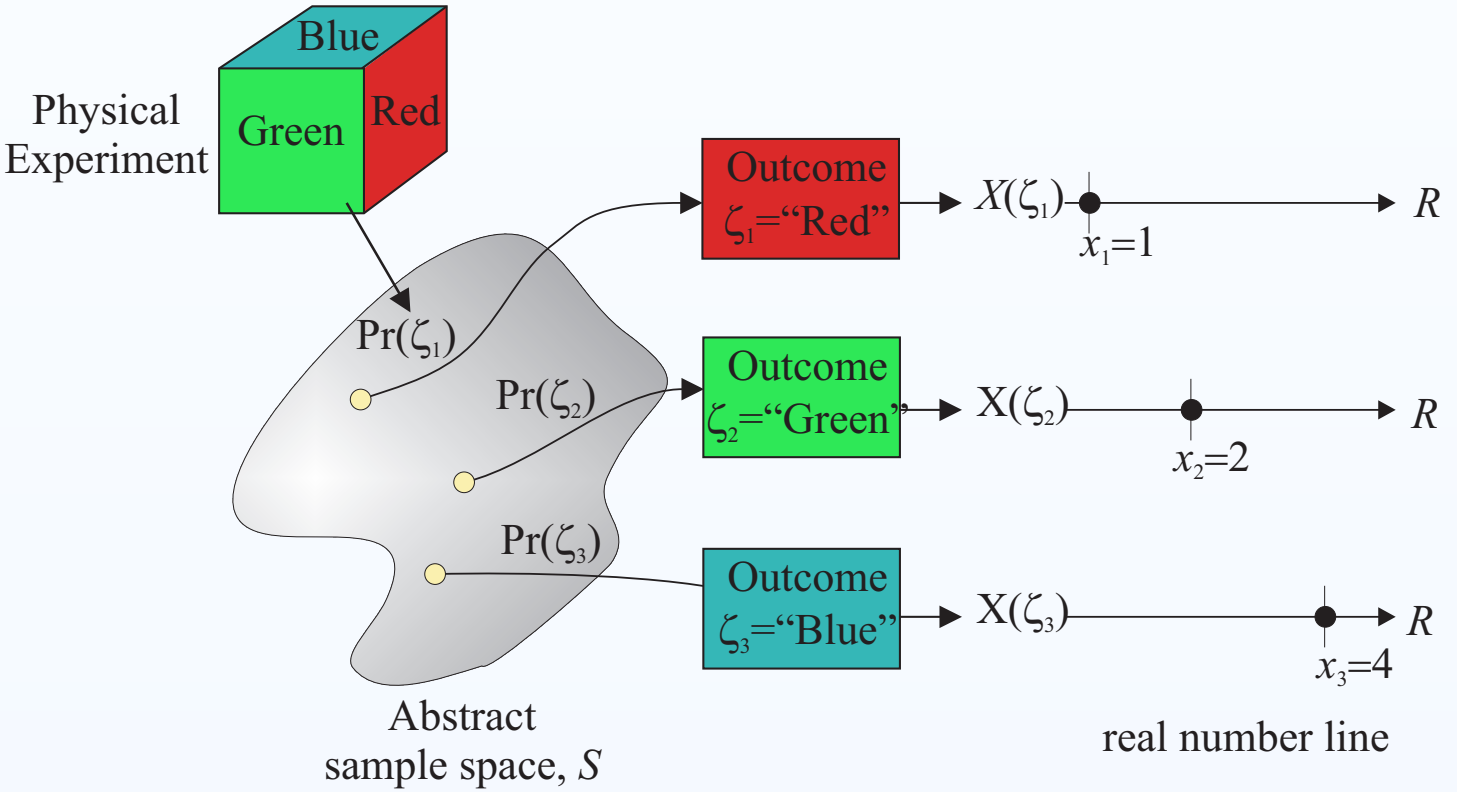
Example (Two Children). A family has two children. What is the probability that both are boys, given that at least one is a boy?

SOLUTION. The younger and older children may each be male or female, and it is assumed that each is equally likely.

Conditional Probability

Handout 3
Scalar Random Variables

Definition



A graphical representation of a random variable for a more specific example.

Definition

A **random variable (RV)** $X(\zeta)$ is a mapping that assigns a real number $X \in (-\infty, \infty)$ to every outcome ζ from an abstract probability space.

1. the interval $\{X(\zeta) \leq x\}$ is an event in the abstract probability space for every $x \in \mathbb{R}$;

2. $\Pr(X(\zeta) = \infty) = 0$ and $\Pr(X(\zeta) = -\infty) = 0$.

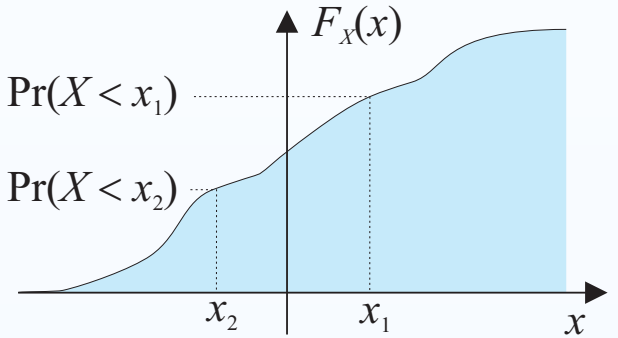
Definition

Example (Rolling die). Consider rolling a die, with six outcomes $\{\zeta_i,\, i \in \{1,\ldots,6\}\}$. In this experiment, assign the number 1 to every *even* outcome, and the number 0 to every *odd* outcome. Then the **RV** $X(\zeta)$ is given by:

$$X(\zeta_1) = X(\zeta_3) = X(\zeta_5) = 0 \quad \text{and} \quad X(\zeta_2) = X(\zeta_4) = X(\zeta_6) = 1$$

$\bowtie$

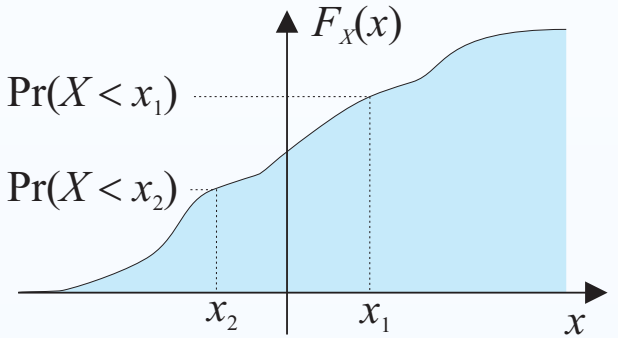
Distribution functions



The cumulative distribution function.

- The **probability set function** $\Pr\left(X\left(\zeta\right) \leq x\right)$ is a function of the set $\{X\left(\zeta\right) \leq x\}$, and therefore of the point $x \in \mathbb{R}$.

Distribution functions



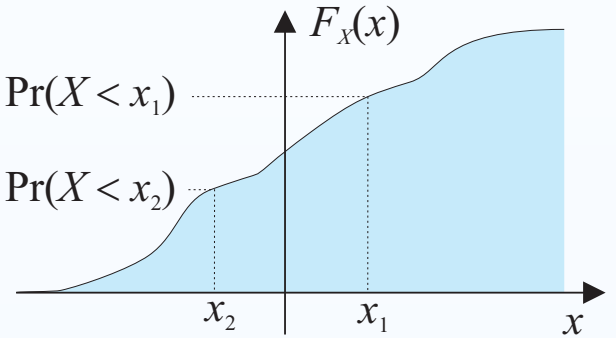
The cumulative distribution function.

- The **probability set function** $\Pr\left(X\left(\zeta\right) \leq x\right)$ is a function of the set $\left\{X\left(\zeta\right) \leq x\right\}$, and therefore of the point $x \in \mathbb{R}$.

- This probability is the **cumulative distribution function (cdf)**, $F_X\left(x\right)$ of a **RV** $X\left(\zeta\right)$, and is defined by:

$$F_X\left(x\right) \triangleq \Pr\left(X\left(\zeta\right) \leq x\right)$$

Distribution functions

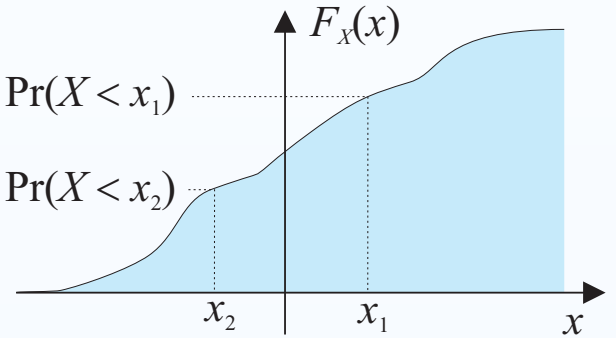


The cumulative distribution function.

- It hence follows that the probability of being within an interval $(x_\ell, x_r]$ is given by:

$$\begin{aligned}\Pr\left(x_\ell < X\left(\zeta\right) \leq x_r\right) &= \Pr\left(X\left(\zeta\right) \leq x_r\right) - \Pr\left(X\left(\zeta\right) \leq x_\ell\right) \\ &= F_X\left(x_r\right) - F_X\left(x_\ell\right)\end{aligned}$$

Distribution functions



The cumulative distribution function.

- It hence follows that the probability of being within an interval $(x_\ell, x_r]$ is given by:

$$\begin{aligned}\Pr\left(x_\ell < X\left(\zeta\right) \leq x_r\right) &= \Pr\left(X\left(\zeta\right) \leq x_r\right) - \Pr\left(X\left(\zeta\right) \leq x_\ell\right) \\ &= F_X\left(x_r\right) - F_X\left(x_\ell\right)\end{aligned}$$

- For small intervals, it is clearly apparent that gradients are important.

Kolmogorov's Axioms

The events $\{X \leq x_1\}$ and $\{x_1 < X \leq x_2\}$ are mutually exclussive events. Therefore, their union equals $\{x \leq x_2\}$, and therefore:

$$\Pr(X \leq x_1) + \Pr(x_1 < X \leq x_2) = \Pr(X \leq x_2)$$

$$\int_{-\infty}^{x_1} p(v)\, dv + \Pr(x_1 < X \leq x_2) = \int_{-\infty}^{x_2} p(v)\, dv$$

$$\Rightarrow \qquad \Pr(x_1 < X \leq x_2) = \int_{x_1}^{x_2} p(v)\, dv$$

Moreover, it follows that $\Pr(-\infty < X \leq \infty) = 1$ and the probability of the impossible event, $\Pr(X \leq -\infty) = 0$. Hence, the cdf satisfies the axiomatic definition of probability.

Density functions

- The **probability density function (pdf)**, $f_X(x)$ of a **RV** $X(\zeta)$, is defined as a formal derivative:

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}$$

Note $f_X(x)$ is not a **probability** on its own; it must be multiplied by a certain interval Δx to obtain a probability:

$$f_X(x)\,\Delta x \approx F_X(x+\Delta x) - F_X(x) \approx \Pr(x < X(\zeta) \leq x + \Delta x)$$

Density functions

- The **probability density function (pdf)**, $f_X(x)$ of a **RV** $X(\zeta)$, is defined as a formal derivative:

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}$$

Note $f_X(x)$ is not a **probability** on its own; it must be multiplied by a certain interval Δx to obtain a probability:

$$f_X(x)\,\Delta x \approx F_X(x+\Delta x) - F_X(x) \approx \Pr\left(x < X(\zeta) \leq x+\Delta x\right)$$

- It directly follows that:

$$F_X(x) = \int_{-\infty}^{x} f_X(v)\,dv$$

Density functions

- The **probability density function (pdf)**, $f_X(x)$ of a **RV** $X(\zeta)$, is defined as a formal derivative:

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}$$

Note $f_X(x)$ is not a **probability** on its own; it must be multiplied by a certain interval Δx to obtain a probability:

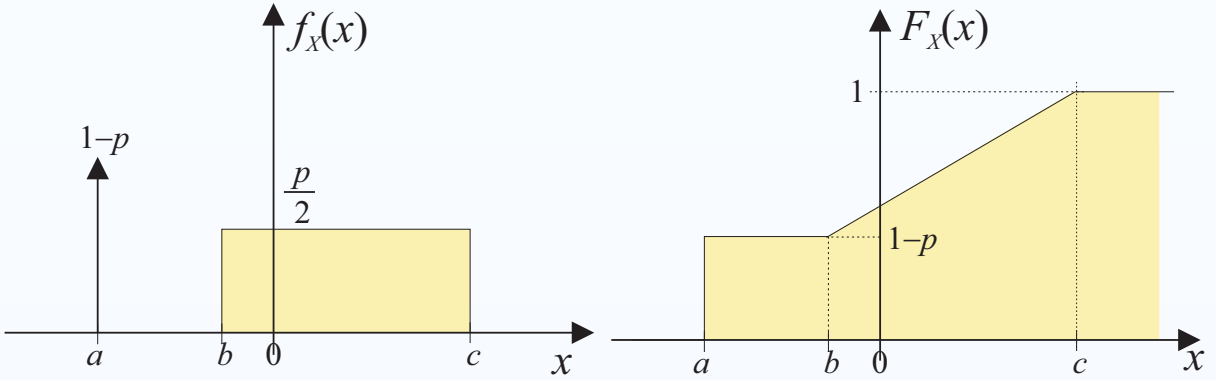
$$f_X(x)\,\Delta x \approx F_X(x+\Delta x) - F_X(x) \approx \Pr\left(x < X(\zeta) \leq x+\Delta x\right)$$

- It directly follows that:

$$F_X(x) = \int_{-\infty}^{x} f_X(v)\,dv$$

- For discrete-valued **RV**, use the **pmf**, p_k, the probability that $X(\zeta)$ takes on a value equal to x_k: $p_k \triangleq \Pr\left(X(\zeta) = x_k\right)$.

Density functions



A probability density function and its corresponding cumulative distribution function for a RV which is a mixture of continuous and discrete components.

Properties: Distributions and Densities

- Properties of **cdf**:

$$0 \leq F_X(x) \leq 1, \quad \lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to \infty} F_X(x) = 1$$

$F_X(x)$ is a monotonically increasing function of x:

$$F_X(a) \leq F_X(b) \quad \text{if} \quad a \leq b$$

Properties: Distributions and Densities

- Properties of **cdf**:

$$0 \leq F_X(x) \leq 1, \quad \lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to \infty} F_X(x) = 1$$

$F_X(x)$ is a monotonically increasing function of x:

$$F_X(a) \leq F_X(b) \quad \text{if} \quad a \leq b$$

- Properties of **pdfs**:

$$f_X(x) \geq 0, \quad \int_{-\infty}^{\infty} f_X(x)\, dx = 1$$

Properties: Distributions and Densities

- Properties of **cdf**:

$$0 \leq F_X(x) \leq 1, \quad \lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to \infty} F_X(x) = 1$$

$F_X(x)$ is a monotonically increasing function of x:

$$F_X(a) \leq F_X(b) \quad \text{if} \quad a \leq b$$

- Properties of **pdfs**:

$$f_X(x) \geq 0, \quad \int_{-\infty}^{\infty} f_X(x)\, dx = 1$$

- Probability of arbitrary events:

$$\Pr(x_1 < X(\zeta) \leq x_2) = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x)\, dx$$

Common Continuous RVs

Uniform distribution

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x \leq b, \\ 0 & \text{otherwise} \end{cases}$$

Normal distribution

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right], \quad x \in \mathbb{R}$$

Cauchy distribution

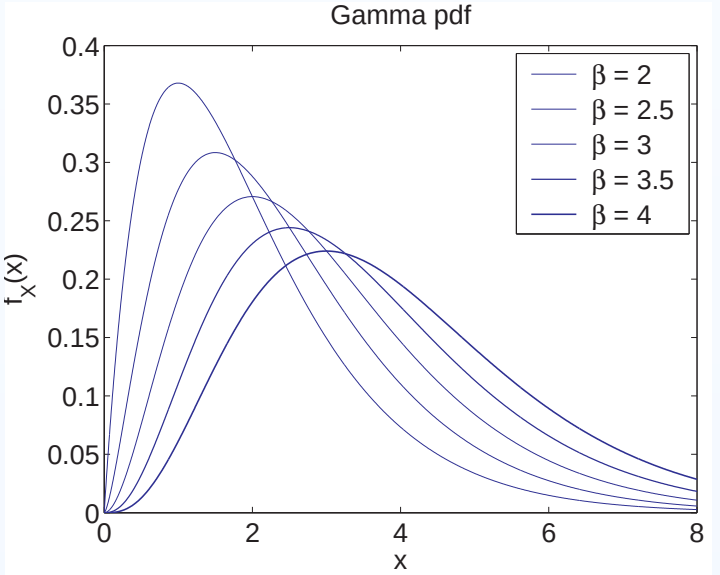
$$f_X(x) = \frac{\beta}{\pi} \frac{1}{(x-\mu_X)^2 + \beta^2}$$

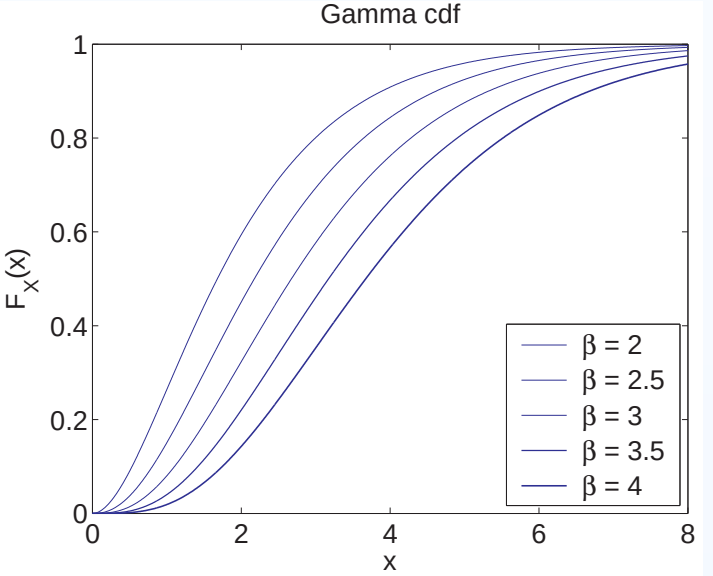
The Cauchy random variable is symmetric around the value $x = \mu_X$, but its mean and variance do not exist.

Common Continuous RVs

Gamma distribution

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{1}{\Gamma(\beta)} \alpha^\beta \, x^{\beta-1} \, e^{-\alpha x} & \text{if } x \geq 0, \end{cases}$$



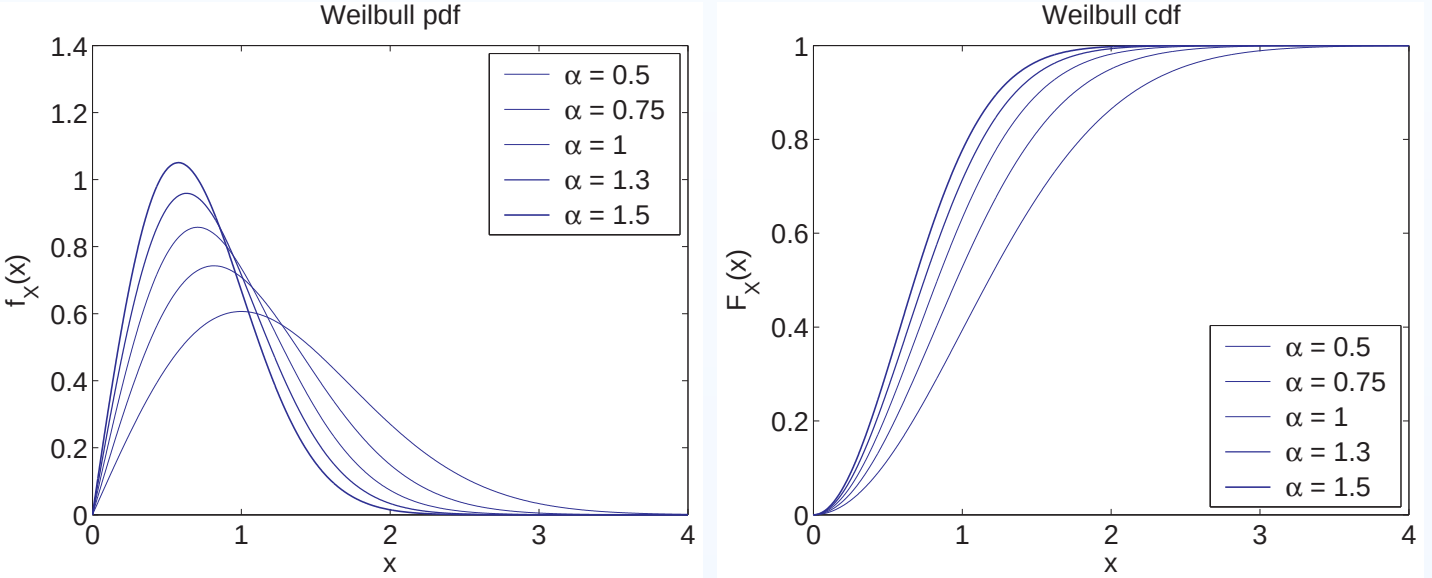


The Gamma density and distribution functions, for the case when $\alpha = 1$ and for various values of β.

Common Continuous RVs

Weibull distribution

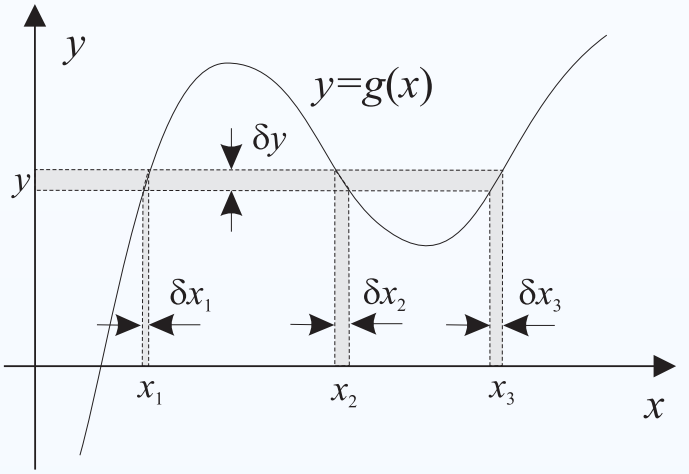
$$f_X(x) = \begin{cases} 0 & x < 0 \\ \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}} & x \geq 0 \end{cases}$$



The Weibull density and distribution functions, for the case when $\alpha = 1$, and for various values of the parameter β.

Probability transformation rule

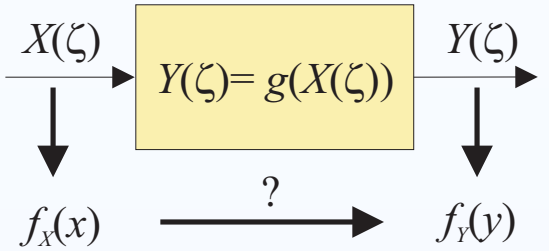
Suppose a random variable $Y(\zeta)$ is a function, g, of a random variable $X(\zeta)$, which has pdf given by $f_X(x)$. What is $f_Y(y)$?



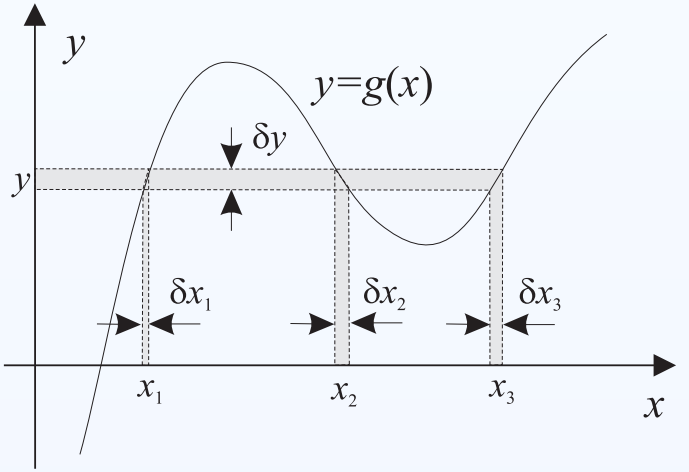
The mapping $y = g(x)$, and the effect of the mapping on intervals.

Probability transformation rule

Suppose a random variable $Y(\zeta)$ is a function, g, of a random variable $X(\zeta)$, which has pdf given by $f_X(x)$. What is $f_Y(y)$?

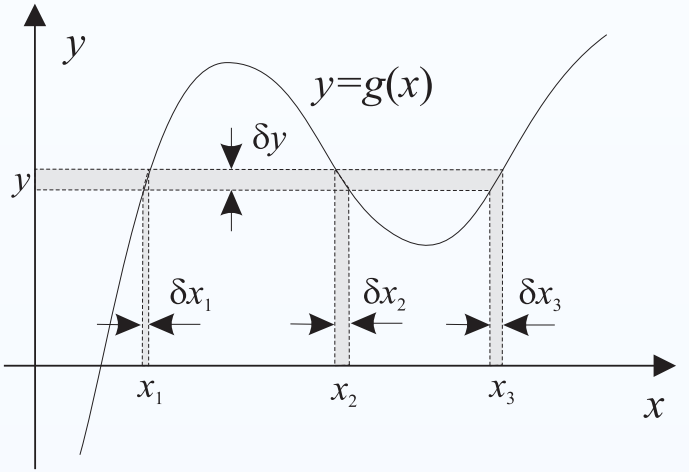


The mapping $y = g(x)$.



The mapping $y = g(x)$, and the effect of the mapping on intervals.

Probability transformation rule



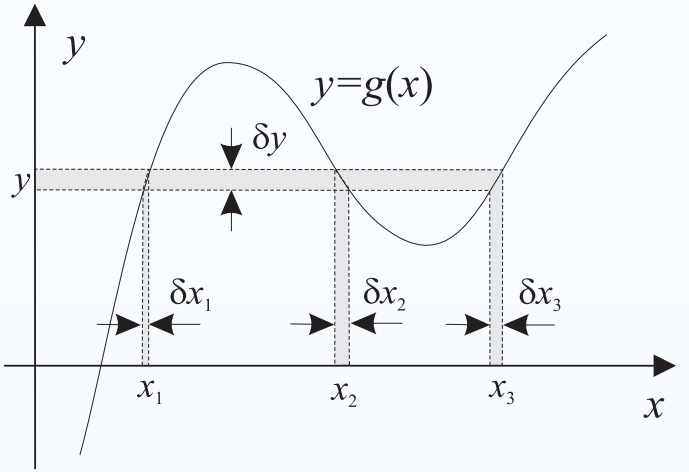
The mapping $y = g(x)$, and the effect of the mapping on intervals.

Theorem (Probability transformation rule). Denote the real roots of $y = g(x)$ by $\{x_n,\, n \in \mathcal{N}\}$, such that

$$y = g(x_1) = \cdots = g(x_N)$$

$$f_Y(y) = \sum_{n=1}^{N} \frac{f_X(x_n)}{|g'(x_n)|}$$

Probability transformation rule



The mapping $y = g(x)$, and the effect of the mapping on intervals.

Theorem (Probability transformation rule). Denote the real roots of $y = g(x)$ by $\{x_n,\, n \in \mathcal{N}\}$, such that

$$y = g(x_1) = \cdots = g(x_N)$$

Then, if the $Y(\zeta) = g[X(\zeta)]$, the pdf of $Y(\zeta)$ in terms of the pdf of $X(\zeta)$ is given by:

$$f_Y(y) = \sum^{N} \frac{f_X(x_n)}{|g'(x_n)|}$$

Probability transformation rule

Example (Log-normal distribution). Let $Y = e^X$, where $X \sim \mathcal{N}(0, 1)$. Find the pdf for the RV Y.

Probability transformation rule

Example (Log-normal distribution). Let $Y = e^X$, where $X \sim \mathcal{N}(0, 1)$. Find the pdf for the RV Y.

SOLUTION. Since $X \sim \mathcal{N}(0, 1)$, then:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

□

Probability transformation rule

Example (Log-normal distribution). Let $Y = e^X$, where $X \sim \mathcal{N}(0, 1)$. Find the pdf for the RV Y.

SOLUTION. Since $X \sim \mathcal{N}(0, 1)$, then:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Considering the transformation $y = g(x) = e^x$, there is one root, given by $x = \ln y$. Therefore, the derivative of this expression is $g'(x) = e^x = y$.

Hence, it follows:

$$f_Y(y) = \frac{f_X(x)}{g'(x)} = \frac{1}{y\sqrt{2\pi}} e^{-\frac{(\ln y)^2}{2}}$$

□

Expectations

To completely characterise a **RV**, the **pdf** must be known. However, it is desirable to summarise key aspects of the **pdf** by using a few parameters rather than having to specify the entire density function.

- The **expected** or **mean value** of a function of a **RV** $X(\zeta)$ is given by:

$$\mathbb{E}[X(\zeta)] = \int_{\mathbb{R}} x\, f_X(x)\, dx$$

Expectations

- If $X(\zeta)$ is discrete, then its corresponding **pdf** may be written in terms of its **pmf** as:

$$f_X(x) = \sum_k p_k \, \delta(x - x_k)$$

where the **Dirac-delta**, $\delta(x - x_k)$, is unity if $x = x_k$, and zero otherwise.

Expectations

- If $X(\zeta)$ is discrete, then its corresponding **pdf** may be written in terms of its **pmf** as:

$$f_X(x) = \sum_k p_k \, \delta(x - x_k)$$

where the **Dirac-delta**, $\delta(x - x_k)$, is unity if $x = x_k$, and zero otherwise.

- Hence, for a discrete **RV**, the **expected** value is given by:

$$\mu_x = \int_{\mathbb{R}} x \, f_X(x) \, dx = \int_{\mathbb{R}} x \sum_k p_k \, \delta(x - x_k) \, dx = \sum_k x_k \, p_k$$

where the order of integration and summation have been interchanged, and the sifting-property applied.

Properties of expectation operator

The expectation operator computes a statistical average by using the density $f_X(x)$ as a weighting function. Hence, the mean μ_x can be regarded as the *center of gravity* of the density.

Properties of expectation operator

The expectation operator computes a statistical average by using the density $f_X(x)$ as a weighting function. Hence, the mean μ_x can be regarded as the *center of gravity* of the density.

- If $f_X(x)$ is an even function, then $\mu_X = 0$. Note that since $f_X(x) \geq 0$, then $f_X(x)$ cannot be an odd function.

- If $f_X(x)$ is symmetrical about $x = a$, such that $f_X(a - x) = f_X(x + a)$, then $\mu_X = a$.

Properties of expectation operator

The expectation operator computes a statistical average by using the density $f_X(x)$ as a weighting function. Hence, the mean μ_x can be regarded as the *center of gravity* of the density.

- If $f_X(x)$ is an even function, then $\mu_X = 0$. Note that since $f_X(x) \geq 0$, then $f_X(x)$ cannot be an odd function.

- If $f_X(x)$ is symmetrical about $x = a$, such that $f_X(a - x) = f_X(x + a)$, then $\mu_X = a$.

- The expectation operator is linear:

$$\mathbb{E}\left[\alpha\, X\left(\zeta\right) + \beta\right] = \alpha\, \mu_X + \beta$$

Properties of expectation operator

The expectation operator computes a statistical average by using the density $f_X(x)$ as a weighting function. Hence, the mean μ_x can be regarded as the *center of gravity* of the density.

- If $f_X(x)$ is an even function, then $\mu_X = 0$. Note that since $f_X(x) \geq 0$, then $f_X(x)$ cannot be an odd function.

- If $f_X(x)$ is symmetrical about $x = a$, such that $f_X(a - x) = f_X(x + a)$, then $\mu_X = a$.

- The expectation operator is linear:

$$\mathbb{E}\left[\alpha\, X\left(\zeta\right) + \beta\right] = \alpha\, \mu_X + \beta$$

- If $Y(\zeta) = g\{X\left(\zeta\right)\}$ is a **RV** obtained by transforming $X\left(\zeta\right)$ through a suitable function, the expectation of $Y(\zeta)$ is:

$$\mathbb{E}\left[Y(\zeta)\right] \triangleq \mathbb{E}\left[g\{X\left(\zeta\right)\}\right] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx$$

Moments

Recall that **mean** and **variance** can be defined as:

$$\mathbb{E}\left[X\left(\zeta\right)\right] = \mu_X = \int_{\mathbb{R}} x\, f_X(x)\, dx$$

$$\text{var}\left[X\left(\zeta\right)\right] = \sigma_X^2 = \int_{\mathbb{R}} x^2\, f_X(x)\, dx - \mu_X^2 = \mathbb{E}\left[X^2(\zeta)\right] - \mathbb{E}^2\left[X\left(\zeta\right)\right]$$

Thus, key characteristics of the **pdf** of a **RV** can be calculated if the expressions $\mathbb{E}\left[X^m(\zeta)\right],\, m \in \{1, 2\}$ are known.

Moments

Recall that **mean** and **variance** can be defined as:

$$\mathbb{E}\left[X\left(\zeta\right)\right] = \mu_X = \int_{\mathbb{R}} x\, f_X(x)\, dx$$

$$\text{var}\left[X\left(\zeta\right)\right] = \sigma_X^2 = \int_{\mathbb{R}} x^2\, f_X(x)\, dx - \mu_X^2 = \mathbb{E}\left[X^2(\zeta)\right] - \mathbb{E}^2\left[X\left(\zeta\right)\right]$$

Thus, key characteristics of the **pdf** of a **RV** can be calculated if the expressions $\mathbb{E}\left[X^m(\zeta)\right],\, m \in \{1, 2\}$ are known.

Further aspects of the **pdf** can be described by defining various **moments** of $X\left(\zeta\right)$: the m-th moment of $X\left(\zeta\right)$ is given by:

$$r_X^{(m)} \triangleq \mathbb{E}\left[X^m(\zeta)\right] = \int_{\mathbb{R}} x^m\, f_X(x)\, dx$$

Note, of course, that in general: $\mathbb{E}\left[X^m(\zeta)\right] \neq \mathbb{E}^m\left[X\left(\zeta\right)\right]$.

