Probability and Random Variables; and Classical Estimation Theory
UDRC Summer School, 20th July 2015

Course lecture notes
(Full Version)

Dr James R. Hopgood

Room 2.05
Alexander Graham Bell Building
The King's Buildings
Mayfield Road
Edinburgh
EH9 3JL
Scotland, UK
James.Hopgood@ed.ac.uk
Telephone: +44 (0)131 650 5571
Fax: +44 (0)131 650 6554.

Major revision, Wednesday 16th July, 2014.
Last printed revision with minor corrections, 16 July, 2015.

Typeset by the author with the LaTeX 2_ε Documentation System, with $\mathcal{AMS}$-LaTeX Extensions, in 12/18 pt Times and Euler fonts.

INSTITUTE FOR DIGITAL COMMUNICATIONS,
School of Engineering,
College of Science and Engineering,
Kings's Buildings,
Edinburgh, EH9 3JL. U.K.

Copyright Statement

This document does not contain copyright material.

The author of this document

1. holds the copyright for all lecture and course materials in this module;
2. holds the copyright for students notes, summaries, or recordings that substantially reflect the lecture content or materials;
3. makes these materials available only for personal use by students studying this module;
4. reserves the right that no part of the notes, tutorials, solutions, or other course materials may be distributed or reproduced for commercial purposes without express written consent from the author; this does not prevent students from sharing notes on an individual basis for personal use.

These lecture notes consist of entirely original work, where all material has been written and typeset by the author. No figures or substantial pieces of text has been reproduced verbatim from other texts.

However, there is some material that has been based on work in a number of previous textbooks, and therefore some sections and paragraphs have strong similarities in structure and wording. These texts have been referenced and include, amongst a number of others, in order of contributions:

- Manolakis D. G., V. K. Ingle, and S. M. Kogon, *Statistical and Adaptive Signal Processing*: *Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing*, McGraw Hill, Inc., 2000.

  IDENTIFIERS – *Paperback*, ISBN10: 0070400512, ISBN13: 9780070400511

- Therrien C. W., *Discrete Random Signals and Statistical Signal Processing*, Prentice-Hall, Inc., 1992.

  IDENTIFIERS – *Paperback*, ISBN10: 0130225452, ISBN13: 9780130225450
  Hardback, ISBN10: 0138521123, ISBN13: 9780138521127

- Kay S. M., *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Inc., 1993.

  IDENTIFIERS – *Hardback*, ISBN10: 0133457117, ISBN13: 9780133457117
  Paperback, ISBN10: 0130422681, ISBN13: 9780130422682

- Papoulis A. and S. Pillai, *Probability, Random Variables, and Stochastic Processes*, Fourth edition, McGraw Hill, Inc., 2002.

  IDENTIFIERS – *Paperback*, ISBN10: 0071226613, ISBN13: 9780071226615
  Hardback, ISBN10: 0072817259, ISBN13: 9780072817256

- Proakis J. G. and D. G. Manolakis, *Digital Signal Processing*: *Principles, Algorithms, and Applications*, Pearson New International Edition, Fourth edition, Pearson Education, 2013.

  IDENTIFIERS – *Paperback*, ISBN10: 1292025735, ISBN13: 9781292025735

- Mulgew B., P. M. Grant, and J. S. Thompson, *Digital Signal Processing*: *Concepts and Applications*, Palgrave, Macmillan, 2003.

  IDENTIFIERS – *Paperback*, ISBN10: 0333963563, ISBN13: 9780333963562

  See `http://www.see.ed.ac.uk/~{}pmg/SIGPRO`

- Therrien C. W. and M. Tummala, *Probability and Random Processes for Electrical and Computer Engineers*, Second edition, CRC Press, 2011.

  IDENTIFIERS – *Hardback*, ISBN10: 1439826986, ISBN13: 978-1439826980

- Press W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Receipes in C*: *The Art of Scientific Computing*, Second edition, Cambridge University Press, 1992.

  IDENTIFIERS – *Paperback*, ISBN10: 0521437202, ISBN13: 9780521437202
  Hardback, ISBN10: 0521431085, ISBN13: 9780521431088

The material in [Kay:1993] is mainly covered in Handout 5; material in [Therrien:1992] and [Papoulis:1991] is covered throughout the course. The following labelling convention is used for numbering equations that are taken from the various recommended texts. Equations labelled as:

M:v.w.xyz are similar to those in [Manolakis:2001] with the corresponding label;

T:w.xyz are similar to those in [Therrien:1992] with the corresponding label;

K:w.xyz are similar to those in [Kay:1993] with the corresponding label;

P:v.w.xyz are used in chapters referring to basic digital signal processing (DSP), and are references made to [Proakis:1996].

Contents

Acronyms

2-D	two-dimensional
3-D	three-dimensional
AIC	Akaike’s information criterion
AWGN	additive white Gaussian noise
BFGS	Broyden-Fletcher-Goldfarb-Shannon
BIC	B-Information criterion
BSS	blind source separation
CAT	Parzen’s criterion autoregressive transfer function
CCTV	closed-circuit television
CD	compact disc
CLT	central limit theorem
CRLB	Cramér-Rao lower-bound
DAT	digital audio tape
DC	“direct current”
DNA	deoxyribonucleic acid
DSP	digital signal processing
DVD	digital versitile disc
DVD-A	digital versitile disc-audio
EEG	electroencephalogram
FPE	final prediction error
FS	Fourier series
FT	Fourier transform

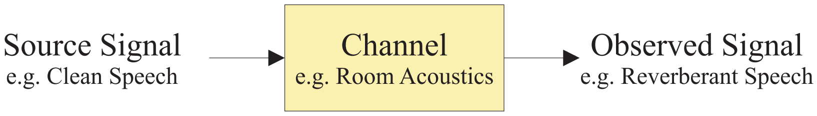
ICA	independent component analysis
LHS	left hand side
LITP	linear in the parameters
LS	least-squares
LSE	least-squares estimate
LTI	linear time-invariant
MAP	maximum *a posteriori*
MDL	minimum description length
ML	maximum-likelihood
MLE	maximum-likelihood estimate
MMAP	maximum marginal *a posteriori*
MMSE	minimum mean-square error
MRI	magnetic resonance imaging
MS	mean-square
MSE	mean-squared error
MVU	minimum variance unbiased
MVUE	minimum variance unbiased estimator
NMRI	nuclear magnetic resonance imaging
RHS	right hand side
SACD	super-audio CD
SSP	statistical signal processing
WGN	white Gaussian noise
cdf	cumulative distribution function
iff	if, and only if,
i. i. d.	independent and identically distributed
i. t. o.	in terms of
pdf	probability density function
pmf	probability mass function
RV	random variable
w. r. t.	with respect to

Acronyms

PET	Probability, Random Variables, and Estimation Theory
SSP	Statistical Signal Processing

1

Module Overview, Aims and Objectives

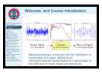


> Everything that needs to be said has already been said. But since no one was listening, everything must be said again.
>
> André Gide

> If you can't explain it simply, you don't understand it well enough.
>
> Albert Einsten

This handout also provides an introduction to signals and systems, and an overview of statistical signal processing applications.

1.1 Obtaining the Latest Version of these Handouts



New slide

- This research tutorial is intended to cover a wide range of aspects which cover the fundamentals of statistical signal processing. It is written at a level which assumes knowledge of undergraduate mathematics and signal processing nomenclature, but otherwise should be accessible to most technical graduates.

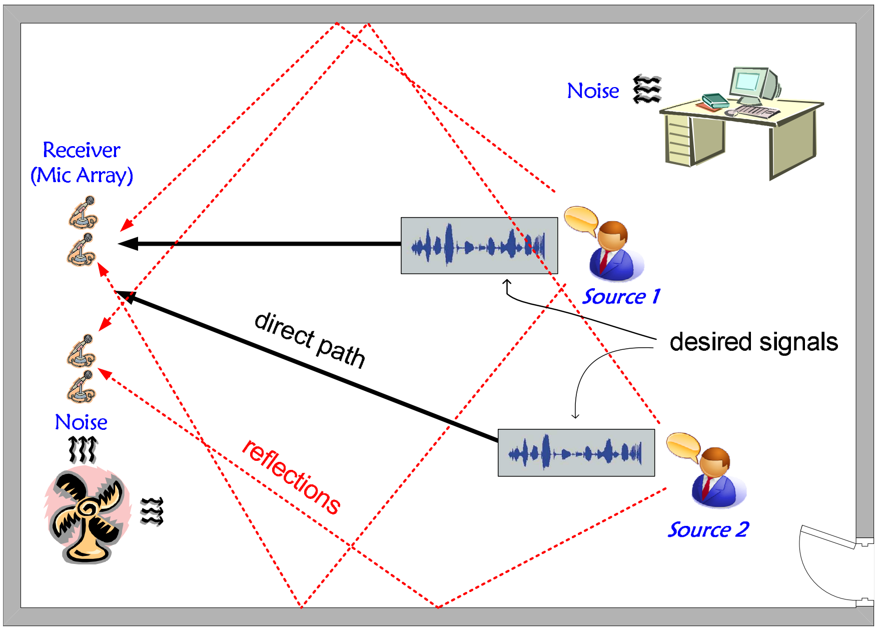


Figure 1.1: Source localisation and BSS. An example of topics using statistical signal processing.

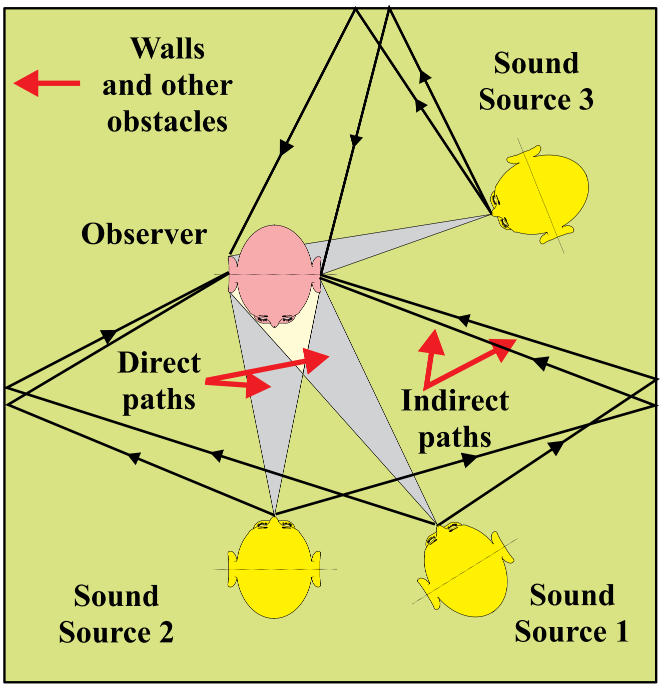


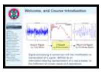
Figure 1.2: Humans turn their head in the direction of interest in order to reduce inteference from other directions; *joint detection, localisation, and enhancement*. An application of probability and estimation theory, and statistical signal processing.

KEYPOINT! (Latest Slides). Please note the following:

- 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. In particular, there are likely to be a few typos in the document, so if there is something that isn't clear, please feel free to email me so I can correct it (or make it clearer).
- The latest version of this document can be found online and downloaded at:

  `http://www.mod-udrc.org/events/2015-summer-school`
- Extended thanks are given to the many MSc students over the past 11 years who have helped proof-read and improve these documents.

1.2 Module Abstract



New slide

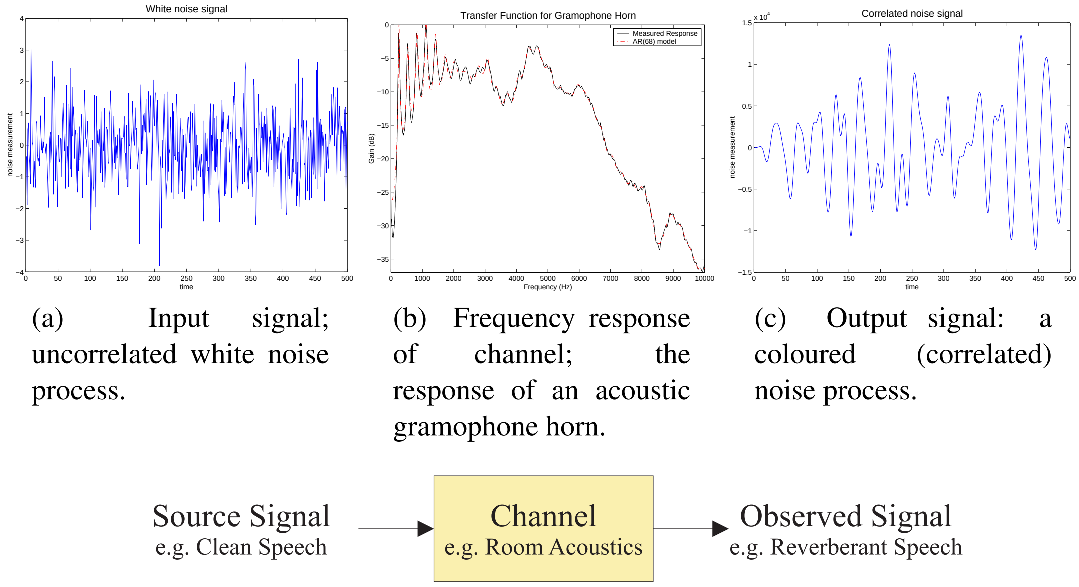
The notion of **random** or **stochastic** quantities is an extremely powerful concept that can be constructively used to model observations that result from real-world processes. These quantities could be scalar measurements, such as an instantaneous measurement of distance, or they could be vector-measurements such as a coordinate. They could be random signals either in one-dimension, or in higher-dimensions, such as images. Stochastic quantities such as random signals, by their very nature, are described using the mathematics of probability and statistics. By making assumptions such as the availability of an infinite number of observations or data samples, time-invariant statistics, and known signal or observation models, it is possible to estimate the properties of these random quantities or signals and, consequently, use them in *signal processing* algorithms.

In practice, of course, these statistical properties must be estimated from finite-length data signals observed in noise. In order to understand both the concept of stochastic processes and the inherent **uncertainty** of signal estimates from finite-length sequences, it is first necessary to understand the fundamentals of **probability**, **random variables**, and **estimation theory**.

This topic is covered in two related lecture modules:

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

together introduce the subject of statistical signal modelling and estimation. In particular, the module *Statistical Signal Processing* investigates which statistical properties are relevant for dealing with signal processing problems, how these properties can be estimated from real-world signals, and how they can be used in signal processing algorithms to achieve a particular goal.



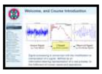
(a) Input signal; uncorrelated white noise process.

(b) Frequency response of channel; the response of an acoustic gramophone horn.

(c) Output signal: a coloured (correlated) noise process.

(d) Block diagram of system representing convolution.

Figure 1.3: Solutions to the so-called *blind deconvolution problem* require statistical signal processing methods.

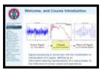


New slide

1.3 Introduction and Overview

> 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.

Gone is the era where information in the form of electrical signals are processed through analogue devices. For the foreseeable future, processing of digital, sampled, or discrete-time signals is the definitive approach to analysing data and extracting information. In this course, it is assumed that the reader already has a grounding in digital signal processing (DSP), and this module will take you to the next level; a tour of the exciting, fascinating, and active research area of *statistical signal processing (SSP)*.



New slide

1.3.1 Description and Learning Outcomes

Module Aims The aims of the two modules *Probability, Random Variables, and Estimation Theory (PET)*, and *statistical signal processing (SSP)*, are similar to those of the text book [Manolakis:2000, page xvii]. The principle aim of the modules are:

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

Pre-requisites It is strongly recommended that the student has previously attended an undergraduate level course in either signals and systems, digital signal processing, automatic control, or an equivalent course.

Section 1.3.2 provides further details regarding the material a student should have previously covered.

Short Description The **Probability, Random Variables, and Estimation Theory** module introduces the fundamental statistical tools that are required to analyse and describe advanced signal processing algorithms. It provides a unified mathematical framework which is the basis for describing random events and signals, and how to describe key characteristics of random processes.

The module covers probability theory, considers the notion of random variables and vectors, how they can be manipulated, and provides an introduction to estimation theory. It is demonstrated that many estimation problems, and therefore signal processing problems, can be reduced to an exercise in either *optimisation* or *integration*. While these problems can be solved using deterministic numerical methods, the module introduces **Monte Carlo** techniques which are the basis of powerfull stochastic optimisation and integration algorithms. These methods rely on being able to sample numbers, or variates, from arbitrary distributions. This module will therefore discuss the various techniques which are necessary to understand these methods and, if time permits, techniques for random number generation are considered.

The **Statistical Signal Processing** module then consider representing real-world signals by stochastic or random processes. The tools for analysing these random signals are developed in the **Probability, Random Variables, and Estimation Theory** module, and this module extends them to deal with time series. The notion of statistical quantities such as autocorrelation and auto-covariance are extended from random vectors to random processes, and a frequency-domain analysis framework is developed. This module also investigates the affect of systems and transformations on time-series, and how they can be used to help design powerful signal processing algorithms to achieve a particular task.

The module introduces the notion of representing signals using parametric models; it extends the broad topic of statistical estimation theory covered in the **Probability, Random Variables, and Estimation Theory** module for determining optimal model parameters. In particular, the **Bayesian paradigm** for statistical parameter estimation is introduced. Emphasis is placed on relating these concepts to state-of-the-art applications and signals.

Keywords Probability, scalar and multiple random variables, stochastic processes, power spectral densities, linear systems theory, linear signal models, estimation theory, and Monte Carlo methods.

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

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**.

Intended Learning Outcomes At the end of the **Probability, Random Variables, and Estimation Theory** module, a student should be able to:

1. define, understand and manipulate scalar and multiple random variables, using the theory of probability; this should include the tools of probability transformations and characteristic functions;
2. explain the notion of characterising random variables and random vectors using moments, and be able to manipulate them; understand the relationship between random variables within a random vector;
3. understand the central limit theorem (CLT) and explain its use in estimation theory and the sum of random variables;
4. understand the principles of estimation theory; understand and be apply to apply estimation techniques such as maximum-likelihood, least squares, and Bayesian estimation;
5. be able to characterise the uncertainty in an estimator, as well as characterise the performance of an estimator (bias, variance, and so forth); understand the Cramér-Rao lower-bound (CRLB) and minimum variance unbiased estimator (MVUE) estimators.
6. if time permits, explain and apply methods for generating random numbers, or random variates, from an arbitrary distribution, using methods such as accept-reject and Gibbs sampling; understand the notion of stochastic numerical methods for solving *integration* and *optimisation* problems.