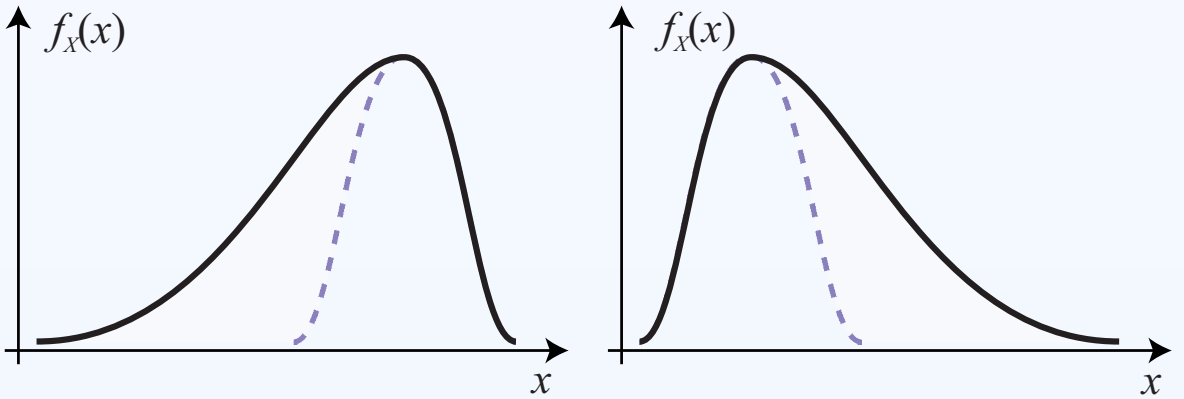
Higher-order statistics

Two important and commonly used higher-order statistics that are useful for characterising a random variable are:

Skewness characterises the degree of asymmetry of a distribution. It is a normalised third-order central moment:

$$\tilde{\kappa}_X^{(3)} \triangleq \mathbb{E}\left[\left\{\frac{X(\zeta)-\mu_X}{\sigma_X}\right\}^3\right] = \frac{1}{\sigma_X^3}\gamma_X^{(3)}$$

and is a *dimensionless* quantity.



Higher-order statistics

Two important and commonly used higher-order statistics that are useful for characterising a random variable are:

Skewness characterises the degree of asymmetry of a distribution. It is a normalised third-order central moment:

$$\tilde{\kappa}_X^{(3)} \triangleq \mathbb{E}\left[\left\{\frac{X(\zeta)-\mu_X}{\sigma_X}\right\}^3\right] = \frac{1}{\sigma_X^3}\gamma_X^{(3)}$$

and is a *dimensionless* quantity.

The **skewness** is:

$$\tilde{\kappa}_X^{(3)} = \begin{cases} < 0 & \text{if the density leans or stretches out towards the left} \\ 0 & \text{if the density is symmetric about } \mu_X \\ > 0 & \text{if the density leans or stretches out towards the right} \end{cases}$$

Higher-order statistics

Kurtosis measures relative flatness or *peakedness* of a distribution about its mean value.

Higher-order statistics

Kurtosis measures relative flatness or *peakedness* of a distribution about its mean value.

It is defined based on a normalised fourth-central moment:

$$\tilde{\kappa}_X^{(4)} \triangleq \mathbb{E}\left[\left\{\frac{X(\zeta)-\mu_X}{\sigma_X}\right\}^4\right] - 3 = \frac{1}{\sigma_X^4}\gamma_X^{(4)} - 3$$

Higher-order statistics

Kurtosis measures relative flatness or *peakedness* of a distribution about its mean value.

It is defined based on a normalised fourth-central moment:

$$\tilde{\kappa}_X^{(4)} \triangleq \mathbb{E}\left[\left\{\frac{X(\zeta)-\mu_X}{\sigma_X}\right\}^4\right] - 3 = \frac{1}{\sigma_X^4}\gamma_X^{(4)} - 3$$

This measure is relative with respect to a normal distribution, which has the property $\gamma_X^{(4)} = 3\sigma_X^4$, therefore having zero kurtosis.

Handout 4
Multiple Random Variables

Abstract

A *group* of signal observations can be modelled as a collection of random variables (RVs) that can be grouped to form a **random vector**, or **vector RV**.

- This is an extension of the concept of a RV, and generalises many of the results presented for scalar RVs.

Abstract

A *group* of signal observations can be modelled as a collection of random variables (RVs) that can be grouped to form a **random vector**, or **vector RV**.

- This is an extension of the concept of a RV, and generalises many of the results presented for scalar RVs.
- Note that each element of a **random vector** is not necessarily generated independently from a separate *experiment*.

Abstract

A *group* of signal observations can be modelled as a collection of random variables (RVs) that can be grouped to form a **random vector**, or **vector RV**.

- This is an extension of the concept of a RV, and generalises many of the results presented for scalar RVs.
- Note that each element of a **random vector** is not necessarily generated independently from a separate *experiment*.
- Random vectors also lead to the notion of the relationship between the random elements.

Abstract

A *group* of signal observations can be modelled as a collection of random variables (RVs) that can be grouped to form a **random vector**, or **vector RV**.

- This is an extension of the concept of a RV, and generalises many of the results presented for scalar RVs.
- Note that each element of a **random vector** is not necessarily generated independently from a separate *experiment*.
- Random vectors also lead to the notion of the relationship between the random elements.
- This course mainly deals with real-valued random vectors, although the concept can be extended to complex-valued random vectors.

Definition of Random Vectors

A real-valued random vector $\mathbf{X}(\zeta)$ containing N real-valued RVs, each denoted by $X_n(\zeta)$ for $n \in \mathcal{N} = \{1, \ldots, N\}$, is denoted by the column-vector:

$$\mathbf{X}(\zeta) = \begin{bmatrix} X_1(\zeta) & X_2(\zeta) & \cdots & X_N(\zeta) \end{bmatrix}^T$$

Definition of Random Vectors

A real-valued random vector $\mathbf{X}(\zeta)$ containing N real-valued RVs, each denoted by $X_n(\zeta)$ for $n \in \mathcal{N} = \{1, \ldots, N\}$, is denoted by the column-vector:

$$\mathbf{X}(\zeta) = \begin{bmatrix} X_1(\zeta) & X_2(\zeta) & \cdots & X_N(\zeta) \end{bmatrix}^T$$

A real-valued random vector can be thought as a mapping from an abstract probability space to a vector-valued, real space $\mathbb{R}^N$.

Definition of Random Vectors

A real-valued random vector $\mathbf{X}(\zeta)$ containing N real-valued RVs, each denoted by $X_n(\zeta)$ for $n \in \mathcal{N} = \{1, \ldots, N\}$, is denoted by the column-vector:

$$\mathbf{X}(\zeta) = \begin{bmatrix} X_1(\zeta) & X_2(\zeta) & \cdots & X_N(\zeta) \end{bmatrix}^T$$

A real-valued random vector can be thought as a mapping from an abstract probability space to a vector-valued, real space $\mathbb{R}^N$.

Denote a specific value for a random vector as:

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix}^T$$

Then the notation $\mathbf{X}(\zeta) \leq \mathbf{x}$ is equivalent to the event $\{X_n(\zeta) \leq x_n,\, n \in \mathcal{N}\}$.

Distribution and Density Functions

The **joint cdf** completely characterises a random vector, and is defined by:

$$F_{\mathbf{X}}(\mathbf{x}) \triangleq \Pr\left(\{X_n(\zeta) \leq x_n,\, n \in \mathcal{N}\}\right) = \Pr\left(\mathbf{X}(\zeta) \leq \mathbf{x}\right)$$

Distribution and Density Functions

The **joint cdf** completely characterises a random vector, and is defined by:

$$F_{\mathbf{X}}(\mathbf{x}) \triangleq \Pr\left(\{X_n(\zeta) \leq x_n,\, n \in \mathcal{N}\}\right) = \Pr\left(\mathbf{X}(\zeta) \leq \mathbf{x}\right)$$

A random vector can also be characterised by its **joint pdf**, which is defined by

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \lim_{\Delta\mathbf{x}\to\mathbf{0}} \frac{\Pr\left(\{x_n < X_n(\zeta) \leq x_n + \Delta x_n,\, n \in \mathcal{N}\}\right)}{\Delta x_1 \cdots \Delta x_N} \\ &= \frac{\partial}{\partial x_1}\frac{\partial}{\partial x_2}\cdots\frac{\partial}{\partial x_N} F_{\mathbf{X}}(\mathbf{x}) \end{aligned}$$

Distribution and Density Functions

The **joint cdf** completely characterises a random vector, and is defined by:

$$F_{\mathbf{X}}(\mathbf{x}) \triangleq \Pr\left(\{X_n(\zeta) \leq x_n,\, n \in \mathcal{N}\}\right) = \Pr\left(\mathbf{X}(\zeta) \leq \mathbf{x}\right)$$

A random vector can also be characterised by its **joint pdf**, which is defined by

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \lim_{\Delta\mathbf{x}\to\mathbf{0}} \frac{\Pr\left(\{x_n < X_n(\zeta) \leq x_n + \Delta x_n,\, n \in \mathcal{N}\}\right)}{\Delta x_1 \cdots \Delta x_N} \\ &= \frac{\partial}{\partial x_1}\frac{\partial}{\partial x_2}\cdots\frac{\partial}{\partial x_N} F_{\mathbf{X}}(\mathbf{x}) \end{aligned}$$

Hence, it follows:

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_N} f_{\mathbf{X}}(\mathbf{v})\, dv_N \cdots dv_1 = \int_{-\infty}^{\mathbf{x}} f_{\mathbf{X}}(\mathbf{v})\, d\mathbf{v}$$

Distribution and Density Functions

- Properties of **joint-cdf**:

$$0 \leq F_{\mathbf{X}}(\mathbf{x}) \leq 1, \quad \lim_{\mathbf{x} \to -\infty} F_{\mathbf{X}}(\mathbf{x}) = 0, \quad \lim_{\mathbf{x} \to \infty} F_{\mathbf{X}}(\mathbf{x}) = 1$$

$F_{\mathbf{X}}(\mathbf{x})$ is a monotonically increasing function of $\mathbf{x}$:

$$F_{\mathbf{X}}(\mathbf{a}) \leq F_{\mathbf{X}}(\mathbf{b}) \quad \text{if} \quad \mathbf{a} \leq \mathbf{b}$$

Distribution and Density Functions

- Properties of **joint-cdf**:

$$0 \leq F_{\mathbf{X}}(\mathbf{x}) \leq 1, \quad \lim_{\mathbf{x} \to -\infty} F_{\mathbf{X}}(\mathbf{x}) = 0, \quad \lim_{\mathbf{x} \to \infty} F_{\mathbf{X}}(\mathbf{x}) = 1$$

$F_{\mathbf{X}}(\mathbf{x})$ is a monotonically increasing function of $\mathbf{x}$:

$$F_{\mathbf{X}}(\mathbf{a}) \leq F_{\mathbf{X}}(\mathbf{b}) \quad \text{if} \quad \mathbf{a} \leq \mathbf{b}$$

- Properties of **joint-pdfs**:

$$f_{\mathbf{X}}(\mathbf{x}) \geq 0, \quad \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} = 1$$

Distribution and Density Functions

- Properties of **joint-cdf**:

$$0 \leq F_{\mathbf{X}}(\mathbf{x}) \leq 1, \quad \lim_{\mathbf{x}\to-\infty} F_{\mathbf{X}}(\mathbf{x}) = 0, \quad \lim_{\mathbf{x}\to\infty} F_{\mathbf{X}}(\mathbf{x}) = 1$$

$F_{\mathbf{X}}(\mathbf{x})$ is a monotonically increasing function of $\mathbf{x}$:

$$F_{\mathbf{X}}(\mathbf{a}) \leq F_{\mathbf{X}}(\mathbf{b}) \quad \text{if} \quad \mathbf{a} \leq \mathbf{b}$$

- Properties of **joint-pdfs**:

$$f_{\mathbf{X}}(\mathbf{x}) \geq 0, \quad \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} = 1$$

- Probability of arbitrary events; note that

$$\Pr(\mathbf{x}_1 < \mathbf{X}(\zeta) \leq \mathbf{x}_2) \neq F_{\mathbf{X}}(\mathbf{x}_2) - F_{\mathbf{X}}(\mathbf{x}_1) = \int_{\mathbf{x}_1}^{\mathbf{x}_2} f_{\mathbf{X}}(\mathbf{v})\, d\mathbf{v}$$

Distribution and Density Functions

Example ([Therrien:1992, Example 2.1, Page 20]). The joint-pdf of a random vector $\mathbf{Z}(\zeta)$ which has two elements and therefore two random variables given by $X(\zeta)$ and $Y(\zeta)$ is given by:

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

⋈

Calculate the joint-cumulative distribution function, $F_{\mathbf{Z}}(\mathbf{z})$.

Distribution and Density Functions

Example ([Therrien:1992, Example 2.1, Page 20]). The joint-pdf of a random vector $\mathbf{Z}(\zeta)$ which has two elements and therefore two random variables given by $X(\zeta)$ and $Y(\zeta)$ is given by:

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the joint-cumulative distribution function, $F_{\mathbf{Z}}(\mathbf{z})$.

SOLUTION. First note that the pdf integrates to unity since:

$$\int_{-\infty}^{\infty} f_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z} = \int_0^1 \int_0^1 \frac{1}{2}(x+3y)\, dx\, dy = \int_0^1 \frac{1}{2}\left[\frac{1}{2}x^2 + 3xy\right]_0^1 dy$$

□

Distribution and Density Functions

Example ([Therrien:1992, Example 2.1, Page 20]). The joint-pdf of a random vector $\mathbf{Z}(\zeta)$ which has two elements and therefore two random variables given by $X(\zeta)$ and $Y(\zeta)$ is given by:

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the joint-cumulative distribution function, $F_{\mathbf{Z}}(\mathbf{z})$.

SOLUTION. First note that the pdf integrates to unity since:

$$\int_{-\infty}^{\infty} f_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z} = \int_0^1 \int_0^1 \frac{1}{2}(x+3y)\, dx\, dy = \int_0^1 \frac{1}{2}\left[\frac{1}{2}x^2 + 3xy\right]_0^1 dy$$

$$= \int_0^1 \frac{1}{4} + \frac{3}{2}y\, dy = \left[\frac{y}{4} + \frac{3y^2}{4}\right]_0^1 = \frac{1}{4} + \frac{3}{4} = 1$$

□

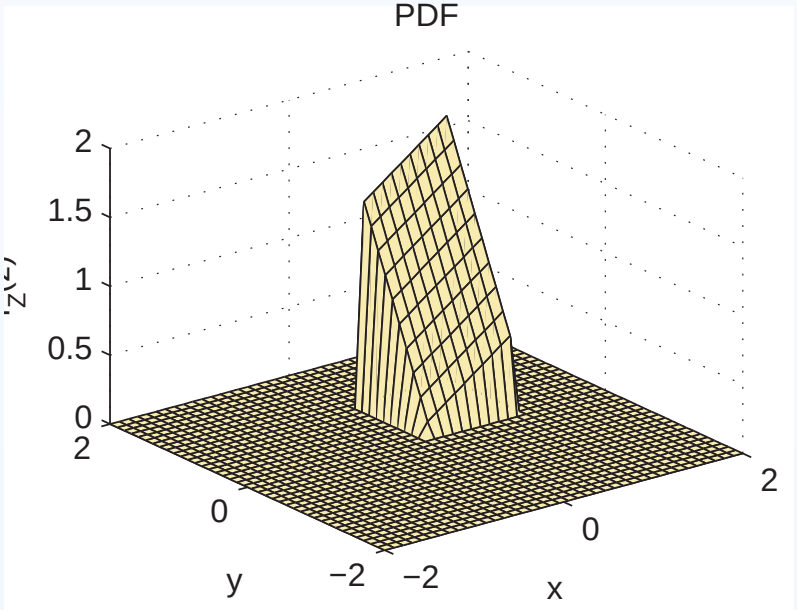
Distribution and Density Functions

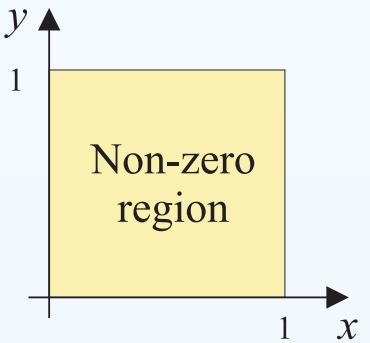
Example ([Therrien:1992, Example 2.1, Page 20]).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the joint-cumulative distribution function, $F_{\mathbf{Z}}(\mathbf{z})$.

SOLUTION. The pdf is shown here:





Region of support for pdf.

Distribution and Density Functions

Example ([Therrien:1992, Example 2.1, Page 20]).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the joint-cumulative distribution function, $F_{\mathbf{Z}}(\mathbf{z})$.

SOLUTION. For $x \leq 0$ or $y \leq 0$, $f_{\mathbf{Z}}(\mathbf{z}) = 0$, and thus $F_{\mathbf{Z}}(\mathbf{z}) = 0$.

Distribution and Density Functions

Example ([Therrien:1992, Example 2.1, Page 20]).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the joint-cumulative distribution function, $F_{\mathbf{Z}}(\mathbf{z})$.

SOLUTION. For $x \leq 0$ or $y \leq 0$, $f_{\mathbf{Z}}(\mathbf{z}) = 0$, and thus $F_{\mathbf{Z}}(\mathbf{z}) = 0$.

If $0 < x \leq 1$ *and* $0 < y \leq 1$, the cdf is given by:

$$F_{\mathbf{Z}}(\mathbf{z}) = \int_{-\infty}^{\mathbf{z}} f_{\mathbf{Z}}(\bar{\mathbf{z}})\, d\bar{\mathbf{z}} = \int_0^y \int_0^x \frac{1}{2}(\bar{x}+3\bar{y})\, d\bar{x}\, d\bar{y}$$

□

Distribution and Density Functions

Example ([Therrien:1992, Example 2.1, Page 20]).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the joint-cumulative distribution function, $F_{\mathbf{Z}}(\mathbf{z})$.

SOLUTION. For $x \leq 0$ or $y \leq 0$, $f_{\mathbf{Z}}(\mathbf{z}) = 0$, and thus $F_{\mathbf{Z}}(\mathbf{z}) = 0$.

If $0 < x \leq 1$ *and* $0 < y \leq 1$, the cdf is given by:

$$F_{\mathbf{Z}}(\mathbf{z}) = \int_{-\infty}^{\mathbf{z}} f_{\mathbf{Z}}(\bar{\mathbf{z}})\, d\bar{\mathbf{z}} = \int_0^y \int_0^x \frac{1}{2}(\bar{x}+3\bar{y})\, d\bar{x}\, d\bar{y}$$

$$= \int_0^y \frac{1}{2}\left(\frac{x^2}{2}+3x\bar{y}\right) d\bar{y} = \frac{1}{2}\left(\frac{x^2}{2}y + \frac{3xy^2}{2}\right) = \frac{xy}{4}(x+3y)$$

□

Distribution and Density Functions

Example ([Therrien:1992, Example 2.1, Page 20]).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the joint-cumulative distribution function, $F_{\mathbf{Z}}(\mathbf{z})$.

SOLUTION. For $x \leq 0$ or $y \leq 0$, $f_{\mathbf{Z}}(\mathbf{z}) = 0$, and thus $F_{\mathbf{Z}}(\mathbf{z}) = 0$.

If $0 < x \leq 1$ *and* $0 < y \leq 1$, the cdf is given by:

$$F_{\mathbf{Z}}(\mathbf{z}) = \int_{-\infty}^{\mathbf{z}} f_{\mathbf{Z}}(\bar{\mathbf{z}})\, d\bar{\mathbf{z}} = \int_0^y \int_0^x \frac{1}{2}(\bar{x}+3\bar{y})\, d\bar{x}\, d\bar{y}$$

$$= \int_0^y \frac{1}{2}\left(\frac{x^2}{2} + 3x\bar{y}\right) d\bar{y} = \frac{1}{2}\left(\frac{x^2}{2}y + \frac{3xy^2}{2}\right) = \frac{xy}{4}(x+3y)$$

Finally, if $x > 1$ or $y > 1$, the upper limit of integration for the corresponding variable becomes equal to 1.

Distribution and Density Functions

Example ([Therrien:1992, Example 2.1, Page 20]).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the joint-cumulative distribution function, $F_{\mathbf{Z}}(\mathbf{z})$.

SOLUTION. Hence, in summary, it follows:

$$F_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} 0 & x \leq 0 \quad \text{or} \quad y \leq 0 \\ \frac{xy}{4}(x+3y) & 0 < x,\, y \leq 1 \\ \frac{x}{4}(x+3) & 0 < x \leq 1,\, 1 < y \\ \frac{y}{4}(1+3y) & 0 < y \leq 1,\, 1 < x \\ 1 & 1 < x,\, y < \infty \end{cases}$$

□

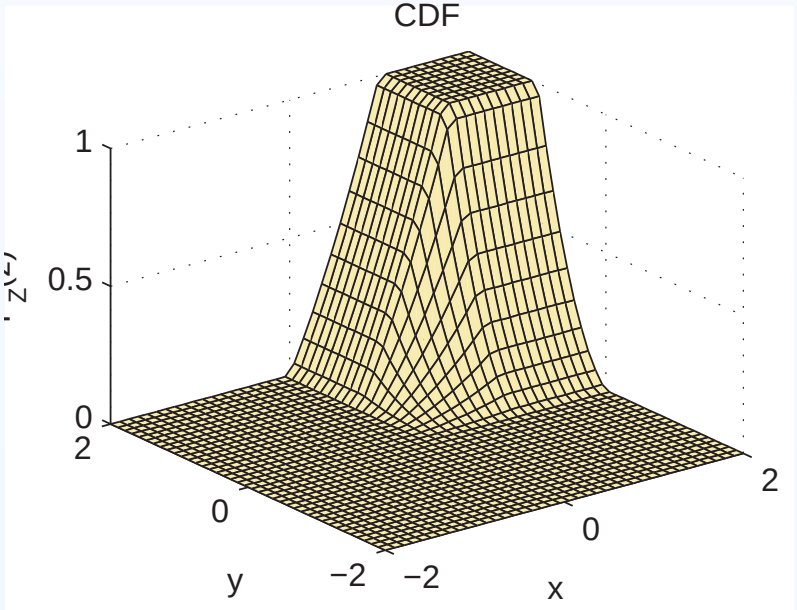
Distribution and Density Functions

Example ([Therrien:1992, Example 2.1, Page 20]).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the joint-cumulative distribution function, $F_{\mathbf{Z}}(\mathbf{z})$.

SOLUTION. The cdf is plotted here:



A plot of the cumulative distribution function.

Marginal Density Function

The joint pdf characterises the random vector; the so-called **marginal pdf** describes a subset of RVs from the random vector.

Marginal Density Function

The joint pdf characterises the random vector; the so-called **marginal pdf** describes a subset of RVs from the random vector.

Let $\mathbf{k}$ be an M-dimensional vector containing unique indices to elements in the N-dimensional random vector $\mathbf{X}(\zeta)$,

$$\mathbf{k} = \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_M \end{bmatrix}$$

Marginal Density Function

The joint pdf characterises the random vector; the so-called **marginal pdf** describes a subset of RVs from the random vector.

Let $\mathbf{k}$ be an M-dimensional vector containing unique indices to elements in the N-dimensional random vector $\mathbf{X}(\zeta)$,

Now define a M-dimensional random vector, $\mathbf{X}_{\mathbf{k}}(\zeta)$, that contains the M random variables which are components of $\mathbf{X}(\zeta)$ and indexed by the elements of $\mathbf{k}$. In other-words, if

$$\mathbf{k} = \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_M \end{bmatrix} \quad \text{then} \quad \mathbf{X}_{\mathbf{k}}(\zeta) = \begin{bmatrix} X_{k_1}(\zeta) \\ X_{k_2}(\zeta) \\ \vdots \\ X_{k_M}(\zeta) \end{bmatrix}$$

Marginal Density Function

The **marginal pdf** is then given by:

$$f_{\mathbf{X_k}}\left(\mathbf{x_k}\right) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{N-M \text{ integrals}} f_{\mathbf{X}}\left(\mathbf{x}\right) d\mathbf{x_{-k}}$$

Marginal Density Function

The **marginal pdf** is then given by:

$$f_{\mathbf{X_k}}\left(\mathbf{x_k}\right) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{N-M \text{ integrals}} f_{\mathbf{X}}\left(\mathbf{x}\right) d\mathbf{x_{-k}}$$

A special case is the **marginal pdf** describing the individual RV X_j:

$$f_{X_j}\left(x_j\right) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{N-1 \text{ integrals}} f_{\mathbf{X}}\left(\mathbf{x}\right) dx_1 \cdots dx_{j-1} dx_{j+1} \cdots dx_N$$

Marginal Density Function

The **marginal pdf** is then given by:

$$f_{\mathbf{X_k}}\left(\mathbf{x_k}\right) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{N-M \text{ integrals}} f_{\mathbf{X}}\left(\mathbf{x}\right) d\mathbf{x_{-k}}$$

A special case is the **marginal pdf** describing the individual RV X_j:

$$f_{X_j}\left(x_j\right) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{N-1 \text{ integrals}} f_{\mathbf{X}}\left(\mathbf{x}\right) dx_1 \cdots dx_{j-1} dx_{j+1} \cdots dx_N$$

Marginal pdfs will become particular useful when dealing with Bayesian parameter estimation later in the course.

Marginal Density Function

Example (Marginalisation). The joint-pdf of a random vector $\mathbf{Z}(\zeta)$ which has two elements and therefore two random variables given by $X(\zeta)$ and $Y(\zeta)$ is given by:

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

⋈

Calculate the marginal-pdfs, $f_X(x)$ and $f_Y(y)$, and their corresponding marginal-cdfs, $F_X(x)$ and $F_Y(y)$.

Marginal Density Function

Example (Marginalisation). The joint-pdf of a random vector $\mathbf{Z}(\zeta)$ which has two elements and therefore two random variables given by $X(\zeta)$ and $Y(\zeta)$ is given by:

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the marginal-pdfs, $f_X(x)$ and $f_Y(y)$, and their corresponding marginal-cdfs, $F_X(x)$ and $F_Y(y)$.

SOLUTION. By definition:

$$f_X(x) = \int_{\mathbb{R}} f_{\mathbf{Z}}(\mathbf{z})\, dy$$

$$f_Y(y) = \int_{\mathbb{R}} f_{\mathbf{Z}}(\mathbf{z})\, dx$$

Marginal Density Function

Example (Marginalisation).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the marginal-pdfs, $f_X(x)$ and $f_Y(y)$, and their corresponding marginal-cdfs, $F_X(x)$ and $F_Y(y)$.

SOLUTION. Taking $f_X(x)$, then:

$$f_X(x) = \begin{cases} \frac{1}{2}\int_0^1 (x+3y)\,dy & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

□

Marginal Density Function

Example (Marginalisation).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the marginal-pdfs, $f_X(x)$ and $f_Y(y)$, and their corresponding marginal-cdfs, $F_X(x)$ and $F_Y(y)$.

SOLUTION. Taking $f_X(x)$, then:

$$f_X(x) = \begin{cases} \frac{1}{2}\int_0^1 (x+3y)\,dy & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

which after a simple integration gives:

$$f_X(x) = \begin{cases} \frac{1}{2}\left(x+\frac{3}{2}\right) & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Marginal Density Function

Example (Marginalisation).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the marginal-pdfs, $f_X(x)$ and $f_Y(y)$, and their corresponding marginal-cdfs, $F_X(x)$ and $F_Y(y)$.

SOLUTION. The cdf, $F_X(x)$, is thus given by:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du = \begin{cases} 0 & x \leq 0 \\ \frac{1}{2}\int_0^x \left(u+\frac{3}{2}\right) du & 0 \leq x \leq 1 \\ \frac{1}{2}\int_0^1 \left(u+\frac{3}{2}\right) du & x > 1 \end{cases}$$

□

Marginal Density Function

Example (Marginalisation).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the marginal-pdfs, $f_X(x)$ and $f_Y(y)$, and their corresponding marginal-cdfs, $F_X(x)$ and $F_Y(y)$.

SOLUTION. The cdf, $F_X(x)$, is thus given by:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du = \begin{cases} 0 & x \leq 0 \\ \frac{1}{2}\int_0^x \left(u + \frac{3}{2}\right) du & 0 \leq x \leq 1 \\ \frac{1}{2}\int_0^1 \left(u + \frac{3}{2}\right) du & x > 1 \end{cases}$$

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ \frac{x}{4}(x+3) & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$$

Marginal Density Function

Example (Marginalisation).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the marginal-pdfs, $f_X(x)$ and $f_Y(y)$, and their corresponding marginal-cdfs, $F_X(x)$ and $F_Y(y)$.

SOLUTION. Similarly, it can be shown that:

$$f_Y(y) = \begin{cases} \frac{1}{2}\left(\frac{1}{2}+3y\right) & 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and

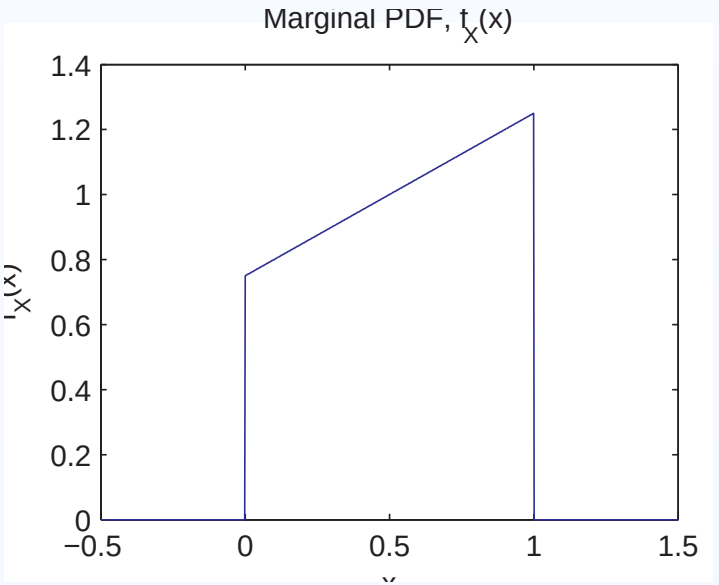
$$F_Y(y) = \begin{cases} 0 & y \leq 0 \\ \frac{y}{4}(1+3y) & 0 \leq y \leq 1 \\ 1 & y > 1 \end{cases}$$

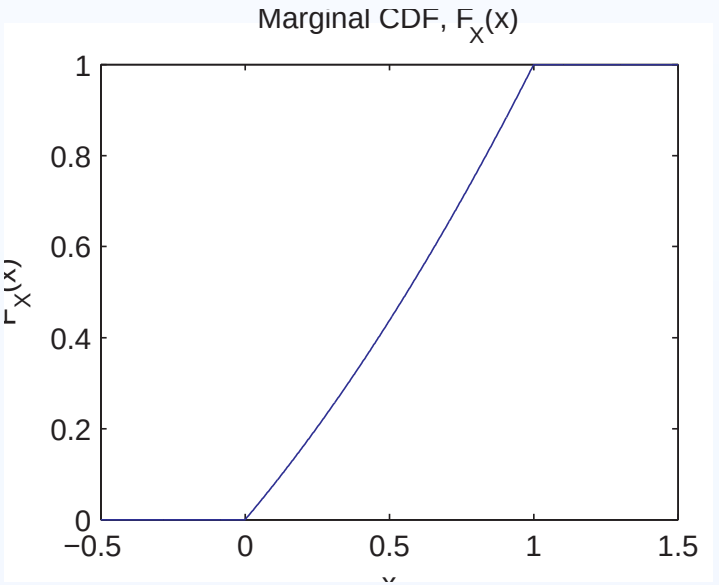
Marginal Density Function

Example (Marginalisation).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

SOLUTION. The marginal-pdfs and cdfs are shown below.





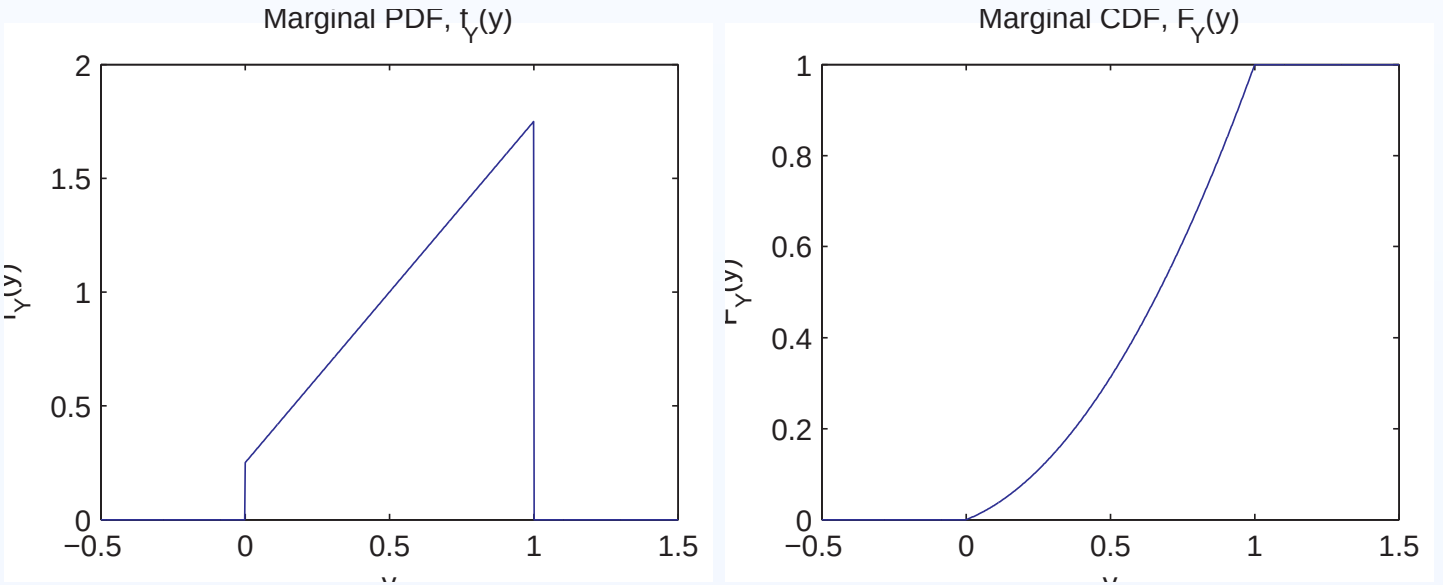
The marginal-pdf, $f_X(x)$, and cdf, $F_X(x)$, for the RV, $X(\zeta)$.

Marginal Density Function

Example (Marginalisation).

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \leq x,\, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

SOLUTION. The marginal-pdfs and cdfs are shown below.



The marginal-pdf, $f_Y(y)$, and cdf, $F_Y(y)$, for the RV, $Y(\zeta)$.

Independence

Two random variables, $X_1(\zeta)$ and $X_2(\zeta)$ are **independent** if the events $\{X_1(\zeta) \leq x_1\}$ and $\{X_2(\zeta) \leq x_2\}$ are jointly independent; that is, the events do not influence one another, and

$$\Pr\left(X_1(\zeta) \leq x_1,\, X_2(\zeta) \leq x_2\right) = \Pr\left(X_1(\zeta) \leq x_1\right) \Pr\left(X_2(\zeta) \leq x_2\right)$$

Independence

Two random variables, $X_1(\zeta)$ and $X_2(\zeta)$ are **independent** if the events $\{X_1(\zeta) \leq x_1\}$ and $\{X_2(\zeta) \leq x_2\}$ are jointly independent; that is, the events do not influence one another, and

$$\Pr\left(X_1(\zeta) \leq x_1,\, X_2(\zeta) \leq x_2\right) = \Pr\left(X_1(\zeta) \leq x_1\right) \Pr\left(X_2(\zeta) \leq x_2\right)$$

This then implies that

$$F_{X_1,X_2}\left(x_1,\, x_2\right) = F_{X_1}\left(x_1\right) F_{X_2}\left(x_2\right)$$
$$f_{X_1,X_2}\left(x_1,\, x_2\right) = f_{X_1}\left(x_1\right) f_{X_2}\left(x_2\right)$$

Conditionals and Bayes's

The notion of joint probabilities and pdf also leads to the notion of conditional probabilities; what is the probability of a random vector $\mathbf{Y}(\zeta)$, given the random vector $\mathbf{X}(\zeta)$.

Conditionals and Bayes's

The notion of joint probabilities and pdf also leads to the notion of conditional probabilities; what is the probability of a random vector $\mathbf{Y}(\zeta)$, given the random vector $\mathbf{X}(\zeta)$.

The **conditional pdf** of $\mathbf{Y}(\zeta)$ given $\mathbf{X}(\zeta)$ is defined as:

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) = \frac{f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})}$$

Conditionals and Bayes's

The notion of joint probabilities and pdf also leads to the notion of conditional probabilities; what is the probability of a random vector $\mathbf{Y}(\zeta)$, given the random vector $\mathbf{X}(\zeta)$.

The **conditional pdf** of $\mathbf{Y}(\zeta)$ given $\mathbf{X}(\zeta)$ is defined as:

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) = \frac{f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})}$$

If the random vectors $\mathbf{X}(\zeta)$ and $\mathbf{Y}(\zeta)$ are independent, then the conditional pdf must be identical to the unconditional pdf: $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) = f_{\mathbf{Y}}(\mathbf{y})$. Hence, it follows that:

$$f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}}(\mathbf{y})$$

Conditionals and Bayes's

Since

$$f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} \mid \mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{YX}}(\mathbf{y}, \mathbf{x})$$

it follows

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} \mid \mathbf{y}) = \frac{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{f_{\mathbf{Y}}(\mathbf{y})}$$

Conditionals and Bayes's

Since

$$f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} \mid \mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{YX}}(\mathbf{y}, \mathbf{x})$$

it follows

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} \mid \mathbf{y}) = \frac{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{f_{\mathbf{Y}}(\mathbf{y})}$$

Since $f_{\mathbf{Y}}(\mathbf{y})$ can be expressed as:

$$f_{\mathbf{Y}}(\mathbf{y}) = \int_{\mathbb{R}} f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})\, d\mathbf{x} = \int_{\mathbb{R}} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$$

then it follows

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} \mid \mathbf{y}) = \frac{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{\int_{\mathbb{R}} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}}$$

Statistical Description

Statistical averages are more manageable, but less of a complete description of random vectors.

Statistical Description

Statistical averages are more manageable, but less of a complete description of random vectors.

With care, it is possible to extend many of the statistical descriptors for scalar RVs to random vectors.

Statistical Description

Statistical averages are more manageable, but less of a complete description of random vectors.

With care, it is possible to extend many of the statistical descriptors for scalar RVs to random vectors.

However, it is important to understand that multiple RVs leads to the notion of measuring their interaction or dependence. This concept is useful in abstract, but also when dealing with stochastic processes or time-series.

Statistical Description

Mean vector The **mean vector** is the first-moment of the random vector, and is given by:

$$\boldsymbol{\mu}_{\mathbf{X}} = \mathbb{E}\left[\mathbf{X}\left(\zeta\right)\right] = \begin{bmatrix} \mathbb{E}\left[X_1(\zeta)\right] \\ \vdots \\ \mathbb{E}\left[X_N(\zeta)\right] \end{bmatrix} = \begin{bmatrix} \mu_{X_1} \\ \vdots \\ \mu_{X_N} \end{bmatrix}$$

Statistical Description

Mean vector The **mean vector** is the first-moment of the random vector, and is given by:

$$\boldsymbol{\mu}_{\mathbf{X}} = \mathbb{E}\left[\mathbf{X}\left(\zeta\right)\right] = \begin{bmatrix} \mathbb{E}\left[X_1(\zeta)\right] \\ \vdots \\ \mathbb{E}\left[X_N(\zeta)\right] \end{bmatrix} = \begin{bmatrix} \mu_{X_1} \\ \vdots \\ \mu_{X_N} \end{bmatrix}$$

Correlation Matrix The second-order moments of the random vector describe the spread of the distribution. The **autocorrelation matrix** is defined by:

$$\mathbf{R}_{\mathbf{X}} \triangleq \mathbb{E}\left[\mathbf{X}\left(\zeta\right)\mathbf{X}^{H}\left(\zeta\right)\right] = \begin{bmatrix} r_{X_1X_1} & \cdots & r_{X_1X_N} \\ \vdots & \ddots & \vdots \\ r_{X_NX_1} & \cdots & r_{X_NX_N} \end{bmatrix}$$

Statistical Description

Correlation Matrix The diagonal terms

$$r_{X_i X_i} \triangleq \mathbb{E}\left[\left|X_i(\zeta)\right|^2\right], \quad i \in \{1, \ldots, N\}$$

are the second-order moments of each of the RVs, $X_i(\zeta)$.

Statistical Description

Correlation Matrix The diagonal terms

$$r_{X_i X_i} \triangleq \mathbb{E}\left[|X_i(\zeta)|^2\right], \quad i \in \{1, \ldots, N\}$$

are the second-order moments of each of the RVs, $X_i(\zeta)$.

The off-diagonal terms

$$r_{X_i X_j} \triangleq \mathbb{E}\left[X_i(\zeta) X_j^*(\zeta)\right] = r_{X_j X_i}^*, \quad i \neq j$$

measure the **correlation**, or statistical similarity between the RVs $X_i(\zeta)$ and $X_j(\zeta)$.

Statistical Description

Correlation Matrix The diagonal terms

$$r_{X_i X_i} \triangleq \mathbb{E}\left[|X_i(\zeta)|^2\right], \quad i \in \{1, \ldots, N\}$$

are the second-order moments of each of the RVs, $X_i(\zeta)$.

The off-diagonal terms

$$r_{X_i X_j} \triangleq \mathbb{E}\left[X_i(\zeta) X_j^*(\zeta)\right] = r_{X_j X_i}^*, \quad i \neq j$$

measure the **correlation**, or statistical similarity between the RVs $X_i(\zeta)$ and $X_j(\zeta)$.

If the $X_i(\zeta)$ and $X_j(\zeta)$ are **orthogonal** then their **correlation** is zero:

$$r_{X_i X_j} = \mathbb{E}\left[X_i(\zeta) X_j^*(\zeta)\right] = 0, \quad i \neq j$$

Statistical Description

Covariance Matrix The **autocovariance matrix** is defined by:

$$\mathbf{\Gamma_X} \triangleq \mathbb{E}\left[\left(\mathbf{X}\left(\zeta\right)-\boldsymbol{\mu}_{\mathbf{X}}\right)\left(\mathbf{X}\left(\zeta\right)-\boldsymbol{\mu}_{\mathbf{X}}\right)^{H}\right] = \begin{bmatrix} \gamma_{X_1X_1} & \cdots & \gamma_{X_1X_N} \\ \vdots & \ddots & \cdots \\ \gamma_{X_NX_1} & \cdots & \gamma_{X_NX_N} \end{bmatrix}$$

Example (Valid correlation matrix). Determine whether the following is a valid correlation matrix:

$$\mathbf{R}_X = \begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix}$$

SOLUTION. This is not a valid correlation matrix as it is not symmetric, which is a requirement of a valid correlation matrix. In otherwords, $\mathbf{R}_X^T \neq \mathbf{R}_X$.

Statistical Description

Covariance Matrix The **autocovariance matrix** is defined by:

$$\mathbf{\Gamma_X} \triangleq \mathbb{E}\left[\left(\mathbf{X}\left(\zeta\right)-\boldsymbol{\mu}_{\mathbf{X}}\right)\left(\mathbf{X}\left(\zeta\right)-\boldsymbol{\mu}_{\mathbf{X}}\right)^{H}\right] = \begin{bmatrix} \gamma_{X_1X_1} & \cdots & \gamma_{X_1X_N} \\ \vdots & \ddots & \dots \\ \gamma_{X_NX_1} & \cdots & \gamma_{X_NX_N} \end{bmatrix}$$

The diagonal terms

$$\gamma_{X_iX_i} \triangleq \sigma^2_{X_i} = \mathbb{E}\left[\left|X_i(\zeta)-\mu_{X_i}\right|^2\right], \quad i \in \{1,\dots,N\}$$

are the **variances** of each of the RVs, $X_i(\zeta)$.

Example (Valid correlation matrix). Determine whether the following is a valid correlation matrix:

$$\mathbf{R}_X = \begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix}$$

Statistical Description

Covariance Matrix The off-diagonal terms

$$\begin{aligned}\gamma_{X_iX_j} &\triangleq \mathbb{E}\left[\left(X_i(\zeta)-\mu_{X_i}\right)\left(X_j(\zeta)-\mu_{X_j}\right)^*\right]\\ &= r_{X_iX_j}-\mu_{X_i}\mu^*_{X_j}=\gamma^*_{X_jX_i}, \quad i\neq j\end{aligned}$$

measure the **covariance** $X_i(\zeta)$ and $X_j(\zeta)$.

Example (Valid correlation matrix). Determine whether the following is a valid correlation matrix:

$$\mathbf{R}_X=\begin{bmatrix}0 & 1\\ 2 & 3\end{bmatrix}$$

SOLUTION. This is not a valid correlation matrix as it is not symmetric, which is a requirement of a valid correlation matrix. In otherwords, $\mathbf{R}_X^T\neq\mathbf{R}_X$.

Statistical Description

Covariance Matrix The off-diagonal terms

$$\begin{aligned}\gamma_{X_i X_j} &\triangleq \mathbb{E}\left[\left(X_i(\zeta)-\mu_{X_i}\right)\left(X_j(\zeta)-\mu_{X_j}\right)^*\right] \\ &= r_{X_i X_j} - \mu_{X_i}\mu^*_{X_j} = \gamma^*_{X_j X_i}, \quad i \neq j\end{aligned}$$

measure the **covariance** $X_i(\zeta)$ and $X_j(\zeta)$.

It should also be noticed that the **covariance** and **correlation** matrices are positive semidefinite; that is, they satisfy the relations:

$$\mathbf{a}^H \mathbf{R_X a} \geq 0$$

$$\mathbf{a}^H \mathbf{\Gamma_X a} \geq 0$$

for any complex vector $\mathbf{a}$.

Example (Valid correlation matrix). Determine whether the

Statistical Description

Covariance Matrix Theorem (Positive semi-definiteness of correlation matrix). PRO

There are various methods to demonstrate this, but one is as follows. Consider the sum of RVs:

$$Y(\zeta) = \sum_{n=1}^{N} a_n X_n(\zeta) = \mathbf{a}^T \mathbf{X}(\zeta)$$

The variance of $Y(\zeta)$ must, by definition, be positive, as must its second moment. Considering the second moment, then:

$$\begin{aligned}\mathbb{E}\left[Y^2(\zeta)\right] &= \mathbb{E}\left[\mathbf{a}^T \mathbf{X}(\zeta) \mathbf{X}(\zeta)^T \mathbf{a}\right] \\ &= \mathbf{a}^T \mathbb{E}\left[\mathbf{X}(\zeta) \mathbf{X}(\zeta)^T\right] \mathbf{a} = \mathbf{a}^T \mathbf{R}_X \mathbf{a} \geq 0\end{aligned}$$

Example (Valid correlation matrix). Determine whether the following is a valid correlation matrix:

$$\mathbf{R}_X = \begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix}$$

Statistical Description

The autocorrelation and autocovariance matrices are related, and it can easily be seen that:

$$\mathbf{\Gamma_X} \triangleq \mathbb{E}\left[\left[\mathbf{X}\left(\zeta\right) - \boldsymbol{\mu}_\mathbf{X}\right]\left[\mathbf{X}\left(\zeta\right) - \boldsymbol{\mu}_\mathbf{X}\right]^H\right] = \mathbf{R_X} - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^H$$

In fact, if $\boldsymbol{\mu}_\mathbf{X} = 0$, then $\mathbf{\Gamma_X} = \mathbf{R_X}$.

If the random variables $X_i(\zeta)$ and $X_j(\zeta)$ are **independent**, then they are also **uncorrelated** since:

$$\begin{aligned} r_{X_iX_j} &= \mathbb{E}\left[X_i(\zeta)\, X_j(\zeta)^*\right] = \mathbb{E}\left[X_i(\zeta)\right]\mathbb{E}\left[X_j^*(\zeta)\right] \\ &= \mu_{X_i}\mu_{X_j}^* \quad \Rightarrow \quad \gamma_{X_iX_j} = 0 \end{aligned}$$

Statistical Description

The autocorrelation and autocovariance matrices are related, and it can easily be seen that:

$$\mathbf{\Gamma_X} \triangleq \mathbb{E}\left[\left[\mathbf{X}\left(\zeta\right) - \boldsymbol{\mu}_\mathbf{X}\right]\left[\mathbf{X}\left(\zeta\right) - \boldsymbol{\mu}_\mathbf{X}\right]^H\right] = \mathbf{R_X} - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^H$$

In fact, if $\boldsymbol{\mu}_\mathbf{X} = 0$, then $\mathbf{\Gamma_X} = \mathbf{R_X}$.

If the random variables $X_i(\zeta)$ and $X_j(\zeta)$ are **independent**, then they are also **uncorrelated** since:

$$\begin{aligned} r_{X_iX_j} &= \mathbb{E}\left[X_i(\zeta)\, X_j(\zeta)^*\right] = \mathbb{E}\left[X_i(\zeta)\right]\mathbb{E}\left[X_j^*(\zeta)\right] \\ &= \mu_{X_i}\mu_{X_j}^* \quad \Rightarrow \quad \gamma_{X_iX_j} = 0 \end{aligned}$$

Note, however, that uncorrelatedness does not imply independence, unless the RVs are jointly-Gaussian.

Statistical Description

Cross-correlation is defined as

$$\mathbf{R_{XY}} \triangleq \mathbb{E}\left[\mathbf{X}\left(\zeta\right)\mathbf{Y}^{H}\left(\zeta\right)\right] = \begin{bmatrix} \mathbb{E}\left[X_1(\zeta)Y_1^*(\zeta)\right] & \cdots & \mathbb{E}\left[X_1(\zeta)Y_M^*(\zeta)\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}\left[X_N(\zeta)Y_1^*(\zeta)\right] & \cdots & \mathbb{E}\left[X_N(\zeta)Y_M^*(\zeta)\right] \end{bmatrix}$$

Statistical Description

Cross-correlation is defined as

$$\mathbf{R_{XY}} \triangleq \mathbb{E}\left[\mathbf{X}\left(\zeta\right)\mathbf{Y}^{H}\left(\zeta\right)\right] = \begin{bmatrix} \mathbb{E}\left[X_1(\zeta)Y_1^*(\zeta)\right] & \cdots & \mathbb{E}\left[X_1(\zeta)Y_M^*(\zeta)\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}\left[X_N(\zeta)Y_1^*(\zeta)\right] & \cdots & \mathbb{E}\left[X_N(\zeta)Y_M^*(\zeta)\right] \end{bmatrix}$$

Cross-covariance is defined as

$$\begin{aligned} \mathbf{\Gamma_{XY}} &\triangleq \mathbb{E}\left[\left\{\mathbf{X}\left(\zeta\right)-\boldsymbol{\mu}_{\mathbf{X}}\right\}\left\{\mathbf{Y}\left(\zeta\right)-\boldsymbol{\mu}_{\mathbf{Y}}\right\}^{H}\right] \\ &= \mathbf{R_{XY}} - \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\mu}_{\mathbf{Y}}^{H} \end{aligned}$$

Statistical Description

Cross-correlation is defined as

$$\mathbf{R_{XY}} \triangleq \mathbb{E}\left[\mathbf{X}\left(\zeta\right)\mathbf{Y}^{H}\left(\zeta\right)\right] = \begin{bmatrix} \mathbb{E}\left[X_1(\zeta)Y_1^*(\zeta)\right] & \cdots & \mathbb{E}\left[X_1(\zeta)Y_M^*(\zeta)\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}\left[X_N(\zeta)Y_1^*(\zeta)\right] & \cdots & \mathbb{E}\left[X_N(\zeta)Y_M^*(\zeta)\right] \end{bmatrix}$$

Cross-covariance is defined as

$$\begin{aligned} \mathbf{\Gamma_{XY}} &\triangleq \mathbb{E}\left[\left\{\mathbf{X}\left(\zeta\right)-\boldsymbol{\mu}_{\mathbf{X}}\right\}\left\{\mathbf{Y}\left(\zeta\right)-\boldsymbol{\mu}_{\mathbf{Y}}\right\}^{H}\right] \\ &= \mathbf{R_{XY}} - \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\mu}_{\mathbf{Y}}^{H} \end{aligned}$$

- Uncorrelated if $\mathbf{\Gamma_{XY}} = 0 \quad \Rightarrow \quad \mathbf{R_{XY}} = \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\mu}_{\mathbf{Y}}^{H}.$
- Orthogonal if $\mathbf{R_{XY}} = 0.$

Probability Transformation Rule

Theorem (Probability Transformation Rule). The set of random variables $\mathbf{X}(\zeta) = \{X_n(\zeta),\ n \in \mathcal{N}\}$ are transformed to a new set of RVs, $\mathbf{Y}(\zeta) = \{Y_n(\zeta),\ n \in \mathcal{N}\}$, using the transformations:

$$Y_n(\zeta) = g_n(\mathbf{X}(\zeta)), \quad n \in \mathcal{N}$$

Probability Transformation Rule

Theorem (Probability Transformation Rule). The set of random variables $\mathbf{X}(\zeta) = \{X_n(\zeta),\ n \in \mathcal{N}\}$ are transformed to a new set of RVs, $\mathbf{Y}(\zeta) = \{Y_n(\zeta),\ n \in \mathcal{N}\}$, using the transformations:

$$Y_n(\zeta) = g_n(\mathbf{X}(\zeta)), \quad n \in \mathcal{N}$$

Assuming M-real vector-roots of the equation $\mathbf{y} = \mathbf{g}(\mathbf{x})$ by $\{\mathbf{x}_m,\ m \in \mathcal{M}\}$,

$$\mathbf{y} = \mathbf{g}(\mathbf{x}_1) = \cdots = \mathbf{g}(\mathbf{x}_M)$$

then the joint-pdf of $\mathbf{Y}(\zeta)$ in terms of (i. t. o.) the joint-pdf of $\mathbf{X}(\zeta)$is:

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{m=1}^{M} \frac{f_{\mathbf{X}}(\mathbf{x}_m)}{|J(\mathbf{x}_m)|}$$

The Jacobian is defined in the notes, but is the usual definition!

Polar Transformation

Consider the transformation from the random vector $\mathbf{C}(\zeta) = [X(\zeta), Y(\zeta)]^T$ to $\mathbf{P}(\zeta) = [r(\zeta), \theta(\zeta)]^T$, where

$$r(\zeta) = \sqrt{X^2(\zeta) + Y^2(\zeta)}$$

$$\theta(\zeta) = \arctan \frac{Y(\zeta)}{X(\zeta)}$$

Polar Transformation

Consider the transformation from the random vector $\mathbf{C}(\zeta) = [X(\zeta), Y(\zeta)]^T$ to $\mathbf{P}(\zeta) = [r(\zeta), \theta(\zeta)]^T$, where

$$r(\zeta) = \sqrt{X^2(\zeta) + Y^2(\zeta)}$$

$$\theta(\zeta) = \arctan \frac{Y(\zeta)}{X(\zeta)}$$

The Jacobian is given by:

$$J_{\mathbf{g}}(\mathbf{c}) = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix}^{-1} = \frac{1}{r}$$

Thus, it follows that:

$$f_{R,\Theta}(r,\theta) = r f_{XY}(r\cos\theta, r\sin\theta)$$

Auxiliary Variables

The density of a RV that is *one* function $Z(\zeta) = g(X(\zeta), Y(\zeta))$ of two RVs can be determined by choosing a **auxiliary variable.**

Auxiliary Variables

The density of a RV that is *one* function $Z(\zeta) = g(X(\zeta), Y(\zeta))$ of two RVs can be determined by choosing a **auxiliary variable.**

$$f_Z(z) = \int_{\mathbb{R}} f_{WZ}(w, z)\, dw = \sum_{m=1}^{M} \int_{\mathbb{R}} \frac{f_{\mathbf{XY}}(x_m, y_m)}{|J(x_m, y_m)|}\, dw$$

Auxiliary Variables

The density of a RV that is *one* function $Z(\zeta) = g(X(\zeta), Y(\zeta))$ of two RVs can be determined by choosing a **auxiliary variable**.

$$f_Z(z) = \int_{\mathbb{R}} f_{WZ}(w, z)\, dw = \sum_{m=1}^{M} \int_{\mathbb{R}} \frac{f_{\mathbf{XY}}(x_m, y_m)}{|J(x_m, y_m)|}\, dw$$

Example (Sum of two RVs). If $X(\zeta)$ and $Y(\zeta)$ have joint-pdf $f_{XY}(x, y)$, find the pdf of the RV $Z(\zeta) = aX(\zeta) + bY(\zeta)$.

Auxiliary Variables

The density of a RV that is *one* function $Z(\zeta) = g(X(\zeta), Y(\zeta))$ of two RVs can be determined by choosing a **auxiliary variable**.

$$f_Z(z) = \int_{\mathbb{R}} f_{WZ}(w, z)\, dw = \sum_{m=1}^{M} \int_{\mathbb{R}} \frac{f_{\mathbf{XY}}(x_m, y_m)}{|J(x_m, y_m)|}\, dw$$

Example (Sum of two RVs). If $X(\zeta)$ and $Y(\zeta)$ have joint-pdf $f_{XY}(x, y)$, find the pdf of the RV $Z(\zeta) = aX(\zeta) + bY(\zeta)$.

SOLUTION. Use as the auxiliary variable the function $W(\zeta) = Y(\zeta)$. The system $z = ax + by$, $w = y$ has a single solution at $x = \frac{z-bw}{a}$, $y = w$.

Thus:

$$f_Z(z) = \frac{1}{|a|} \int_{\mathbb{R}} f_{XY}\left(\frac{z - bw}{a}, w\right) dw$$

Multivariate Gaussian Density Function

Gaussian random vectors play a very important role in the design and analysis of signal processing systems. A Gaussian random vector is characterised by a multivariate Normal or Gaussian density function.

Multivariate Gaussian Density Function

Gaussian random vectors play a very important role in the design and analysis of signal processing systems. A Gaussian random vector is characterised by a multivariate Normal or Gaussian density function.

For a *real* random vector, this density function has the form:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{\Gamma_X}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})^T \mathbf{\Gamma}_{\mathbf{X}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})\right]$$

where N is the dimension of $\mathbf{X}(\zeta)$, and $\mathbf{X}(\zeta)$ has mean $\boldsymbol{\mu}_{\mathbf{X}}$ and covariance $\mathbf{\Gamma_X}$. It is often denoted as:

$$f_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \middle| \boldsymbol{\mu}_{\mathbf{X}}, \mathbf{\Gamma_X}\right)$$

Multivariate Gaussian Density Function

The normal distribution is a useful model of a random vector because of its many important properties.

1. $f_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \middle| \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Gamma}_{\mathbf{X}}\right)$ is completely specified by its mean $\boldsymbol{\mu}_{\mathbf{X}}$ and covariance $\boldsymbol{\Gamma}_{\mathbf{X}}$.

2. If the components of $\mathbf{X}(\zeta)$ are mutually uncorrelated, then they are also independent.

Multivariate Gaussian Density Function

The normal distribution is a useful model of a random vector because of its many important properties.

1. $f_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \,\middle|\, \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Gamma}_{\mathbf{X}}\right)$ is completely specified by its mean $\boldsymbol{\mu}_{\mathbf{X}}$ and covariance $\boldsymbol{\Gamma}_{\mathbf{X}}$.

2. If the components of $\mathbf{X}(\zeta)$ are mutually uncorrelated, then they are also independent.

3. A linear transformation of a normal random vector is also normal.

   This is a particularly useful, since the output of a linear system subject to a Gaussian input is also Gaussian.

Multivariate Gaussian Density Function

The normal distribution is a useful model of a random vector because of its many important properties.

1. $f_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \,\middle|\, \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Gamma}_{\mathbf{X}}\right)$ is completely specified by its mean $\boldsymbol{\mu}_{\mathbf{X}}$ and covariance $\boldsymbol{\Gamma}_{\mathbf{X}}$.

2. If the components of $\mathbf{X}(\zeta)$ are mutually uncorrelated, then they are also independent.

3. A linear transformation of a normal random vector is also normal.

   This is a particularly useful, since the output of a linear system subject to a Gaussian input is also Gaussian.

4. If $\mathbf{X}(\zeta)$ and $\mathbf{Y}(\zeta)$ are *jointly*-Gaussian, then so are their *marginal*-distributions, and their *conditional*-distributions.

Handout 5
Estimation Theory

Introduction

- Thus far, have assumed that either the pdf or statistical values, such as mean, covariance, or higher order statistics, associated with a problem are fully known.

Introduction

- Thus far, have assumed that either the pdf or statistical values, such as mean, covariance, or higher order statistics, associated with a problem are fully known.

- In most practical applications, this is the exception rather than the rule.

Introduction

- Thus far, have assumed that either the pdf or statistical values, such as mean, covariance, or higher order statistics, associated with a problem are fully known.
- In most practical applications, this is the exception rather than the rule.
- The properties and parameters of random events must be obtained by collecting and analysing finite set of measurements.

Introduction

- Thus far, have assumed that either the pdf or statistical values, such as mean, covariance, or higher order statistics, associated with a problem are fully known.
- In most practical applications, this is the exception rather than the rule.
- The properties and parameters of random events must be obtained by collecting and analysing finite set of measurements.
- This handout will consider the problem of **Parameter Estimation**. This refers to the estimation of a parameter that is fixed, but is unknown.

Properties of Estimators

Consider the set of N observations, $\mathcal{X} = \{x[n]\}_0^{N-1}$, from a *random experiment*; suppose they are used to estimate a parameter θ of the process using some function:

$$\hat{\theta} = \hat{\theta}\left[\mathcal{X}\right] = \hat{\theta}\left[\{x[n]\}_0^{N-1}\right]$$

Properties of Estimators

Consider the set of N observations, $\mathcal{X} = \{x[n]\}_0^{N-1}$, from a *random experiment*; suppose they are used to estimate a parameter θ of the process using some function:

$$\hat{\theta} = \hat{\theta}\left[\mathcal{X}\right] = \hat{\theta}\left[\{x[n]\}_0^{N-1}\right]$$

The function $\hat{\theta}\left[\mathcal{X}\right]$ is known as an **estimator** whereas the value taken by the estimator, using a particular set of observations, is called a **point-estimate**.

Properties of Estimators

Consider the set of N observations, $\mathcal{X} = \{x[n]\}_0^{N-1}$, from a *random experiment*; suppose they are used to estimate a parameter θ of the process using some function:

$$\hat{\theta} = \hat{\theta}\left[\mathcal{X}\right] = \hat{\theta}\left[\{x[n]\}_0^{N-1}\right]$$

The function $\hat{\theta}\left[\mathcal{X}\right]$ is known as an **estimator** whereas the value taken by the estimator, using a particular set of observations, is called a **point-estimate**.

An aim is to design an estimator, $\hat{\theta}$, that should be as close to the true value of the parameter, θ, as possible.

Properties of Estimators

Consider the set of N observations, $\mathcal{X} = \{x[n]\}_0^{N-1}$, from a *random experiment*; suppose they are used to estimate a parameter θ of the process using some function:

$$\hat{\theta} = \hat{\theta}\left[\mathcal{X}\right] = \hat{\theta}\left[\{x[n]\}_0^{N-1}\right]$$

The function $\hat{\theta}\left[\mathcal{X}\right]$ is known as an **estimator** whereas the value taken by the estimator, using a particular set of observations, is called a **point-estimate**.

An aim is to design an estimator, $\hat{\theta}$, that should be as close to the true value of the parameter, θ, as possible.

Since $\hat{\theta}$ is a function of a number of particular realisations of a random outcome (or experiment), then it is itself a RV, and thus has a mean and variance.

Bias of estimator

The **bias** of an estimator $\hat{\theta}$ of a parameter θ is defined as:

$$B(\hat{\theta}) \triangleq \mathbb{E}\left[\hat{\theta}\right] - \theta$$

Bias of estimator

The **bias** of an estimator $\hat{\theta}$ of a parameter θ is defined as:

$$B(\hat{\theta}) \triangleq \mathbb{E}\left[\hat{\theta}\right] - \theta$$

If θ is large, then a small deviation would give what would appear to be a large bias. Thus, the **normalised bias** is often used instead:

$$\epsilon_b(\hat{\theta}) \triangleq \frac{B(\hat{\theta})}{\theta} = \frac{\mathbb{E}\left[\hat{\theta}\right]}{\theta} - 1, \quad \theta \neq 0$$

Bias of estimator

The **bias** of an estimator $\hat{\theta}$ of a parameter θ is defined as:

$$B(\hat{\theta}) \triangleq \mathbb{E}\left[\hat{\theta}\right] - \theta$$

If θ is large, then a small deviation would give what would appear to be a large bias. Thus, the **normalised bias** is often used instead:

$$\epsilon_b(\hat{\theta}) \triangleq \frac{B(\hat{\theta})}{\theta} = \frac{\mathbb{E}\left[\hat{\theta}\right]}{\theta} - 1, \quad \theta \neq 0$$

Example (Biasness of sample mean estimator). Is the sample mean, $\hat{\mu}_x = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$ biased?

SOLUTION. No, since

$$\mathbb{E}\left[\hat{\mu}_x\right] = \mathbb{E}\left[\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] = \frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left[x[n]\right] = \frac{N\mu_X}{N} = \mu_X.$$

Variance of estimator

The **variance** of the estimator $\hat{\theta}$ is defined by:

$$\text{var}\left[\hat{\theta}\right] = \sigma^2_{\hat{\theta}} \triangleq \mathbb{E}\left[\left|\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right|^2\right]$$

However, a minimum variance criterion is not always compatible with the minimum bias requirement; reducing the variance may result in an increase in bias.

Variance of estimator

The **variance** of the estimator $\hat{\theta}$ is defined by:

$$\text{var}\left[\hat{\theta}\right] = \sigma^2_{\hat{\theta}} \triangleq \mathbb{E}\left[\left|\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right|^2\right]$$

However, a minimum variance criterion is not always compatible with the minimum bias requirement; reducing the variance may result in an increase in bias.

Therefore, a compromise or balance between these two conflicting criteria is required, and this is provided by the mean-squared error (MSE) measure described below.

Variance of estimator

The **variance** of the estimator $\hat{\theta}$ is defined by:

$$\text{var}\left[\hat{\theta}\right] = \sigma^2_{\hat{\theta}} \triangleq \mathbb{E}\left[\left|\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right|^2\right]$$

However, a minimum variance criterion is not always compatible with the minimum bias requirement; reducing the variance may result in an increase in bias.

Therefore, a compromise or balance between these two conflicting criteria is required, and this is provided by the mean-squared error (MSE) measure described below.

The **normalised standard deviation** is defined by:

$$\epsilon_r \triangleq \frac{\sigma_{\hat{\theta}}}{\theta}, \quad \theta \neq 0$$

Mean square error

Minimising estimator variance can increase bias. A compromise criterion is the mean-squared error (MSE) of the estimator, which is given by:

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\left[\left|\hat{\theta} - \theta\right|^2\right] = \sigma_{\hat{\theta}}^2 + |B(\hat{\theta})|^2$$

Mean square error

Minimising estimator variance can increase bias. A compromise criterion is the mean-squared error (MSE) of the estimator, which is given by:

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[\left|\hat{\theta} - \theta\right|^2\right] = \sigma_{\hat{\theta}}^2 + |B(\hat{\theta})|^2$$

The estimator $\hat{\theta}_{\text{MSE}} = \hat{\theta}_{\text{MSE}}[\mathcal{X}]$ which minimises $\text{MSE}(\hat{\theta})$ is known as the minimum mean-square error:

$$\hat{\theta}_{MSE} = \arg_{\hat{\theta}} \min \text{MSE}(\hat{\theta})$$

Mean square error

Minimising estimator variance can increase bias. A compromise criterion is the mean-squared error (MSE) of the estimator, which is given by:

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\left[\left|\hat{\theta} - \theta\right|^2\right] = \sigma_{\hat{\theta}}^2 + |B(\hat{\theta})|^2$$

The estimator $\hat{\theta}_{\mathrm{MSE}} = \hat{\theta}_{\mathrm{MSE}}[\mathcal{X}]$ which minimises $\mathrm{MSE}(\hat{\theta})$ is known as the minimum mean-square error:

$$\hat{\theta}_{MSE} = \arg_{\hat{\theta}} \min \mathrm{MSE}(\hat{\theta})$$

This measures the average mean squared deviation of the estimator from its true value.

Unfortunately, adoption of this natural criterion leads to unrealisable estimators; ones which cannot be written solely as a function of the data.

Cramer-Rao Lower Bound

If the MSE can be minimised when the bias is zero, then clearly the variance is also minimised. Such estimators are called minimum variance unbiased estimators (MVUEs).

Cramer-Rao Lower Bound

If the MSE can be minimised when the bias is zero, then clearly the variance is also minimised. Such estimators are called MVUEs.