At the end of the **Statistical Signal Processing** module, a student should be able to:

1. explain, describe, and understand the notion of a random process and statistical time series;
2. characterise random processes in terms of its statistical properties, including the notion of stationarity and ergodicity;
3. define, describe, and understand the notion of the power spectral density of stationary random processes; analyse and manipulate power spectral densities;

4. analyse in both time and frequency the affect of transformations and linear systems on random processes, both in terms of the density functions, and statistical moments;
5. explain the notion of parametric signal models, and describe common regression-based signal models in terms of its statistical characteristics, and in terms of its affect on random signals;
6. apply least squares, maximum-likelihood, and Bayesian estimators to model based signal processing problems;

1.3.2 Prerequisites

The mathematical treatment throughout this module is kept at a level that is within the grasp of final-year undergraduate and graduate students, with a background in **digital signal processing (DSP)**, **linear system and control** theory, basic **probability theory**, **calculus**, **linear algebra**, and a competence in Engineering mathematics.

In summary, it is assumed that the reader has knowledge of:

1. Engineering mathematics, including linear algebra, manipulation of vectors and matrices, complex numbers, linear transforms including Fourier series and Fourier transforms, z-transforms, and Laplace transforms;

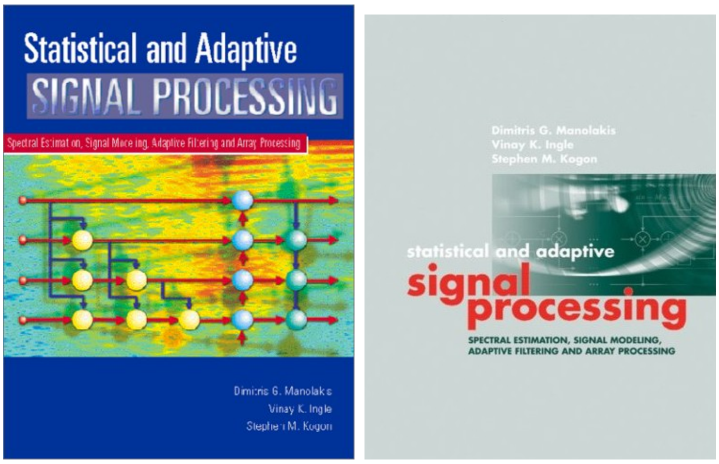
2. basic probability and statistics, albeit with a solid understanding;

3. differential and integral calculus, including differentiating products and quotients, functions of functions, integration by parts, integration by substitution;

4. basic digital signal processing (DSP), including:

   - the notions of deterministic continuous-time signals, discrete-time signals and digital (quantised) signals;
   - filtering and inverse filtering of signals; convolution;
   - the response of linear systems to harmonic inputs; analysing the time and frequency domain properties of signals and systems;
   - sampling of continuous time processes, Nyquist's sampling theorem and signal reconstruction;
   - and analysing discrete-time signals and systems.

Note that while the reader should have been exposed to the idea of a **random variable**, it is **not** assumed that the reader has been introduced to *random signals* in any form. A list of recommended texts covering these prerequisites is given in Section 1.3.3.



(a) Cover of *paperback* version.

(b) Cover of *hardback* version.

Figure 1.4: The main **course text** for this module: [Manolakis:2000].

1.3.3 Recommended Texts for Module Content

The **recommended text** for this module is cited throughout this document as [Manolakis:2000]. The full reference is:

> Manolakis D. G., V. K. Ingle, and S. M. Kogon, *Statistical and Adaptive Signal Processing*: *Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing*, McGraw Hill, Inc., 2000.

IDENTIFIERS – *Paperback*, ISBN10: 0070400512, ISBN13: 9780070400511

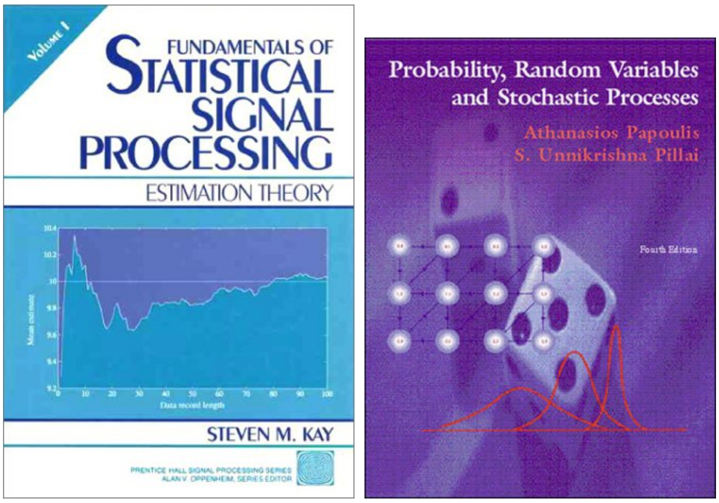
It is recommended that, if you wish to purchase a hard-copy of this book, you try and find this paperback version; it should be possible to order a copy relatively cheaply through the US version of Amazon (check shipping costs). However, please note that this book is now available, at great expense, in hard-back from an alternative publisher. The full reference is:

> Manolakis D. G., V. K. Ingle, and S. M. Kogon, *Statistical and Adaptive Signal Processing*: *Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing*, Artech House, 2005.

IDENTIFIERS – *Hardback*, ISBN10: 1580536107, ISBN13: 9781580536103

Images of the book covers are shown in Figure 1.4. For further reading, or an alternative perspective on the subject matter, other recommended text books for this module include:

1. Therrien C. W., *Discrete Random Signals and Statistical Signal Processing*, Prentice-Hall, Inc., 1992.


Volume I
FUNDAMENTALS OF STATISTICAL SIGNAL PROCESSING
ESTIMATION THEORY
STEVEN M. KAY
PRENTICE HALL SIGNAL PROCESSING SERIES
ALAN V. OPPENHEIM, SERIES EDITOR
Probability, Random Variables and Stochastic Processes
Athanasios Papoulis
S. Unnikrishna Pillai
Fourth Edition

(a) **Recommended text**: [Kay:1993].

(b) **Recommended text**: [Papoulis:1991].

Figure 1.5: Additional recommended texts for the course.

IDENTIFIERS – *Paperback*, ISBN10: 0130225452, ISBN13: 9780130225450
Hardback, ISBN10: 0138521123, ISBN13: 9780138521127

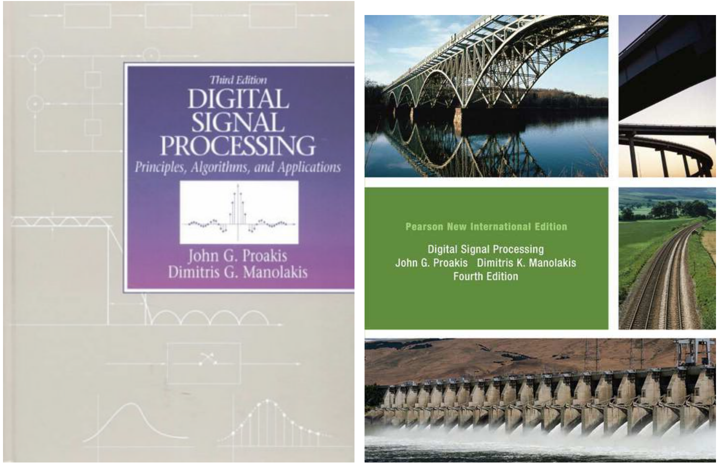
2. Kay S. M., *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Inc., 1993.

IDENTIFIERS – *Hardback*, ISBN10: 0133457117, ISBN13: 9780133457117
Paperback, ISBN10: 0130422681, ISBN13: 9780130422682

3. Papoulis A. and S. Pillai, *Probability, Random Variables, and Stochastic Processes*, Fourth edition, McGraw Hill, Inc., 2002.

IDENTIFIERS – *Paperback*, ISBN10: 0071226613, ISBN13: 9780071226615
Hardback, ISBN10: 0072817259, ISBN13: 9780072817256

These are referenced throughout as [Therrien:1992], [Kay:1993], and [Papoulis:1991], respectively. Images of the book covers are shown in Figure 1.5. The material in [Kay:1993] is mainly covered in Handout 5 on Estimation Theory of the PET module. The material in [Therrien:1992] and [Papoulis:1991] is covered throughout the course, with the former primarily in the Statistical Signal Processing (SSP) module.


Third Edition
DIGITAL SIGNAL PROCESSING
Principles, Algorithms, and Applications
John G. Proakis
Dimitris G. Manolakis
Pearson New International Edition
Digital Signal Processing
John G. Proakis Dimitris K. Manolakis
Fourth Edition

(a) Third Edition cover. (b) Fourth Edition cover.

Figure 1.6: **Course text**: further reading for digital signal processing and mathematics, [Proakis:1996].

KEYPOINT! **(Proposed Recommended Text Book for Future Years).** Finally, Therrien has also published a recent book which covers much of this course extremely well, and therefore comes thoroughly recommended. It has a number of excellent examples, and covers the material in good detail.

> Therrien C. W. and M. Tummala, *Probability and Random Processes for Electrical and Computer Engineers*, Second edition, CRC Press, 2011.

IDENTIFIERS – *Hardback*, ISBN10: 1439826986, ISBN13: 978-1439826980

1.3.4 Recommended Texts: Prerequisite Material

As mentioned in Section 1.3.2 above, regarding the prerequisites, it is assumed that the reader has a basic knowledge of digital signal processing. If not, or if the reader wishes to revise the topic, the following book which is *highly* recommended:

> Proakis J. G. and D. G. Manolakis, *Digital Signal Processing*: *Principles, Algorithms, and Applications*, Third edition, Prentice-Hall, Inc., 1996.

IDENTIFIERS – *Paperback*, ISBN10: 0133942899, ISBN13: 9780133942897
Hardback, ISBN10: 0133737624, ISBN13: 9780133737622

This is cited throughout as [Proakis:1996] and is referred to in the second handout. This is the *third edition* to the book, and a *fourth edition has recently been released*:

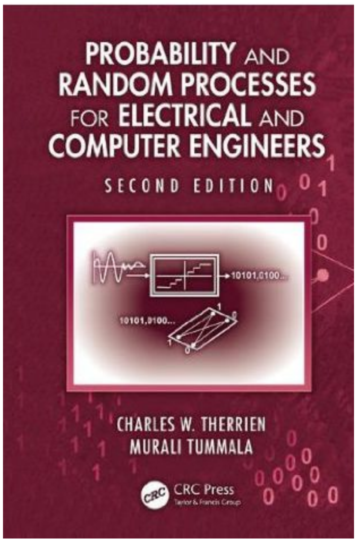


Figure 1.7: Further reading for statistical signal processing, [Therrien:2011].

Proakis J. G. and D. G. Manolakis, *Digital Signal Processing*: *Principles, Algorithms, and Applications*, Pearson New International Edition, Fourth edition, Pearson Education, 2013.

IDENTIFIERS – *Paperback*, ISBN10: 1292025735, ISBN13: 9781292025735

Although it is best to purchase the *fourth edition*, please bear in mind that the equation references throughout the lecture notes correspond to the third edition. For an undergraduate level text book covering an introduction to signals and systems theory, which it is assumed you have covered, the following is recommended [Mulgrew:2002]:

Mulgew B., P. M. Grant, and J. S. Thompson, *Digital Signal Processing*: *Concepts and Applications*, Palgrave, Macmillan, 2003.

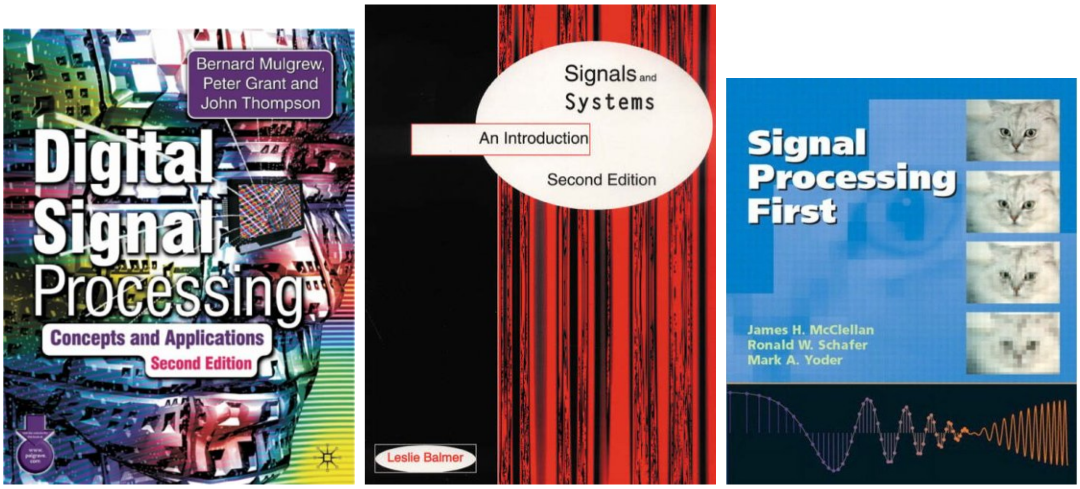
IDENTIFIERS – *Paperback*, ISBN10: 0333963563, ISBN13: 9780333963562

See `http://www.see.ed.ac.uk/~{}pmg/SIGPRO`

The latest edition was printed in 2003, but any of the book edition will do. An alternative presentation of roughly the same material is provided by the following book [Balmer:1997]:

Balmer L., *Signals and Systems*: *An Introduction*, Second edition, Prentice-Hall, Inc., 1997.

IDENTIFIERS – *Paperback*, ISBN10: 0134954729, ISBN13: 9780134956725



(a) [Mulgrew:2002]. (b) [Balmer:1997]. (c) [McClennan:2003].

Figure 1.8: Undergraduate texts on Signals and Systems.

The Appendix on complex numbers may prove useful.

For an excellent and gentle introduction to signals and systems, with an elegant yet thorough overview of the mathematical framework involved, have a look at the following book, if you can get hold of a copy (but don't go spending money on it):

> McClellan J. H., R. W. Schafer, and M. A. Yoder, *Signal Processing First*, Pearson Education, Inv, 2003.

IDENTIFIERS – *Paperback*, ISBN10: 0131202650, ISBN13: 9780131202658
Hardback, ISBN10: 0130909998, ISBN13: 9780130909992

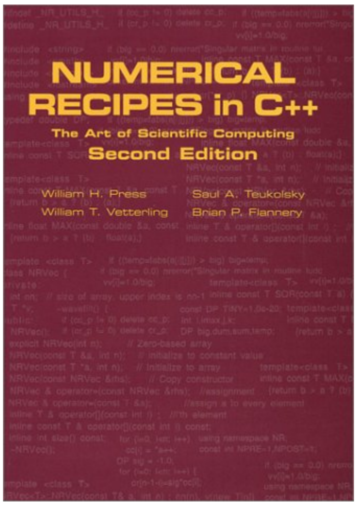
1.3.5 Further Recommended Reading

For additional reading, and for guides to the implementation of numerical algorithms used for some of the actual calculations in this lecture course, the following book is also strongly recommended:

> Press W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Receipes in C*: *The Art of Scientific Computing*, Second edition, Cambridge University Press, 1992.

IDENTIFIERS – *Paperback*, ISBN10: 0521437202, ISBN13: 9780521437202
Hardback, ISBN10: 0521431085, ISBN13: 9780521431088

Please note that there are many versions of the *numerical recipes* book, and that any version will do. So it would be worth getting the latest version.



NUMERICAL RECIPES in C++

The Art of Scientific Computing

Second Edition

William H. Press
William T. Vetterling
Saul A. Teukolsky
Brian P. Flannery

(a) **Recommended text**: [Press:1992].

Figure 1.9: Further reading for numerical methods and mathematics.

1.3.6 Additional Resources

Other useful resources include:

- The extremely comprehensive and interactive mathematics encyclopedia:

  Weisstein E. W., *MathWorld*, From MathWorld - A Wolfram Web Resource, 2008.

  See `http://mathworld.wolfram.com`

- Connexions is an environment for collaboratively developing, freely sharing, and rapidly publishing scholarly content on the Web. A wide variety of technical lectures can be found at:

  Connexions, The Connexions Project, 2008.

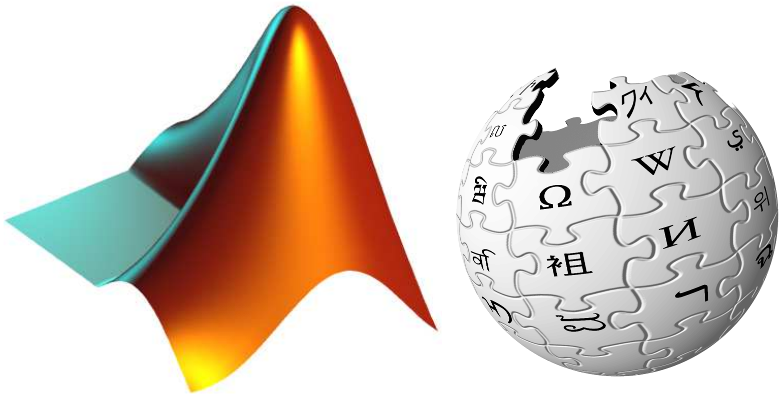
  See `http://cnx.org`

- The Wikipedia online encyclopedia is very useful, although beware that there is no guarantee that the technical articles are either correct, or comprehensive. However, there are some excellent articles available on the site, so it is worth taking a look.

  *Wikipedia, The Free Encyclopedia*Wikipedia, The Free Encyclopedia, 2008.

  See `http://en.wikipedia.org/`

- The Mathworks website, the creators of MATLAB, contains much useful information:



(a) The MATLAB logo. MATLAB is a useful utility to experiment with.