MVUE possess the important property that they attain a minimum bound on the variance of the estimator, called the Cramér-Rao lower-bound (CRLB).

Cramer-Rao Lower Bound

Theorem (CRLB - scalar parameter). If $\mathbf{X}(\zeta) = [x[0,\zeta], \cdots, x[N-1,\zeta]]^T$ and $f_{\mathbf{X}}(\mathbf{x} \mid \theta)$ is the joint density of $\mathbf{X}(\zeta)$ which depends on fixed but unknown parameter θ, then the variance of the estimator $\hat{\theta}$ is bounded by:

$$\text{var}\left[\hat{\theta}\right] \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta}\right)^2\right]}$$

Cramer-Rao Lower Bound

Theorem (CRLB - scalar parameter). If $\mathbf{X}(\zeta) = [x[0,\zeta], \cdots, x[N-1,\zeta]]^T$ and $f_{\mathbf{X}}(\mathbf{x} \mid \theta)$ is the joint density of $\mathbf{X}(\zeta)$ which depends on fixed but unknown parameter θ, then the variance of the estimator $\hat{\theta}$ is bounded by:

$$\text{var}\left[\hat{\theta}\right] \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta}\right)^2\right]}$$

Alternatively, it may also be expressed as:

$$\text{var}\left[\hat{\theta}\right] \geq -\frac{1}{\mathbb{E}\left[\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta^2}\right]}$$

Cramer-Rao Lower Bound

Theorem (CRLB - scalar parameter). If $\mathbf{X}(\zeta) = [x[0,\zeta], \cdots, x[N-1,\zeta]]^T$ and $f_{\mathbf{X}}(\mathbf{x} \mid \theta)$ is the joint density of $\mathbf{X}(\zeta)$ which depends on fixed but unknown parameter θ, then the variance of the estimator $\hat{\theta}$ is bounded by:

$$\text{var}\left[\hat{\theta}\right] \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta}\right)^2\right]}$$

Alternatively, it may also be expressed as:

$$\text{var}\left[\hat{\theta}\right] \geq -\frac{1}{\mathbb{E}\left[\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta^2}\right]}$$

The function $\ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)$ is called the **log-likelihood** of θ.

Cramer-Rao Lower Bound

Theorem (CRLB - scalar parameter). If $\mathbf{X}(\zeta) = [x[0,\zeta], \cdots, x[N-1,\zeta]]^T$ and $f_{\mathbf{X}}(\mathbf{x} \mid \theta)$ is the joint density of $\mathbf{X}(\zeta)$ which depends on fixed but unknown parameter θ, then the variance of the estimator $\hat{\theta}$ is bounded by:

$$\text{var}\left[\hat{\theta}\right] \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta}\right)^2\right]} \qquad \diamondsuit$$

Alternatively, it may also be expressed as:

$$\text{var}\left[\hat{\theta}\right] \geq -\frac{1}{\mathbb{E}\left[\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta^2}\right]}$$

Furthermore, an unbiased estimator may be found that attains the bound for all θ if, and only if, (iff)

$$\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta} = I(\theta)\left(\hat{\theta} - \theta\right)$$

Consistency of an Estimator

If the MSE of the estimator,

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[|\hat{\theta} - \theta|^2\right] = \sigma^2_{\hat{\theta}} + |B(\hat{\theta})|^2$$

approaches zero as the sample size N becomes large, then both the bias and the variance tends toward zero.

Consistency of an Estimator

If the MSE of the estimator,

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[|\hat{\theta} - \theta|^2\right] = \sigma^2_{\hat{\theta}} + |B(\hat{\theta})|^2$$

approaches zero as the sample size N becomes large, then both the bias and the variance tends toward zero.

Thus, the sampling distribution tends to concentrate around θ, and as $N \to \infty$, it will become an impulse at θ.

Consistency of an Estimator

If the MSE of the estimator,

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[|\hat{\theta} - \theta|^2\right] = \sigma^2_{\hat{\theta}} + |B(\hat{\theta})|^2$$

approaches zero as the sample size N becomes large, then both the bias and the variance tends toward zero.

Thus, the sampling distribution tends to concentrate around θ, and as $N \to \infty$, it will become an impulse at θ.

This is a very important and desirable property, and such an estimator is called a **consistent estimator**.

Maximum Likelihood Estimation

The joint density of the RVs $\mathbf{X}(\zeta) = \{x[n,\zeta]\}_0^{N-1}$, which depends on fixed but unknown parameter $\boldsymbol{\theta}$, is $f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$.

Maximum Likelihood Estimation

The joint density of the RVs $\mathbf{X}(\zeta) = \{x[n,\zeta]\}_0^{N-1}$, which depends on fixed but unknown parameter $\boldsymbol{\theta}$, is $f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$.

This same quantity, viewed as a function of the parameter $\boldsymbol{\theta}$ when a particular set of observations, $\hat{\mathbf{x}}$ is given, is known as the **likelihood function**.

Maximum Likelihood Estimation

The joint density of the RVs $\mathbf{X}(\zeta) = \{x[n,\zeta]\}_0^{N-1}$, which depends on fixed but unknown parameter $\boldsymbol{\theta}$, is $f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$.

This same quantity, viewed as a function of the parameter $\boldsymbol{\theta}$ when a particular set of observations, $\hat{\mathbf{x}}$ is given, is known as the **likelihood function**.

The **maximum-likelihood estimate (MLE)** of the parameter $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}_{ml}$, is defined as that value of $\boldsymbol{\theta}$ that maximises $f_{\mathbf{X}}\left(\hat{\mathbf{x}} \mid \boldsymbol{\theta}\right)$.

Maximum Likelihood Estimation

The joint density of the RVs $\mathbf{X}(\zeta) = \{x[n,\zeta]\}_0^{N-1}$, which depends on fixed but unknown parameter $\boldsymbol{\theta}$, is $f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$.

This same quantity, viewed as a function of the parameter $\boldsymbol{\theta}$ when a particular set of observations, $\hat{\mathbf{x}}$ is given, is known as the **likelihood function**.

The **maximum-likelihood estimate (MLE)** of the parameter $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}_{ml}$, is defined as that value of $\boldsymbol{\theta}$ that maximises $f_{\mathbf{X}}\left(\hat{\mathbf{x}} \mid \boldsymbol{\theta}\right)$.

The MLE for $\boldsymbol{\theta}$ is defined by:

$$\hat{\boldsymbol{\theta}}_{ml}(\mathbf{x}) = \arg_{\boldsymbol{\theta}} \max f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$$

Note that since $\hat{\boldsymbol{\theta}}_{ml}(\mathbf{x})$ depends on the random observation vector $\mathbf{x}$, and so is *itself a RV*.

Properties of the MLE

1. The MLE satisfies

$$\nabla_{\boldsymbol{\theta}} f_{\mathbf{x}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ml}} = \mathbf{0}_{P\times 1}$$
$$\nabla_{\boldsymbol{\theta}} \ln f_{\mathbf{x}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ml}} = \mathbf{0}_{P\times 1}$$

These results assume that the MLE does not occur at a boundary, and that in the set of stationary points of the function, one of them corresponds to a global maximum. Note that minimising the likelihood is equivalent to minimising the log-likelihood, since the likelihood function is always positive, and the logarithm is a monotonic function. It is also necessary to verify which of the stationary points corresponds to the global maximum.

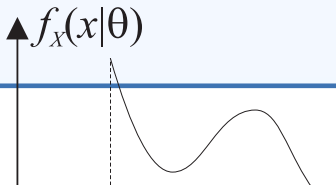
Properties of the MLE

1. The MLE satisfies

$$\nabla_{\boldsymbol{\theta}} f_{\mathbf{x}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ml}} = \mathbf{0}_{P\times 1}$$
$$\nabla_{\boldsymbol{\theta}} \ln f_{\mathbf{x}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ml}} = \mathbf{0}_{P\times 1}$$

These results assume that the MLE does not occur at a boundary, and that in the set of stationary points of the function, one of them corresponds to a global maximum. Note that minimising the likelihood is equivalent to minimising the log-likelihood, since the likelihood function is always positive, and the logarithm is a monotonic function. It is also necessary to verify which of the stationary points corresponds to the global maximum.

2. If an MVUE exists and the MLE does not occur at a boundary, then the MLE *is* the MVUE.



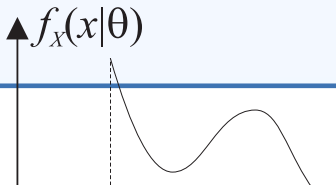
Properties of the MLE

1. The MLE satisfies

$$\nabla_{\boldsymbol{\theta}} f_{\mathbf{x}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ml}} = \mathbf{0}_{P\times 1}$$
$$\nabla_{\boldsymbol{\theta}} \ln f_{\mathbf{x}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ml}} = \mathbf{0}_{P\times 1}$$

These results assume that the MLE does not occur at a boundary, and that in the set of stationary points of the function, one of them corresponds to a global maximum. Note that minimising the likelihood is equivalent to minimising the log-likelihood, since the likelihood function is always positive, and the logarithm is a monotonic function. It is also necessary to verify which of the stationary points corresponds to the global maximum.

2. If an MVUE exists and the MLE does not occur at a boundary, then the MLE *is* the MVUE.



DC Level in white Gaussian noise

Example ([Therrien:1991, Example 6.1, Page 282]). A constant but unknown signal is observed in additive white Gaussian noise (WGN). That is,

$$x[n] = A + w[n] \quad \text{where} \quad w[n] \sim \mathcal{N}\left(0, \sigma_w^2\right)$$

for $n \in \mathcal{N} = \{0, \ldots, N-1\}$. Calculate the MLE of the unknown signal A.

DC Level in white Gaussian noise

Example ([Therrien:1991, Example 6.1, Page 282]). A constant but unknown signal is observed in additive WGN. That is,

$$x[n] = A + w[n] \quad \text{where} \quad w[n] \sim \mathcal{N}\left(0, \sigma_w^2\right)$$

for $n \in \mathcal{N} = \{0, \ldots, N-1\}$. Calculate the MLE of the unknown signal A.

SOLUTION. Since this is a memoryless system, and $w(n)$ are independent and identically distributed (i. i. d.), then so is $x[n]$, and

the log-likelihood is given by:

$$\ln f_{\mathbf{x}}\left(\mathbf{x} \mid A\right) = -\frac{N}{2}\ln(2\pi\sigma_w^2) - \frac{\sum_{n\in\mathcal{N}}\left(x[n]-A\right)^2}{2\sigma_w^2}$$

Differentiating this expression w. r. t. A

and setting to zero :

MLE for Transformed Parameter

Theorem (Invariance Property of the MLE). The MLE of the parameter $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$, where $\mathbf{g}$ is an r-dimensional function of the $P \times 1$ parameter $\boldsymbol{\theta}$, and the pdf, $f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})$ is parameterised by $\boldsymbol{\theta}$, is given by

$$\hat{\boldsymbol{\alpha}}_{ml} = \mathbf{g}(\hat{\boldsymbol{\theta}}_{ml})$$

where $\hat{\boldsymbol{\theta}}_{ml}$ is the MLE of $\boldsymbol{\theta}$.

MLE for Transformed Parameter

Theorem (Invariance Property of the MLE). The MLE of the parameter $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$, where $\mathbf{g}$ is an r-dimensional function of the $P \times 1$ parameter $\boldsymbol{\theta}$, and the pdf, $f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$ is parameterised by $\boldsymbol{\theta}$, is given by

$$\hat{\boldsymbol{\alpha}}_{ml} = \mathbf{g}(\hat{\boldsymbol{\theta}}_{ml})$$

where $\hat{\boldsymbol{\theta}}_{ml}$ is the MLE of $\boldsymbol{\theta}$.

The MLE of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{ml}$, is obtained by maximising $f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$. If the function $\mathbf{g}$ is not an invertible function, then $\hat{\boldsymbol{\alpha}}$ maximises the modified likelihood function $\bar{p}_T\left(\mathbf{x} \mid \boldsymbol{\alpha}\right)$ defined as:

$$\bar{p}_T\left(\mathbf{x} \mid \boldsymbol{\alpha}\right) = \max_{\boldsymbol{\theta}:\boldsymbol{\alpha}=\mathbf{g}(\boldsymbol{\theta})} f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right) \qquad \diamondsuit$$

Least Squares

The estimators discussed so far have attempted to find an optimal or nearly optimal (for large data records) estimator for example, the MVUE.

Least Squares

The estimators discussed so far have attempted to find an optimal or nearly optimal (for large data records) estimator for example, the MVUE.

An alternate philosophy is a class of estimators that in general have no optimality properties associated with them, but make *good sense* for many problems of interest: the **principle of least squares**.

Least Squares

The estimators discussed so far have attempted to find an optimal or nearly optimal (for large data records) estimator for example, the MVUE.

An alternate philosophy is a class of estimators that in general have no optimality properties associated with them, but make *good sense* for many problems of interest: the **principle of least squares**.

A salient feature of the method is that *no probabilistic assumptions* are made about the data; only a *signal model* is assumed.

Least Squares

The estimators discussed so far have attempted to find an optimal or nearly optimal (for large data records) estimator for example, the MVUE.

An alternate philosophy is a class of estimators that in general have no optimality properties associated with them, but make *good sense* for many problems of interest: the **principle of least squares**.

A salient feature of the method is that *no probabilistic assumptions* are made about the data; only a *signal model* is assumed.

As will be seen, it turns out that the LSE can be calculated when just the first and second moments are known, and through the solution of *linear* equations.

The Least Squares Approach

In the least-squares (LS) approach, it is sought to minimise the squared difference between the given, or observed, data $x[n]$ and the assumed, or hidden, signal or noiseless data.

The Least Squares Approach

In the LS approach, it is sought to minimise the squared difference between the given, or observed, data $x[n]$ and the assumed, or hidden, signal or noiseless data.

Here it is assumed that the hidden or unobserved signal is generated by some model which, in turn, depends on some unknown parameter $\boldsymbol{\theta}$.

The Least Squares Approach

In the LS approach, it is sought to minimise the squared difference between the given, or observed, data $x[n]$ and the assumed, or hidden, signal or noiseless data.

Here it is assumed that the hidden or unobserved signal is generated by some model which, in turn, depends on some unknown parameter $\boldsymbol{\theta}$.

The LSE of $\boldsymbol{\theta}$ chooses the value that makes $s[n]$ closest to the observed data $x[n]$, and this *closeness* is measured by the LS error criterion:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} \left(x[n] - s[n]\right)^2$$

The Least Squares Approach

In the LS approach, it is sought to minimise the squared difference between the given, or observed, data $x[n]$ and the assumed, or hidden, signal or noiseless data.

Here it is assumed that the hidden or unobserved signal is generated by some model which, in turn, depends on some unknown parameter $\boldsymbol{\theta}$.

The LSE of $\boldsymbol{\theta}$ chooses the value that makes $s[n]$ closest to the observed data $x[n]$, and this *closeness* is measured by the LS error criterion:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} \left(x[n] - s[n]\right)^2$$

The LSE is given by:

$$\hat{\boldsymbol{\theta}}_{LSE} = \arg_{\boldsymbol{\theta}} \min J(\boldsymbol{\theta})$$

DC Level

Example ([Kay:1993, Example 6.1, Page 221]). It is assumed that an observed signal, $x[n]$, is a perturbed version of an unknown signal, $s[n]$, which is modelled as $s[n] = A$, for $n \in \mathcal{N} = \{0, \ldots, N-1\}$. Calculate the LSE of the unknown signal A.

DC Level

Example ([Kay:1993, Example 6.1, Page 221]). It is assumed that an observed signal, $x[n]$, is a perturbed version of an unknown signal, $s[n]$, which is modelled as $s[n] = A$, for $n \in \mathcal{N} = \{0, \ldots, N-1\}$. Calculate the LSE of the unknown signal A.

SOLUTION. According to the LS approach, then:

$$\hat{A}_{LSE} = \arg_A \min J(A) \quad \text{where} \quad J(A) = \sum_{n=0}^{N-1} (x[n] - A)^2$$

Differentiating w. r. t. A and setting the result to zero produces

$$\hat{A}_{LSE} = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$$

which is the sample mean estimator.

Linear Least Squares

Thus, the unknown random-vector $\mathbf{s}$ is linear in the unknown parameter vector $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_P]$,

$$\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$$

The LSE is found by minimising:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} |x[n] - s[n]|^2 = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$$

Linear Least Squares

Thus, the unknown random-vector $\mathbf{s}$ is linear in the unknown parameter vector $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_P]$,

$$\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$$

The LSE is found by minimising:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} |x[n] - s[n]|^2 = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$$

Setting the gradient of $J(\boldsymbol{\theta})$ to zero yields the LSE:

$$\hat{\boldsymbol{\theta}}_{LSE} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$$

Linear Least Squares

Thus, the unknown random-vector $\mathbf{s}$ is linear in the unknown parameter vector $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_P]$,

$$\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$$

The LSE is found by minimising:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} |x[n] - s[n]|^2 = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$$

Setting the gradient of $J(\boldsymbol{\theta})$ to zero yields the LSE:

$$\hat{\boldsymbol{\theta}}_{LSE} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$$

The equations $\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} = \mathbf{H}^T\mathbf{x}$, to be solved for $\hat{\boldsymbol{\theta}}$, are termed the **normal equation**.

Handout 6
MonteCarlo

Introduction

Many signal processing problems can be reduced to either an *optimisation* problem or an *integration* problem:

Introduction

Many signal processing problems can be reduced to either an *optimisation* problem or an *integration* problem:

Optimisation: involves finding the solution to

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} h(\boldsymbol{\theta})$$

where $h(\cdot)$ is a scalar function of a multi-dimensional vector of parameters, $\boldsymbol{\theta}$.

Typically, $h(\cdot)$ might represent some **cost function**, and it is implicitly assumed that the optimisation cannot be calculated explicitly.

Introduction

Many signal processing problems can be reduced to either an *optimisation* problem or an *integration* problem:

Integration: involves evaluating an integral,

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

that cannot explicitly be calculated in *closed form*.

Introduction

Many signal processing problems can be reduced to either an *optimisation* problem or an *integration* problem:

Integration: involves evaluating an integral,

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

that cannot explicitly be calculated in *closed form*.

For example, the Gaussian-error function:

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}}\, d\theta$$

Again, the integral may be multi-dimensional, and in general $\boldsymbol{\theta}$ is a vector.

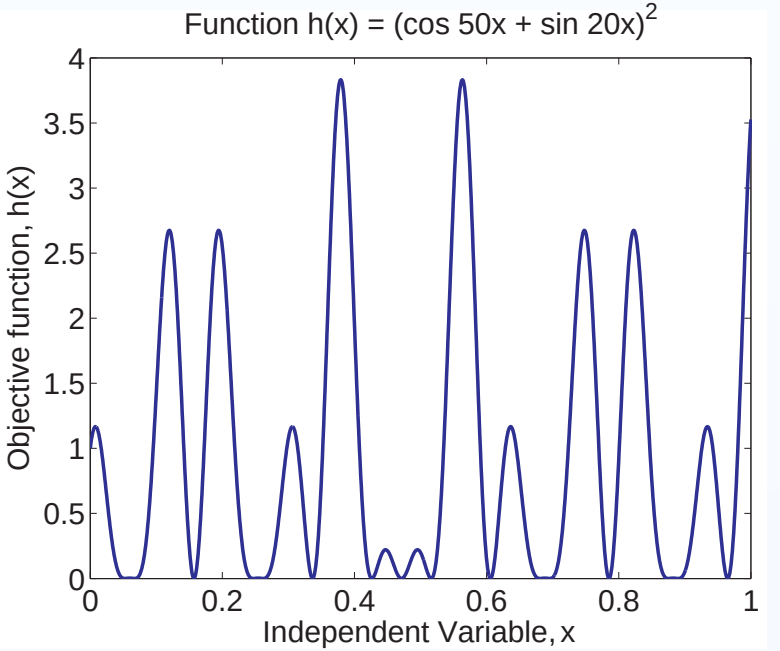
Introduction

Many signal processing problems can be reduced to either an *optimisation* problem or an *integration* problem:

Optimisation and Integration Some problems involve both integration and optimisation: a fundamental problem is the maximisation of a marginal distribution:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \int_{\Omega} f(\boldsymbol{\theta}, \boldsymbol{\omega})\, d\boldsymbol{\omega}$$

Deterministic Numerical Methods



Plot of the function $h(x) = (\cos 50x + \sin 20x)^2$, $0 \leq x \leq 1$.

There are various deterministic solutions to the optimisation and integration problems.

Deterministic Numerical Methods

Optimisation: 1. Golden-section search and Brent's Method in one dimension;

2. Nelder and Mead Downhill Simplex method in multi-dimensions;
3. Gradient and Variable-Metric methods in multi-dimensions, typically an extension of Newton-Raphson methods.

Deterministic Numerical Methods

Integration: Most deterministic integration rely on classic formulas for equally spaced abscissas:

1. simple Riemann integration;
2. standard and extended Simpson's and Trapezoidal rules;
3. refinements such as Romberg Integration.

Unfortunately, these methods are not easily extended to multi-dimensions.

Deterministic Numerical Methods

Integration: Most deterministic integration rely on classic formulas for equally spaced abscissas:

1. simple Riemann integration;
2. standard and extended Simpson's and Trapezoidal rules;
3. refinements such as Romberg Integration.

More sophisticated approaches allow non-uniformally spaced abscissas at which the function is evaluated.

These methods tend to use Gaussian quadratures and orthogonal polynomials. Splines are also used.

Unfortunately, these methods are not easily extended to multi-dimensions.

Deterministic Optimisation

The **Nelder-Mead Downhill Simplex method** simply crawls downhill in a straightforward fashion that makes almost no special assumptions about your function.

This can be extremely slow, but it can be robust.

Deterministic Optimisation

Gradient methods are typically based on the Newton-Raphson algorithm which solves $\nabla h(\boldsymbol{\theta}) = \mathbf{0}$.

For a scalar function, $h(\boldsymbol{\theta})$, of a vector of independent variables $\boldsymbol{\theta}$, a sequence $\boldsymbol{\theta}_n$ is produced such that:

Deterministic Optimisation

Gradient methods are typically based on the Newton-Raphson algorithm which solves $\nabla h(\boldsymbol{\theta}) = \mathbf{0}$.

For a scalar function, $h(\boldsymbol{\theta})$, of a vector of independent variables $\boldsymbol{\theta}$, a sequence $\boldsymbol{\theta}_n$ is produced such that:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \left(\nabla \nabla^T h\left(\boldsymbol{\theta}_n\right)\right)^{-1} \nabla h\left(\boldsymbol{\theta}_n\right)$$

Numerous variants of Newton-Raphson-type techniques exist, and include the **steepest descent method**, or the **Levenberg-Marquardt method**.

Deterministic Integration

The integral

$$\mathcal{I} = \int_a^b f(\theta)\, d\theta,$$

where θ is a scalar, and $b > a$, can be solved with the trapezoidal rule using

$$\hat{I} = \frac{1}{2} \sum_{k=0}^{N-1} (\theta_{k+1} - \theta_k)\left(f(\theta_k) + f(\theta_{k+1})\right)$$

where the θ_k's constitute an ordered partition of $[a, b]$.

Deterministic Integration

The integral

$$\mathcal{I} = \int_a^b f(\theta)\, d\theta,$$

where θ is a scalar, and $b > a$, can be solved with the trapezoidal rule using

$$\hat{I} = \frac{1}{2} \sum_{k=0}^{N-1} (\theta_{k+1} - \theta_k)\left(f(\theta_k) + f(\theta_{k+1})\right)$$

where the θ_k's constitute an ordered partition of $[a, b]$.

Another formula is **Simpson's rule**:

$$\hat{I} = \frac{\delta}{3}\left\{ f(a) + 4\sum_{k=1}^{N} f(\theta_{2k-1}) + 2\sum_{k=1}^{N} h(\theta_{2k}) + f(b) \right\}$$

in the case of equally spaced samples with $\delta = \theta_{k+1} - \theta_k$.

Monte Carlo Numerical Methods

Monte Carlo methods are stochastic techniques, in which random numbers are generated and use to examine some problem.

Monte Carlo Integration

Consider the integral,

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Monte Carlo Integration

Consider the integral,

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Defining a function $\pi(\boldsymbol{\theta})$ which is non-zero and positive for all $\boldsymbol{\theta} \in \Theta$, this integral can be expressed in the alternate form:

$$\mathcal{I} = \int_{\Theta} \frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

where the function $\pi(\boldsymbol{\theta}) > 0$, $\boldsymbol{\theta} \in \Theta$ is a pdf which satisfies

$$\int_{\Theta} \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} = 1$$

Monte Carlo Integration

Consider the integral,

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Defining a function $\pi(\boldsymbol{\theta})$ which is non-zero and positive for all $\boldsymbol{\theta} \in \Theta$, this integral can be expressed in the alternate form:

$$\mathcal{I} = \int_{\Theta} \frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

where the function $\pi(\boldsymbol{\theta}) > 0$, $\boldsymbol{\theta} \in \Theta$ is a pdf which satisfies

$$\int_{\Theta} \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} = 1$$

This may be written as an expectation:

$$\mathcal{I} = \mathbb{E}_{\pi}\left[\frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right]$$

Monte Carlo Integration

This expectation can be estimated using the idea of the **sample expectation**, and leads to the idea behind Monte Carlo integration:

1. Sample N random variates from a density function $\pi(\boldsymbol{\theta})$,

$$\boldsymbol{\theta}^{(k)} \sim \pi(\boldsymbol{\theta}), \quad k \in \mathcal{N} = \{0, \ldots, N-1\}$$

2. Calculate the sample average of the expectation using

$$\hat{\mathcal{I}} = \frac{1}{N} \sum_{k=0}^{N-1} \frac{f(\boldsymbol{\theta}^{(k)})}{\pi(\boldsymbol{\theta}^{(k)})} \approx \mathbb{E}_{\pi} \left[\frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \right]$$

Stochastic Optimisation

There are two distinct approaches to the Monte Carlo optimisation of the objective function $h(\boldsymbol{\theta})$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} h(\boldsymbol{\theta})$$

The first method is broadly known as an **exploratory approach**, while the second approach is based on a **probabilistic approximation** of the objective function.

Stochastic Optimisation

Exploratory approach This approach is concerned with fast *explorations* of the sample space rather than working with the objective function directly.

Stochastic Optimisation

Exploratory approach This approach is concerned with fast *explorations* of the sample space rather than working with the objective function directly.

For example, maximisation can be solved by sampling a large number, N, of independent random variables, $\{\boldsymbol{\theta}^{(k)}\}$, from a pdf $\pi(\boldsymbol{\theta})$, and taking the estimate:

$$\hat{\boldsymbol{\theta}} \approx \underset{\{\boldsymbol{\theta}^{(k)}\}}{\arg\max}\, h\left(\boldsymbol{\theta}^{(k)}\right)$$

Typically, when no specific features regarding the function $h(\boldsymbol{\theta})$, are taken into account, $\pi(\boldsymbol{\theta})$ will take on a uniform distribution over Θ.

Stochastic Optimisation

Exploratory approach This approach is concerned with fast *explorations* of the sample space rather than working with the objective function directly.

For example, maximisation can be solved by sampling a large number, N, of independent random variables, $\{\boldsymbol{\theta}^{(k)}\}$, from a pdf $\pi(\boldsymbol{\theta})$, and taking the estimate:

$$\hat{\boldsymbol{\theta}} \approx \underset{\{\boldsymbol{\theta}^{(k)}\}}{\arg\max}\, h\left(\boldsymbol{\theta}^{(k)}\right)$$

Typically, when no specific features regarding the function $h(\boldsymbol{\theta})$, are taken into account, $\pi(\boldsymbol{\theta})$ will take on a uniform distribution over Θ.

Stochastic Approximation  The Monte Carlo EM algorithm

Generating Random Variables

This section discusses a variety of techniques for generating random variables from a different distributions.

Uniform Variates

The foundation underpinning all stochastic simulations is the ability to generate a sequence of i. i. d. uniform random variates over the range $(0, 1]$.

Uniform Variates

The foundation underpinning all stochastic simulations is the ability to generate a sequence of i. i. d. uniform random variates over the range $(0, 1]$.

Random variates are *pseudo* or *synthetic* and not truly random since they are usually generated using a recurrence of the form:

$$x_{n+1} = (a\, x_n + b) \mod m$$

Uniform Variates

The foundation underpinning all stochastic simulations is the ability to generate a sequence of i. i. d. uniform random variates over the range $(0, 1]$.

Random variates are *pseudo* or *synthetic* and not truly random since they are usually generated using a recurrence of the form:

$$x_{n+1} = (a\, x_n + b) \mod m$$

This is known as the linear congruential generator.

However, suitable values of a, b and m can be chosen such that the random variates pass all statistical tests of randomness.

Transformation Methods

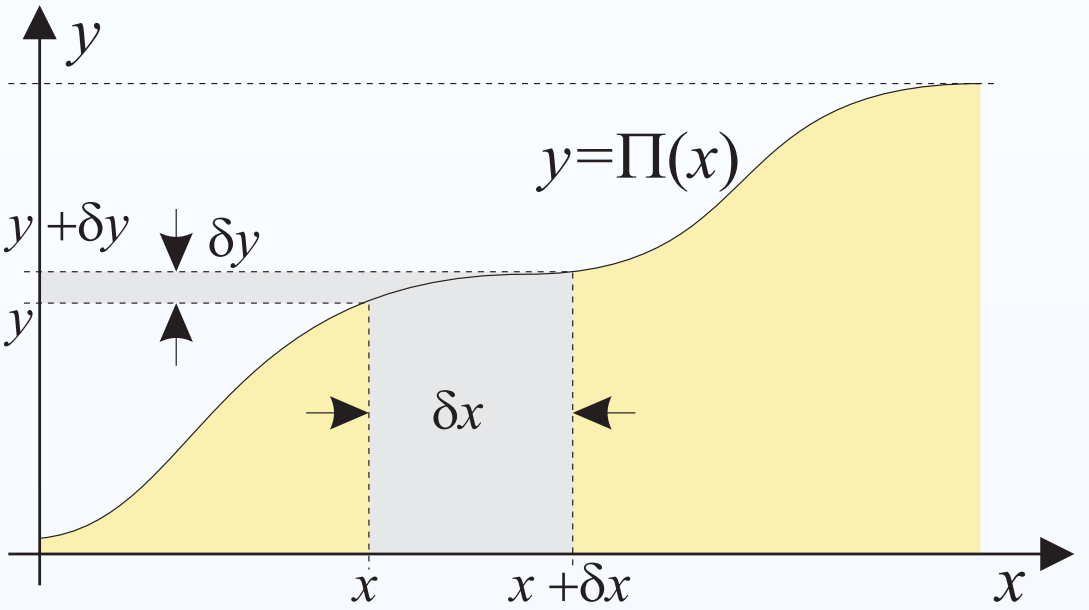
It is possible to sample from a number of extremely important probability distributions by applying various probability transformation methods.

Theorem (Probability transformation rule). Denote the real roots of $y = g(x)$ by $\{x_n,\, n \in \mathcal{N}\}$, such that

$$y = g(x_1) = \cdots = g(x_N)$$

PROOF. The proof is given in the handout on scalar random variables.

Inverse Transform Method

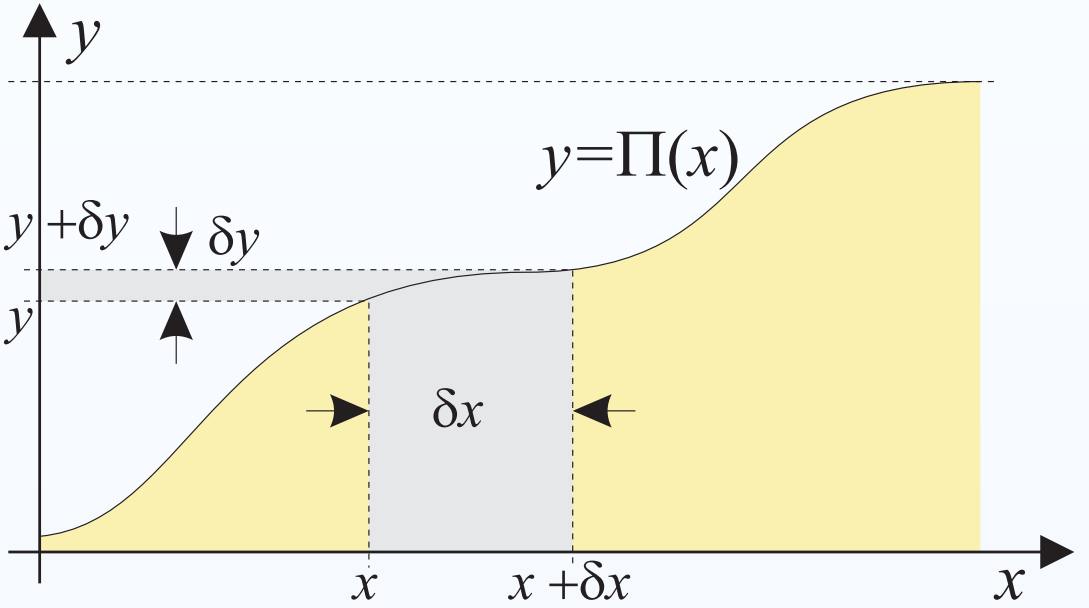


A simple derivation of the inverse transform method

$X(\zeta)$ and $Y(\zeta)$ are RVs related by the function $Y(\zeta) = \Pi(X(\zeta))$.

$\Pi(\zeta)$ is monotonically increasing so that there is only one solution to the equation $y = \Pi(x)$, $x = \Pi^{-1}(y)$.

Inverse Transform Method



A simple derivation of the inverse transform method

$$f_X(x) = \frac{d\Pi(x)}{dx} f_Y(y)$$

Now, suppose $Y(\zeta) \sim \mathcal{U}_{[0,1]}$ is a uniform random variable. If $\Pi(x)$ is the cdf corresponding to a desired pdf $\pi(x)$, then

$$f_X(x) = \pi(x), \quad \text{where} \quad x = \Pi^{-1}(y)$$

Inverse Transform Method

In otherwords, if

$$U(\zeta) \sim \mathcal{U}_{[0,\,1]},\; X(\zeta) = \Pi^{-1} U(\zeta) \sim \pi\left(x\right)$$

Inverse Transform Method

In otherwords, if

$$U(\zeta) \sim \mathcal{U}_{[0,\,1]},\; X(\zeta) = \Pi^{-1} U(\zeta) \sim \pi\left(x\right)$$

Example (Exponential variable generation). If $X(\zeta) \sim \mathcal{E}xp(1)$, such that $\pi(x) = e^{-x}$ and $\Pi(x) = 1 - e^{-x}$, then solving for x in terms of $u = 1 - e^{-x}$ gives $x = -\log(1-u)$.

Inverse Transform Method

In otherwords, if

$$U(\zeta) \sim \mathcal{U}_{[0,\,1]},\; X(\zeta) = \Pi^{-1} U(\zeta) \sim \pi\left(x\right)$$

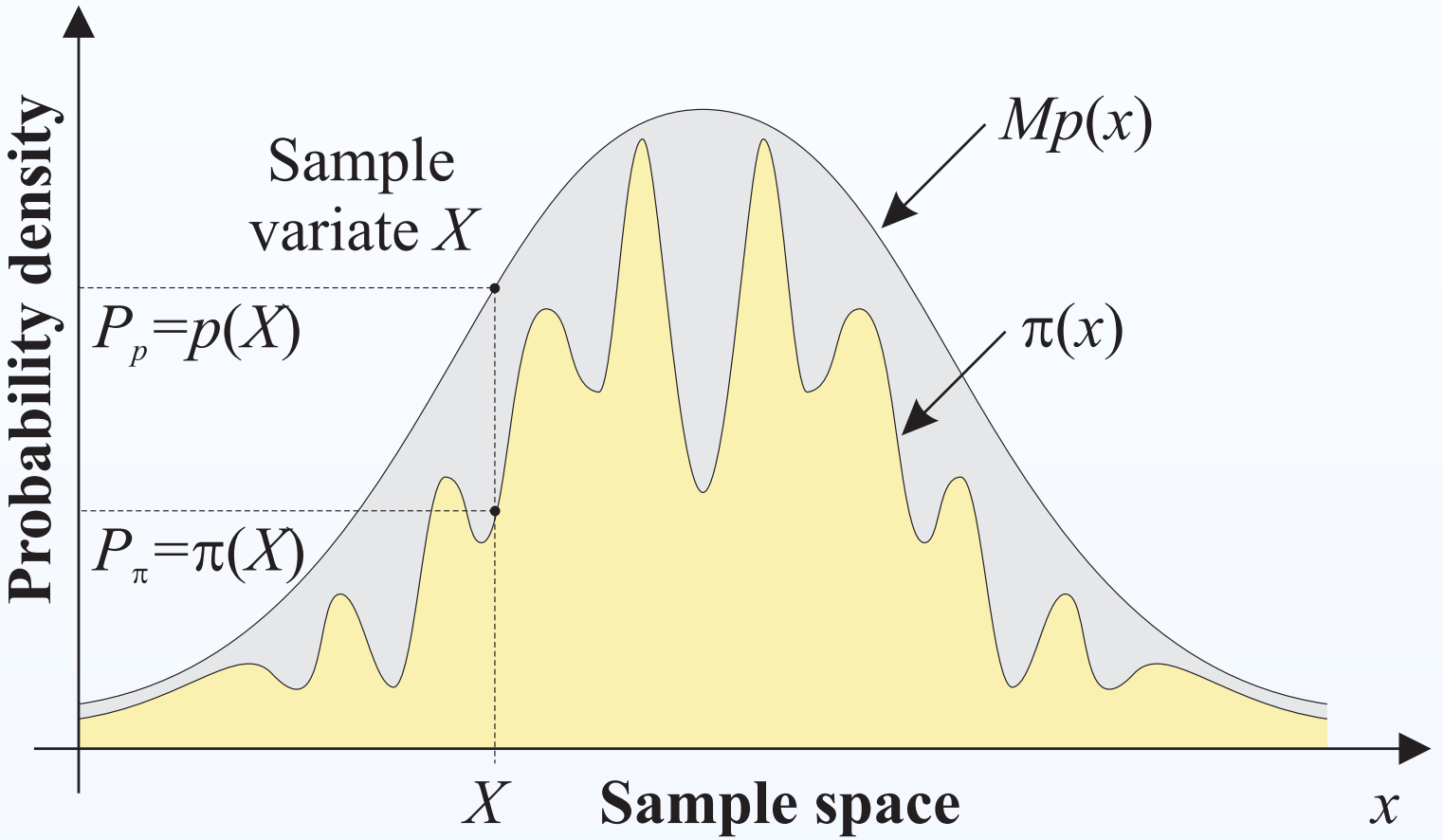
Example (Exponential variable generation). If $X(\zeta) \sim \mathcal{E}xp(1)$, such that $\pi(x) = e^{-x}$ and $\Pi(x) = 1 - e^{-x}$, then solving for x in terms of $u = 1 - e^{-x}$ gives $x = -\log(1-u)$.

Therefore, if $U(\zeta) \sim \mathcal{U}_{[0,\,1]}$, then the RV from the transformation $X(\zeta) = -\log U(\zeta)$ has the exponential distribution (since $U(\zeta)$ and $1 - U(\zeta)$ are both uniform).

Acceptance-Rejection Sampling

For most distributions, it is often difficult or even impossible to directly simulate using either the inverse transform or probability transformations.

Acceptance-Rejection Sampling



On average, you would expect to have too many variates that take on the value X by a factor of

$$u(X) = \frac{P_p}{P_\pi} = \frac{p(X)}{\pi(X)}$$

Acceptance-Rejection Sampling

Thus, to reduce the number of variates that take on a value of X, simply throw away a number of samples in proportion to the amount of *over sampling*.

Acceptance-Rejection Sampling

Thus, to reduce the number of variates that take on a value of X, simply throw away a number of samples in proportion to the amount of *over sampling*.

1. Generate the random variates $X \sim p(x)$ and $U \sim \mathcal{U}_{[0,\,1]}$;
2. Accept X if $U \leq P_a = \frac{\pi(X)}{Mp(x)}$;
3. Otherwise, reject and return to first step.

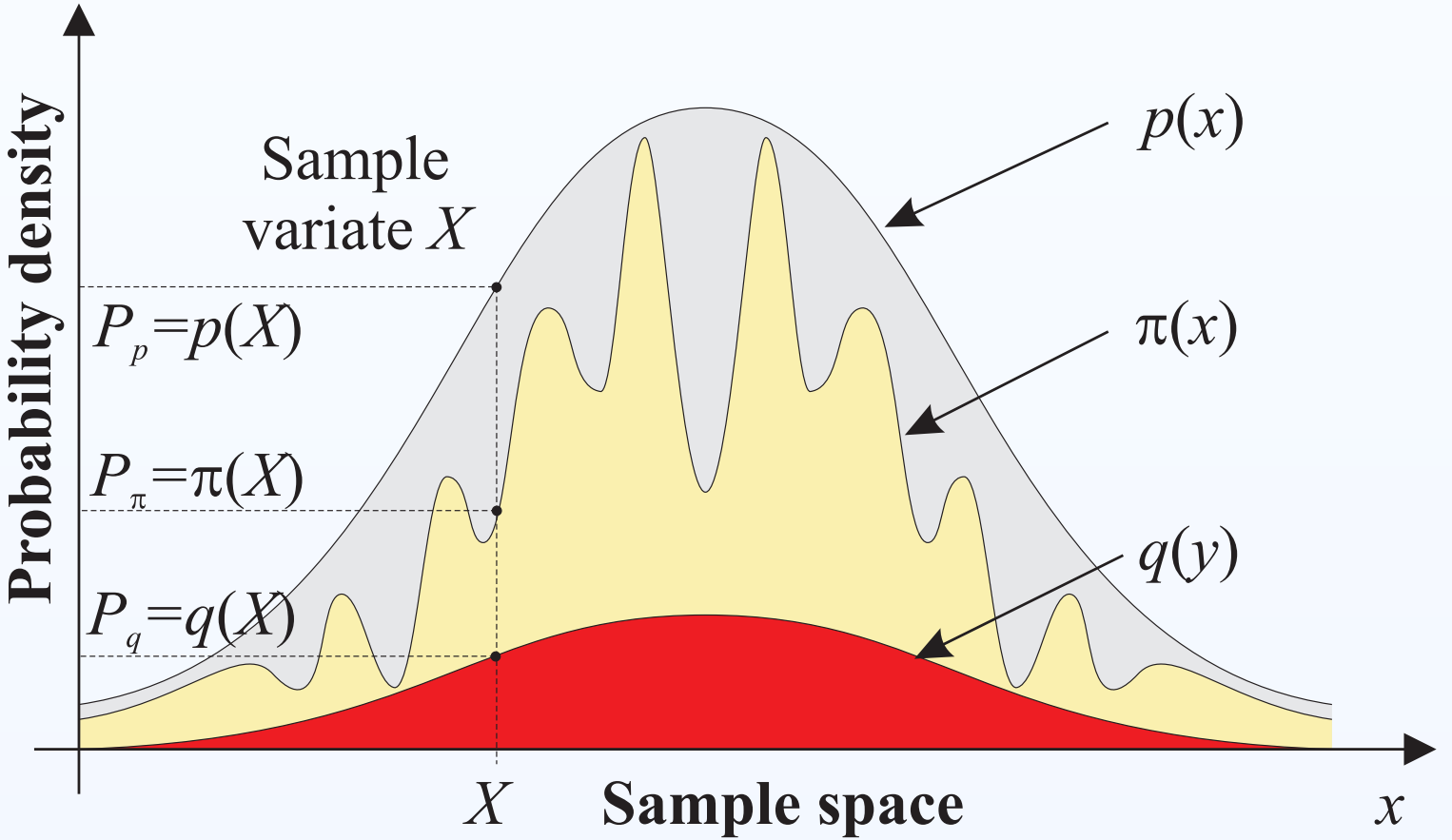
Envelope and Squeeze Methods

A problem with many sampling methods, which can make the density $\pi(x)$ difficult to simulate, is that the function may require substantial computing time at each evaluation.

It is possible to reduce the algorithmic complexity by looking for another computationally simple function, $q(x)$ which *bounds* $\pi(x)$ *from below*.

Envelope and Squeeze Methods

If X satisfies $q(X) \leq \pi(X)$, then it should be accepted when $U \leq \frac{q(X)}{Mp(x)}$, since this also satisfies $U \leq \frac{\pi(X)}{Mp(x)}$.



Envelope and Squeeze Methods

This leads to the **envelope accept-reject algorithm**:

1. Generate the random variates $X \sim p(x)$ and $U \sim \mathcal{U}_{[0,\,1]}$;
2. Accept X if $U \leq \frac{q(X)}{Mp(x)}$;
3. Otherwise, accept X if $U \leq \frac{\pi(X)}{Mp(x)}$;
4. Otherwise, reject and return to first step.

Envelope and Squeeze Methods

This leads to the **envelope accept-reject algorithm**:

1. Generate the random variates $X \sim p(x)$ and $U \sim \mathcal{U}_{[0,\,1]}$;
2. Accept X if $U \leq \frac{q(X)}{Mp(x)}$;
3. Otherwise, accept X if $U \leq \frac{\pi(X)}{Mp(x)}$;
4. Otherwise, reject and return to first step.

By construction of a lower envelope on $\pi\left(x\right)$, the number of function evaluations is potentially decreased by a factor of

$$P_{\bar{\pi}} = \frac{1}{M} \int q\left(x\right)\, dx$$

which is the probability that $\pi\left(x\right)$ is not evaluated.

Importance Sampling

The problem with accept-reject sampling methods is finding the envelope functions and the constant M.

Importance Sampling

The problem with accept-reject sampling methods is finding the envelope functions and the constant M.

The simplest application of **importance sampling** is in Monte Carlo integration. Suppose that is is desired to evaluate the function:

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Importance Sampling

The problem with accept-reject sampling methods is finding the envelope functions and the constant M.

The simplest application of **importance sampling** is in Monte Carlo integration. Suppose that is is desired to evaluate the function:

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Approximate by empirical average:

$$\hat{\mathcal{I}} = \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{I}_{\Theta}\left(\boldsymbol{\theta}^{(k)}\right), \quad \text{where} \quad \boldsymbol{\theta}^{(k)} \sim f(\boldsymbol{\theta})$$

where $\mathbb{I}_{\mathcal{A}}(a)$ is the indicator function, and is equal to one if $a \in \mathcal{A}$ and zero otherwise.

Importance Sampling

Defining an *easy-to-sample-from* density $\pi(\boldsymbol{\theta}) > 0$, $\forall \boldsymbol{\theta} \in \Theta$:

$$\mathcal{I} = \int_{\Theta} \frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \mathbb{E}_{\pi}\left[\frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right],$$

Importance Sampling

Defining an *easy-to-sample-from* density $\pi(\boldsymbol{\theta}) > 0,\ \forall \boldsymbol{\theta} \in \Theta$:

$$\mathcal{I} = \int_{\Theta} \frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \mathbb{E}_{\pi}\left[\frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right],$$

leads to an estimator based on the **sample expectation**;

$$\hat{\mathcal{I}} = \frac{1}{N} \sum_{k=0}^{N-1} \frac{f(\boldsymbol{\theta}^{(k)})}{\pi(\boldsymbol{\theta}^{(k)})}$$

Other Methods

Include:

- representing pdfs as mixture of distributions;
- algorithms for log-concave densities, such as the adaptive rejection sampling scheme;
- generalisations of accept-reject;
- method of composition (similar to Gibbs sampling);
- ad-hoc methods, typically based on probability transformations and order statistics (for example, generating Beta distributions with integer parameters).

Markov chain Monte Carlo Methods

A **Markov chain** is the first generalisation of an independent process, where each *state* of a Markov chain depends on the previous state only.

The Metropolis-Hastings algorithm

The **Metropolis-Hastings algorithm** is an extremely flexible method for producing a random sequence of samples from a given density.

1. Generate a random sample from a **proposal distribution**: $Y \sim g\left(y \mid X^{(k)}\right)$.

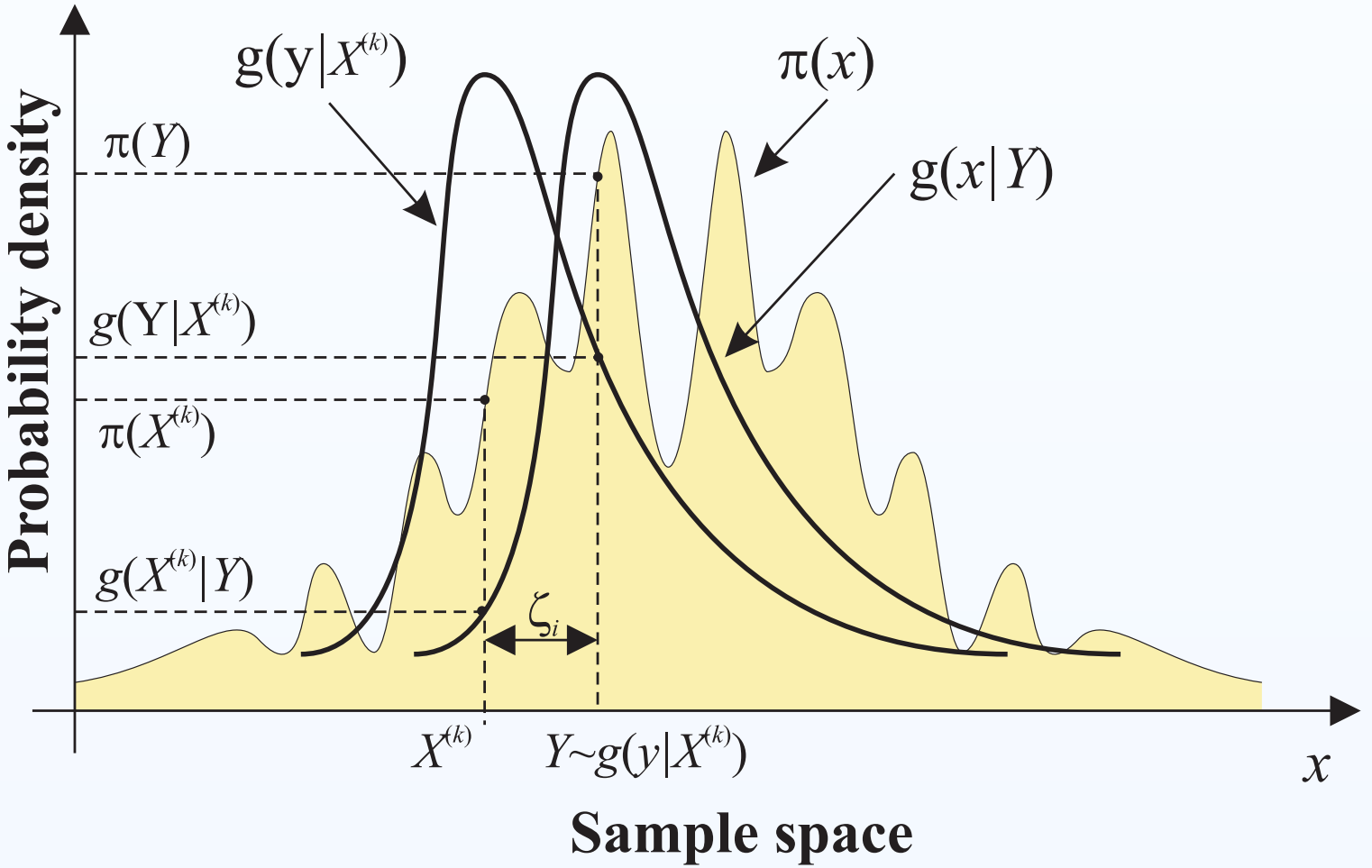
2. Set the new random variate to be:

$$X^{(k+1)} = \begin{cases} Y & \text{with probability } \rho(X^{(k)}, Y) \\ X^{(k)} & \text{with probability } 1 - \rho(X^{(k)}, Y) \end{cases}$$

where the acceptance ratio function $\rho(x, y)$ is given by:

$$\rho(x, y) = \min\left\{\frac{\pi(y)}{g(y \mid x)}\left(\frac{\pi(x)}{g(x \mid y)}\right)^{-1}, 1\right\} \equiv \min\left\{\frac{\pi(y)}{\pi(x)}\frac{g(x \mid y)}{g(y \mid x)}, 1\right\}$$

The Metropolis-Hastings algorithm



Gibbs Sampling

Gibbs sampling is a Monte Carlo method that facilitates sampling from a multivariate density function, $\pi(\theta_0, \theta_1, \ldots, \theta_M)$ by drawing successive samples from marginal densities of smaller dimensions.

Gibbs Sampling

Gibbs sampling is a Monte Carlo method that facilitates sampling from a multivariate density function, $\pi\left(\theta_0, \theta_1, \ldots, \theta_M\right)$ by drawing successive samples from marginal densities of smaller dimensions.

Using the probability chain rule,

$$\pi\left(\{\theta_m\}_{m=1}^{M}\right) = \pi\left(\theta_\ell \mid \{\theta_m\}_{m=1,m\neq\ell}^{M}\right) \pi\left(\{\theta_m\}_{m=1,m\neq\ell}^{M}\right)$$

The Gibbs sampler works by drawing random variates from the marginal densities $\pi\left(\theta_\ell \mid \{\theta_m\}_{m=1,m\neq\ell}^{M}\right)$ in a cyclic iterative pattern.

Gibbs Sampling

First iteration:

$$\theta_1^{(1)} \sim \pi\left(\theta_1 \mid \theta_2^{(0)}, \theta_3^{(0)}, \theta_4^{(0)}, \ldots, \theta_M^{(0)}\right)$$

$$\theta_2^{(1)} \sim \pi\left(\theta_2 \mid \theta_1^{(1)}, \theta_3^{(0)}, \theta_4^{(0)}, \ldots, \theta_M^{(0)}\right)$$

$$\theta_3^{(1)} \sim \pi\left(\theta_3 \mid \theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(0)}, \ldots, \theta_M^{(0)}\right)$$

$$\vdots \qquad\qquad \vdots$$

$$\theta_M^{(1)} \sim \pi\left(\theta_M \mid \theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(1)}, \ldots, \theta_{M-1}^{(1)}\right)$$

Gibbs Sampling

Second iteration:

$$\theta_1^{(2)} \sim \pi\left(\theta_1 \mid \theta_2^{(1)}, \theta_3^{(1)}, \theta_4^{(1)}, \ldots, \theta_M^{(1)}\right)$$

$$\theta_2^{(2)} \sim \pi\left(\theta_2 \mid \theta_1^{(2)}, \theta_3^{(1)}, \theta_4^{(1)}, \ldots, \theta_M^{(1)}\right)$$

$$\theta_3^{(2)} \sim \pi\left(\theta_3 \mid \theta_1^{(2)}, \theta_2^{(2)}, \theta_4^{(1)}, \ldots, \theta_M^{(1)}\right)$$

$$\vdots \qquad\qquad \vdots$$

$$\theta_M^{(2)} \sim \pi\left(\theta_M \mid \theta_1^{(2)}, \theta_2^{(2)}, \theta_4^{(2)}, \ldots, \theta_{M-1}^{(2)}\right)$$

Gibbs Sampling

$k+1$-th iteration:

$$\theta_1^{(k+1)} \sim \pi\left(\theta_1 \mid \theta_2^{(k)}, \theta_3^{(k)}, \theta_4^{(k)}, \ldots, \theta_M^{(k)}\right)$$

$$\theta_2^{(k+1)} \sim \pi\left(\theta_2 \mid \theta_1^{(k+1)}, \theta_3^{(k)}, \theta_4^{(k)}, \ldots, \theta_M^{(k)}\right)$$

$$\theta_3^{(k+1)} \sim \pi\left(\theta_3 \mid \theta_1^{(k+1)}, \theta_2^{(k+1)}, \theta_4^{(k)}, \ldots, \theta_M^{(k)}\right)$$

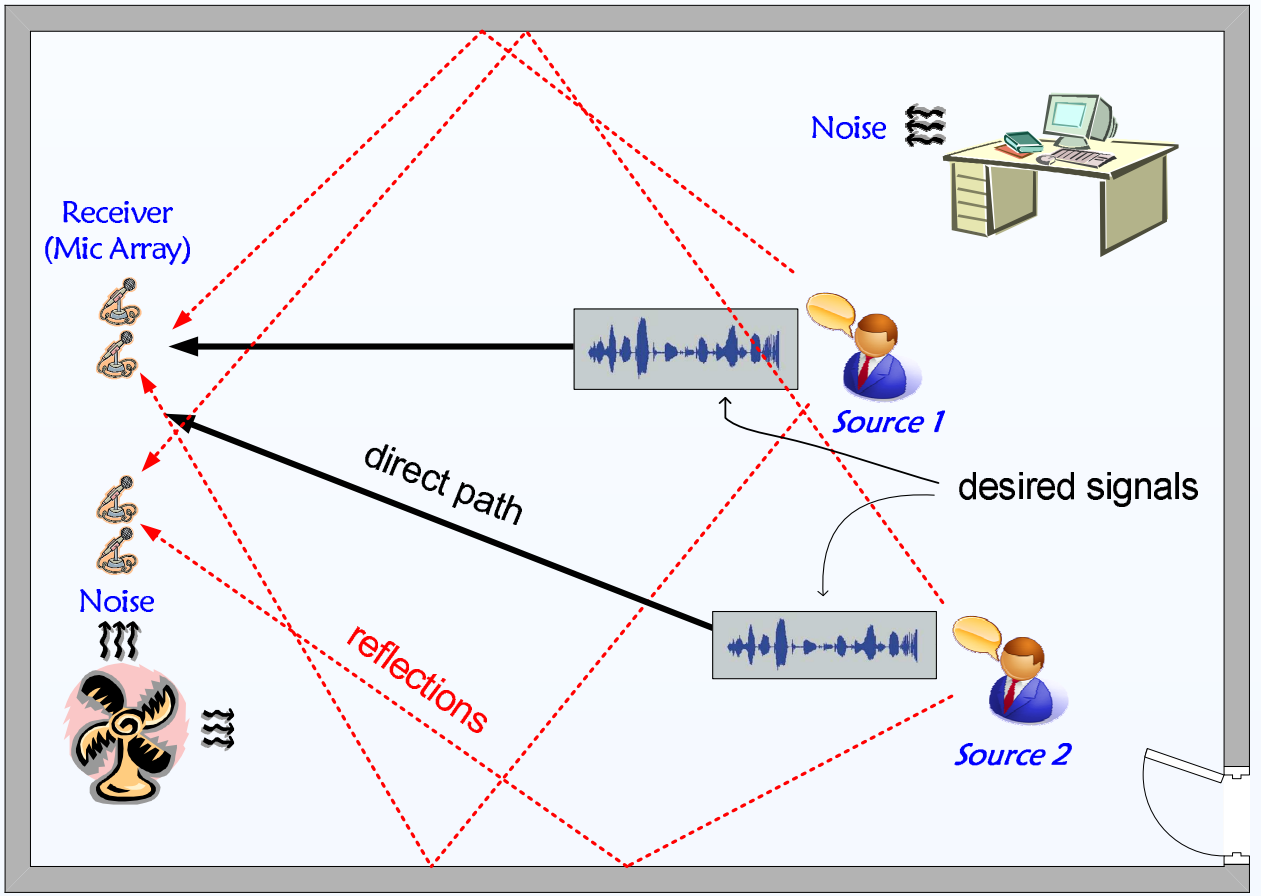
$$\vdots \qquad\qquad \vdots$$

$$\theta_M^{(k+1)} \sim \pi\left(\theta_M \mid \theta_1^{(k)}, \theta_2^{(k)}, \theta_4^{(k)}, \ldots, \theta_{M-1}^{(k)}\right)$$

At the end of the j-th iteration, the samples $\theta_0^{(j)}, \theta_1^{(j)}, \ldots, \theta_M^{(j)}$ are considered to be drawn from the joint-density $\pi(\theta_0, \theta_1, \ldots, \theta_M)$.

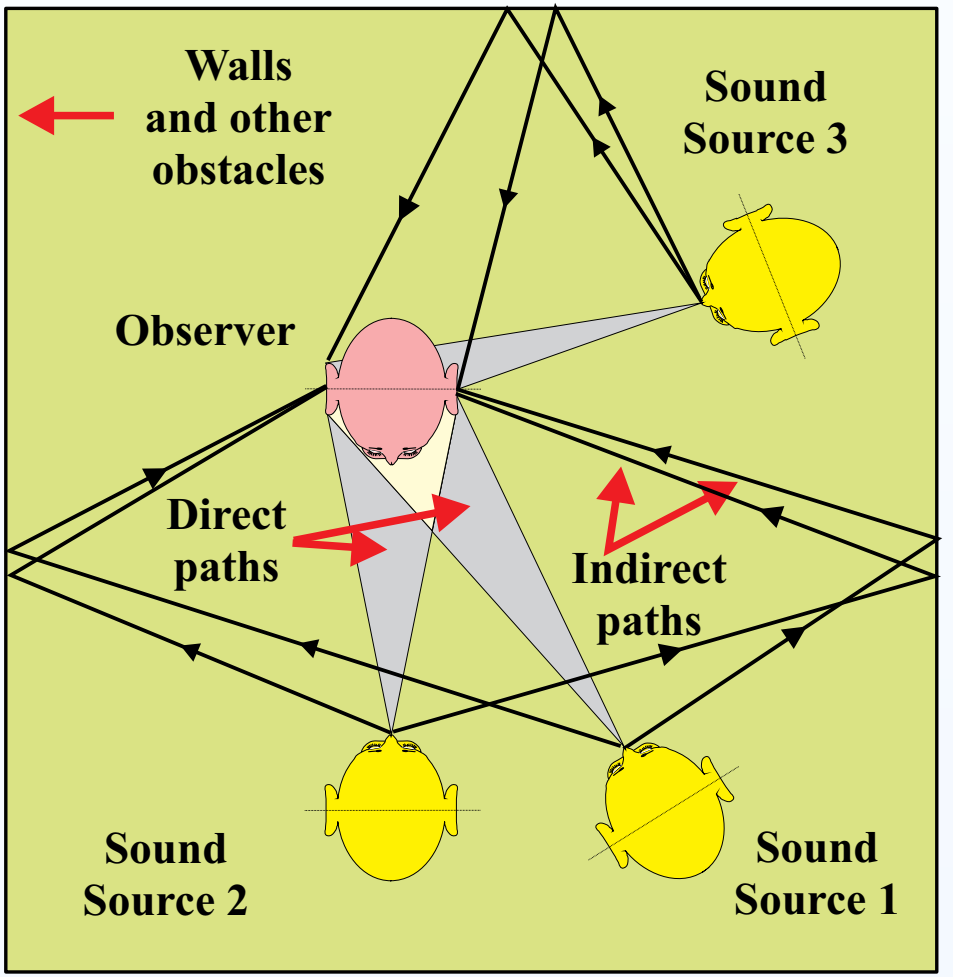
Handout 7
Passive Target Localisation

Introduction



Source localisation and BSS.

Introduction



Humans turn their head in the direction of interest in order to reduce interference from other directions; *joint detection, localisation, and enhancement*.

Introduction

- This research tutorial is intended to cover a wide range of aspects which link acoustic source localisation (ASL) and blind source separation (BSS).
- This tutorial is being continually updated, and feedback is welcomed. The documents published on the USB stick may differ to the slides presented on the day.
- The latest version of this document can be found online and downloaded at:

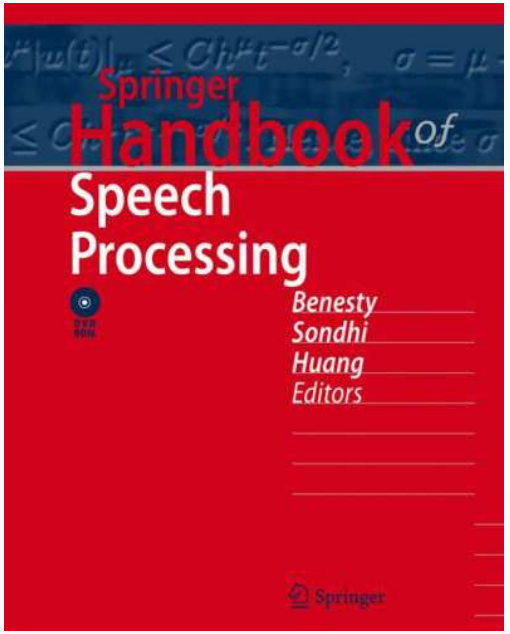
  `http://mod-udrc.org/events/2016-summer-school`

- Thanks to Xionghu Zhong and Ashley Hughes for borrowing some of their diagrams from their dissertations.

Structure of the Tutorial

- Recommended Texts
- Conceptual link between ASL and BSS.
- Geometry of source localisation.
- Spherical and hyperboloidal localisation.
- Estimating TDOAs.
- Steered beamformer response function.
- Multiple target localisation using BSS.
- Conclusions.

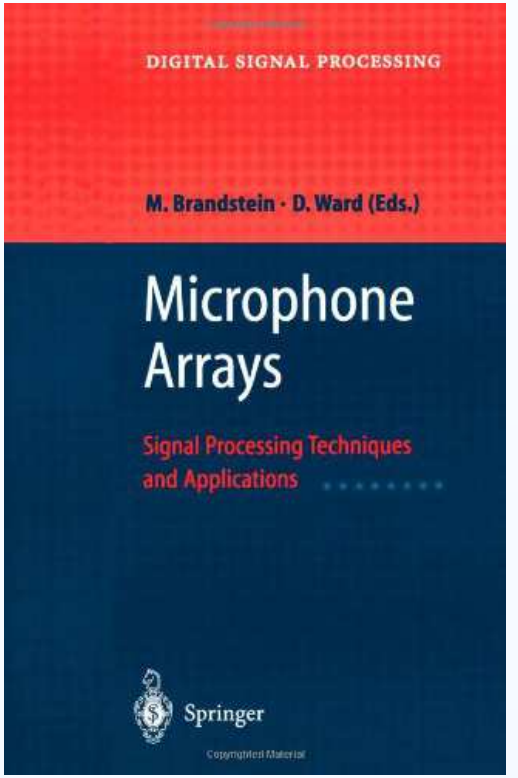
Recommended Texts



Recommended book chapters and the references therein.

- Huang Y., J. Benesty, and J. Chen, "Time Delay Estimation and Source Localization," in *Springer Handbook of Speech Processing* by J. Benesty, M. M. Sondhi, and Y. Huang, pp. 1043–1063, , Springer, 2008.

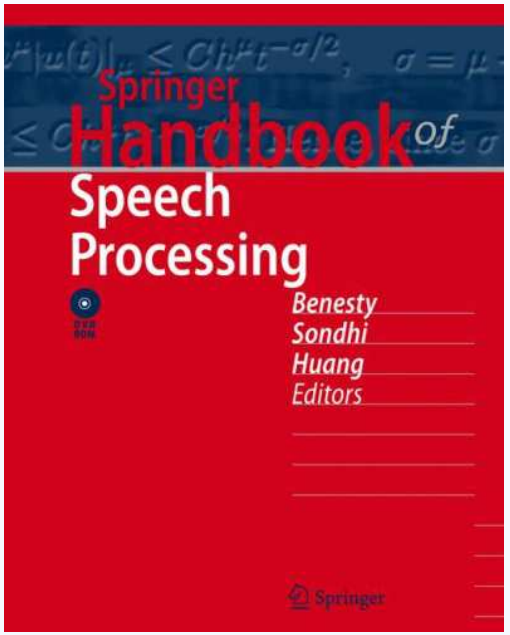
Recommended Texts

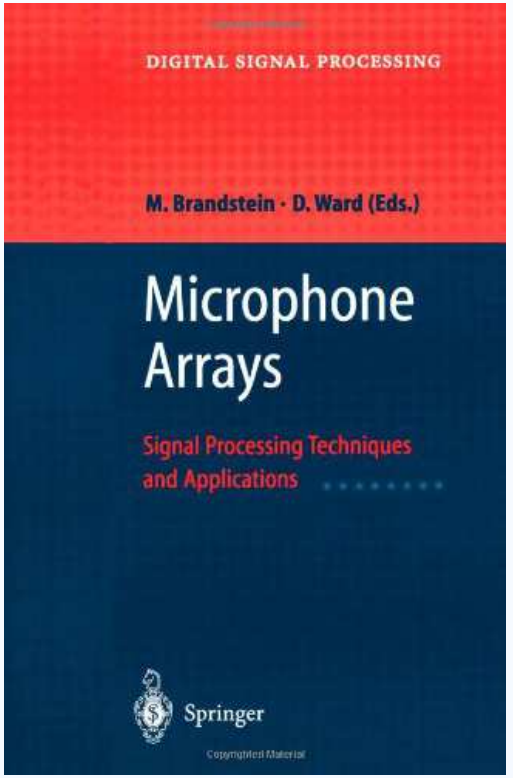


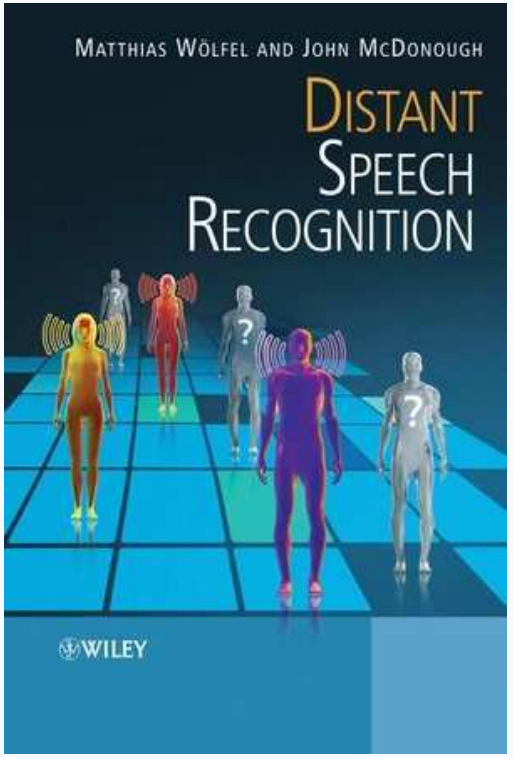
Recommended book chapters and the references therein.