(b) Wikipedia, The Free Encyclopedia.

Figure 1.10: Some useful resources.

MATLAB: The language of technical computing, The MathWorks, Inc., 2008.

See `http://www.mathworks.com/`

- And, of course, the one website to rule them all:

  Google Search Engine, Google, Inc., 2008.

  See `http://www.google.co.uk`

1.3.7 Convention for Equation Numbering

In this handout, the following labelling convention is used for numbering equations that are taken from the various recommended texts. This labelling should be helpful for locating the relevant sections in the books for further reading. Equations labelled as:

M:v.w.xyz are similar to those with the same equation reference in the core recommended text book, namely [Manolakis:2001];

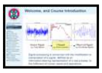
T:w.xyz are similar to those in [Therrien:1992] with the corresponding label;

K:w.xyz are similar to those in [Kay:1993] with the corresponding label;

P:v.w.xyz are used in chapters referring to basic DSP, and are references made to [Proakis:1996].

All other equation labeling refers to intra-cross-referencing for these handouts. Most equations are numbered for ease of referencing the equations, should you wish to refer to them in tutorials or email communications, and so forth.

1.4 What are Signals and Systems?



New slide

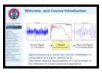
Common usage and understanding of the word *signal* is actually correct from an Engineering perspective within some rather broad definitions: a signal is thought of as *something* that carries information. Usually, that *something* is a pattern of variations of a physical quantity that can be manipulated, stored, or transmitted by a physical process. Examples include speech signals, general audio signals, video or image signals, biomedical signals, radar signals, and seismic signals, to name but a few.

So formally, a **signal** is defined as an information-bearing representation of a real physical process. It is important to recognise that signals can take many equivalent forms or representations. For example, a speech signal is produced as an acoustic signal, but it can be converted to an electrical signal by a microphone, or a pattern of magnetization on a magnetic tape, or even as a string of numbers as in digital audio recording.

The term *system* is a little more ambiguous, and can be subject to interpretation. The word *system* can correctly be understood as a process, but often the word *system* is used to refer to a large organisation that administers or implements some process.

In Engineering terminology, a **system** is something that can manipulate, change, record, or transmit **signal**s. In general, **system**s operate on **signal**s to produce new signals or new signal representations. For example, an audio compact disc (CD) stores or represents a music signal as a sequence of numbers. A CD player is a system for converting the numerical representation of the signal stored on the disk to an acoustic signal that can be heard.

1.4.1 Mathematical Representation of Signals

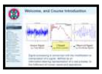


New slide

A *signal* is defined as an information-bearing representation of a real process. It is a pattern of variations, commonly referred to as a waveform, that encodes, represents, and carries information.

Many signals are naturally thought of as a pattern of variations with time. For example, a speech signal arises as a pattern of changing air pressure in the vocal tract, creating a sound wave, which is then converted into electrical energy using a microphone. This electrical signal can then be plotted as a time-waveform, and an example is shown in Figure 1.11. The vertical axis denotes air pressure or microphone voltage, and the horizontal axis represents time. This particular plot shows four contiguous segments of the speech waveform. The second plot is a continuation of the first, and so on, and each plot is vertically offset with the starting time of each segment shown on the left vertical axis.

1.4.1.1 Continuous-time and discrete-time signals



New slide

The signal shown in Figure 1.11 is an example of a one-dimensional **continuous-time signal**. Such signals can be represented mathematically as a function of a single independent variable, t, which represents time and can take on any real-valued number.

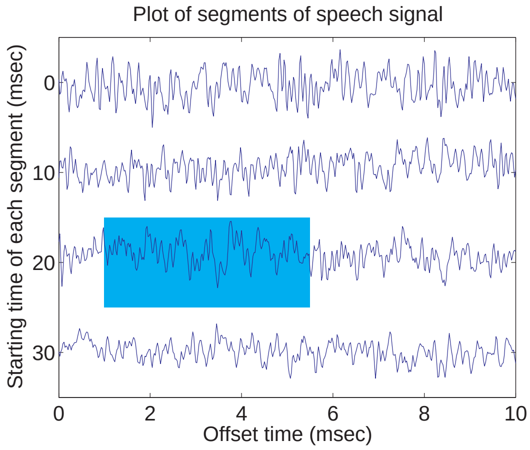


Figure 1.11: Plot of part of a speech signal. This signal can be represented by the function $s(t)$, where t is the independent variable representing time. The shaded region is shown in more detail in Figure 1.12.

Hence, each segment of the speech waveform can be associated with a function $s(t)$. In some cases, the function $s(t)$ might be a simple function, such as a sinusoid, but for real signals, it will be a complicated function.

Generally, most *real world* signals are continuous in time and analogue: this means they exist for all time-instances, and can assume any value, within a predefined range, at these time instances. Although most signals originate as continuous-time signals, digital processors and devices can only deal with **discrete-time signals**. A discrete-time representation of a signal can be obtained from a continuous-time signal by a process known as **sampling**. There is an elegant theoretical foundation to the process of sampling, although it suffices to say that the result of sampling a continuous-time signal at isolated, equally spaced points in time is a sequence of numbers that can be represented as a function of an index variable that can take on only discrete integer values.

The sampling points are spaced by the **sampling period**, denoted by T_s. Hence, the continuous-time signal, $s(t)$, is *sampled* at times $t = nT_s$ resulting in the discrete-time waveform denoted by:

$$s[n] = s(nT_s), \quad n \in \{0,\, 1,\, 2,\, \dots\}. \tag{1.1}$$

where n is the index variable. A discrete-time signal is sometimes referred to as a discrete-time sequence, since the waveform $s[n]$ is a sequence of numbers. Note, the convention that parenthesis $(\,)$ are used to enclose the independent variable of a continuous-time function, and square brackets $[\,]$ enclose the index variable of a discrete-time signal. Unfortunately, this notation is not always adhered too (and is not yet consistent in these notes either).

Figure 1.12 shows an example of a short segment of the speech waveform from Figure 1.11, with a sampling period of $T_s = \frac{1}{44100}$ seconds, or a sampling frequency of $f_s = \frac{1}{T_s} = 44.1$ kHz. It is not possible to evaluate the continuous-time function

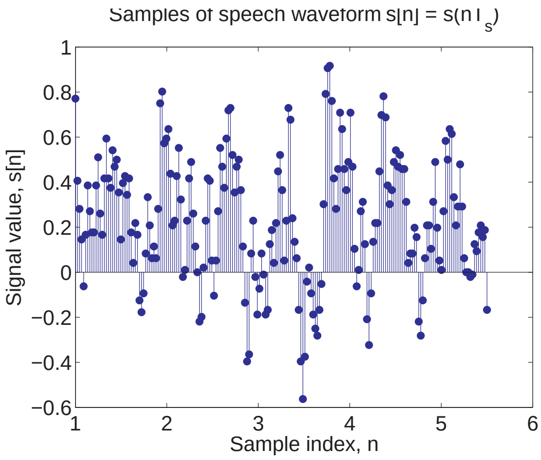


Figure 1.12: Example of a discrete-time signal. This is a sampled version of the shaded region shown in Figure 1.11.

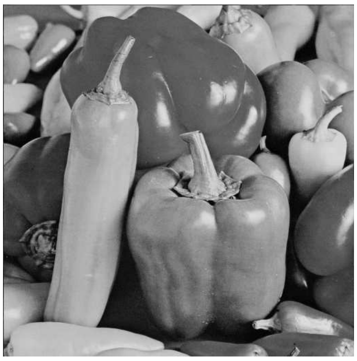
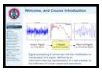


Figure 1.13: Example of a signal that can be represented by a function of two spatial variables.

$s(t)$ for every value of t, only at a finite-set of points, which will take a finite time to evaluate. Intuitively, however, it is known that the closer the spacing of the sampled points, the more the sequence retains the shape of the original continuous-time signal. The question arises, then, regarding what is the smallest **sampling period** that can be used to retain all or most of the information about the original signal.

1.4.1.2 Other types of signals



New slide

While many signals can be considered as evolving patterns in time, many other signals are not time-varying patterns at all. For example, an image formed by focusing light through a lens onto an imaging array is a spatial pattern. Thus, an image is represented mathematically as a function of two independent spatial variables, x and y; thus, a picture might be denoted as $p(x, y)$. An example of a **gray-scale image** is shown in Figure 1.13; thus, the value $p(x_0, y_0)$ represents the particular shade of gray at position (x_0, y_0) in the image.

Although images such as that shown in Figure 1.13 represents a quantity from a physical two-dimensional (2-D) spatial continuum, digital images are usually discrete-variable 2-D signals obtained by sampling a continuous-variable 2-D signal. Such a 2-D discrete-variable signal would be represented by a 2-D sequence or array of numbers, and is denoted by:

$$p[m,\, n] = p(m\Delta_x,\, n\Delta_y), \quad m, n \in \{0,\, 1,\, \dots\, N-1\}. \tag{1.2}$$

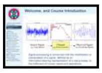
where m and n take on integer values, and Δ_x and Δ_y are the horizontal and vertical sampling spacing or periods, respectively.

Two-dimensional functions are appropriate mathematical representations of still images that do not change with time; on the other hand, a sequence of images that creates a video requires a third independent variable for time. Thus, a video sequence is represented by the three-dimensional (3-D) function $v(x,\, y,\, t)$.

The purpose of this section is to introduce the idea that signals can:

- be represented by mathematical functions in one or more dimensions;
- be functions of continuous or discrete variables.

The connection between functions and signals is key to signal processing and, at this point, functions serve as abstract symbols for signals. This is an important, but very simple, concept for using mathematics to describe signals and systems in a systematic way.



1.4.2 Mathematical Representation of Systems

New slide

A **system** manipulates, chances, records, or transmits **signal**s. To be more specific, a one-dimensional continuous-time system takes an input signal $x(t)$ and produces a corresponding output signal $y(t)$. This can be represented mathematically by the expression

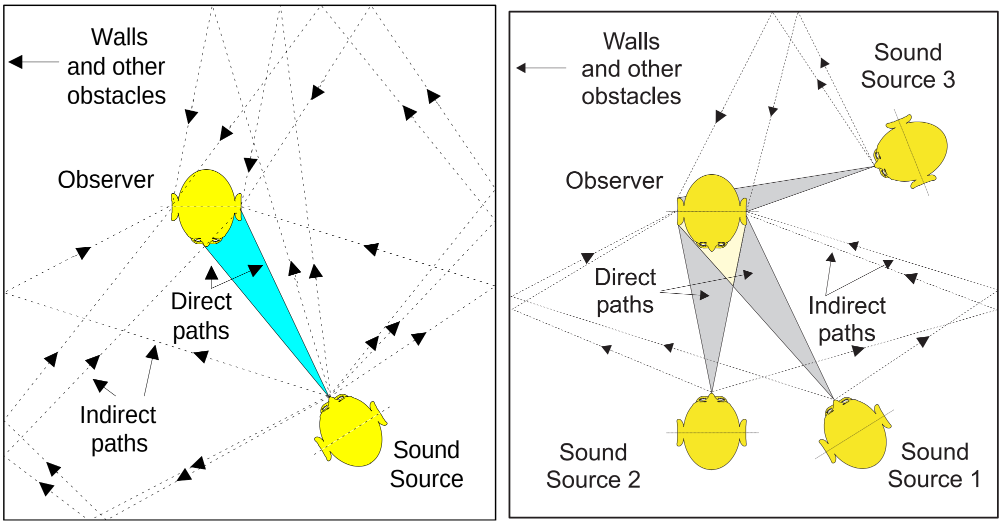
$$y(t) = \mathcal{T}\left\{x(t)\right\} \tag{1.3}$$

which means that the input signal, $x(t)$, be it a waveform or an image, is operated on by the system, which is symbolised by the operator $\mathcal{T}$ to produce the output $y(t)$. So, for example, consider a signal that is the square of the input signal; this is represented by the equation

$$y(t) = \left[x(t)\right]^2 \tag{1.4}$$

Figure 1.14 and Figure 1.16 show how signals can be generated and observed in a real application. In Figure 1.14, the sound source and the information received by the observer, or microphone, are the **signals**; the room acoustics represent the **system**. Figure 1.15 shows the **input signal** to the system, a *characterisation of the system*, and the resulting **output signal**. In Figure 1.16, the blurred images are the result of the original image being passed through a **linear system**; the linear system represents the physical process of a camera, for example, being out-of-focus, or in motion relative to the object of interest.

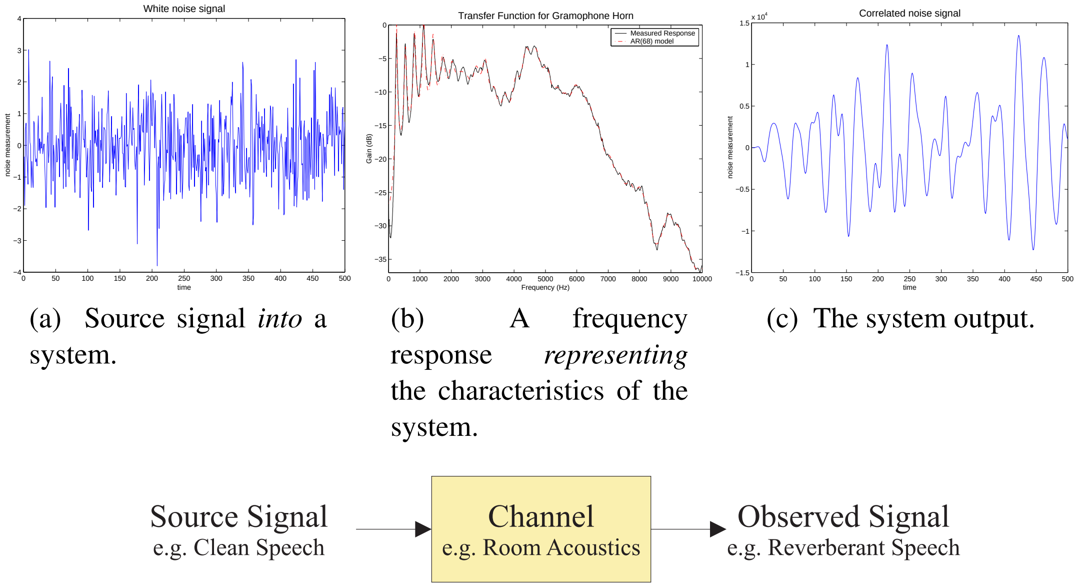
The subject of signals and systems is the basis of a branch of Engineering known as signal processing; this area is formally defined as follows:



(a) Acoustic path from a sound source to a microphone.

(b) Many sound sources within a room.

Figure 1.14: Observed signals in room acoustics.



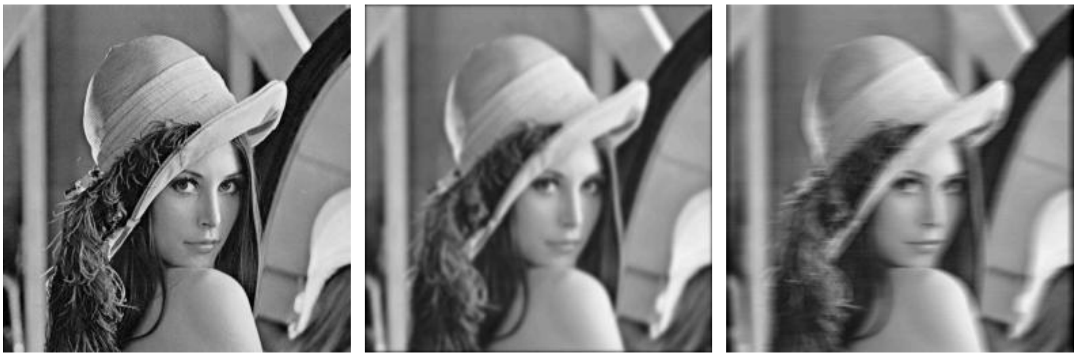
(a) Source signal *into* a system.

(b) A frequency response *representing* the characteristics of the system.

(c) The system output.

(d) Block diagram representation of signal paths.

Figure 1.15: The result of passing a signal through a system.



(a) An original unblurred noiseless image.

(b) An image distorted by an out-of-focus blur.

(c) Image distorted by motion blur.

Figure 1.16: A blind image deconvolution problem; restoration of natural photographic images.

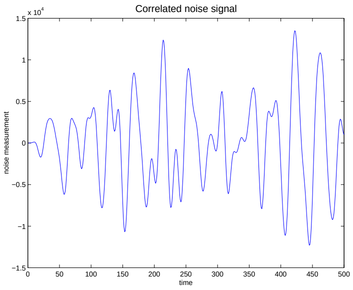
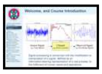


Figure 1.17: Amplitude-verses-time plot.

> *Signal processing* is concerned with the modification or manipulation of a signal, defined as an information-bearing representation of a real process, that has been passed through a *system*, to the fulfillment of human needs and aspirations.

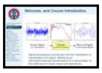


New slide

1.4.3 Deterministic Signals

The deterministic signal model assumes that signals are explicitly known for all time from time $t = -\infty$ to $t = +\infty$, where $t \in \mathbb{R}$, the set of all real numbers. There is absolutely no uncertainty whatsoever regarding their past, present, or future signal values. The simplest description of such signals is an amplitude-verses-time plot, such as that shown in Figure 1.17; this *time history* helps in the identification of specific patterns, which can subsequently be used to extract information from the signal. However, quite often, information present in a signal becomes more evident by transformation of the signal into another domain, and one of the most nature examples is the frequency domain.

1.5 Motivation for Signal Modelling



Some state-of-the-art applications of statistical signal processing include the following: *New slide*

Biomedical From medical imaging to analysis and diagnosis, signal processing is now dominant in patient monitoring, preventive health care, and tele-medicine. From analysing electroencephalogram (EEG) scans to magnetic resonance imaging (MRI) (or nuclear magnetic resonance imaging (NMRI)), to classification and analysis of deoxyribonucleic acid (DNA) from micro-arrays, signal processing is required to make sense of the analogue signals to then provide information to clinicians and doctors.

Surveillance and homeland security From fingerprint analysis, voice transcription and communication monitoring, to the analysis of closed-circuit television (CCTV) footage, digital signal processing is applied in many areas of homeland security. It is an especially well-funded area at the moment.

Target tracking and navigation Although radar and sonar principally use analogue signals for *illuminating* an object with either an electromagnetic or acoustic wave, discrete-time signal processing is the primary method for analysing the received data. Typical features for estimation include detecting targets, estimating the position, orientation, and velocity of the object, target tracking and target identification.

Of recent interest is tracking groups of targets, such as a convey of vehicles, or a flock of birds. Attempting to track each individual target is an overly complicated problem, and by considering the group dynamics of a particular scenario, the multi-target tracking problem is substantially simplified.

Mobile communications New challenges in mobile communications include next-generation networks; users demand higher data-rates which, in-turn, requires higher bandwidth. Typically, higher-bandwidth communication systems have shorter range. Rather than have more and more base stations for the mobile network, there is substantial research into mobile ad-hoc networks.

A mobile ad-hoc network is a self-configuring network of mobile routers connected by wireless links, forming an arbitrary topology. The routers are free to move randomly and organize themselves arbitrarily; thus, the network's wireless topology may change rapidly and unpredictably. The challenge is to design a system that can cope with this changing topology, and is a very active area of research in communication theory.

A testament to the change in mobile communications is the availability of cheap mobile broadband modems which provide broadband Internet access which is comparable with fixed-line technologies that were available only a few years ago.

Speech enhancement and recognition Whether for the analysis of a black-box recording, for enhancing speech recognition in noisy and reverberant environments, or for the improved acoustic clarity of mobile phone conversations, the enhancement of acoustic signals is still a major aspect of signal processing research.

Many signal processing systems are designed to extract information for some purpose. They share the common problem of needing to estimate the values of a group of parameters. Such algorithms involve signal modelling and spectral estimation. Some typical applications and the desired parameter include:

Radar Radar is primarily used in determining the position of an aircraft or other moving object; for example, in airport surveillance. It is desirable to estimate the range of the aircraft, as determined by the time for an electromagnetic pulse to be reflected by the aircraft.

Sonar Sonar is also interested in the position of a target, such as a submarine. However, whereas radar is, mostly, an *active* device in the sense that it transmits an electromagnetic pulse to *illuminate* the target, sonar listens for noise radiated by the target. This radiated noise includes sounds generated by machinery, or the propeller action. Then, by using a **sensor array** where the relative positions of each sensor are known, the time delay between the arrival of the pulse at each sensor can be measured and this can be used to determine the bearing of the target.

Image analysis It might be desirable to estimate the position and orientation of an object from a camera image. This would be useful, for example, in guiding a robot to pick up an object. Alternatively, it might be desirable to remove various forms of blur from an image, as shown in Figure 1.16; this blur might be characterised by a parametric function.

Biomedicine A parameter of interest might be the heart rate of a fetus.

Communications Estimate the carrier frequency of a signal such that the signal can be demodulated to baseband.

Control Estimate the position of a boat such that corrective navigational action can be taken.

Seismology Estimate the underground distance of an oil deposit based on sound reflections due to different densities of oil and rock layers.

And the list can go on, with a multitude of applications stemming from the analysis of data from physical experiments through to economic analysis. To gain some motivation for looking at various aspects of statistical signal processing, some specific applications will be considered that require the tools this module will introduce. These applications include:

- Speech Modelling and Recognition

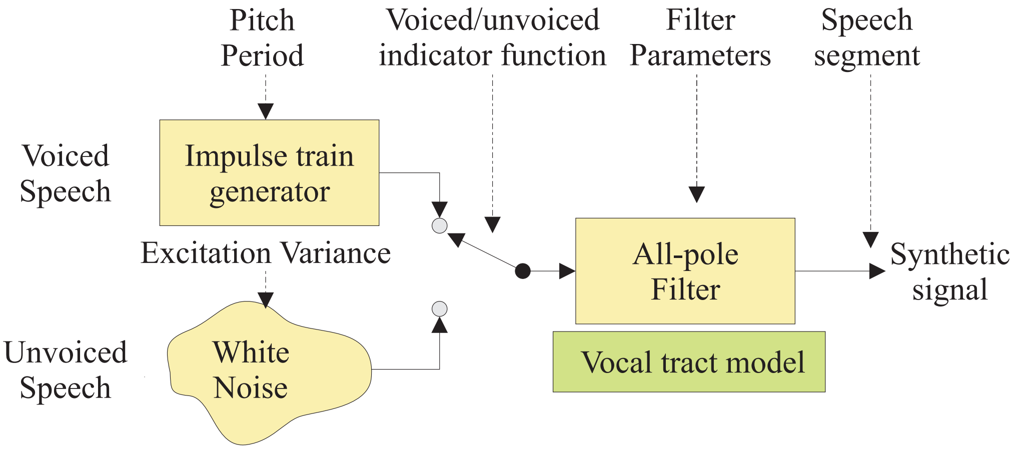
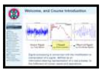


Figure 1.18: The speech synthesis model.

- Single Channel Blind System Identification
- Blind Signal Separation
- Data Compression
- Enhancement of Signals in Noise

1.5.1 Speech Modelling and Recognition



New slide

Statistical parametric modelling can be used to characterise the speech production system, and therefore can be applied in the analysis and synthesis of speech. In the analysis of speech, the waveform is sampled at a rate of about 8 to 20 kHz, and broken up into short segments whose duration is typically 10 to 20 ms; this results in consecutive segments containing about 80 to 400 time samples.

Most speech sounds, generally, are classified as either *voiced* or *unvoiced* speech:

- voiced speech is characteristic of vowels;
- unvoiced speech is characteristic of consonants at the beginning of syllables, fricatives (/f/, /s/ sounds), and a combination of these.

Thinking of the types of sound fields created by vowels, it is apparent that *voiced speech* has a harmonic quality. In fact, it is sometimes known as frequency-modulated speech. A commonly used model for voiced speech exploits this harmonic characteristic, and uses the so-called *sum-of-sinusoids* decomposition. *Unvoiced speech*, on the other hand, does not exhibit such a harmonic structure, although it does possesses a form that can be modelled using the statistical models introduced in later lectures.

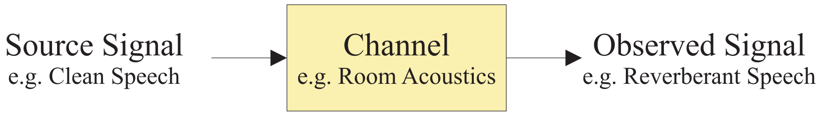
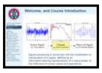


Figure 1.19: Solutions to the blind deconvolution problem requires advanced statistical signal processing.

For both of these types of speech, the production is modelled by driving or exciting a linear system, representing the vocal tract, with an excitation having a flat (or constant) spectrum.

The vocal tract, in turn, is modelled by using a pole-zero system, with the poles modelling the vocal tract resonances and the zeros serving the purpose of dampening the spectral response between pole frequencies. In the case of voiced speech, the input to the vocal tract model is a quasi-periodic pulse waveform, whereas for unvoiced speech, the source is modelled as random noise. Thus, the complete set of parameters for this model include an indicator variable as to whether the speech is voiced or unvoiced, the pitch period for voiced sounds, the gain or variance parameter for unvoiced sounds, and the coefficients for the all-pole filter modelling the vocal tract filter. The model is shown in Figure 1.18. This model is widely used for low-bit-rate (less than 2.4 kbits/s) **speech coding**, **synthetic speech generation**, and extraction of features for speaker and **speech recognition**.



1.5.2 Single Channel Blind System Identification

New slide

Consider the following abstract problem that is shown in Figure 1.19:

- The output only of a system is observed, and it is desirable to estimate the source signal that is applied to the input of the system without knowledge of the system itself. In other-words, the output observation, $\mathbf{x} = \{x[n],\, n \in \mathbb{Z}\}$,[1] is modelled as a function of the unknown source signal, $\mathbf{s} = \{s[n],\, n \in \mathbb{Z}\}$, with an unknown, possibly nonlinear, distortion denoted by $\mathcal{F}$; more formally, $\mathbf{x} = \mathcal{F}(\mathbf{s})$.

- When the function $\mathcal{F}$ is linear time-invariant (LTI), and defined by the impulse response $h[n]$, then:

$$x[n] = h[n] * s[n] = \sum_{k \in \mathbb{Z}} h[n-k]\, s[n] \tag{1.5}$$

- **Problem**: Given only $\{x[n]\}$, estimate either the channel function, $\mathcal{F}$, which in the LTI case will be represented by the impulse response $h[n]$, *or* a scaled shifted version of the source signal, $\{s[n]\}$; i.e. $\hat{s}[n] = a\, s[n-l]$ for some l.

[1] The notation $n \in \mathbb{Z}$ means that n belongs to, or is an element of, the set of integers: $\{-\infty, \dots, -2, -1, 0, 1, 2, \dots, \infty\}$. In otherwords, it may take on any integer value.

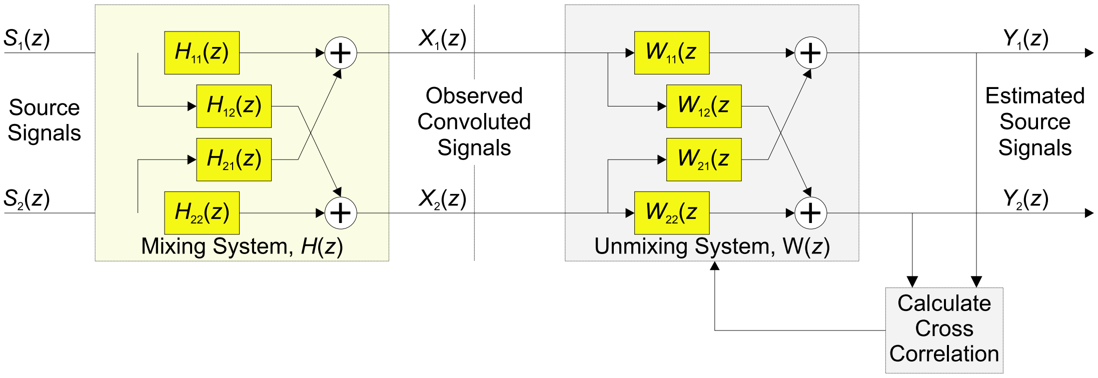


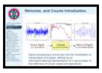
Figure 1.20: Standard signal separation using the independent component assumption.

The distortion operator, $\mathcal{F}$, could represent the:

- acoustical properties of a room (with applications in **hands free telephones**, **hearing aids**, **archive restoration**, and **automatic speech recognition**);
- effect of multi-path radio propagation (with applications in **communication channels**);
- non-impulsive excitation in seismic applications (with applications in **seismology**);
- blurring functions in **image processing**; in this case, the signals are 2-D.

This problem can only be solved by parametrically modelling the source signal and channel, and using **parameter estimation** techniques to determine the appropriate parameter values.

1.5.3 Blind Signal Separation



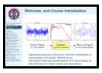
New slide

An extremely broad and fundamental problem in signal processing is BSS, and an important special case is the separation of a mixture of audio signals in an acoustic environment. Typical applications include the separation of overlapping speech signals, the separation of musical instruments, enhancement of speech recordings in the presence of background sounds, or any variation of the three. In general, a number of sounds at discrete locations within a room are filtered due to room acoustics and then mixed at the observation points; for example, a microphone will pick up a number of reverberant sounds simultaneously (see Figure 1.14).

A very powerful paradigm within which signal separation can be achieved is the assumption that the source signals are statistically independent of one another; this is known as independent component analysis (ICA). Figure 1.20 demonstrates a separation algorithm based on ICA; an "unmixing" system is chosen that has minimal statistical correlation (a sufficient but not necessary condition for statistical independence, as will be seen later in this course) of the hypothesised separated signals, thereby matching the statistical characteristics of the original signals. This algorithm then uses standard convex optimisation algorithms to solve the minimisation problem.

It is clear, then, that this approach to ICA requires good estimates of the correlation functions from a limited amount of data.

1.5.4 Data Compression



Welcome, and Course Introduction

New slide

Three basis principles of data compression for communication systems include:

Mathematically Lossless Compression This principle looks for mathematical coding schemes that reduce the *bits* required to represent a signal. For example, long runs of 0's might be replaced by a shorter representation. This method of compression is used in computer file compression systems.

Lossy compression by removing redundant information This approach is often performed in a transform domain, such as the frequency domain. There might be many Fourier coefficients that are small, and do not significantly contribute to the representation of the signal. If these small coefficients are not transmitted, then compression is achieved.

Lossless compression by linear prediction If it is possible to *predict* the current data sample from previous data samples, then it would not be necessary to transmit the current data symbol. Typically, however, the prediction is not completely accurate. However, by only transmitting the *difference* between the prediction and the actual value, which is typically a lot smaller than the actual value, then it turns out a fewer number of bits need to be transmitted, and thus compression achieved. The trick is to design a good *predictor*, and this is where statistical signal processing comes in handy.

1.5.5 Enhancement of Signals in Noise

High quality digital audio has in recent years dramatically raised expectations about sound quality. For example, high quality media such as:

- compact disc
- digital audio tape
- digital versitile disc-audio and super-audio CD.

Audio degradation is any undesirable modification to an audio signal occurring as the result of, or subsequent to, the recording process. Disturbances or distortions such as

(a) The digital versitile disc-audio (DVD-A) logo.

(b) The super-audio CD (SACD) logo.

Figure 1.21: High-quality audio formats.

1. background noise,
2. echoes and reverberation,
3. and media noise.

must be reduced to adequately low levels. Ideal restoration reconstructs the original sound exactly as would be received by transducers (microphone etc.,) in the absence of noise and acoustic distortion. Interest in historical material led to restoration of degraded sources including

1. wax cylinders recordings,
2. disc recordings (78rpm, etc.),
3. and magnetic tape recordings.

Restoration is also required in contemporary digital recordings if distortion too intrusive. **Note** that noise present in recording environment, such as audience noise at a musical performance, considered part of *performance*. Statistical signal processing is required in such applications.

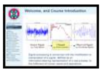
2

Review of Basic Probability Theory

> All knowledge degenerates into probability.
>
> David Hume

This handout gives a review of the fundamentals of probability theory.

2.1 Introduction



New slide

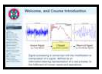
The theory of probability deals with averages of mass phenomena occurring sequentially or simultaneously; in signal processing and communications, this might include radar detection, signal detection, anomaly detection, parameter estimation, and so forth.

How does one start considering the notion and meaning of probability? It has been *observed* in many fields that certain averages approach a constant value as the number of observations increases, and this value remains the same if the averages are evaluated over any subsequence (of observations) specified before the experiment is performed. In a coin experiment, for example, the percentage of heads approaches 0.5 or some other constant, and the same average is obtained if every fourth, sixth, or arbitrary selection of tosses is chosen. Note that the notion of an average is not in-itself a probabilistic term.

The purpose of the theory of probability is to describe and predict these averages in terms of probabilities of events. The probability of an event A is a number $\Pr(A)$ assigned to this event. This number *could* be interpreted as follows:

> 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:
>
> $$\Pr(A) \approx \frac{n_A}{n} \tag{2.1}$$
>
> provided that n is *sufficiently large*.

Note that this interpretation and the language used is all very imprecise, and phrases such as *high degree of certainty*, *close to*, and *sufficiently large* has no clear meaning. These terms will be more precisely defined as concepts are introduced throughout this course.



New slide

2.2 Classical Definition of Probability

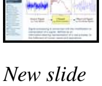
For several centuries, the theory of probability was based on the *classical definition*, which states that the probability $\Pr(A)$ of an event A is determine *a priori* without actual experimentation. It is given by the ratio:

$$\Pr(A) = \frac{N_A}{N} \tag{2.2}$$

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*.

This definition, however, has some difficulties when the number of possible outcomes is infinite, as illustrated in the following example in Section 2.2.1.



New slide

2.2.1 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?

KEYPOINT! (Recalling Geometry!). To fully appreciate this problem, it is perhaps worth being aware of the geometry of this problem. The idea of the geometry is to keep simple geometric shapes, rather than to play on some obscure geometric properties. Therefore, note that if three tangents to a circle of radius ${}^{r}\!/_{2}$ are drawn at angular intervals of $120\,\mathrm{degs}$, then the resulting equilateral triangle fits inside a larger circle of radius r, as shown in Figure 2.1. The length of the sides of one of this equilateral triangle is $r\sqrt{3}$.

Using the classical definition of probability, three reasonable solutions can be obtained:

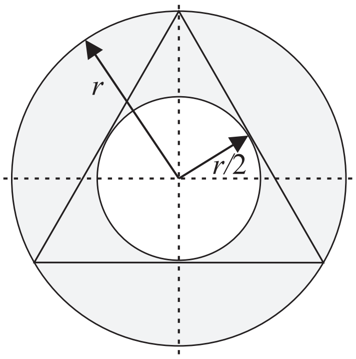
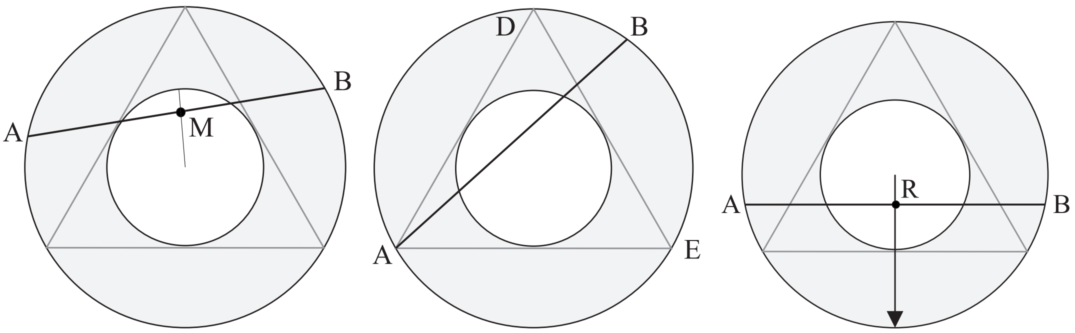


Figure 2.1: Bertrand's paradox, problem definition.



(a) The *midpoint method*. (b) The *endpoint method*. (c) The *radius method*.