- Chapter 8: DiBiase J. H., H. F. Silverman, and M. S. Brandstein, "Robust Localization in Reverberant Rooms," in *Microphone Arrays* by M. Brandstein and D. Ward, pp. 157–180, , Springer Berlin Heidelberg, 2001.

Recommended Texts







Recommended book chapters and the references therein.

- Chapter 10 of Wolfel M. and J. McDonough, *Distant Speech Recognition*, Wiley, 2009.

IDENTIFIERS – *Hardback*, ISBN13: 978-0-470-51704-8

Recommended Texts

Some recent PhD thesis on the topic include:

- Zhong X., *"Bayesian framework for multiple acoustic source tracking,"* Ph.D. thesis, University of Edinburgh, 2010.
- Pertila P., *"Acoustic Source Localization in a Room Environment and at Moderate Distances,"* Ph.D. thesis, Tampere University of Technology, 2009.
- Fallon M., *"Acoustic Source Tracking using Sequential Monte Carlo,"* Ph.D. thesis, University of Cambridge, 2008.

Why Source Localisation?

A number of blind source separation (BSS) techniques rely on knowledge of the desired source position:

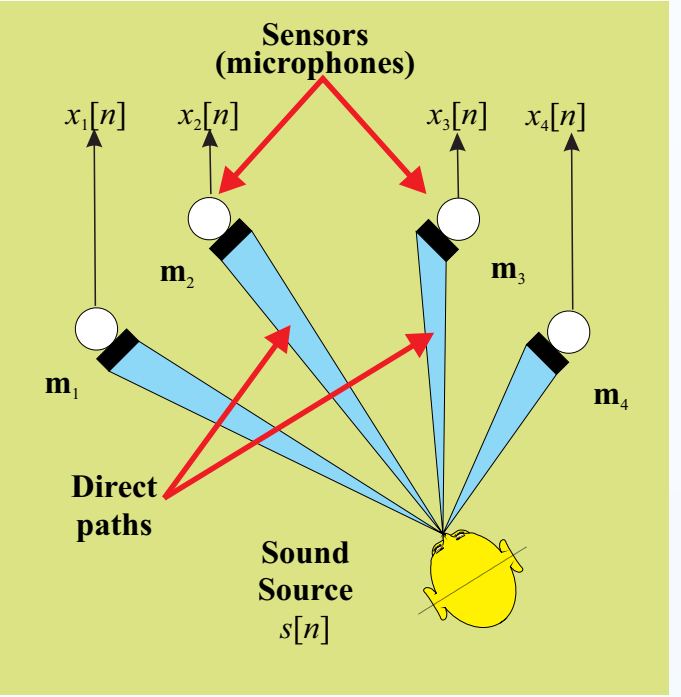
1. Look-direction in beamforming techniques.

2. Camera steering for audio-visual BSS (including Robot Audition).

3. Parametric modelling of the mixing matrix.

Equally, a number of multi-target acoustic source localisation (ASL) techniques rely on BSS.

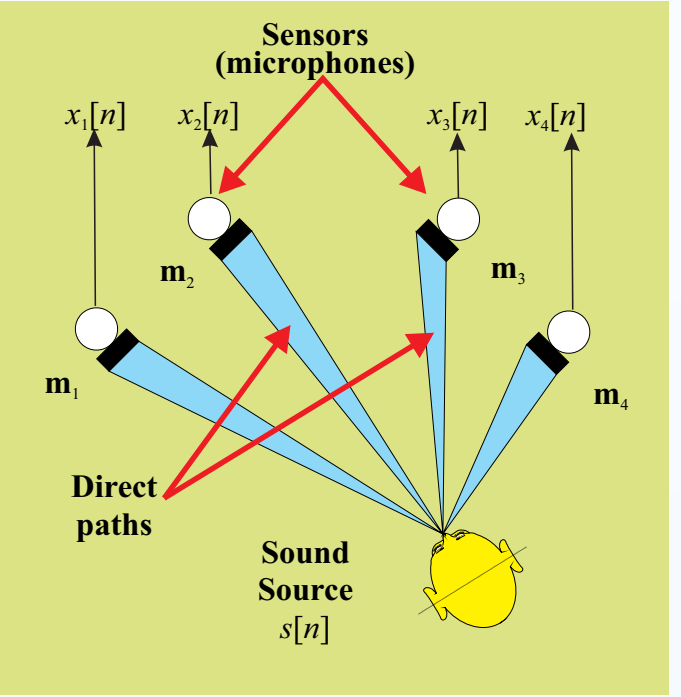
ASL Methodology



Ideal free-field model.

- Most ASL techniques rely on the fact that an impinging wavefront reaches one sensor before it reaches another.

ASL Methodology



Ideal free-field model.

- Most ASL techniques rely on the fact that an impinging wavefront reaches one sensor before it reaches another.
- Most ASL algorithms are designed assuming there is no reverberation present, the *free-field assumption*.

ASL Methodology

An uniform linear array (ULA) of microphones.

- Typically, this acoustic sensor is a microphone; will primarily consider *omni-directional pressure sensors*, and rely on the TDOA between the signals at different microphones.

ASL Methodology

An ULA of microphones.

- Typically, this acoustic sensor is a microphone; will primarily consider *omni-directional pressure sensors*, and rely on the TDOA between the signals at different microphones.
- Other measurement types include:
  - range difference measurements;
  - interaural level difference;
  - joint TDOA and vision techniques.

ASL Methodology

- Another sensor modality might include acoustic vector sensors (AVSs) which measure both air pressure and air velocity. Useful for applications such as sniper localisation.

An acoustic vector sensor.

Source Localization Strategies

Existing source localisation methods can loosely be divided into three generic strategies:

1. those based on maximising the SRP of a beamformer;

   - location estimate derived directly from a filtered, weighted, and sum version of the signal data.

Source Localization Strategies

Existing source localisation methods can loosely be divided into three generic strategies:

1. those based on maximising the SRP of a beamformer;
   - location estimate derived directly from a filtered, weighted, and sum version of the signal data.
2. techniques adopting high-resolution spectral estimation concepts (see Stephan Weiss's talk);
   - any localisation scheme relying upon an application of the signal correlation matrix.

Source Localization Strategies

Existing source localisation methods can loosely be divided into three generic strategies:

1. those based on maximising the SRP of a beamformer;
   - location estimate derived directly from a filtered, weighted, and sum version of the signal data.
2. techniques adopting high-resolution spectral estimation concepts (see Stephan Weiss's talk);
   - any localisation scheme relying upon an application of the signal correlation matrix.
3. approaches employing TDOA information.
   - source locations calculated from a set of TDOA estimates measured across various combinations of microphones.

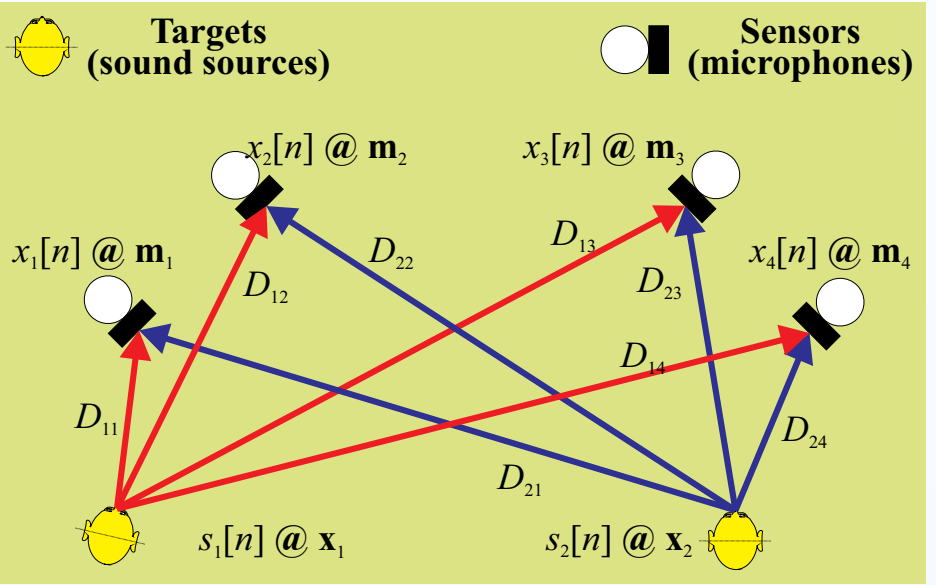
Source Localization Strategies

Spectral-estimation approaches See Stephan Weiss's talk :-)

TDOA-based estimators Computationally cheap, but suffers in the presence of noise and reverberation.

SBF approaches Computationally intensive, superior performance to TDOA-based methods. However, possible to dramatically reduce computational load.

Geometric Layout

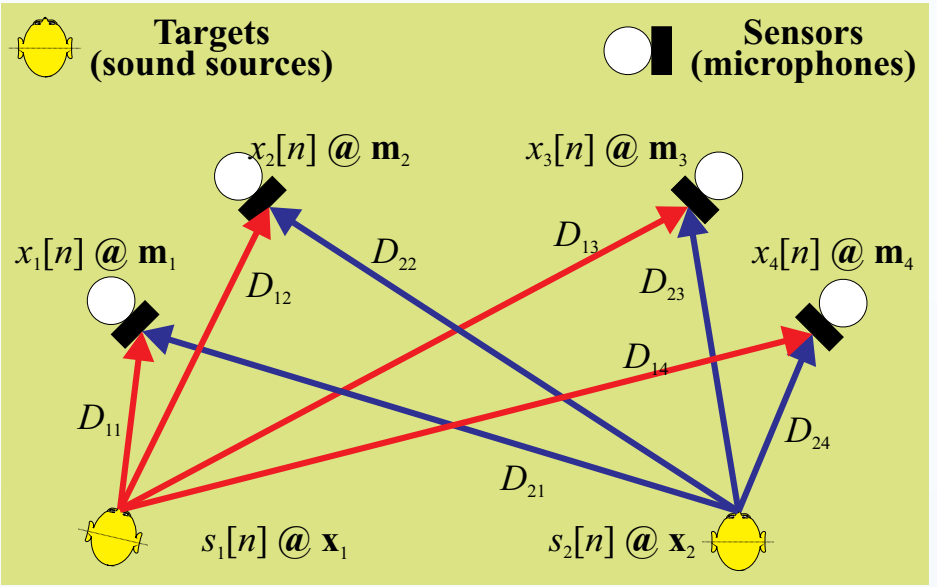


Geometry assuming a free-field model.

Suppose there is a:

- sensor array consisting of N microphones located at positions $\mathbf{m}_i \in \mathbb{R}^3$, for $i \in \{0, \ldots, N-1\}$,
- M talkers (or targets) at positions $\mathbf{x}_k \in \mathbb{R}^3$, for $k \in \{0, \ldots, M-1\}$.

Geometric Layout



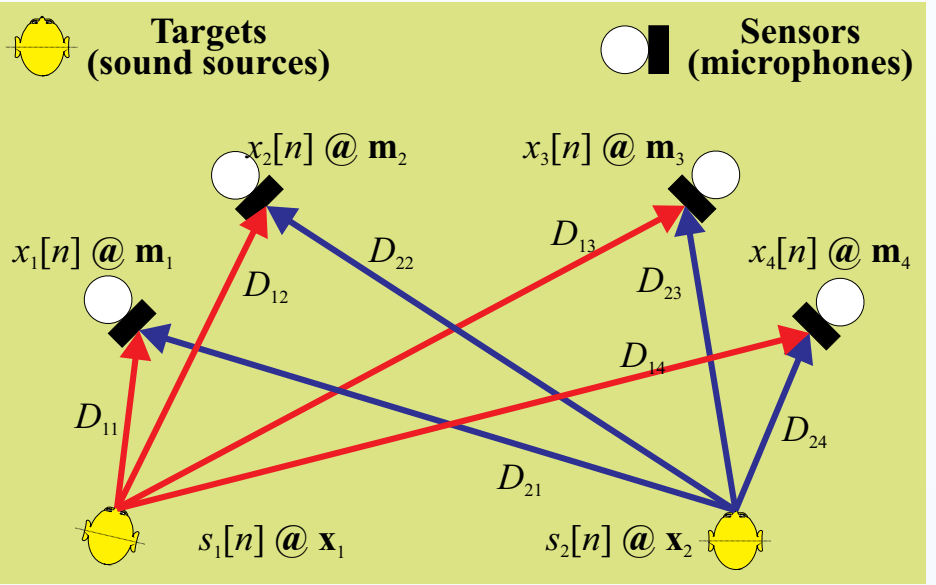
Geometry assuming a free-field model.

The TDOA between the microphones at position $\mathbf{m}_i$ and $\mathbf{m}_j$ due to a source at $\mathbf{x}_k$ can be expressed as:

$$T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right) \triangleq T_{ij}\left(\mathbf{x}_k\right) = \frac{\left|\mathbf{x}_k - \mathbf{m}_i\right| - \left|\mathbf{x}_k - \mathbf{m}_j\right|}{c}$$

where c is the speed of sound, which is approximately 344 m/s.

Geometric Layout



Geometry assuming a free-field model.

The distance from the target at $\mathbf{x}_k$ to the sensor located at $\mathbf{m}_i$ will be defined by D_{ik}, and is called the range.

$$T_{ij}\left(\mathbf{x}_k\right) = \frac{1}{c}\left(D_{ik} - D_{jk}\right)$$

Ideal Free-field Model

- In an anechoic free-field acoustic environment, the signal from source k, denoted by $s_k(t)$, propagates to the i-th sensor at time t according to the expression:

$$x_{ik}(t) = \alpha_{ik}\, s_k(t - \tau_{ik}) + b_{ik}(t)$$

where $b_{ik}(t)$ denotes additive noise. Note that, in the frequency domain, this expression is given by:

$$X_{ik}(\omega) = \alpha_{ik}\, S_k(\omega)\, e^{-j\omega\,\tau_{ik}} + B_{ik}(\omega)$$

- The additive noise source is assumed to be uncorrelated with the source signal, as well as the noise signals at the other microphones.

Ideal Free-field Model

- In an anechoic free-field acoustic environment, the signal from source k, denoted by $s_k(t)$, propagates to the i-th sensor at time t according to the expression:

$$x_{ik}(t) = \alpha_{ik}\, s_k(t - \tau_{ik}) + b_{ik}(t)$$

where $b_{ik}(t)$ denotes additive noise. Note that, in the frequency domain, this expression is given by:

$$X_{ik}(\omega) = \alpha_{ik}\, S_k(\omega)\, e^{-j\omega\,\tau_{ik}} + B_{ik}(\omega)$$

- The additive noise source is assumed to be uncorrelated with the source signal, as well as the noise signals at the other microphones.

- The TDOA between the i-th and j-th microphone is given by:

$$\tau_{ijk} = \tau_{ik} - \tau_{jk} = T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k)$$

TDOA and Hyperboloids

It is important to be aware of the geometrical properties that arise from the TDOA relationship

$$T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right) = \frac{\left|\mathbf{x}_k - \mathbf{m}_i\right| - \left|\mathbf{x}_k - \mathbf{m}_j\right|}{c}$$

TDOA and Hyperboloids

It is important to be aware of the geometrical properties that arise from the TDOA relationship

$$T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right) = \frac{\left|\mathbf{x}_k - \mathbf{m}_i\right| - \left|\mathbf{x}_k - \mathbf{m}_j\right|}{c}$$

- This defines one half of a hyperboloid of two sheets, centered on the midpoint of the microphones, $\mathbf{v}_{ij} = \frac{\mathbf{m}_i + \mathbf{m}_j}{2}$.

$$\left(\mathbf{x}_k - \mathbf{v}_{ij}\right)^T \mathbf{V}_{ij} \left(\mathbf{x}_k - \mathbf{v}_{ij}\right) = 1$$

TDOA and Hyperboloids

It is important to be aware of the geometrical properties that arise from the TDOA relationship

$$T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right) = \frac{\left|\mathbf{x}_k - \mathbf{m}_i\right| - \left|\mathbf{x}_k - \mathbf{m}_j\right|}{c}$$

- This defines one half of a hyperboloid of two sheets, centered on the midpoint of the microphones, $\mathbf{v}_{ij} = \frac{\mathbf{m}_i + \mathbf{m}_j}{2}$.

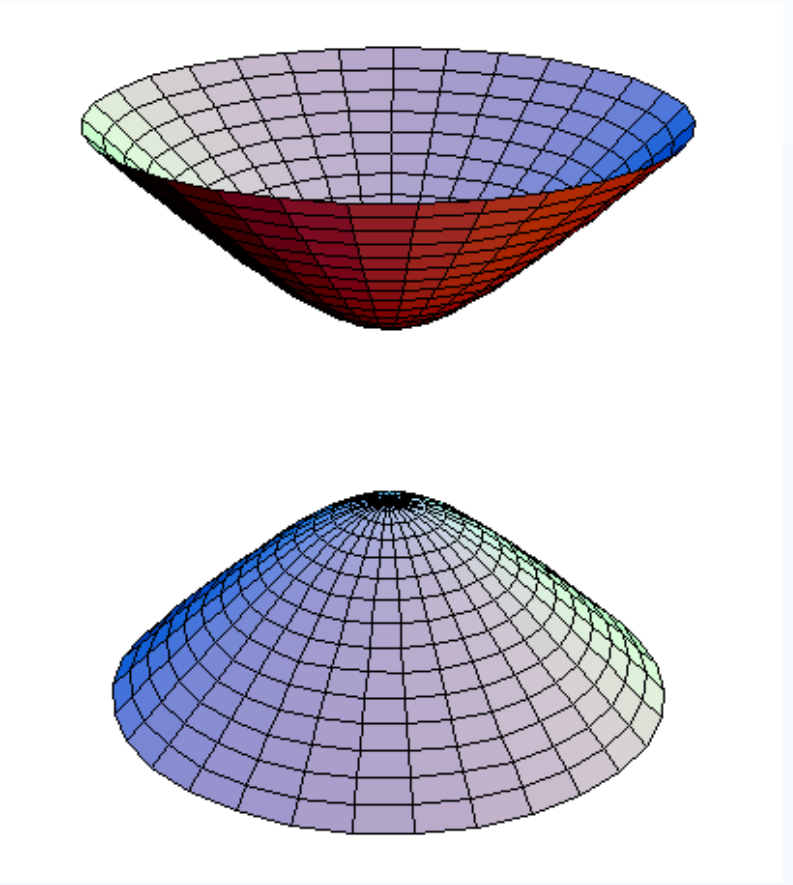
$$\left(\mathbf{x}_k - \mathbf{v}_{ij}\right)^T \mathbf{V}_{ij} \left(\mathbf{x}_k - \mathbf{v}_{ij}\right) = 1$$

- For source with a large source-range to microphone-separation ratio, the hyperboloid may be well-approximated by a cone with a constant direction angle relative to the axis of symmetry.

$$\phi_{ij} = \cos^{-1}\left(\frac{c\,T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right)}{\left|\mathbf{m}_i - \mathbf{m}_j\right|}\right)$$

TDOA and Hyperboloids

$$T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k) = \frac{|\mathbf{x}_k - \mathbf{m}_i| - |\mathbf{x}_k - \mathbf{m}_j|}{c}$$

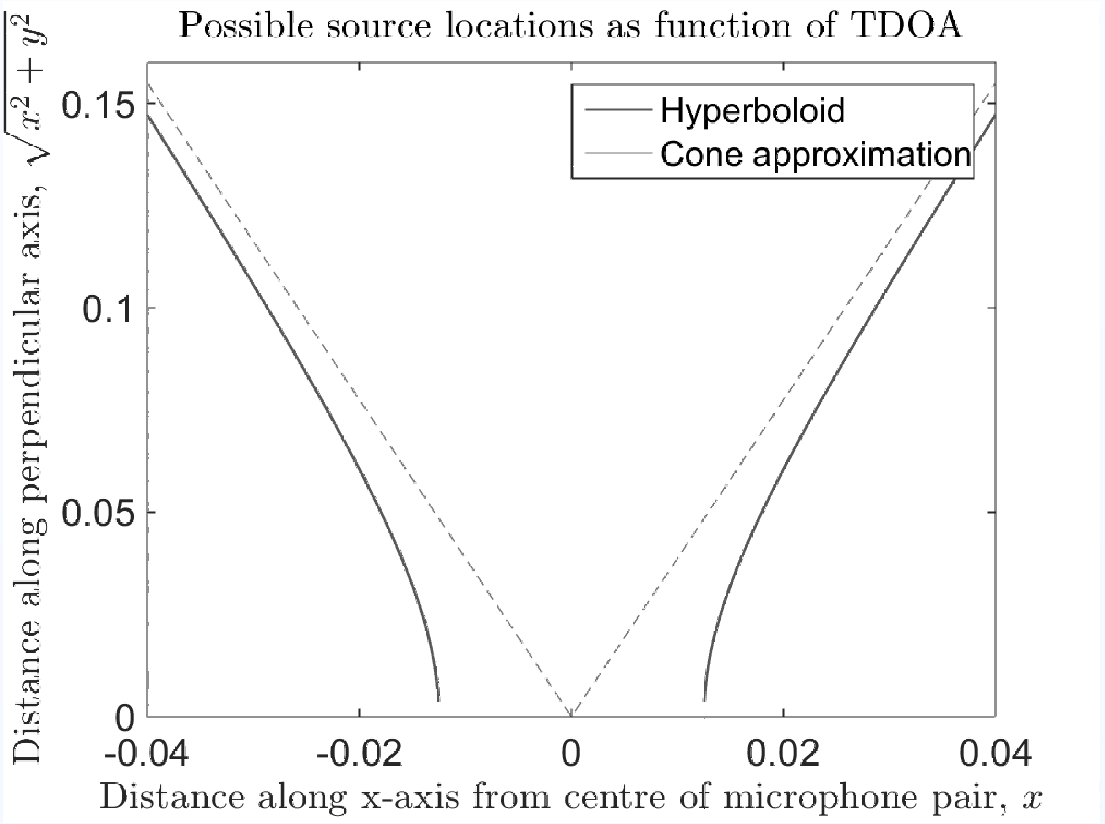


Hyperboloid of two sheets

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = -1$$

TDOA and Hyperboloids

$$T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right) = \frac{\left|\mathbf{x}_k - \mathbf{m}_i\right| - \left|\mathbf{x}_k - \mathbf{m}_j\right|}{c}$$



Hyperboloid, for a microphone separation of $d = 0.1$, and a time-delay of $\tau_{ij} = \frac{d}{4c}$.

Indirect TDOA-based Methods

This is typically a two-step procedure in which:

- Typically, TDOAs are extracted using the GCC function, or an AED algorithm.

Indirect TDOA-based Methods

This is typically a two-step procedure in which:

- Typically, TDOAs are extracted using the GCC function, or an AED algorithm.
- A hypothesised spatial position of the target can be used to predict the expected TDOAs (or corresponding range) at the microphone.

Indirect TDOA-based Methods

This is typically a two-step procedure in which:

- Typically, TDOAs are extracted using the GCC function, or an AED algorithm.
- A hypothesised spatial position of the target can be used to predict the expected TDOAs (or corresponding range) at the microphone.
- The error between the measured and hypothesised TDOAs is then minimised.

Indirect TDOA-based Methods

This is typically a two-step procedure in which:

- Typically, TDOAs are extracted using the GCC function, or an AED algorithm.
- A hypothesised spatial position of the target can be used to predict the expected TDOAs (or corresponding range) at the microphone.
- The error between the measured and hypothesised TDOAs is then minimised.
- Accurate and robust TDOA estimation is the key to the effectiveness of this class of ASL methods.

Indirect TDOA-based Methods

This is typically a two-step procedure in which:

- Typically, TDOAs are extracted using the GCC function, or an AED algorithm.
- A hypothesised spatial position of the target can be used to predict the expected TDOAs (or corresponding range) at the microphone.
- The error between the measured and hypothesised TDOAs is then minimised.
- Accurate and robust TDOA estimation is the key to the effectiveness of this class of ASL methods.
- An alternative way of viewing these solutions is to consider what **spatial positions** of the target could lead to the estimated TDOA.

Spherical Least Squares Error Function

- Suppose the first microphone is located at the origin of the coordinate system, such that $\mathbf{m_0} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T$.

- The range from target k to sensor i can be expressed as :

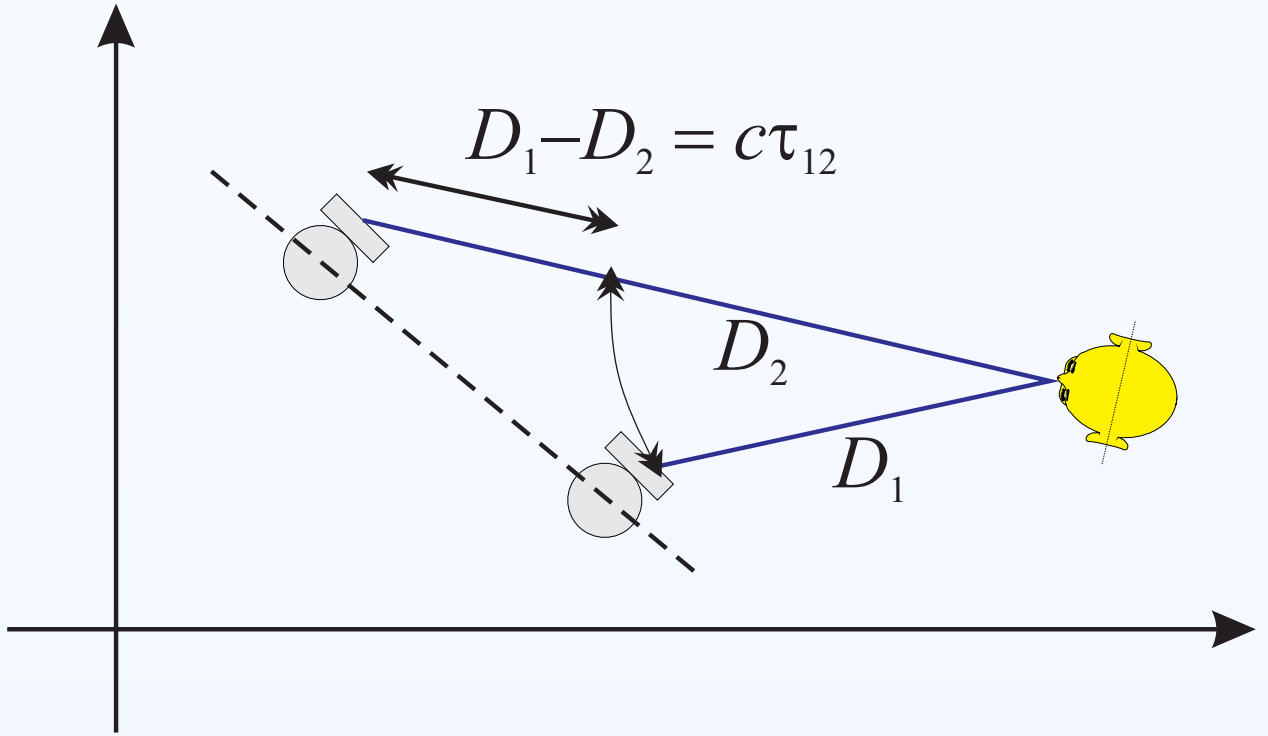
$$\begin{aligned} D_{ik} &= D_{0k} + D_{ik} - D_{0k} \\ &= R_s + c\,T_{i0}\left(\mathbf{x}_k\right) \end{aligned}$$

where $R_{sk} = |\mathbf{x}_k|$ is the range to the first microphone which is at the origin.

Spherical Least Squares Error Function

- In practice, the observations are the TDOAs and, given R_{sk}, these ranges can be considered the **measurement ranges**.

Of course, knowing R_{sk} is half the solution, but it is just one unknown at this stage.



Range and TDOA relationship.

Spherical Least Squares Error Function

- The source-sensor geometry states that the target lies on a sphere centered on the corresponding sensor. Hence,

$$\begin{aligned} D_{ik}^2 &= |\mathbf{x}_k - \mathbf{m}_i|^2 \\ &= \mathbf{x}_k^T \mathbf{x}_k - 2\mathbf{m}_i^T \mathbf{x}_k + \mathbf{m}_i^T \mathbf{m}_i \\ &= R_s^2 - 2\mathbf{m}_i^T \mathbf{x}_k + R_i^2 \end{aligned}$$

$R_i = |\mathbf{m}_i|$ is the distance of the i-th microphone to the origin.

Spherical Least Squares Error Function

- The source-sensor geometry states that the target lies on a sphere centered on the corresponding sensor. Hence,

$$\begin{aligned} D_{ik}^2 &= |\mathbf{x}_k - \mathbf{m}_i|^2 \\ &= \mathbf{x}_k^T \mathbf{x}_k - 2\mathbf{m}_i^T \mathbf{x}_k + \mathbf{m}_i^T \mathbf{m}_i \\ &= R_s^2 - 2\mathbf{m}_i^T \mathbf{x}_k + R_i^2 \end{aligned}$$

- Define the **spherical error function** as:

$$\epsilon_{ik} \triangleq \frac{1}{2}\left(\hat{D}_{ik}^2 - D_{ik}^2\right)$$

Spherical Least Squares Error Function

- The source-sensor geometry states that the target lies on a sphere centered on the corresponding sensor. Hence,

$$\begin{aligned} D_{ik}^2 &= |\mathbf{x}_k - \mathbf{m}_i|^2 \\ &= \mathbf{x}_k^T \mathbf{x}_k - 2\mathbf{m}_i^T \mathbf{x}_k + \mathbf{m}_i^T \mathbf{m}_i \\ &= R_s^2 - 2\mathbf{m}_i^T \mathbf{x}_k + R_i^2 \end{aligned}$$

- Define the **spherical error function** as:

$$\begin{aligned} \epsilon_{ik} &\triangleq \frac{1}{2}\left(\hat{D}_{ik}^2 - D_{ik}^2\right) \\ &= \frac{1}{2}\left\{\left(R_s + c\hat{T}_{i0}\right)^2 - \left(R_s^2 - 2\mathbf{m}_i^T \mathbf{x}_k + R_i^2\right)\right\} \end{aligned}$$

Spherical Least Squares Error Function

- The source-sensor geometry states that the target lies on a sphere centered on the corresponding sensor. Hence,

$$\begin{aligned} D_{ik}^2 &= |\mathbf{x}_k - \mathbf{m}_i|^2 \\ &= \mathbf{x}_k^T \mathbf{x}_k - 2\mathbf{m}_i^T \mathbf{x}_k + \mathbf{m}_i^T \mathbf{m}_i \\ &= R_s^2 - 2\mathbf{m}_i^T \mathbf{x}_k + R_i^2 \end{aligned}$$

- Define the **spherical error function** as:

$$\begin{aligned} \epsilon_{ik} &\triangleq \frac{1}{2}\left(\hat{D}_{ik}^2 - D_{ik}^2\right) \\ &= \frac{1}{2}\left\{\left(R_s + c\,\hat{T}_{i0}\right)^2 - \left(R_s^2 - 2\mathbf{m}_i^T \mathbf{x}_k + R_i^2\right)\right\} \\ &= \mathbf{m}_i^T \mathbf{x_k} + c\,R_s\,\hat{T}_{i0} + \frac{1}{2}\left(c^2\hat{T}_{i0}^2 - R_i^2\right) \end{aligned}$$

Spherical Least Squares Error Function

- Concatenating the error functions for each microphone gives the expression:

$$\begin{aligned}\boldsymbol{\epsilon}_{ik} &= \mathbf{A}\,\mathbf{x}_k - \underbrace{(\mathbf{b}_k - R_{sk}\mathbf{d}_k)}_{\mathbf{v}_k} \\ &\equiv \underbrace{\begin{bmatrix}\mathbf{A} & \mathbf{d}_k\end{bmatrix}}_{\mathbf{S}_k} \underbrace{\begin{bmatrix}\mathbf{x}_k \\ R_{sk}\end{bmatrix}}_{\boldsymbol{\theta}_k} - \mathbf{b}_k\end{aligned}$$

where

$$\mathbf{A} = \begin{bmatrix}\mathbf{m}_0^T \\ \vdots \\ \mathbf{m}_{N-1}^T\end{bmatrix}, \mathbf{d} = c\begin{bmatrix}\hat{T}_{00} \\ \vdots \\ \hat{T}_{(N-1)0}\end{bmatrix}, \quad \mathbf{b}_k = \frac{1}{2}\begin{bmatrix}c^2\hat{T}_{00}^2 - R_0^2 \\ \vdots \\ c^2\hat{T}_{(N-1)0}^2 - R_{N-1}^2\end{bmatrix}$$

Spherical Least Squares Error Function

- The LSE can then be obtained by using $J = \boldsymbol{\epsilon}_i^T \boldsymbol{\epsilon}_i$:

$$J(\mathbf{x}_k) = \left(\mathbf{A}\mathbf{x}_k - \left(\mathbf{b}_k - R_{sk}\,\mathbf{d}_k\right)\right)^T \left(\mathbf{A}\mathbf{x}_k - \left(\mathbf{b}_k - R_{sk}\,\mathbf{d}_k\right)\right)$$

$$J\left(\mathbf{x}_k,\, \boldsymbol{\theta}_k\right) = \left(\mathbf{S}_k\boldsymbol{\theta}_k - \mathbf{b}_k\right)^T \left(\mathbf{S}_k\boldsymbol{\theta}_k - \mathbf{b}_k\right)$$

Spherical Least Squares Error Function

- The LSE can then be obtained by using $J = \boldsymbol{\epsilon}_i^T \boldsymbol{\epsilon}_i$:

$$J(\mathbf{x}_k) = (\mathbf{A}\mathbf{x}_k - (\mathbf{b}_k - R_{sk}\,\mathbf{d}_k))^T (\mathbf{A}\mathbf{x}_k - (\mathbf{b}_k - R_{sk}\,\mathbf{d}_k))$$

$$J(\mathbf{x}_k, \boldsymbol{\theta}_k) = (\mathbf{S}_k\boldsymbol{\theta}_k - \mathbf{b}_k)^T (\mathbf{S}_k\boldsymbol{\theta}_k - \mathbf{b}_k)$$

- Note that as $R_{sk} = |\mathbf{x}_k|$, these parameters aren't independent. Therefore, the problem can either be formulated as:
  - a nonlinear least-squares problem in $\mathbf{x}_k$;
  - a linear minimisation subject to quadratic constraints:

$$\hat{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta}_k} (\mathbf{S}_k\boldsymbol{\theta}_k - \mathbf{b}_k)^T (\mathbf{S}_k\boldsymbol{\theta}_k - \mathbf{b}_k)$$

subject to the constraint

$$\boldsymbol{\theta}_k\,\Delta\,\boldsymbol{\theta}_k = 0 \qquad \text{where} \qquad \Delta = \text{diag}\,[1,\,1,\,1,\,-1]$$

Spherical Least Squares Error Function

Two-step Spherical LSE Approaches

To avoid solving either a nonlinear or a constrained least-squares problem, it is possible to solve the problem in two steps, namely:

1. solving a LLS problem in $\mathbf{x}_k$ *assuming* the range to the target, R_{sk}, is known;

2. and then solving for R_{sk} given an estimate of $\mathbf{x}_k$ i. t. o. R_{sk}.