Figure 2.2: Different selection methods.

1. In the **random midpoints** method, a cord is selected by choosing a point M anywhere in the circle, an end-point A on the circumference of the circle, and constructing a chord AB through these chosen points. This is shown graphically in Figure 2.2a.

   It is reasonable, therefore, to consider as *favourable outcomes* all points inside the inner-circle of radius ${}^{r}/_{2}$, and to consider *all possible outcomes* as points inside the outer-circle of radius r. Therefore, using as a measure of these outcomes the corresponding areas, it follows that:

   $$p = \frac{\pi \left(\frac{r}{2}\right)^2}{\pi r^2} = \frac{1}{4} \tag{2.3}$$

2. 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. This is shown in Figure 2.2b, where the point A has been drawn to coincide with the particular triangle drawn. If B lies on the arc between the two other vertices, D and E, of the triangle whose first vertex coincides with A, then AB will be longer than the length of the side of the triangle.

   The *favourable outcomes* are now the points on this arc, and since the angle of the arc DE is $\frac{2\pi}{3}$ radians, a measure of this outcome is the arc length $\frac{2\pi r}{3}$. Moreover, the total outcomes are all the points on the circumference of the main circle, and therefore it follows:

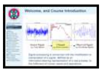
   $$p = \frac{\frac{2\pi r}{3}}{2\pi r} = \frac{1}{3} \tag{2.4}$$

3. 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. The construction of this chord is shown in Figure 2.2c.

   The *favourable outcomes* are the points on the radius that lie *inside* of the inner-circle, or a measure of this outcome is given by the diameter of the inner-circle, r. The total outcomes are the points on the diameter of the outer-circle, and a measure of that respective length is $2r$. Therefore, the probability is given by

   $$p = \frac{r}{2r} = \frac{1}{2} \tag{2.5}$$

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



New slide

2.2.2 Using the Classical Definition

The difficulty with the classical definition in Equation 2.2, as seen in Bertrand's Paradox, is in determining N and N_A.

Example 2.1 (Rolling two dice). Two dice are rolled; find the probability, p, that the sum of the numbers shown equals 7. Consider three possibilities:

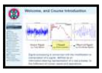
1. The *possible outcomes* total 11 which are the sums $\{2,\, 3,\, \ldots,\, 12\}$. Of these, only one (the sum 7) is favourable. Hence, $p = \frac{1}{11}$.

   This is, of course, wrong, and the reason is that each of the 11 possible outcomes are *not* equally probable.

2. Similarly, writing down the possible pairs of shown numbers, without distinguishing between the first and second die. There are then 21 pairs, $(1, 1)$, $(1, 2)$, $\ldots$, $(1, 6)$, $(2, 1)$, $\ldots$, $(6, 6)$, of which there are three favourable pairs $(3, 4)$, $(5,\ 2)$ and $(6, 1)$. However, gain, the pairs $(3, 4)$ and $(6, 6)$, for example, are not equally likely.

3. Therefore, to count all possible outcomes which are equally probable, it is necessary to could all pairs of numbers distinguishing between the first and second die. This will give the correct probability.

2.2.3 Difficulties with the Classical Definition



New slide

The classical definition in Equation 2.2 can be questioned on several grounds, namely:

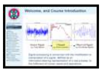
1. The term **equally probable** in the definition of probability is making use of a concept still to be defined!

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 in Equation 2.2.

3. If the number of possible outcomes is infinite, then some other measure of infinity for determining the classical probability ration in Equation 2.2 is needed, such as length, or area. This leads to difficulties, as discussed in Bertrand's paradox.

2.3 Axiomatic Definition



New slide

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:

$$\Pr(A) \geq 0 \tag{2.6}$$

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:

$$\Pr(S) = 1 \tag{2.7}$$

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

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) \tag{2.8}$$

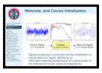
or more generally, if $A_1, A_2, \dots$ is a collection of disjoint events, such that $A_i \cap A_j = \emptyset$ for all pairs i, j satisfying $i \neq j$, then:

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i) \tag{2.9}$$

Note that Equation 2.9 does not directly follow from Equation 2.8, even though it may appear to. Dealing with infinitely many sets requires further insight, and here the result of Equation 2.9 is actually an additional condition known as the **axiom of infinite additivity**.

These axioms can be formalised by defining measures and fields as appropriate, but the level of detail is beyond this course.

These axioms, once formalised, are known as the **Kolmogorov Axioms**, named after the Russian mathematician. Note that an alternative approach to deriving the laws of probability theory from a certain set of postulates was developed by Cox. However, this won't be considered in this course.



New slide

2.3.1 Set Theory

Since the classical definition of probability details in total number of outcomes, as well as events, it is necessary to utilise the mathematical language of sets to formulise precise definitions.

A **set** is a collection of objects called **elements**. For example, "*car*, *apple*, *apple*" is a set with three elements whose elements are a car, an apple, and a pencil. The set "*heads*, *tails*" has two elements, while the set "1, 2, 3, 5", has four. It is assumed that most readers will have come across **set theory** to some extent, and therefore, it will be used throughout the document as and when needed.

Some basic notation, however, includes the following:

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

$$A \cup B = B \cup A, \quad (A \cup B) \cup C = A \cup (B \cup C) \tag{2.10}$$

$$AB = BA, \quad (AB)C = A(BC), \quad A(B \cup C) = AB \cup AC \tag{2.11}$$

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:

$$A \cup \overline{A} = S \quad \text{and} \quad A \cap \overline{A} \equiv A\overline{A} = \{\emptyset\} \tag{2.12}$$

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

$$\bigcup_{i=1}^{\infty} A_i = S, \quad A_i \cap A_j = \{\emptyset\}, \quad i \neq j \quad \Rightarrow \quad U = [A_1, \dots, A_n] \tag{2.13}$$

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

$$\overline{A \cup B} = \overline{A} \cap \overline{B} \equiv \overline{A}\,\overline{B} \quad \text{and} \quad \overline{A \cap B} \equiv \overline{AB} = \overline{A} \cup \overline{B} \tag{2.14}$$

As an application of this, note that:

$$\overline{A \cup BC} = \overline{A}\,\overline{BC} = \overline{A}\left(\overline{B} \cup \overline{C}\right) \tag{2.15}$$

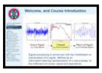
$$= \left(\overline{A}\,\overline{B}\right) \cup \left(\overline{A}\,\overline{C}\right) \tag{2.16}$$

$$= \overline{A \cup B} \cup \overline{A \cup C} \tag{2.17}$$

$$\Rightarrow \quad A \cup BC = (A \cup B)\,(A \cup C) \tag{2.18}$$

This result can easily be derived by using Venn diagrams, and it is worth checking this result yourself. This latter identity will also be used later in Section 2.3.2.

2.3.2 Properties of Axiomatic Probability



New slide

Some simple consequences of the definition of probability defined in Section 2.3 follow immediately:

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

$$\Pr(\emptyset) = 0 \tag{2.19}$$

Complements Since $A \cup \overline{A} = S$ and $A\overline{A} = \{\emptyset\}$, then using Equation 2.8, $\Pr\left(A \cup \overline{A}\right) = \Pr(A) + \Pr\left(\overline{A}\right) = \Pr(S) = 1$, such that:

$$\Pr\left(\overline{A}\right) = 1 - \Pr(A) \tag{2.20}$$

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

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) \tag{2.21}$$

Example 2.2 (Proof of the Sum Rule). Prove the result in Equation 2.21.

SOLUTION. To prove this, separately write $A \cup B$ and B as the union of two mutually exclusive events (using Equation 2.18 and the fact $A \cup \overline{A} = S$ and $S\,B = B$).

- First, note that

  $$A \cup \left(\overline{A}\,B\right) = \left(A \cup \overline{A}\right)(A \cup B) = A \cup B \tag{2.22}$$

  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:

  $$B = \left(A \cup \overline{A}\right)B = (A\,B) \cup \left(\overline{A}\,B\right) \tag{2.23}$$

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

Using these two disjoint unions, then:

$$\Pr(A \cup B) = \Pr\left(A \cup \left(\overline{A}\,B\right)\right) = \Pr(A) + \Pr\left(\overline{A}\,B\right) \tag{2.24}$$

$$\Pr(B) = \Pr\left((A\,B) \cup \left(\overline{A}\,B\right)\right) = \Pr(A\,B) + \Pr\left(\overline{A}\,B\right) \tag{2.25}$$

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

$$\Pr(A \cup B) - \Pr(B) = \Pr\left(A \cup \left(\overline{A}\,B\right)\right) = \Pr(A) - \Pr(A\,B) \tag{2.26}$$

□

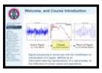
Example 2.3 (Sum Rule). Let A and B be events with probabilities $\Pr(A) = 3/4$ and $\Pr(B) = 1/3$. Show that $1/12 \leq \Pr(A\,B) \leq 1/3$.

SOLUTION. Using the sum rule, that:

$$\Pr(A\,B) = \Pr(A) + \Pr(B) - \Pr(A \cup B) \geq \Pr(A) + \Pr(B) - 1 = \frac{1}{12} \tag{2.27}$$

□

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 $\Pr(A\,B) \leq \min\{\Pr(A), \Pr(B)\} = 1/3$.



New slide

2.3.3 Countable Spaces

If the **certain event**, S, consists of N outcomes, and N is a finite number, then the probabilities of all events can be expressed in terms of the probabilities $\Pr(\zeta_i) = p_i$ of the elementary events $\{\zeta_i\}$.

Example 2.4 (Cups and Saucers). Six cups and saucers come in pairs: there are two cups and saucers which are red, two which are while, and two which are blue. If the cups are placed randomly onto the saucers (one each), find the probability that no cup is upon a saucer of the same pattern.

SOLUTION.
- Lay the saucers in order, say as $RRWWBB$.
- The cups may be arranged in $6!$ ways, but since each pair of a given colour may be switched without changing the appearance, there are $6!/(2!)^3 = 90$ distinct arrangements.

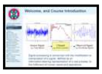
  By assumption, each of these are equally likely.
- The arrangements in which cups never match their saucers are:

$$\begin{array}{llll} \underline{WW}BBRR, & \underline{WB}RBWR, & \underline{BW}BRRW, & \underline{BB}RRWW \\ & \underline{WB}BRWR, & \underline{BW}RBRW & \\ & \underline{WB}RBRW, & \underline{BW}RBWR & \\ & \underline{WB}BRWR, & \underline{BW}BRRW & \end{array} \tag{2.28}$$

□

- Hence, the required probability is $^{10}/_{90} = 1/9$.

2.3.4 The Real Line



New slide

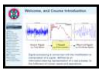
If the **certain event**, S, consists of a non-countable infinity of elements, then its probabilities cannot be determined in terms of the probabilities of elementary events. This is the case if S is the set of points in an n-dimensional space.

Suppose that S is the set of all real numbers. Its subsets can be considered as sets of points on the real line. 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.

2.4 Conditional Probability



New slide

To introduce conditional probability, consider the discussion about proportions in Section 2.1. 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:

$$\Pr\left(A \mid B\right) \approx \frac{n_{AB}}{n_B} = \frac{{}^{n_{AB}}/_{n}}{{}^{n_B}/_{n}} = \frac{\Pr\left(AB\right)}{\Pr\left(B\right)} \tag{2.29}$$

provided that n is sufficiently large.

The **conditional probability** of an event A assuming another event B, denoted by $\Pr\left(A \mid B\right)$, is defined by the ratio:

$$\Pr\left(A \mid B\right) = \frac{\Pr\left(A \cap B\right)}{\Pr\left(A\right)} \tag{2.30}$$

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

Example 2.5 (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.

There are four possibilities for the gender of the children, namely:

$$S = \{GG,\, GB,\, BG,\, BB\} \tag{2.31}$$

where the four possibilities are equally probable:

$$\Pr\left(GG\right) = \Pr\left(GB\right) = \Pr\left(BG\right) = \Pr\left(BB\right) = \frac{1}{4} \tag{2.32}$$

The subset of S which contains the possibilities of one child being a boy is at $S_B = \{GB,\, BG,\, BB\}$, and therefore the conditional probability:

$$\Pr\left(BB \mid S_B\right) = \frac{\Pr\left(BB \cap \left(GB \cup BG \cup BB\right)\right)}{\Pr\left(S_B\right)} \tag{2.33}$$

Note that $\{BB \cap \left(GB \cup BG \cup BB\right)\} = \{BB\}$, and that $\Pr\left(S_B\right) = 1 - \Pr\left(S_B\right) = 1 - \Pr\left(GG\right) = \frac{3}{4}$. Therefore:

$$\Pr\left(BB \mid S_B\right) = \frac{\Pr\left(BB\right)}{1 - \Pr\left(GG\right)} = \frac{{}^{1}/_{4}}{{}^{3}/_{4}} = \frac{1}{3} \tag{2.34}$$

□

Note that the question is completely different if it were *what is the probability that both are boys, given that the youngest child is a boy*, in which case the solution is $^{1}/_{2}$. This is since information has been provided about one of the children, thereby distinguishing between the children.

The example in Example 2.5 might seem a little abstract to signal processing, but there are other ways of phrasing exactly the same problem. Using an example taken from [Therrien:2011], it could be phrased as follows:

> A compact disc (CD) selected from the *bins* at Simon's Surplus are as likely to be good as they are bad. Simon decides to sell these CDs in packages of two, but guarantees that in each package, at least one CD will be good. What is the probability that when you buy a signle package, you get two good CDs?

Example 2.6 (Prisoner's Paradox). Three prisoners, A, B and C, are in separate cells and sentenced to death. The governor has selected one of them at random to be pardoned. The warden knows which one is pardoned, but is not allowed to tell. Prisoner A begs the warden to let him know the identity of one of the *others* who is going to be executed.

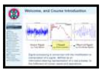
> *If B is to be pardoned, give me C's name, and vice-versa. And if I'm to be pardoned, flip a coin to decide whether to name B or C.*

The warden tells A that B is to be executed. Prisoner A is pleased because he believes that his probability of surviving has gone up from $1/3$ to $1/2$, as it is now between him and C. Prisoner A secretly tells C the news, who is also pleased, because he reasons that A still has a chance of $1/3$ to be the pardoned one, but his chance has gone up to $2/3$. What is the correct answer?

3 Scalar Random Variables

This handout introduces the concept of a random variable, its probabilistic description in terms of pdfs and cdfs, and characteristic features such as mean, variance, and other moments. It covers the probability transformation rule and characteristic functions.

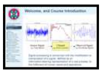
3.1 Abstract



New slide

- Deterministic signals are interesting from an analytical perspective since their *signal value* or *amplitude* are uniquely and completely specified by a functional form, albeit that function might be very complicated. Thus, a deterministic signal is some function of time: $x = x(t)$.
- In practice, this precise description cannot be obtained for real-world signals and, moreover, it can be argued philosophically that real-world signals are not deterministic but, rather, they are inherently random or *stochastic* in nature.
- Although random signals evolve in time stochastically, their average properties are often deterministic, and thus can be specified by an explicit functional form.
- This part of the course looks at the properties of stochastic processes, both in terms of an exact probabilistic description, and also characteristic features such as mean, variance, and other moments.

3.2 Definition Random Variables



New slide

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. This mapping from

ζ to X should satisfy the following two conditions:

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

The second condition states that, although X is allowed to take the values $x = \pm\infty$, the outcomes form a set with zero probability.

KEYPOINT! (Nature of Outcomes). Note that the outcomes of events are not necessarily numbers themselves, although they should be distinct in nature. Hence, examples of outcomes might be:

- outcomes of tossing coins (head/tails); card drawn from a deck (King, Queen, 8-of-Hearts);
- characters or words (A-Z); symbols used in deoxyribonucleic acid (DNA) sequencing (A, T, G, C);
- a numerical result, such as the number rolled on a die.

A more graphical representation of a discrete RV is shown in Figure 3.1. In this model, a physical experiment can lead to a number of possible events representing the outcomes of the experiment. These outcomes may be values, or they may be symbols, or some other representation of the event. Each outcome (or event), ζ_k, has a probability $\Pr(\zeta_k)$ assigned to it. Each outcome ζ_k then a real number assigned to that outcome, x_k. The RV is then defined as the collection of these three values; an outcome index, the probability of the outcome, and the real value assigned to that outcome, thus $X(\zeta) = \{\zeta_k, \Pr(\zeta_k), x_k$.

A more specific example is shown in Figure 3.2 in which the **experiment** is that of rolling a die, the **outcomes** are the colors of the dies, each **event** is simply each **outcome**, and the specific user-defined values assigned are the numbers shown.

Example 3.1 (Rolling die). Consider rolling a die, with six outcomes $\{\zeta_i, i \in \{1, \dots, 6\}\}$. 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:

$$X(\zeta_1) = X(\zeta_3) = X(\zeta_5) = 0 \quad \text{and} \quad X(\zeta_2) = X(\zeta_4) = X(\zeta_6) = 1 \tag{3.1}$$

$\bowtie$

Example 3.2 (Letters of the alphabet). Suppose the outcome of an experiment is a letter A to Z, such that $X(A) = 1$, $X(B) = 2$, ..., $X(Z) = 26$. Then the event $X(\zeta) \leq 5$ corresponds to the letters A, B, C, D, or E.

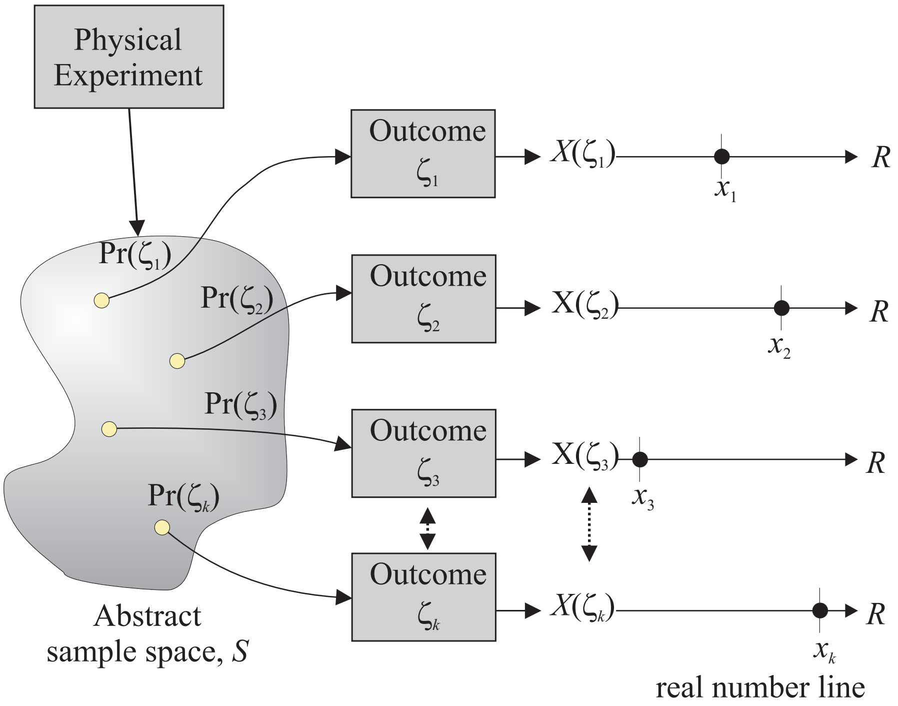


Figure 3.1: A graphical representation of a random variable.

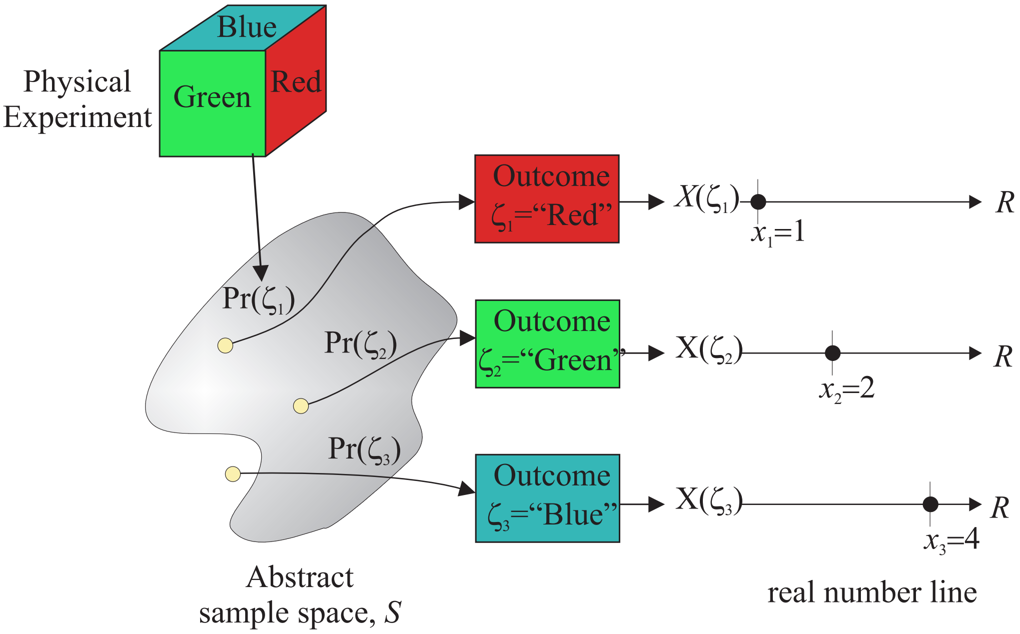
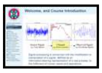


Figure 3.2: A graphical representation of a random variable.

3.2.1 Distribution functions



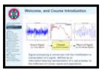
New slide

Random variables are fundamentally characterised by their distribution and density functions. These concepts are considered in this and the next section.

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

$$F_X\left(x\right) \triangleq \Pr\left(X\left(\zeta\right) \leq x\right) \tag{M:3.1.1}$$

3.2.2 Density functions



New slide

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

$$f_X\left(x\right) \triangleq \frac{dF_X\left(x\right)}{dx} \tag{M:3.1.2}$$

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

$$f_X\left(x\right)\,\Delta x \approx \Delta F_X\left(x\right) \triangleq F_X\left(x + \Delta x\right) - F_X\left(x\right) \approx \Pr\left(x < X\left(\zeta\right) \leq x + \Delta x\right) \tag{3.2}$$

  This can be written, more formally, as:

$$f_X\left(x\right) = \lim_{\Delta x \to 0} \frac{F_X(x + \Delta x) - F_X(x)}{\Delta x} \tag{3.3}$$

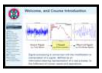
$$= \lim_{\Delta x \to 0} \frac{\Pr\left(x < X\left(\zeta\right) \leq x + \Delta x\right)}{\Delta x} \tag{3.4}$$

- It directly follows that:

$$F_X(x) = \int_{-\infty}^{x} f_X(v)\,dv \tag{M:3.1.4}$$

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

3.2.3 Properties of Distribution and Density Functions



New slide

The following properties are for *continuous* **RVs**. Similar properties follow, *mutatis mutandis*, for discrete **RVs**.

Sidebar 1 Probability of $X(\zeta)$ taking on a specific value

The simplest way to consider why the probability of a RV, $X(\zeta)$, taking on a specific value, x_0, is zero for a continuous RV, but not a discrete one, is to consider the limiting case:

$$\Pr(X(\zeta) = x_0) = \lim_{\Delta x_0 \to 0} \Pr(x_0 - \Delta x_0 \leq X(\zeta) \leq x_0 + \delta x_0) \tag{3.5}$$

which can be expressed in terms of its probability density function (pdf), $f_X(x)$, as:

$$\Pr(X(\zeta) = x_0) = \lim_{\Delta x_0 \to 0} \int_{x_0 - \Delta x_0}^{x_0 + \Delta x_0} f_X(u)\, du \tag{3.6}$$

Suppose that around the region $\mathcal{R} = [x_0 - \Delta x_0,\, x_0 + \Delta x_0]$, the pdf $f_X(x)$ can be expressed as:

$$f_X(x) = p_0\, \delta(x - x_0) \tag{3.7}$$

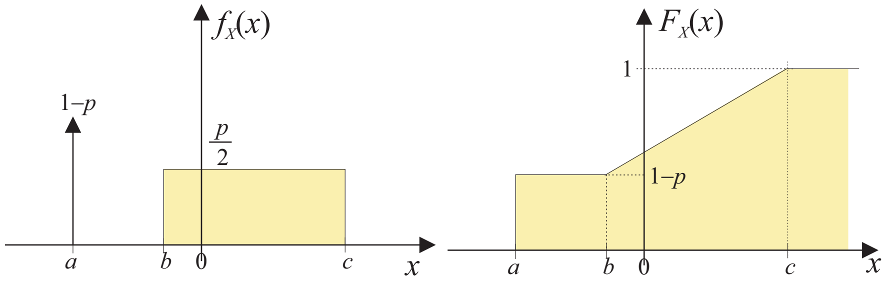
then using the **sifting theorem**, which states that

$$\int_{\mathcal{R}} \phi(t)\; \delta(t - T)\; dt = \begin{cases} \phi(T) & \text{if } T \in \mathcal{R} \\ 0 & \text{otherwise} \end{cases}, \tag{3.8}$$

then it becomes clear that

$$\Pr(X(\zeta) = x_0) = \lim_{\Delta x_0 \to 0} \int_{x_0 - \Delta x_0}^{x_0 + \Delta x_0} p_0\, \delta(x - x_0)\; du = p_0 \tag{3.9}$$

whereas for the continuous time case, the limit in Equation 3.6 tends to zero. In otherwords, only in the case when the pdf of $X(\zeta)$, $f_X(x)$, contains a delta function at a specific value, will the probability of that specific value be non-zero. A delta function in a pdf corresponds to a discrete-component of the RV. An example of a mixture of discrete and continous random variables is shown in Figure 3.3. Note the step function in the cumulative distribution function (cdf).



(a) The pdf. (b) The cdf.

Figure 3.3: A probability density function and its corresponding cumulative distribution function for a RV which is a mixture of continuous and discrete components.

- Properties of **cdf**:

$$0 \leq F_X(x) \leq 1, \quad \lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to \infty} F_X(x) = 1 \tag{M:3.1.6}$$

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

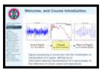
$$F_X(a) \leq F_X(b) \quad \text{if} \quad a \leq b \tag{3.10}$$

- Properties of **pdfs**:

$$f_X(x) \geq 0, \quad \int_{-\infty}^{\infty} f_X(x)\, dx = 1 \tag{M:3.1.7}$$

- Probability of arbitrary events:

$$\Pr(x_1 < X(\zeta) \leq x_2) = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x)\, dx \tag{M:3.1.8}$$



3.2.4 Kolmogorov's Axioms

New slide

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:

$$\Pr(x \leq x_1) + \Pr(x_1 < x \leq x_2) = \Pr(x \leq x_2) \tag{3.11}$$

$$\int_{-\infty}^{x_1} p(v)\, dv + \Pr(x_1 < x \leq x_2) = \int_{-\infty}^{x_2} p(v)\, dv \tag{3.12}$$

$$\Rightarrow \qquad \Pr(x_1 < x \leq x_2) = \int_{x_1}^{x_2} p(v)\, dv \tag{3.13}$$

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.

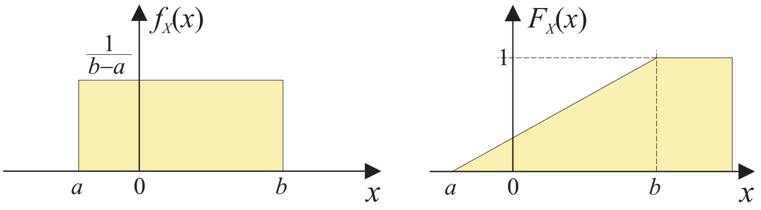
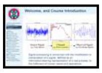


Figure 3.4: The uniform probability density function and cumulative distribution function.

3.3 Examples of Continuous random variables



New slide

Uniform distribution The RV $X(\zeta)$ is *uniform* on $[a, b]$ if it has pdf:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x \leq b, \\ 0 & \text{otherwise} \end{cases} \tag{M:3.1.33}$$

The pdf is plotted in Figure 3.4.

Consequently, the cdf is given by:

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq a, \\ \frac{x-a}{b-a} & \text{if } a < x \leq b, \\ 1 & \text{if } x > b. \end{cases} \tag{M:3.1.34}$$

The cdf is also shown in Figure 3.4. Roughly speaking, X takes on any value between a and b with equal probability. The mean and variance of this random variable are given by, respectively:

$$\mu_X = \frac{a+b}{2} \quad \text{and} \quad \sigma_X^2 = \frac{(b-a)^2}{12} \tag{M:3.1.35}$$

Exponential distribution The RV $X(\zeta)$ is *exponential* with parameter $\lambda > 0$ if it has pdf:

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \lambda e^{-\lambda x} & \text{if } x \geq 0, \end{cases} \tag{3.14}$$

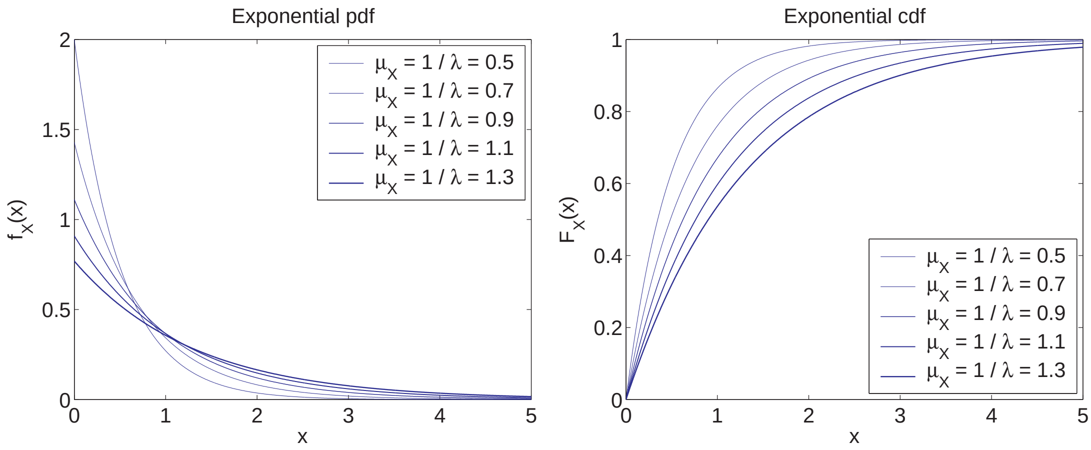
Consequently, the cdf is given by:

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 - e^{-\lambda x} & \text{if } x \geq 0, \end{cases} \tag{3.15}$$

The **exponential distribution** occurs very often in practice as a description of the time elapsing between random events.

The exponential pdf and cdf are shown in Figure 3.5, for various different values of the parameter λ. The mean and variance of this random variable are given by, respectively:

$$\mu_X = \frac{1}{\lambda} \quad \text{and} \quad \sigma_X^2 = \mu_X^2 = \frac{1}{\lambda^2} \tag{3.16}$$



(a) The Exponential pdf.

(b) The Exponential cdf.

Figure 3.5: The exponential density and distribution functions, for various different values of the parameter λ.

Hence, for an exponential distribution, the **mean** and **standard deviation** are identical.

Normal distribution Arguably the most important continuous distribution is the *normal* or **Gaussian distribution**; these terms will be used interchangabally. The pdf of a Gaussian distributed RV, $X(\zeta)$, with mean μ_X and standard deviation σ_X^2, is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right], \quad x \in \mathbb{R} \tag{M:3.1.37}$$

It is common to denote this by:

$$f_X(x) = \mathcal{N}\left(x \,\middle|\, \mu_X, \sigma_X^2\right) \tag{3.17}$$

Note, however, that if $\hat{x}$ is a *sample* of a Gaussian random variable, then it is written:

$$\hat{x} \sim \mathcal{N}\left(\mu_X, \sigma_X^2\right) \tag{3.18}$$

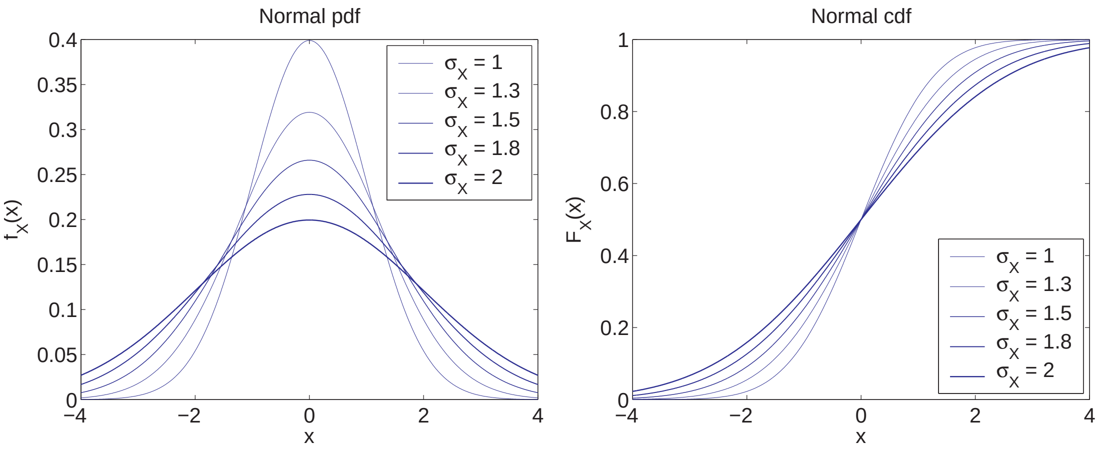
The Gaussian pdf and cdf are shown in Figure 3.6 for a zero-mean RV, and for various variances, σ_X^2.

Gamma distribution The RV $X(\zeta)$ has the **Gamma distribution** with parameters $\alpha > 0$, $\beta > 0$ if it has pdf:

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{1}{\Gamma(\beta)} \alpha^\beta \, x^{\beta-1} \, e^{-\alpha x} & \text{if } x \geq 0, \end{cases} \tag{3.19}$$

where $\Gamma(\beta)$ is the **gamma function** given by:

$$\Gamma(\beta) = \int_0^\infty x^{\beta-1} \, e^{-x} \, dx \tag{3.20}$$



(a) The Normal pdf.

(b) The Normal cdf.

Figure 3.6: The Gaussian density and distribution functions; these plots are for a zero mean normal pdf, and are plotted for various different variances, σ_X^2.

This distribution is often written as $f_X(x) = \mathcal{G}a\left(x \mid \alpha, \beta\right)$. If $\beta = 1$, then X is exponentially distributed with parameter α.

The Gamma pdf and cdf are shown in Figure 3.7, for the case when $\alpha = 1$ and for various values of the parameter β.