Two-step Spherical LSE Approaches

To avoid solving either a nonlinear or a constrained least-squares problem, it is possible to solve the problem in two steps, namely:

1. solving a LLS problem in $\mathbf{x}_k$ *assuming* the range to the target, R_{sk}, is known;

2. and then solving for R_{sk} given an estimate of $\mathbf{x}_k$ i. t. o. R_{sk}.

- Assuming an estimate of R_{sk} this can be solved as

$$\hat{\mathbf{x}}_k = \mathbf{A}^\dagger \mathbf{v}_k = \mathbf{A}^\dagger \left(\mathbf{b}_k - \hat{R}_{sk} \mathbf{d}_k \right) \quad \text{where} \quad \mathbf{A}^\dagger = \left[\mathbf{A}^T \mathbf{A} \right]^{-1} \mathbf{A}^T$$

Note that $\mathbf{A}^\dagger$ is the pseudo-inverse of $\mathbf{A}$.

Spherical Intersection Estimator

This method uses the physical constraint that the range R_{sk} is the Euclidean distance to the target.

- Writing $\hat{R}_{sk}^2 = \hat{\mathbf{x}}_k^T \hat{\mathbf{x}}_k$, it follows that:

$$\hat{R}_{sk}^2 = \left(\mathbf{b}_k - \hat{R}_{sk}\mathbf{d}_k\right)^T \mathbf{A}^{\dagger\,T} \mathbf{A}^{\dagger} \left(\mathbf{b}_k - \hat{R}_{sk}\mathbf{d}_k\right)$$

Spherical Intersection Estimator

This method uses the physical constraint that the range R_{sk} is the Euclidean distance to the target.

- Writing $\hat{R}_{sk}^2 = \hat{\mathbf{x}}_k^T \, \hat{\mathbf{x}}_k$, it follows that:

$$\hat{R}_{sk}^2 = \left(\mathbf{b}_k - \hat{R}_{sk} \mathbf{d}_k \right)^T \mathbf{A}^{\dagger \, T} \mathbf{A}^{\dagger} \left(\mathbf{b}_k - \hat{R}_{sk} \mathbf{d}_k \right)$$

which can be written as the quadratic:

$$a \, \hat{R}_{sk}^2 + b \, \hat{R}_{sk} + c = 0$$

Spherical Intersection Estimator

This method uses the physical constraint that the range R_{sk} is the Euclidean distance to the target.

- Writing $\hat{R}_{sk}^2 = \hat{\mathbf{x}}_k^T \hat{\mathbf{x}}_k$, it follows that:

$$\hat{R}_{sk}^2 = \left(\mathbf{b}_k - \hat{R}_{sk}\mathbf{d}_k\right)^T \mathbf{A}^{\dagger\,T} \mathbf{A}^{\dagger} \left(\mathbf{b}_k - \hat{R}_{sk}\mathbf{d}_k\right)$$

  which can be written as the quadratic:

$$a\,\hat{R}_{sk}^2 + b\,\hat{R}_{sk} + c = 0$$

- The unique, real, positive root is taken as the spherical intersection (SX) estimator of the source range. Hence, the estimator will fail when:

  1. there is no real, positive root, or:

  2. if there are two positive real roots.

Spherical Interpolation Estimator

The spherical interpolation (SI) estimator again uses the spherical least squares error (LSE) function, but this time the range R_{sk} is estimated in the least-squares sense.

Consider again the **spherical error function**:

$$\boldsymbol{\epsilon}_{ik} = \mathbf{A}\mathbf{x}_k - (\mathbf{b}_k - R_{sk}\,\mathbf{d}_k)$$

Spherical Interpolation Estimator

The SI estimator again uses the spherical LSE function, but this time the range R_{sk} is estimated in the least-squares sense.

Consider again the **spherical error function**:

$$\boldsymbol{\epsilon}_{ik} = \mathbf{A}\mathbf{x}_k - (\mathbf{b}_k - R_{sk}\,\mathbf{d}_k)$$

Substituting the LSE gives:

$$\boldsymbol{\epsilon}_{ik} = \mathbf{A}\left[\mathbf{A}^T\mathbf{A}\right]^{-1}\mathbf{A}^T\left(\mathbf{b}_k - \hat{R}_{sk}\mathbf{d}_k\right) - (\mathbf{b}_k - R_{sk}\,\mathbf{d}_k)$$

Spherical Interpolation Estimator

The SI estimator again uses the spherical LSE function, but this time the range R_{sk} is estimated in the least-squares sense.

Consider again the **spherical error function**:

$$\boldsymbol{\epsilon}_{ik} = \mathbf{A}\mathbf{x}_k - (\mathbf{b}_k - R_{sk}\,\mathbf{d}_k)$$

Substituting the LSE gives:

$$\boldsymbol{\epsilon}_{ik} = \mathbf{A}\left[\mathbf{A}^T\mathbf{A}\right]^{-1}\mathbf{A}^T\left(\mathbf{b}_k - \hat{R}_{sk}\mathbf{d}_k\right) - (\mathbf{b}_k - R_{sk}\,\mathbf{d}_k)$$

Defining the projection matrix as $\mathbf{P}_\mathbf{A} = \mathbf{I}_N - \mathbf{A}\left[\mathbf{A}^T\mathbf{A}\right]^{-1}\mathbf{A}^T$,

$$\boldsymbol{\epsilon}_{ik} = R_{sk}\,\mathbf{P}_\mathbf{A}\mathbf{d}_k - \mathbf{P}_\mathbf{A}\mathbf{b}_k$$

Spherical Interpolation Estimator

The SI estimator again uses the spherical LSE function, but this time the range R_{sk} is estimated in the least-squares sense.

Consider again the **spherical error function**:

$$\boldsymbol{\epsilon}_{ik} = \mathbf{A}\mathbf{x}_k - (\mathbf{b}_k - R_{sk}\,\mathbf{d}_k)$$

Defining the projection matrix as $\mathbf{P_A} = \mathbf{I}_N - \mathbf{A}\left[\mathbf{A}^T\mathbf{A}\right]^{-1}\mathbf{A}^T$,

$$\boldsymbol{\epsilon}_{ik} = R_{sk}\,\mathbf{P_A}\mathbf{d}_k - \mathbf{P_A}\mathbf{b}_k$$

Minimising the LSE using the normal equations gives:

$$R_{sk} = \frac{\mathbf{d}_k^T\mathbf{P_A}\mathbf{b}_k}{\mathbf{d}_k^T\,\mathbf{P_A}\mathbf{d}_k}$$

Spherical Interpolation Estimator

The SI estimator again uses the spherical LSE function, but this time the range R_{sk} is estimated in the least-squares sense.

Consider again the **spherical error function**:

$$\boldsymbol{\epsilon}_{ik} = \mathbf{A}\mathbf{x}_k - (\mathbf{b}_k - R_{sk}\,\mathbf{d}_k)$$

Substituting back into the LSE for the target position gives the final estimator:

$$\hat{\mathbf{x}}_k = \mathbf{A}^{\dagger}\left(\mathbf{I}_N - \mathbf{d}_k \frac{\mathbf{d}_k^T \mathbf{P}_{\mathbf{A}}}{\mathbf{d}_k^T\,\mathbf{P}_{\mathbf{A}}\mathbf{d}_k}\right)\mathbf{b}_k$$

This approach is said to perform better, but is computationally slightly more complex than the SX estimator.

Other Approaches

There are several other approaches to minimising the spherical LSE function .

- In particular, the **linear-correction** LSE solves the constrained minimization problem using Lagrange multipliers in a two stage process.

- For further information, see: Huang Y., J. Benesty, and J. Chen, “Time Delay Estimation and Source Localization,” in *Springer Handbook of Speech Processing* by J. Benesty, M. M. Sondhi, and Y. Huang, pp. 1043–1063, , Springer, 2008.

Hyperbolic Least Squares Error Function

- If a TDOA is estimated between two microphones i and j, then the error between this and modelled TDOA is:

$$\epsilon_{ij}(\mathbf{x}_k) = \tau_{ijk} - T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right)$$

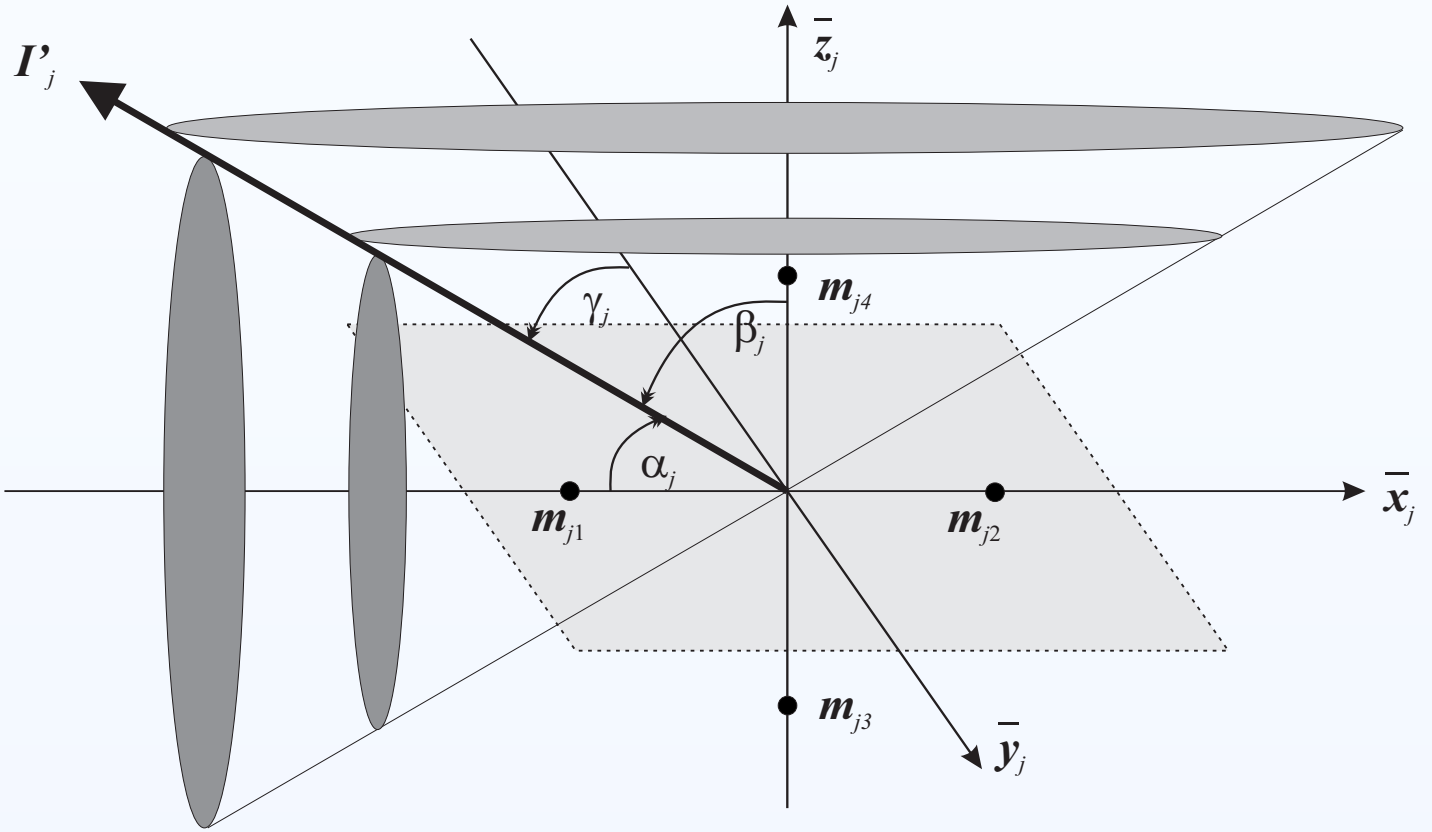
- The total error as a function of target position

$$J(\mathbf{x}_k) = \sum_{i=1}^{N} \sum_{j \neq i=1}^{N} \left(\tau_{ijk} - T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right)\right)^2$$

- Unfortunately, since $T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right)$ is a nonlinear function of $\mathbf{x}_k$, the minimum LSE does not possess a closed-form solution.

Linear Intersection Method

The linear intersection (LI) algorithm works by utilising a *sensor quadruple* with a common midpoint, which allows a bearing line to be deduced from the intersection of two cones.

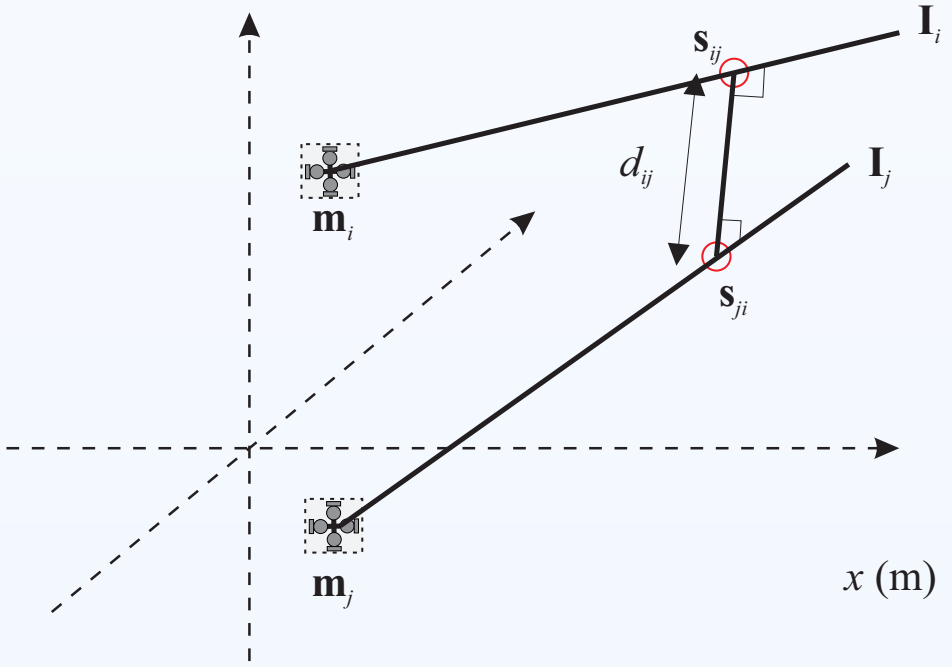


Quadruple sensor arrangement and local Cartesian coordinate system.

Linear Intersection Method

- Given the bearing lines, it is possible to calculate the points $\mathbf{s}_{ij}$ and $\mathbf{s}_{ji}$ on two bearing lines which give the closest intersection. This is basic gemoentry.

- The trick is to note that given these points $\mathbf{s}_{ij}$ and $\mathbf{s}_{ji}$, the theoretical TDOA, $T\left(\mathbf{m}_{1i},\, \mathbf{m}_{2i},\, \mathbf{s}_{ij}\right)$, can be compared with the observed TDOA.



Calculating the points of closest intersection.

TDOA estimation methods

Two key methods for TDOA estimation are using the GCC function and the AED algorithm.

GCC algorithm most popular approach assuming an ideal free-field movel

- computationally efficient, and hence short decision delays;
- perform fairly well in moderately noisy and reverberant environments.

TDOA estimation methods

Two key methods for TDOA estimation are using the GCC function and the AED algorithm.

GCC algorithm most popular approach assuming an ideal free-field movel

- computationally efficient, and hence short decision delays;
- perform fairly well in moderately noisy and reverberant environments.

However, GCC-based methods

- fail when room reverberation is high;
- focus of current research is on combating the effect of room reverberation.

TDOA estimation methods

Two key methods for TDOA estimation are using the GCC function and the AED algorithm.

AED Algorithm Approaches the TDOA estimation approach from a different point of view from the *traditional* GCC method.

- adopts a reverberant rather than free-field model;
- computationally more expensive than GCC;
- can fail when there are common-zeros in the room impulse response (RIR).

GCC TDOA estimation

The GCC algorithm proposed by *Knapp and Carter* is the most widely used approach to TDOA estimation.

- The TDOA estimate between two microphones i and j

$$\hat{\tau_{ij}} = \arg\max_{\ell} r_{x_i\, x_j}[\ell]$$

- The cross-correlation function is given by

$$\begin{aligned} r_{x_i\, x_j}[\ell] &= \mathcal{F}^{-1}\left(\Phi\left(e^{j\omega T_s}\right) P_{x_1 x_2}\left(e^{j\omega T_s}\right)\right) \\ &= \int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}} \Phi\left(e^{j\omega T_s}\right) P_{x_1 x_2}\left(e^{j\omega T_s}\right) e^{j\ell\omega T}\, d\omega \end{aligned}$$

where the CPSD is given by

$$P_{x_1 x_2}\left(e^{j\omega T_s}\right) = \mathbb{E}\left[X_1\left(e^{j\omega T_s}\right) X_2\left(e^{j\omega T_s}\right)\right]$$

CPSD for Free-Field Model

For the free-field model , it follows that for $i \neq j$:

$$
\begin{aligned}
P_{x_i x_j}(\omega) &= \mathbb{E}\left[X_j(\omega)\, X_j(\omega)\right] \\
&= \mathbb{E}\left[\left(\alpha_{ik}\, S_k(\omega)\; e^{-j\omega\,\tau_{ik}} + B_{ik}(\omega)\right)\left(\alpha_{jk}\, S_k(\omega)\; e^{-j\omega\,\tau_{kk}} + B_{jk}(\omega)\right)\right] \\
&= \alpha_{ik}\alpha_{jk} e^{-j\omega\, T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k)} \mathbb{E}\left[\left|S_k(\omega)\right|^2\right]
\end{aligned}
$$

where $\mathbb{E}\left[B_{ik}(\omega)\, B_{jk}(\omega)\right] = 0$ and $\mathbb{E}\left[B_{ik}(\omega)\, S_k(\omega)\right] = 0$.

CPSD for Free-Field Model

For the free-field model , it follows that for $i \neq j$:

$$
\begin{aligned}
P_{x_i x_j}(\omega) &= \mathbb{E}\left[X_j(\omega)\, X_j(\omega)\right] \\
&= \mathbb{E}\left[\left(\alpha_{ik}\, S_k(\omega)\; e^{-j\omega\,\tau_{ik}} + B_{ik}(\omega)\right)\left(\alpha_{jk}\, S_k(\omega)\; e^{-j\omega\,\tau_{kk}} + B_{jk}(\omega)\right)\right] \\
&= \alpha_{ik}\alpha_{jk} e^{-j\omega\, T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k)} \mathbb{E}\left[\left|S_k(\omega)\right|^2\right]
\end{aligned}
$$

where $\mathbb{E}\left[B_{ik}(\omega)\, B_{jk}(\omega)\right] = 0$ and $\mathbb{E}\left[B_{ik}(\omega)\, S_k(\omega)\right] = 0$.

- In particular, note that it follows:

$$\angle P_{x_i x_j}(\omega) = -j\omega\, T(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k)$$

In otherwords, all the TDOA information is conveyed in the phrase rather than the amplitude of the CPSD. This therefore suggests that the weighting function can be chosen to remove the amplitude information.

GCC Processors

Processor Name	Frequency Function
Cross Correlation	1
PHAT	$\frac{1}{\left\|P_{x_1x_2}\left(e^{j\omega T_s}\right)\right\|}$
Roth Impulse Response	$\frac{1}{P_{x_1x_1}\left(e^{j\omega T_s}\right)}$ or $\frac{1}{P_{x_2x_2}\left(e^{j\omega T_s}\right)}$
SCOT	$\frac{1}{\sqrt{P_{x_1x_1}\left(e^{j\omega T_s}\right)P_{x_2x_2}\left(e^{j\omega T_s}\right)}}$
Eckart	$\frac{P_{s_1s_1}\left(e^{j\omega T_s}\right)}{P_{n_1n_1}\left(e^{j\omega T_s}\right)P_{n_2n_2}\left(e^{j\omega T_s}\right)}$
Hannon-Thomson or ML	$\frac{\left\|\gamma_{x_1x_2}\left(e^{j\omega T_s}\right)\right\|^2}{\left\|P_{x_1x_2}\left(e^{j\omega T_s}\right)\right\|\left(1-\left\|\gamma_{x_1x_2}\left(e^{j\omega T_s}\right)\right\|^2\right)}$

where $\gamma_{x_1x_2}\left(e^{j\omega T_s}\right)$ is the normalised CPSD or **coherence function**

GCC Processors

The PHAT-GCC approach can be written as:

$$
\begin{aligned}
r_{x_i\,x_j}[\ell] &= \int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}} \Phi\left(e^{j\omega T_s}\right) P_{x_1x_2}\left(e^{j\omega T_s}\right) e^{j\ell\omega T}\, d\omega \\
&= \int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}} \frac{1}{\left|P_{x_1x_2}\left(e^{j\omega T_s}\right)\right|} \left|P_{x_1x_2}\left(e^{j\omega T_s}\right)\right| e^{j\angle P_{x_1x_2}\left(e^{j\omega T_s}\right)} e^{j\ell\omega T}\, d\omega \\
&= \int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}} e^{j\left(\ell\omega T + \angle P_{x_1x_2}\left(e^{j\omega T_s}\right)\right)}\, d\omega \\
&= \delta\left(\ell\, T_s + \angle P_{x_1x_2}\left(e^{j\omega T_s}\right)\right) \\
&= \delta\left(\ell\, T_s - T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right)\right)
\end{aligned}
$$

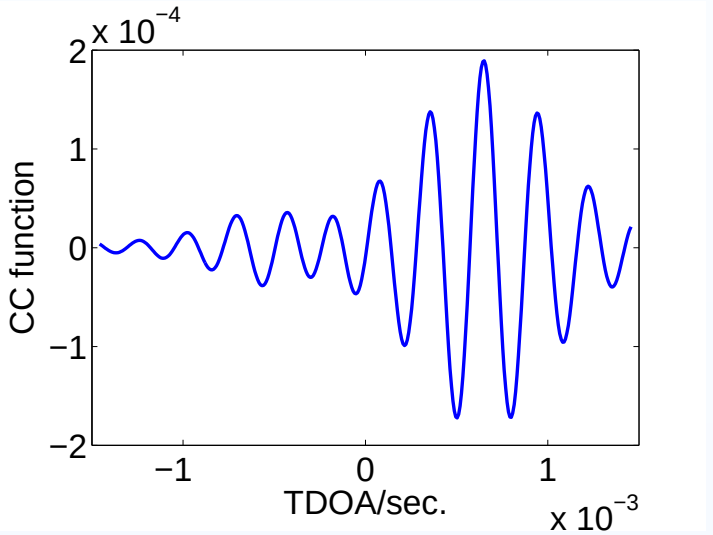
GCC Processors

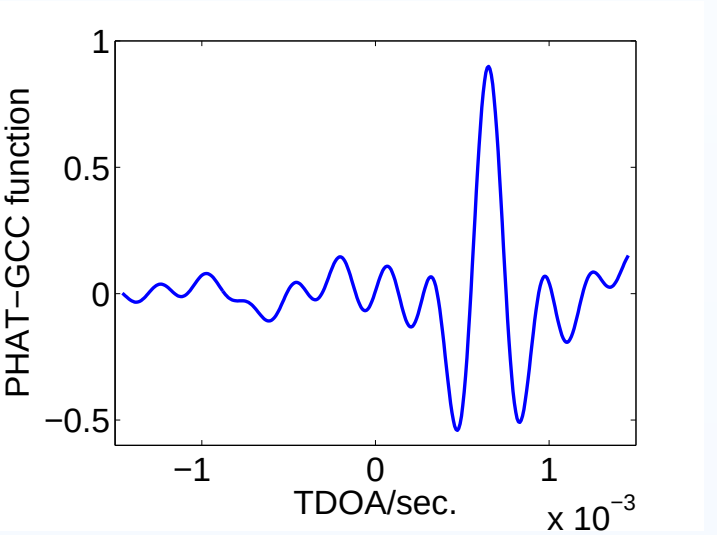
The PHAT-GCC approach can be written as:

$$
\begin{aligned}
r_{x_i\,x_j}[\ell] &= \int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}} \Phi\left(e^{j\omega T_s}\right) P_{x_1x_2}\left(e^{j\omega T_s}\right) e^{j\ell\omega T}\, d\omega \\
&= \int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}} \frac{1}{\left|P_{x_1x_2}\left(e^{j\omega T_s}\right)\right|} \left|P_{x_1x_2}\left(e^{j\omega T_s}\right)\right| e^{j\angle P_{x_1x_2}\left(e^{j\omega T_s}\right)} e^{j\ell\omega T}\, d\omega \\
&= \int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}} e^{j\left(\ell\omega T + \angle P_{x_1x_2}\left(e^{j\omega T_s}\right)\right)}\, d\omega \\
&= \delta\left(\ell\, T_s + \angle P_{x_1x_2}\left(e^{j\omega T_s}\right)\right) \\
&= \delta\left(\ell\, T_s - T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right)\right)
\end{aligned}
$$

- In the absence of reverberation, the GCC-PHAT algorithm gives an impulse at a lag given by the TDOA divided by the sampling period.

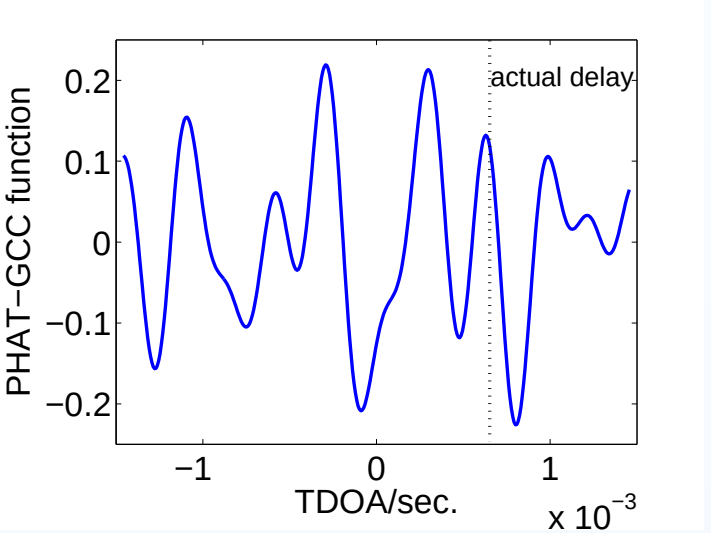
GCC Processors

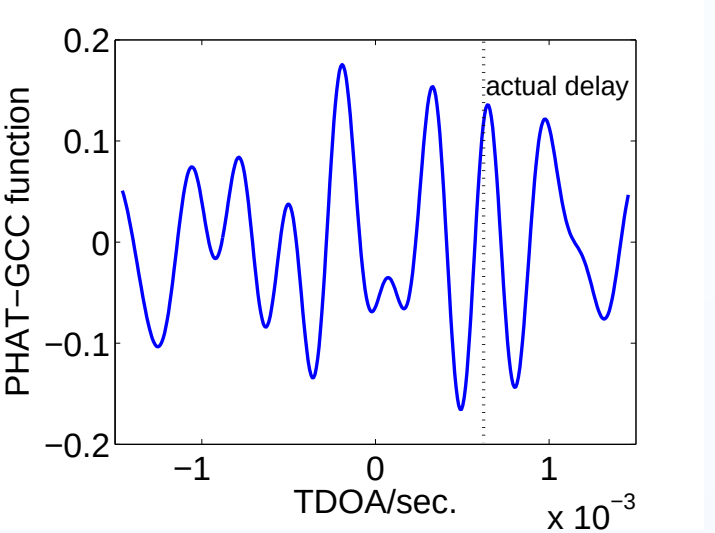




Normal cross-correlation and GCC-PHAT functions for a frame of speech.

GCC Processors





The effect of reverberation and noise on the GCC-PHAT can lead to poor TDOA estimates.

Adaptive Eigenvalue Decomposition

The AED algorithm actually amounts to a **blind channel identification** problem, which then seeks to identify the channel coefficients corresponding to the direct path elements.

Adaptive Eigenvalue Decomposition

The AED algorithm actually amounts to a **blind channel identification** problem, which then seeks to identify the channel coefficients corresponding to the direct path elements.

- Suppose that the acoustic impulse response (AIR) between source k and i is given by $h_{ik}[n]$ such that

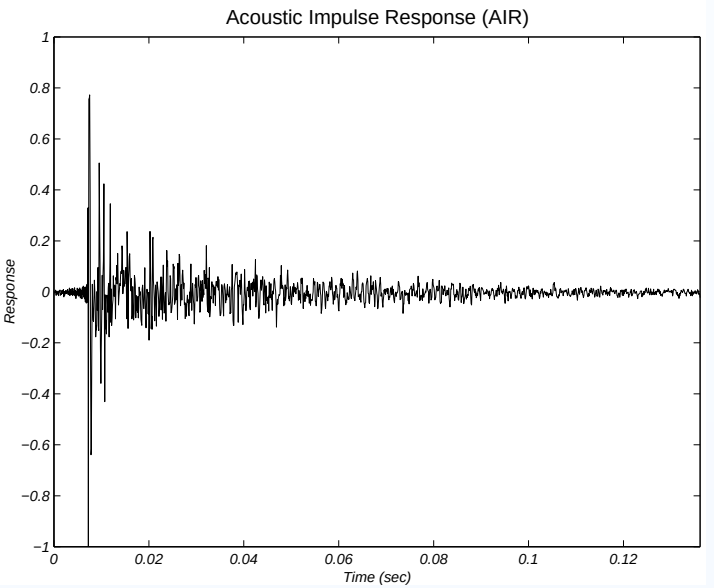
$$x_{ik}[n] = \sum_{m=-\infty}^{\infty} h_{ik}[n-m]\, s_k[m] + b_{ik}[n]$$

then the TDOA between microphones i and j is:

$$\tau_{ijk} = \left\{ \arg\max_{\ell} |h_{ik}[\ell]| \right\} - \left\{ \arg\max_{\ell} |h_{jk}[\ell]| \right\}$$

This assumes a minimum-phase system, but can easily be made robust to a non-minimum-phase system.

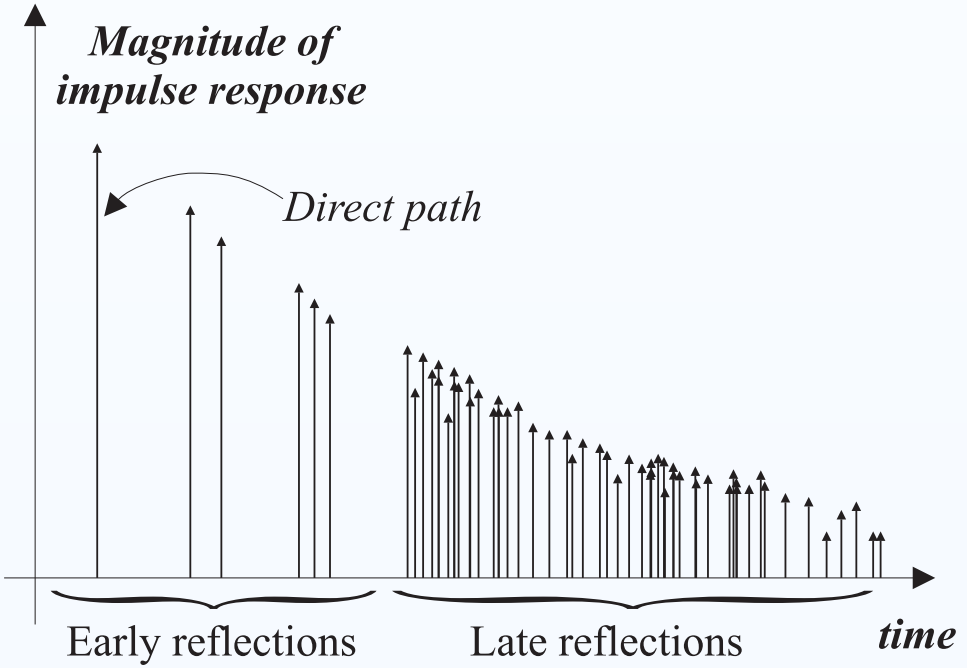
Adaptive Eigenvalue Decomposition



A typical room acoustic impulse response.

- Reverberation plays a major role in ASL and BSS.
- Consider reverberation as the sum total of all sound reflections arriving at a certain point in a room after room has been excited by impulse.

Adaptive Eigenvalue Decomposition

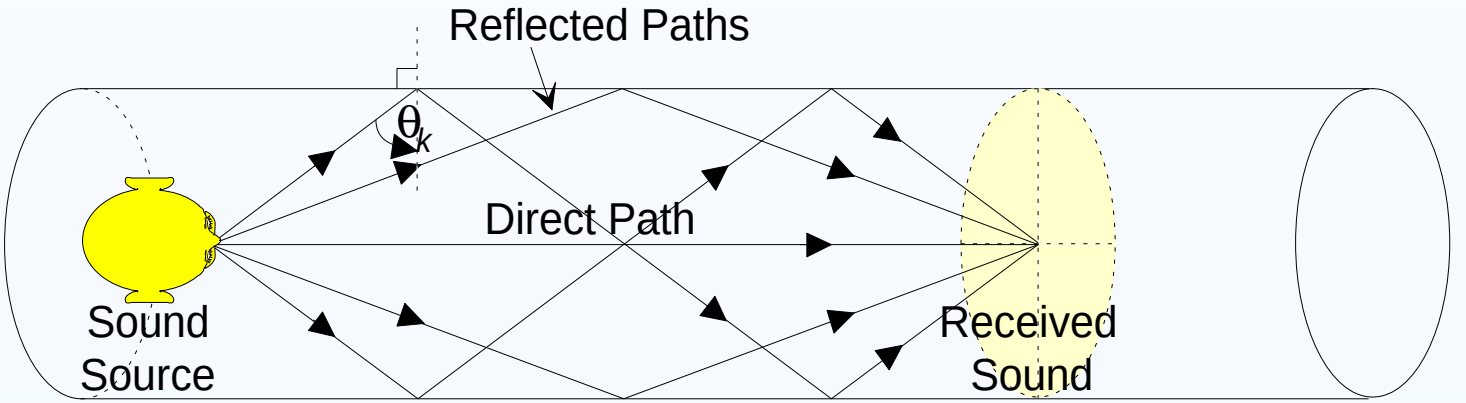


Early and late reflections in an AIR.

Trivia: Perceive early reflections to reinforce direct sound, and can help with speech intelligibility. It can be easier to hold a conversation in a closed room than outdoors

Adaptive Eigenvalue Decomposition

- Room transfer functions are often nonminimum-phase since there is more energy in the reverberant component of the RIR than in the component corresponding to direct path.



Demonstrating nonminimum-phase properties

- Therefore AED will need to consider multiple peaks in the estimated AIR.

Direct Localisation Methods

- Direct localisation methods have the advantage that the relationship between the measurement and the state is linear.
- However, extracting the position measurement requires a multi-dimensional search over the state space and is usually computationally expensive.

Steered Response Power Function

The SBF or SRP function is a measure of correlation across *all pairs* of microphone signals for a set of relative delays that arise from a hypothesised source location.

The frequency domain **delay-and-sum beamformer** steered to a spatial position $\hat{\mathbf{x}}_k$ such that $\hat{\tau}_{pk} = |\hat{\mathbf{x}} - \mathbf{m}_p|$:

$$S(\hat{\mathbf{x}}) = \int_{\Omega} \left| \sum_{p=1}^{N} W_p\left(e^{j\omega T_s}\right) X_p\left(e^{j\omega T_s}\right) e^{j\omega \hat{\tau}_{pk}} \right|^2 d\omega$$

Steered Response Power Function

The SBF or SRP function is a measure of correlation across *all pairs* of microphone signals for a set of relative delays that arise from a hypothesised source location.

The frequency domain **delay-and-sum beamformer** steered to a spatial position $\hat{\mathbf{x}}_k$ such that $\hat{\tau}_{pk} = |\hat{\mathbf{x}} - \mathbf{m}_p|$:

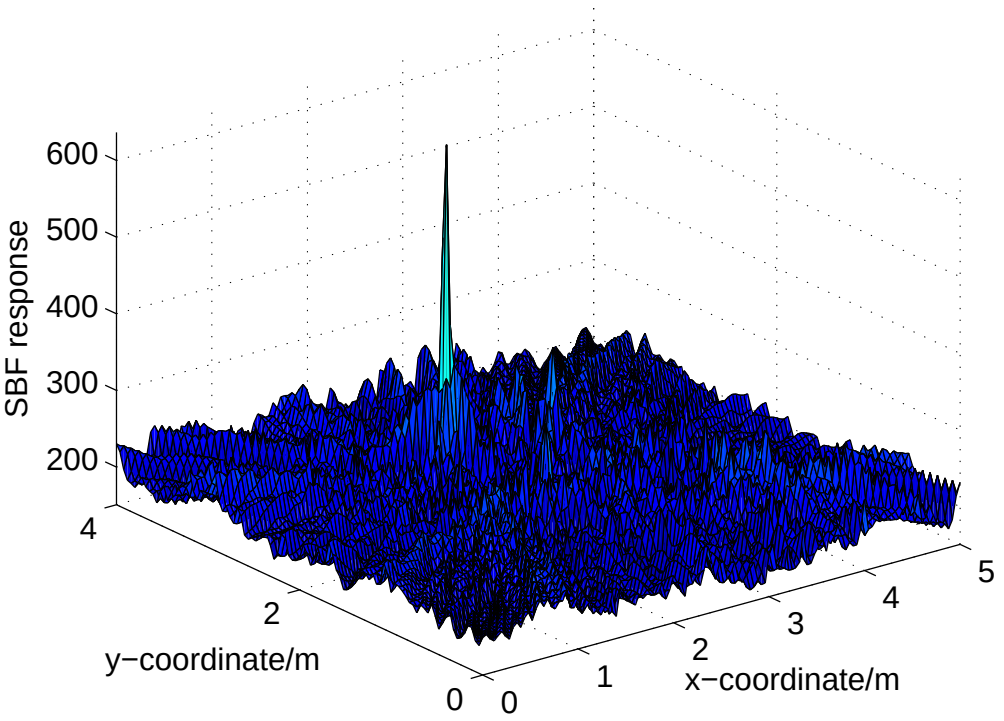
$$S(\hat{\mathbf{x}}) = \int_{\Omega} \left| \sum_{p=1}^{N} W_p\left(e^{j\omega T_s}\right) X_p\left(e^{j\omega T_s}\right) e^{j\omega \hat{\tau}_{pk}} \right|^2 d\omega$$

Taking expectations, $\Phi_{pq}\left(e^{j\omega T_s}\right) = W_p\left(e^{j\omega T_s}\right) W_q^*\left(e^{j\omega T_s}\right)$

$$\mathbb{E}\left[S(\hat{\mathbf{x}})\right] = \sum_{p=1}^{N}\sum_{q=1}^{N} \int_{\Omega} \Phi_{pq}\left(e^{j\omega T_s}\right) P_{x_p x_q}\left(e^{j\omega T_s}\right) e^{j\omega \hat{\tau}_{pqk}}\, d\omega$$

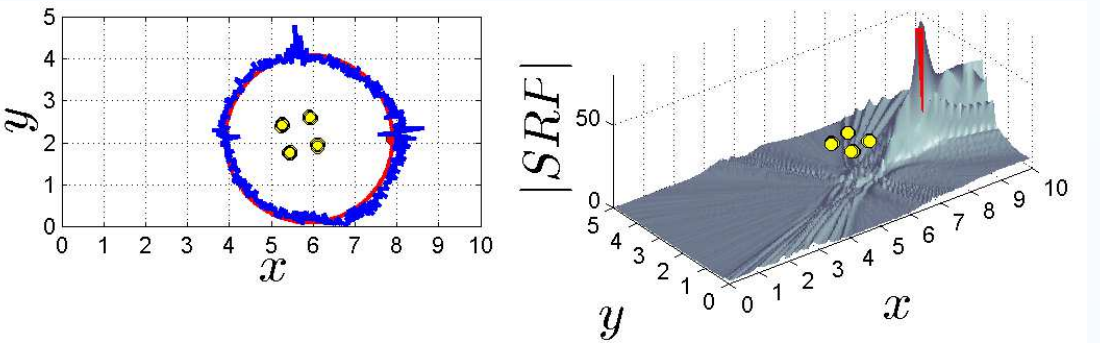
$$= \sum_{p=1}^{N}\sum_{q=1}^{N} r_{x_i x_j}\left[\hat{\tau}_{pqk}\right] \equiv \sum_{p=1}^{N}\sum_{q=1}^{N} r_{x_i x_j}\left[\frac{|\mathbf{x}_k - \mathbf{m}_i| - |\mathbf{x}_k - \mathbf{m}_j|}{c}\right]$$

Steered Response Power Function



SBF response from a frame of speech signal. The integration frequency range is 300 to 3500 Hz. The true source position is at $[2.0, 2.5]m$. The grid density is set to 40 mm.

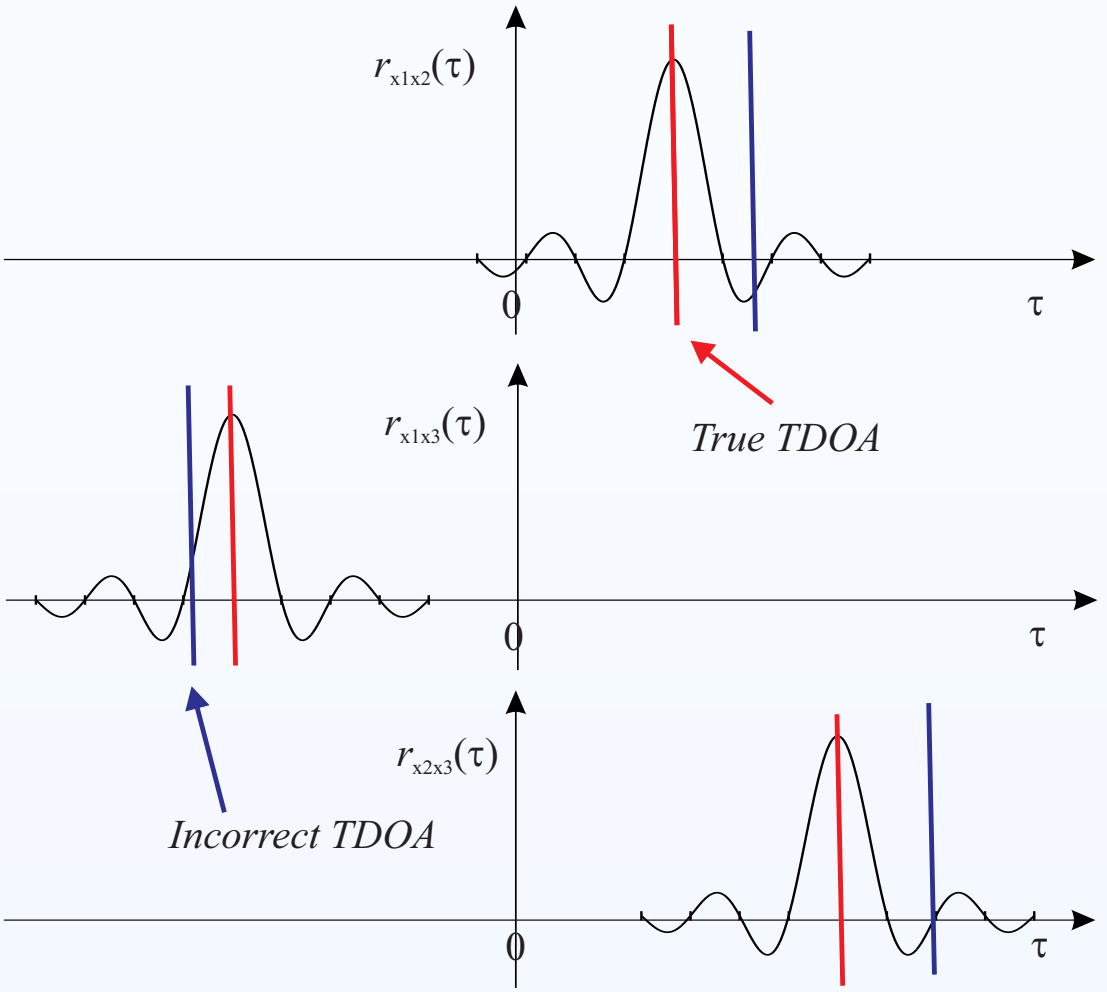
Steered Response Power Function



An example video showing the SBF changing as the source location moves.

- Show video!

Conceptual Intepretation



GCC-PHAT for different microphone pairs.

$$T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \hat{\mathbf{x}}_k\right) = \frac{\left|\hat{\mathbf{x}}_k - \mathbf{m}_i\right| - \left|\hat{\mathbf{x}}_k - \mathbf{m}_j\right|}{c}$$

DUET Algorithm

The degenerate unmixing estimation technique (DUET) algorithm is an approach to BSS that ties in neatly to ASL. Under certain assumptions and circumstances, it is possible to separate more than two sources using only two microphones.

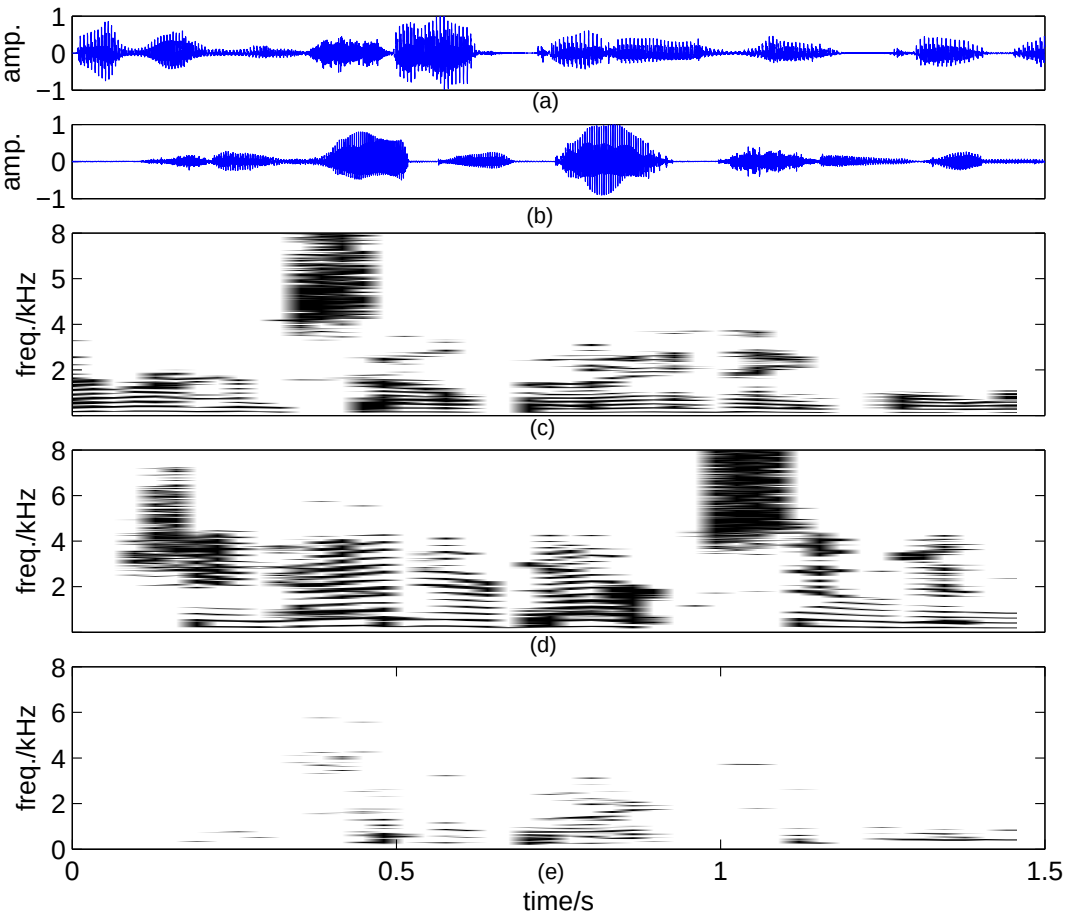
DUET Algorithm

The DUET algorithm is an approach to BSS that ties in neatly to ASL. Under certain assumptions and circumstances, it is possible to separate more than two sources using only two microphones.

- DUET is based on the assumption that for a set of signals $x_k[t]$, their time-frequency representations (TFRs) are predominately non-overlapping. This condition is referred to as W-disjoint orthogonality (WDO):

$$S_p(\omega, t)\, S_q(\omega, t) = 0 \,\forall p \neq q, \forall t, \omega$$

DUET Algorithm



W-disjoint orthogonality of two speech signals. Original speech signal (a) $s_1[t]$ and (b) $s_2[t]$; corresponding STFTs (c) $|S_1(\omega, t)|$ and (d) $|S_2(\omega, t)|$; (e) product $|S_1(\omega, t)\, S_2(\omega, t)|$.

DUET Algorithm

Consider taking a particular time-frequency (TF)-bin, $(\omega,\, t)$, where source p is known to be active. The two received signals in *that TF-bin* can be written as:

$$X_{ip}\left(\omega,\, t\right) = \alpha_{ip}\, e^{-j\omega\,\tau_{ip}}\, S_p\left(\omega,\, t\right) + B_i\left(\omega,\, t\right)$$

$$X_{jp}\left(\omega,\, t\right) = \alpha_{jp}\, e^{-j\omega\,\tau_{jp}}\, S_p\left(\omega,\, t\right) + B_j\left(\omega,\, t\right)$$

DUET Algorithm

Consider taking a particular TF-bin, $(\omega,\, t)$, where source p is known to be active. The two received signals in *that TF-bin* can be written as:

$$X_{ip}\left(\omega,\, t\right) = \alpha_{ip}\, e^{-j\omega\,\tau_{ip}}\, S_p\left(\omega,\, t\right) + B_i\left(\omega,\, t\right)$$

$$X_{jp}\left(\omega,\, t\right) = \alpha_{jp}\, e^{-j\omega\,\tau_{jp}}\, S_p\left(\omega,\, t\right) + B_j\left(\omega,\, t\right)$$

Taking the ratio and ignoring the noise terms gives:

$$H_{ikp}\left(\omega,\, t\right) \triangleq \frac{X_{ip}\left(\omega,\, t\right)}{X_{jp}\left(\omega,\, t\right)} = \frac{\alpha_{ip}}{\alpha_{jp}}\, e^{-j\omega\tau_{ijp}}$$

DUET Algorithm

Consider taking a particular TF-bin, $(\omega,\, t)$, where source p is known to be active. The two received signals in *that TF-bin* can be written as:

$$X_{ip}\left(\omega,\, t\right) = \alpha_{ip}\, e^{-j\omega\,\tau_{ip}}\, S_p\left(\omega,\, t\right) + B_i\left(\omega,\, t\right)$$

$$X_{jp}\left(\omega,\, t\right) = \alpha_{jp}\, e^{-j\omega\,\tau_{jp}}\, S_p\left(\omega,\, t\right) + B_j\left(\omega,\, t\right)$$

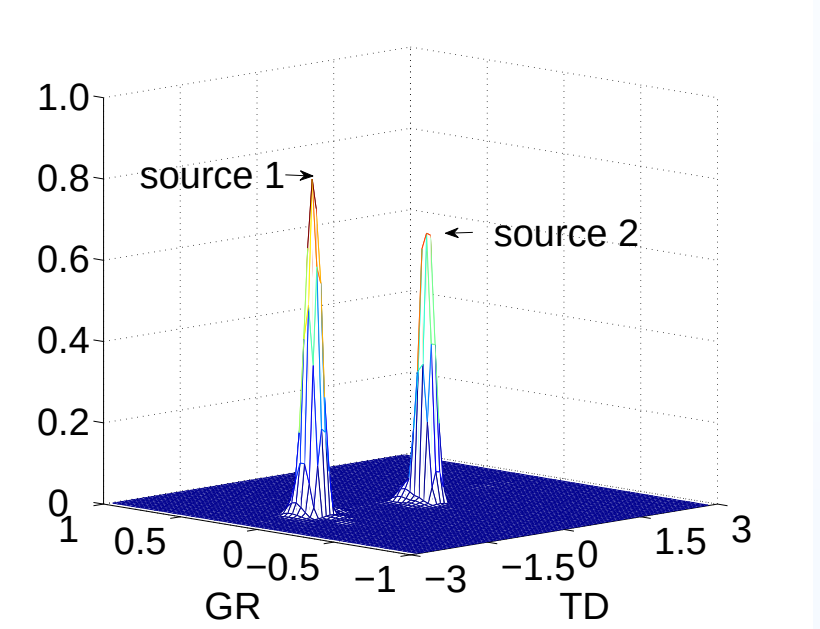
Taking the ratio and ignoring the noise terms gives:

$$H_{ikp}\left(\omega,\, t\right) \triangleq \frac{X_{ip}\left(\omega,\, t\right)}{X_{jp}\left(\omega,\, t\right)} = \frac{\alpha_{ip}}{\alpha_{jp}}\, e^{-j\omega\tau_{ijp}}$$

Hence,

$$\tau_{ijp} = -\frac{1}{\omega}\arg H_{ikp}\left(\omega,\, t\right), \qquad \text{and} \qquad \frac{\alpha_{ip}}{\alpha_{jp}} = \left|H_{ikp}\left(\omega,\, t\right)\right|$$

DUET Algorithm



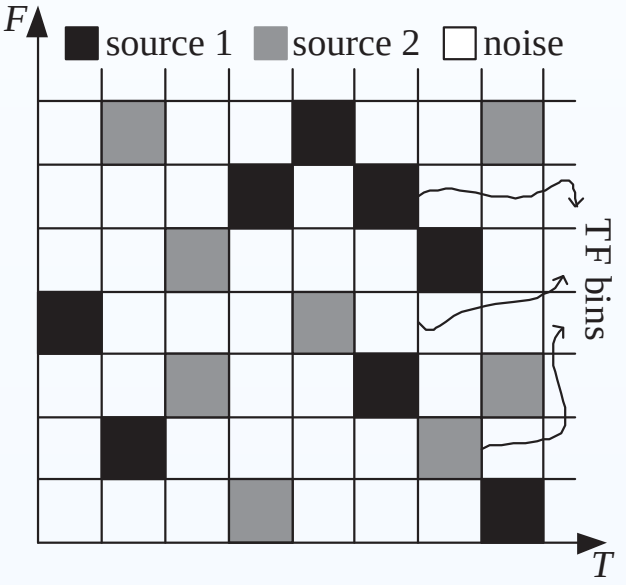


Illustration of the underlying idea in DUET.

DUET Algorithm

This leads to the essentials of the DUET method which are:

1. Construct the TF representation of both mixtures.

2. Take the ratio of the two mixtures and extract local mixing parameter estimates.

DUET Algorithm

This leads to the essentials of the DUET method which are:

1. Construct the TF representation of both mixtures.

2. Take the ratio of the two mixtures and extract local mixing parameter estimates.

3. Combine the set of local mixing parameter estimates into N pairings corresponding to the true mixing parameter pairings.

4. Generate one binary mask for each determined mixing parameter pair corresponding to the TF-bins which yield that particular mixing parameter pair.

DUET Algorithm

This leads to the essentials of the DUET method which are:

1. Construct the TF representation of both mixtures.

2. Take the ratio of the two mixtures and extract local mixing parameter estimates.

3. Combine the set of local mixing parameter estimates into N pairings corresponding to the true mixing parameter pairings.

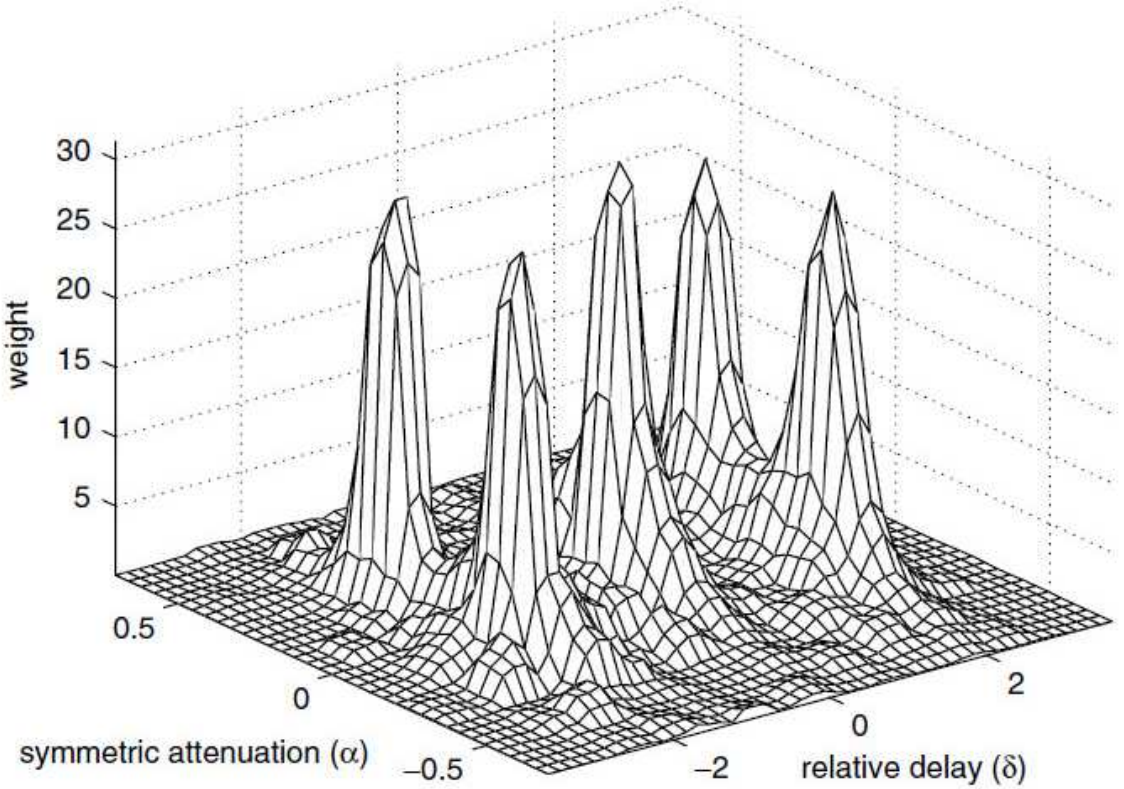
4. Generate one binary mask for each determined mixing parameter pair corresponding to the TF-bins which yield that particular mixing parameter pair.

5. Demix the sources by multiplying each mask with one of the mixtures.

6. Return each demixed TFR to the time domain.

DUET Algorithm

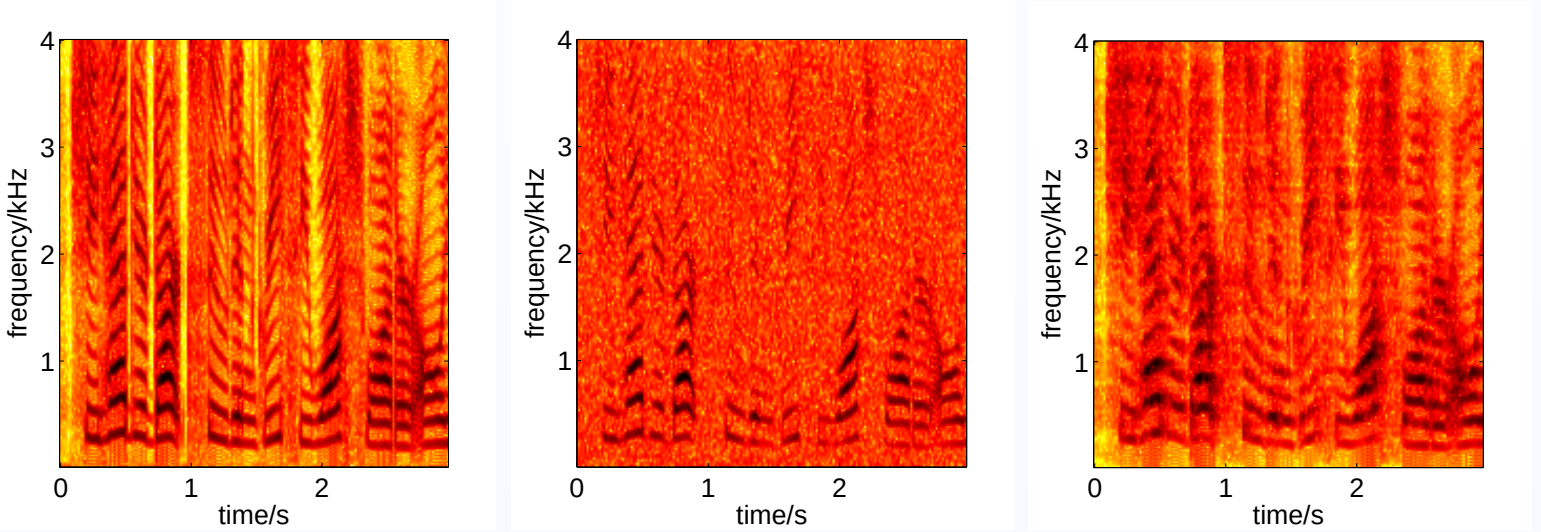
This leads to the essentials of the DUET method which are:



DUET for multiple sources.

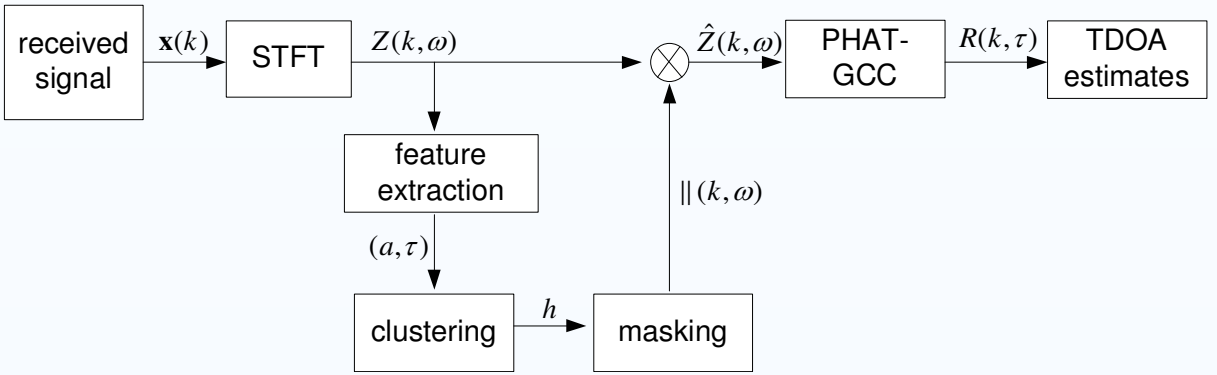
DUET Algorithm

Effect of Reverberation and Noise

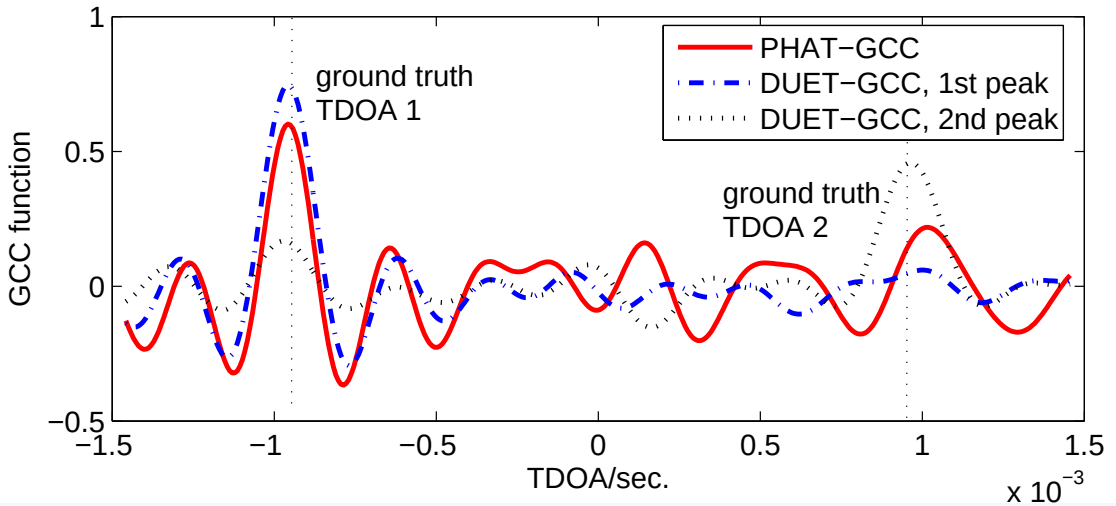


The TFR is very clear in the anechoic environment but smeared around by the reverberation and noise.

Estimating multiple targets



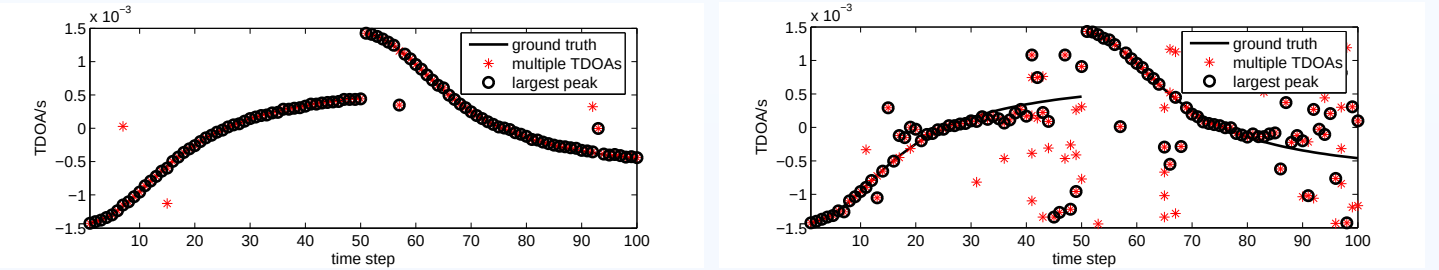
Flow diagram of the DUET-GCC approach. Basically, the speech mixtures are separated by using the DUET in the TF domain, and the PHAT-GCC is then employed for the spectrogram of each source to estimate the TDOAs.



GCC function from DUET approach and traditional PHAT weighting. Two sources are located at $(1.4, 1.2)$m and

Further Topics

- Reduction in complexity of calculating SRP. This includes stochastic region contraction (SRC) and hierarchical searches.
- Multiple-target tracking (see Daniel Clark's Notes)
- Simultaneous (self-)localisation and tracking; estimating sensor and target positions from a moving source.



Acoustic source tracking and localisation.

Further Topics

- Joint ASL and BSS.
- Explicit signal and channel modelling! (None of the material so forth cares whether the signal is speech or music!)
- Application areas such as gunshot localisation; other sensor modalities; diarisation.