Inverse-Gamma distribution The RV $X(\zeta)$ has the **inverse-Gamma distribution** with parameters $\alpha > 0$, $\beta > 0$ is related to a Gamma-distributed RV, say U, through the transformation $X = \frac{1}{U}$. It can be shown using the probability transformation rule that the pdf of X is thus given by:

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{1}{\Gamma(\beta)} \alpha^\beta \, x^{-(\beta+1)} \, e^{-\frac{\alpha}{x}} & \text{if } x \geq 0, \end{cases} \tag{3.21}$$

It is common to denote this by:

$$f_X(x) = \mathcal{IG}\left(x \mid \alpha, \beta\right) \tag{3.22}$$

Note, however, that if $\hat{x}$ is a *sample* of a inverse-gamma distributed variable, then it is written:

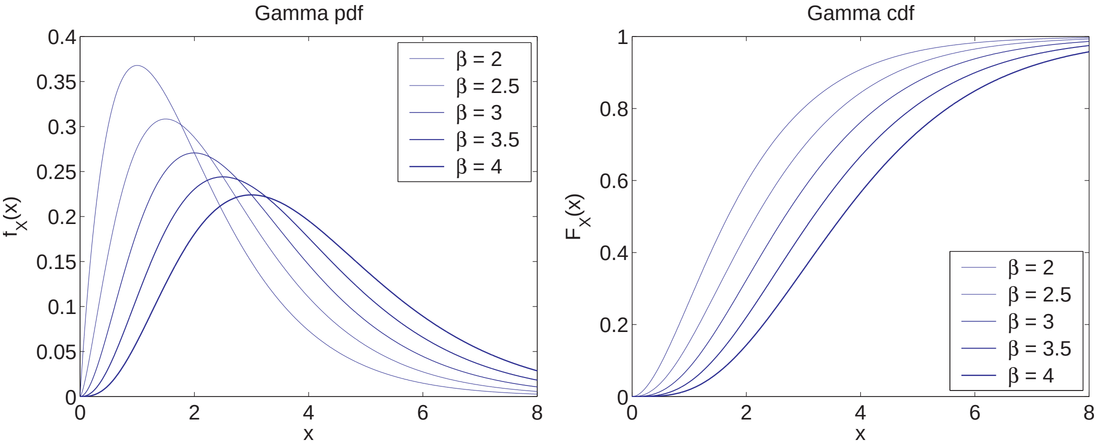
$$\hat{x} \sim \mathcal{IG}(\alpha, \beta) \tag{3.23}$$

Cauchy distribution The RV $X(\zeta)$ has the **Cauchy distribution** with parameters μ_X and β if it has pdf:

$$f_X(x) = \frac{\beta}{\pi} \frac{1}{(x - \mu_X)^2 + \beta^2} \tag{M:3.1.41}$$

The Cauchy random variable has mean μ_X, but its variance does not exist. The corresponding cdf is given by:

$$F_X(x) = \frac{1}{2} + \frac{1}{\pi} \arctan \frac{x - \mu_X}{\beta} \tag{3.24}$$



(a) The Gamma pdf. (b) The Gamma cdf.

Figure 3.7: The Gamma density and distribution functions, for the case when $\alpha = 1$ and for various values of β.

The Cauchy distribution is an appropriate model in which a random variable takes large values with significant probability, and is thus a **heavy-tailed** distribution.

Beta distribution The RV $X(\zeta)$ is *beta*, parameters $a, b > 0$, if it has density function:

$$f_X(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3.25}$$

where the **beta function** is given by

$$B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}\, dx \tag{3.26}$$

If $a = b = 1$, then X is uniform on $[0, 1]$.

Erlang-k distribution The RV $X(\zeta)$ has an **Erlang-k distribution**, with parameters $\gamma > 0$ and $k \in \mathbb{Z}^+$ is a positive integer, if it has density function:

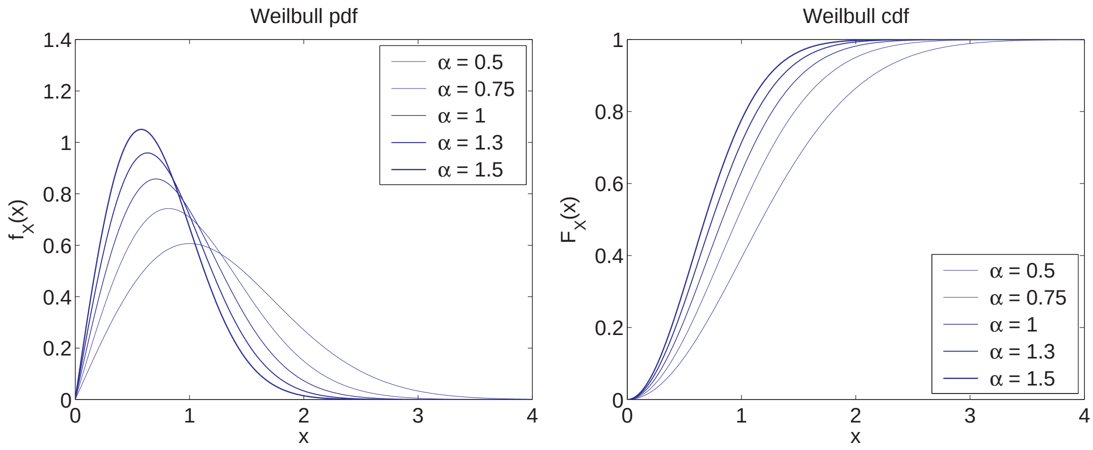
$$f_X(x) = \begin{cases} \frac{\gamma k(\gamma k x)^{k-1}}{(k-1)!} e^{-\gamma k x} & x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.27}$$

The mean and variance of this random variable are given by, respectively:

$$\mu_X = \frac{1}{\gamma} \quad \text{and} \quad \sigma_X^2 = \frac{1}{k\gamma^2} \tag{3.28}$$

Weibull distribution The RV $X(\zeta)$ is *Weibull*, parameters $\alpha, \beta > 0$, if it has density function:

$$f_X(x) = \begin{cases} 0 & x < 0 \\ \alpha\beta x^{\beta-1}\, e^{-\alpha\, x^{\beta}} & x \geq 0 \end{cases} \tag{3.29}$$



(a) The Weibull pdf. (b) The Weibull cdf.

Figure 3.8: The Weibull density and distribution functions, for the case when $\alpha = 1$, and for various values of the parameter β.

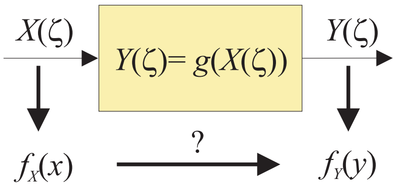


Figure 3.9: The mapping $y = g(x)$.

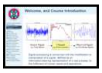
The corresponding the cdf is given by:

$$F_X\left(x\right) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\alpha\, x^{\beta}} & x \geq 0 \end{cases} \tag{3.30}$$

Setting $\beta = 1$ gives the exponential distribution.

The Weibull pdf and cdf are shown in Figure 3.8, for the case when $\alpha = 1$, and for various values of the parameter β.

3.4 Probability transformation rule



New slide

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

This functional relationship is shown diagrammatically in Figure 3.9, and an arbitrary function between $X\left(\zeta\right)$ and $Y\left(\zeta\right)$ is shown in Figure 3.10.

This general question is discussed in detail in, for example, [Papoulis:1991, Chapter 5]. It can be concluded that for $Y(\zeta) = g[X\left(\zeta\right)]$ to be a valid random variable, the function $g(x)$ must have the following properties:

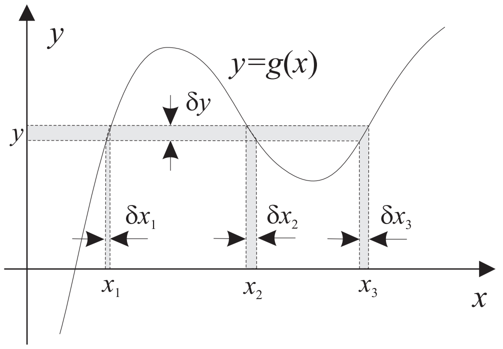


Figure 3.10: The mapping $y = g(x)$, and the effect of the mapping on intervals.

1. Its domain must include the range of the RV $X(\zeta)$.
2. It must be a so-called **Baire function**; that is, for every y, the set $\mathcal{R}_y = \{x : g(x) \le y,\, x \in \mathbb{R}\}$ must consist of the union and intersection of a countable number of intervals. Only then the set $\{Y(\zeta) \le y\}$ is an event.
3. The events $\{g[X(\zeta)] = \pm\infty\}$ must have probability zero.

These properties are usually satisfied.

Consider the set $\mathcal{R} \subset \mathbb{R}$ of the y-axis that is not in the range of the function $g(x)$; that is, $g : \mathbb{R} \nrightarrow \mathbb{R}$. In this case, $\Pr(g[X(\zeta)] \in \mathcal{R}) = 0$. Hence, $f_Y(y) = 0,\, y \in \mathcal{R}$. It suffices, therefore, to consider values of y such that, for some x, $g(x) = y$.

Theorem 3.1 (Probability transformation rule). Denote the real roots of $y = g(x)$ by $\{x_n,\, n \in \mathcal{N}\}$, such that

$$y = g(x_1) = \cdots = g(x_N) \tag{3.31}$$

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

$$f_Y(y) = \sum_{n=1}^{N} \frac{f_X(x_n)}{|g'(x_n)|} \tag{3.32}$$

where $g'(x)$ is the derivative with respect to (w. r. t.) x of $g(x)$.

PROOF. The definition of a **pdf** gives:

$$f_Y(y)\, dy = \Pr(y < Y(\zeta) \le y + dy) \tag{3.33}$$

The set of values x such that $y < g(x) \le y + dy$ consists of the intervals:

$$x_n < x \le x_n + dx_n \tag{3.34}$$

This is shown in Figure 3.10 for the case when there are three solutions to the equation $y = g(x)$. The probability that x lies in this set is, of course,

$$f_X(x_n)\, dx_n = \Pr(x_n < X(\zeta) \le x_n + dx_n) \tag{3.35}$$

and, from the transformation from x to y, then

$$dx_n = \frac{dy}{|g'(x_n)|} \tag{3.36}$$

Since these are mutually exclusive sets, then

$$\Pr\left(y < Y(\zeta) \le y + dy\right) = \sum_{n=1}^{N} \Pr\left(x_n < X\left(\zeta\right) \le x_n + dx_n\right) \tag{3.37}$$

$$= \sum_{n=1}^{N} f_X\left(x_n\right) \frac{dy}{|g'(x_n)|} \tag{3.38}$$

□

and thus the desired result is obtained after minor rearrangement.

Example 3.3 (Log-normal distribution). Let $Y = e^X$, where $X \sim \mathcal{N}\left(0,\, 1\right)$. Find the pdf for the RV Y.

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

$$f_X\left(x\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{3.39}$$

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:

$$f_Y\left(y\right) = \frac{f_X\left(x\right)}{g'(x)} = \frac{1}{y\sqrt{2\pi}} e^{-\frac{(\ln y)^2}{2}} \tag{3.40}$$

□

This distribution is known as the log-normal distribution. It is important for cases where the random variable X might describe the amplitude of a signal in decibels, and where Y is the actual amplitude.

Example 3.4 (Inverse of a random variable). Let $Y = \frac{1}{X}$. Find the pdf for the RV Y, given by $f_Y\left(y\right)$, in terms of the pdf for the RV X, given by $f_X\left(x\right)$. Further, consider the special case when X has a **Cauchy density** with parameter α, such that:

$$f_X\left(x\right) = \frac{\alpha}{\pi} \frac{1}{x^2 + \alpha^2} \tag{3.41}$$

SOLUTION. There is a single solution to the equation $y = \frac{1}{x}$, given by $x = \frac{1}{y}$. Hence, $|g'(x)| = \frac{1}{x^2} = y^2$, and:

$$f_Y\left(y\right) = \frac{1}{y^2} f_X\left(\frac{1}{y}\right) \tag{3.42}$$

In the special case of a **Cauchy density**,

$$f_X\left(x\right) = \frac{\alpha}{\pi} \frac{1}{x^2 + \alpha^2} \tag{3.43}$$

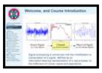
such that:

$$f_Y(y) = \frac{1}{y^2} f_X\left(\frac{1}{y}\right) = \frac{1}{y^2}\frac{\alpha}{\pi}\frac{1}{\frac{1}{y^2}+\alpha^2} \tag{3.44}$$

$$= \frac{1/\alpha}{\pi}\frac{1}{y^2+\frac{1}{\alpha^2}} \tag{3.45}$$

□

which is also a **Cauchy density** with parameter $\frac{1}{\alpha}$.



3.5 Expectations

New slide

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:

$$\mathbb{E}\left[X(\zeta)\right] = \int_{\mathbb{R}} x\, f_X(x)\, dx \tag{3.46}$$

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

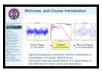
$$f_X(x) = \sum_k p_k\, \delta(x - x_k) \tag{3.47}$$

  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:

$$\mu_x = \int_{\mathbb{R}} x\, f_X(x)\, dx = \int_{\mathbb{R}} x \sum_k p_k\, \delta(x - x_k)\, dx = \sum_k x_k\, p_k \tag{3.48}$$

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



3.5.1 Properties of expectation operator

New slide

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:

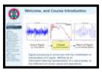
$$\mathbb{E}\left[\alpha\, X\left(\zeta\right) + \beta\right] = \alpha\, \mu_X + \beta \tag{M:3.1.10}$$

- 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:

$$\mathbb{E}\left[Y(\zeta)\right] \triangleq \mathbb{E}\left[g\{X\left(\zeta\right)\}\right] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx \tag{M:3.1.11}$$

  This important property is known as the **invariance of the expectation operator, and is extremely important**. .

3.6 Moments



New slide

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

$$\mathbb{E}\left[X\left(\zeta\right)\right] = \mu_X = \int_{\mathbb{R}} x\, f_X(x)\, dx \tag{3.53}$$

$$\operatorname{var}\left[X\left(\zeta\right)\right] = \sigma_X^2 = \int_{\mathbb{R}} x^2\, f_X(x)\, dx - \mu_X^2 = \mathbb{E}\left[X^2(\zeta)\right] - \mathbb{E}^2\left[X\left(\zeta\right)\right] \tag{3.54}$$

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:

$$r_X^{(m)} \triangleq \mathbb{E}\left[X^m(\zeta)\right] = \int_{\mathbb{R}} x^m\, f_X(x)\, dx \tag{M:3.1.12}$$

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

Example 3.5 (Expectations of non-negative RVs). Let $X\left(\zeta\right)$ be a non-negative RV with pdf $f_X\left(x\right)$. Show that

$$\mathbb{E}\left[X^r(\zeta)\right] = \int_0^{\infty} r\, x^{r-1} \Pr\left(X\left(\zeta\right) > x\right)\, dx \tag{3.55}$$

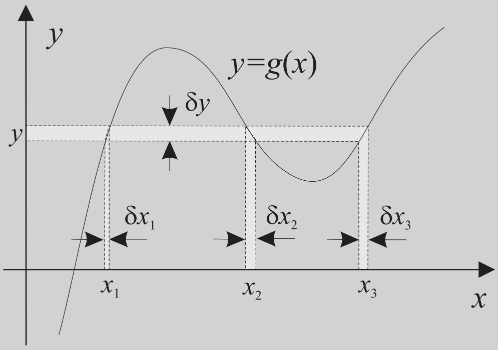
for any $r \geq 1$ for which the expectation is finite.

SOLUTION. In this case, since the question says to *show that*, it is sufficient to manipulate the right hand side (RHS). This proceeds as follows: notice,

$$\int_0^{\infty} r\, x^{r-1} \Pr\left(X\left(\zeta\right) > x\right)\, dx = \int_0^{\infty} r\, x^{r-1} \left\{\int_{y=x}^{\infty} f_X\left(y\right)\, dy\right\} dx \tag{3.56}$$

Sidebar 2 Invariance of Expectation

The invariance of the expectation operator is an extremely important property, and makes statistical analysis of transformed random variables much simpler. It can be explained using similar techniques to those used in deriving the probability transformation rule in Theorem 3.1.



Consider again Figure 3.10 on page 52, which is reproduced above. Let $Y(\zeta) = g\left(X\left(\zeta\right)\right)$. Consider first the approximation for the expectation of $Y(\zeta)$:

$$\mathbb{E}\left[Y\right] = \int_{-\infty}^{\infty} y\, f_Y\left(y\right)\, dy \approx \sum_{\forall k} y_k\, f_Y\left(y_k\right) \delta y \tag{3.49}$$

where $f_Y\left(y_k\right) \delta y = \Pr\left(y_k < Y(\zeta) \leq y_k + \delta y\right)$ is the probability that $Y(\zeta)$ is in the small interval $y_k < Y(\zeta) \leq y_k + \delta y$. This probability, as in Theorem 3.1, can be written as the sum of the probabilities that $X\left(\zeta\right)$ is each of the corresponding small intervals shown in Figure 3.10 above, such that:

$$f_Y\left(y_k\right) \delta y = \sum_{n=1}^{N} \Pr\left(x_{k,n} < X\left(\zeta\right) \leq x_{k,n} + \delta x_{k,n}\right) = \sum_{n=1}^{N} f_X\left(x_{k,n}\right) \delta x_{k,n} \tag{3.50}$$

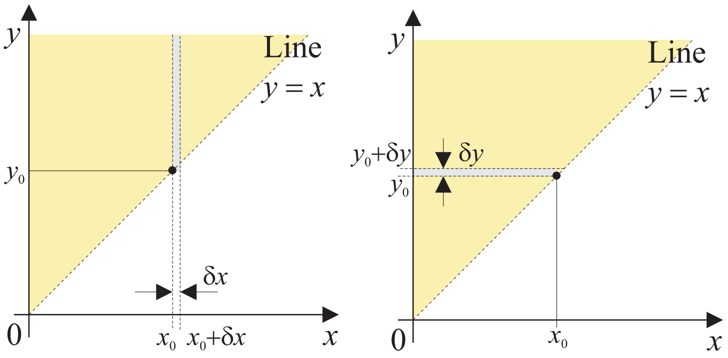
Substituting Equation 3.50 into Equation 3.52 gives:

$$\mathbb{E}\left[Y\right] \approx \sum_{\forall k} y_k \sum_{n=1}^{N} f_X\left(x_{k,n}\right) \delta x_{k,n} = \sum_{\forall k} \sum_{n=1}^{N} g\left(x_{k,n}\right) f_X\left(x_{k,n}\right) \delta x_{k,n} \tag{3.51}$$

Since the double summation merely covers all possible regions of x, this can be reindexed as

$$\mathbb{E}\left[Y\right] \approx \sum_{\forall \ell} g\left(x_\ell\right) f_X\left(x_\ell\right) \delta x_\ell \tag{3.52}$$

which in the limit gives the integral Equation M:3.1.11, page 55. So, in summary, to compute the expectation of $Y(\zeta) = g\left(X\left(\zeta\right)\right)$, it is not necessary to transform and find the pdf of $f_Y\left(y\right)$, but simply use this invariance of expectation property.



(a) Integration w. r. t. x first, and then w. r. t. y.

(b) Integration w. r. t. y first, and then w. r. t. x.

Figure 3.11: The region of integration for the integral in Equation 3.56.

and rearrange the order of integration, noting the region of integration as shown in Figure 3.11, and thus the change in the limits:

$$= \int_0^\infty f_X(y) \left\{ \int_{x=0}^{y} r\, x^{r-1}\, dx \right\} dy \tag{3.57}$$

$$= \int_0^\infty f_X(y) \left[x^r\right]_0^y dy = \int_0^\infty y^r f_X(y)\, dy = \mathbb{E}\left[X^r(\zeta)\right] \tag{3.58}$$

□

3.6.1 Central Moments

Central moments of $X(\zeta)$ can also be defined: the m-th **central moment** of $X(\zeta)$ is given by:

$$\gamma_X^{(m)} \triangleq \mathbb{E}\left[(X(\zeta) - \mu_X)^m\right] = \int_{\mathbb{R}} (x - \mu_X)^m f_X(x)\, dx \tag{M:3.1.14}$$

Some obvious properties that follow from these definitions are:

- The variance of $X(\zeta)$ can be defined as:

$$\text{var}\left[X(\zeta)\right] \triangleq \sigma_X^2 \triangleq \gamma_X^{(2)} = \mathbb{E}\left[(X(\zeta) - \mu_X)^2\right] \tag{3.59}$$

- **Standard deviation** is given by: $\sigma_X = \sqrt{\text{var}\left[X(\zeta)\right]}$.

- Trivial **moments**: $r_X^{(0)} = 1$ and $r_X^{(1)} = \mu_X$.

- Trivial **central moments**: $\gamma_X^{(0)} = 1$, $\gamma_X^{(1)} = 0$, and $\gamma_X^{(2)} = \sigma_X^2$.

3.6.2 Relationship between Moments and Central Moments

Moments and **central moments** are related by the expressions:

$$\gamma_X^{(m)} = \sum_{k=0}^{m} \binom{m}{k} (-1)^k \, \mu_X^k \, r_X^{(m-k)} \tag{M:3.1.16}$$

$$r_X^{(m)} = \sum_{k=0}^{m} \binom{m}{k} \mu_X^k \, \gamma_X^{(m-k)} \tag{3.60}$$

where the general combinatorial term ${}^n C_r = \binom{n}{r}$ is given by

$${}^n C_r = \frac{n!}{r!\,(n-r)!} \tag{3.61}$$

In particular, second-order moments are related as follows:

$$\sigma_X^2 = r_X^{(2)} - \mu_X^2 = \mathbb{E}\left[X^2(\zeta)\right] - \mathbb{E}^2\left[X(\zeta)\right] \tag{M:3.1.17}$$

PROOF. These results are proved by expanding the term $(x - \mu_x)^m$ in the expression for central-moments using the binomial expansion.

Thus, recalling that

$$\gamma_X^{(m)} = \int_{\mathbb{R}} (x - \mu_X)^m \, f_X(x) \, dx \tag{M:3.1.14}$$

then using the binomial:

$$(x+a)^n = \sum_{k=0}^{n} \binom{n}{k} x^k \, a^{n-k} = \sum_{k=0}^{n} \binom{n}{k} a^k \, x^{n-k} \tag{3.62}$$

it follows:

$$\gamma_X^{(m)} = \int_{\mathbb{R}} \sum_{k=0}^{m} \binom{m}{k} x^{m-k} \, (-\mu_X)^k \, f_X(x) \, dx \tag{3.63}$$

$$= \sum_{k=0}^{m} \binom{m}{k} (-1)^k \, \mu_X^k \underbrace{\int_{\mathbb{R}} x^{m-k} \; f_X(x) \, dx}_{r_X^{(m-k)}} \tag{3.64}$$

as required. Similarly, note that

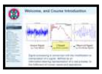
$$r_X^{(m)} = \int_{\mathbb{R}} \left[(x - \mu_X) + \mu_X\right]^m \; f_X(x) \, dx \tag{M:3.1.12}$$

$$= \int_{\mathbb{R}} \sum_{k=0}^{m} \binom{m}{k} \mu_X^k \, (x - \mu_X)^{m-k} \, f_X(x) \, dx \tag{3.65}$$

$$= \sum_{k=0}^{m} \binom{m}{k} \mu_X^k \underbrace{\int_{\mathbb{R}} (x - \mu_X)^{m-k} \, f_X(x) \, dx}_{\gamma_X^{(m-k)}} \tag{3.66}$$

□

giving the desired result.



New slide

3.6.3 Characteristic Functions

The Fourier and Laplace transforms find many uses in probability theory through the concepts of **characteristic functions** and **moment generating functions**.

The **characteristic function** of a rv $X(\zeta)$ is defined by the integral:

$$\Phi_X(\xi) \triangleq \mathbb{E}\left[e^{j\xi\,X(\zeta)}\right] = \int_{-\infty}^{\infty} f_X(x)\, e^{j\xi x}\, dx \tag{M:3.1.21}$$

This can be interpreted as the Fourier transform of $f_X(x)$ with a sign reversal in the complex exponent. To avoid confusion with the pdf, $F_X(x)$ is not used to denote this Fourier transform.

When $j\xi$ is replaced by a complex variable s, the **moment generating function** is obtained, as defined by:

$$\bar{\Phi}_X(s) \triangleq \mathbb{E}\left[e^{sX(\zeta)}\right] = \int_{-\infty}^{\infty} f_X(x)\, e^{sx}\, dx \tag{M:3.1.22}$$

which can be interpreted as the Laplace transform of $f_X(x)$ with a sign reversal in the complex exponent.

Using a series expansion for $e^{sX(\zeta)}$ gives: [1]

$$\bar{\Phi}_X(s) = \mathbb{E}\left[e^{sX(\zeta)}\right] = \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{(sX(\zeta))^n}{n!}\right] \tag{3.68}$$

$$= \sum_{n=0}^{\infty} \frac{s^n}{n!}\,\mathbb{E}\left[X^n(\zeta)\right] \tag{3.69}$$

and noting that $\mathbb{E}\left[X^n(\zeta)\right] = r_X^{(m)}$, this gives:

$$\bar{\Phi}_X(s) = \sum_{n=0}^{\infty} \frac{s^n}{n!}\, r_X^{(n)} \tag{M:3.1.23}$$

provided that every moment $r_X^{(m)}$ exists. Thus, if all moments of $X(\zeta)$ are known and exist, then $\bar{\Phi}_X(s)$ can be assembled, and upon inverse Laplace transformation, the pdf $f_X(x)$ can be determined.

Differentiating $\bar{\Phi}_X(s)$ m-times w. r. t. s, provides the mth-order moment of the RV $X(\zeta)$:

$$r_X^{(m)} = \left.\frac{d^m \bar{\Phi}_X(s)}{ds^m}\right|_{s=0} = (-j)^m \left.\frac{d^m \Phi_X(\xi)}{d\xi^m}\right|_{\xi=0}, \quad m \in \mathbb{Z}^+ \tag{M:3.1.24}$$

[1] It is better if you can work through some of these results for yourself without always having to check every minor step, but just in case you've forgotten, the power series expansion for the exponential function is given by:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \tag{3.67}$$

Theorem 3.2 (Characteristic Functions). The characteristic function $\Phi_X(\xi)$ satisfies:

1. $|\Phi_X(\xi)| \leq \Phi_X(0) = 1$ for all ξ.
2. $\Phi_X(\xi)$ is uniformly continuous on the real axis: $\mathbb{R}$.
3. $\Phi_X(\xi)$ is nonnegative definite, which is to say that:

$$\sum_j \sum_k \Phi_X(\xi_j - \xi_k)\, z_j\, z_k^* \geq 0 \tag{3.70}$$

for all real ξ_i and complex z_i.

PROOF. 1. Clearly, $\Phi_X(0) = \mathbb{E}[1] = 1$. Furthermore, using the Schwartz inequality:

$$\Phi_X(\xi)| \leq \int f_X(x)\, |e^{j\xi x}|\, dx = \int f_X(x)\, dx = 1 \tag{3.71}$$

as required.

2. This is quite a technical property, but for completeness is proved here. Consider:

$$|\Phi_X(\xi + \delta\xi) - \Phi_X(\xi)| = \left|\mathbb{E}\left[e^{j(\xi+\delta\xi)X(\zeta)} - e^{j\xi X(\zeta)}\right]\right| \tag{3.72}$$

using the linearity property of the expectation operator. Using Schwartz's inequality again, where it can be deduced that $|\mathbb{E}[\cdot]| \leq \mathbb{E}[|\cdot|]$, then:

$$|\Phi_X(\xi + \delta\xi) - \Phi_X(\xi)| \leq \mathbb{E}\left[\left|e^{j(\xi+\delta\xi)X(\zeta)} - e^{j\xi X(\zeta)}\right|\right] \tag{3.73}$$

$$\leq \mathbb{E}\left[\left|e^{j\xi X(\zeta)}\left(e^{j\delta\xi X(\zeta)} - 1\right)\right|\right] \tag{3.74}$$

$$\leq \mathbb{E}\left[\left|e^{j\delta\xi X(\zeta)} - 1\right|\right] \tag{3.75}$$

Clearly, the quantity $\left|e^{j\delta\xi X(\zeta)} - 1\right| \to 0$ as $\delta\xi \to 0$, and thus

$$|\Phi_X(\xi + \delta\xi) - \Phi_X(\xi)| \to 0 \quad \text{as } \delta\xi \to 0 \tag{3.76}$$

and therefore $\Phi_X(\xi)$ is uniformally continuous.

3. Finally,

$$\sum_p \sum_q \Phi_X(\xi_p - \xi_q)\, z_p\, z_q^* = \sum_p \sum_q z_p\, z_q^* \int f_X(x)\, e^{j(\xi_p - \xi_q)x}\, dx \tag{3.77}$$

$$= \int f_X(x) \left\{ \sum_p \sum_q z_p e^{j\xi_p x}\, z_q^* e^{-j\xi_q x} \right\} dx \tag{3.78}$$

$$= \int f_X(x) \left| \sum_p z_p e^{j\xi_p x} \right|^2 dx = \mathbb{E}\left[\left| \sum_p z_p e^{j\xi_p x} \right|^2 \right] \geq 0 \tag{3.79}$$

□

Example 3.6 ([Manolakis:2000, Exercise 3.6, Page 144]). Using the **moment generating function**, show that the linear transformation of a Gaussian RV is also Gaussian.

SOLUTION. To answer this question, proceed as follows:

1. Find the moment generating function of a Gaussian RV;
2. Write down $Y(\zeta) = aX(\zeta) + b$, such that:
$$\bar{\Phi}_Y(s) \triangleq \mathbb{E}\left[e^{sY(\zeta)}\right] = \mathbb{E}\left[e^{s(aX(\zeta)+b)}\right] \equiv e^{sb}\mathbb{E}\left[e^{asX(\zeta)}\right] = e^{sb}\bar{\Phi}_X(s\,a) \tag{3.80}$$
where the linearity of the expectation operator has been used.
3. Check to see what distribution this new moment generating function corresponds to.

Thus, start by noting that a **Gaussian random variable** has pdf given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right], \quad x \in \mathbb{R} \tag{M:3.1.37}$$

and the **moment generating function** is given by:

$$\bar{\Phi}_X(s) \triangleq \mathbb{E}\left[e^{sX(\zeta)}\right] = \int_{-\infty}^{\infty} f_X(x)\, e^{sx}\, dx \tag{M:3.1.22}$$

Substituting one into the other gives

$$\begin{aligned}
\bar{\Phi}_X(s) &= \frac{1}{\sqrt{2\pi\sigma_X^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right] e^{sx}\, dx && (3.81)\\
&= \frac{1}{\sqrt{2\pi\sigma_X^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2 - 2(\mu_X + \sigma_X^2 s)x + \mu_X^2}{2\sigma_X^2}\right] dx && (3.82)
\end{aligned}$$

which, by completing the square, can be written as:

$$\bar{\Phi}_X(s) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{(x - \{\mu_X + \sigma_X^2 s\})^2 - (2\mu_X\sigma_X^2 s + \{\sigma_X^2 s\}^2)}{2\sigma_X^2}\right] dx \tag{3.83}$$

$$\bar{\Phi}_X(s) = \exp\left[\mu_X s + \frac{1}{2}\sigma_X^2 s^2\right] \underbrace{\frac{1}{\sqrt{2\pi\sigma_X^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{(x - \{\mu_X + \sigma_X^2 s\})^2}{2\sigma_X^2}\right] dx}_{=1} \tag{3.84}$$

Thus gives the moment generating function for a Gaussian RV as:

$$\bar{\Phi}_X(s) = \exp\left[\mu_X s + \frac{1}{2}\sigma_X^2 s^2\right] \tag{3.85}$$

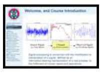
Hence, the moment generating function for the RV $Y(\zeta) = aX(\zeta) + b$ is given by:

$$\bar{\Phi}_Y(s) = e^{sb}\bar{\Phi}_X(s\,a) = e^{sb}\exp\left[a\mu_X s + \frac{1}{2}\sigma_X^2 a^2 s^2\right] \tag{3.86}$$

$$= \exp\left[(a\mu_X + b)s + \frac{1}{2}(\sigma_X^2 a^2)s^2\right] = \exp\left[\mu_Y s + \frac{1}{2}\sigma_Y^2 s^2\right] \tag{3.87}$$

□

where $\mu_Y = a\mu_X + b$ and $\sigma_Y = a\sigma_X$. Thus, the form of the moment generating function for $Y(\zeta)$ is the same as that for a Gaussian RV, and therefore is a Gaussian RV.



3.6.4 Higher-order statistics

New slide

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 about its mean. It is defined as a normalised third-order central moment:

$$\tilde{\kappa}_X^{(3)} \triangleq \mathbb{E}\left[\left\{\frac{X(\zeta) - \mu_X}{\sigma_X}\right\}^3\right] = \frac{1}{\sigma_X^3}\gamma_X^{(3)} \tag{M:3.1.18}$$

and is a *dimensionless* quantity. The **skewness** is:

$$\tilde{\kappa}_X^{(3)} = \begin{cases} < 0 & \text{if the density leans towards the left} \\ 0 & \text{if the density is symmetric about } \mu_X \\ > 0 & \text{if the density leans towards the right} \end{cases} \tag{3.88}$$

In otherwords, if the left side or *left tail* of the distribution is more pronounced than the *right tail*, the function is said to have negative skewness (and leans to the left). If the reverse is true, it has positive skewness (and leans to the right). If the two are equal, it has zero skewness.

Kurtosis measures relative flatness or *peakedness* of a distribution about its mean value. It is defined based on a normalised fourth-central moment:

$$\tilde{\kappa}_X^{(4)} \triangleq \mathbb{E}\left[\left\{\frac{X(\zeta) - \mu_X}{\sigma_X}\right\}^4\right] - 3 = \frac{1}{\sigma_X^4}\gamma_X^{(4)} - 3 \tag{M:3.1.19}$$

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. For this reason, this measure is some times known as **kurtosis excess**, with **kurtosis proper** having the same definition but without the offset of 3.

3.6.5 Cumulants

Cumulants are statistical descriptors that are similar to moments, but provide better information for higher-order moment analysis. Cumulants are derived by considering the **moment generating function**'s natural logarithm. This logarithm is commonly referred to as the **cumulant generating function**. This is given by:

$$\bar{\Psi}_X(s) \triangleq \ln \bar{\Phi}_X(s) = \ln \mathbb{E}\left[e^{sX(\zeta)}\right] \tag{M:3.1.26}$$

When s is replaced by $j\xi$, the resulting function is known as the **second characteristic function**, and is denoted by $\Psi_X(\xi)$.

The **cumulants**, $\kappa_X^{(m)}$, of a RV, $X(\zeta)$, are defined as the derivatives of the **cumulant generating function**; that is:

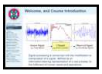
$$\kappa_X^{(m)} \triangleq \left.\frac{d^m \bar{\Psi}_X(s)}{ds^m}\right|_{s=0} = (-j)^m \left.\frac{d^m \Psi_X(\xi)}{d\xi^m}\right|_{\xi=0}, \quad m \in \mathbb{Z}^+ \tag{M:3.1.27}$$

The logarithmic function in the definition of the **cumulant generating function** is useful for dealing with products of characteristic functions, which occurs when dealing with sums of independent RVs.

4
Random Vectors and Multiple Random Variables

This handout extends the concept of a random variable to groups of random variables known as a random vector. The notion of joint, marginal, and conditional probability density functions is introduced. Statistical descriptors of joint random variables is discussed including the notion of correlation. The probability transformation rule and characteristic function is extended to random vectors, and the multivariate Gaussian distribution studied.

4.1 Abstract



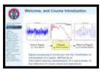
New slide

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*. In other words, the output of a single experiment might be a series of related random variables; for example, biomedical signal analysis, where multiple readings are taken simultaneously.
- Random vectors also lead to the notion of the relationship between the random elements.For example, an experiment might yield multiple outputs that are related somehow. In biomedical Engineering, it might be that electroencephalogram (EEG) signals obtained by taking measurements from

various different positions on the human body are related due to electrical conductance through the body between sensors.

- This course mainly deals with real-valued random vectors, although the concept can be extended to complex-valued random vectors. Details of how to deal with complex-valued random vectors will be discussed in these lecture-notes where they are appropriate and useful, but not specifically as a separate topic. Note that the case of a complex-valued RV, $X(\zeta) = X_R(\zeta) + j\, X_I(\zeta)$ can be considered as a group of $X_R(\zeta)$ and $X_I(\zeta)$, where these are both real-valued RVs.



4.2 Definition of Random Vectors

New slide

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

$$\mathbf{X}(\zeta) = \begin{bmatrix} X_1(\zeta) & X_2(\zeta) & \cdots & X_N(\zeta) \end{bmatrix}^T \tag{M:3.2.1}$$

Hence, the *elements* or *components* of $\mathbf{X}(\zeta)$ are real-valued RVs. The complex-valued RV $X(\zeta) = X_R(\zeta) + j\, X_I(\zeta)$ where $X_R(\zeta)$ and $X_I(\zeta)$ are real-valued RVs can be expressed as the following complex-valued random vector:

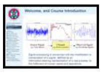
$$\mathbf{X}(\zeta) = \begin{bmatrix} X_R(\zeta) \\ X_I(\zeta) \end{bmatrix} \tag{4.1}$$

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

Denote a specific value for a random vector as:

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix}^T \tag{4.2}$$

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



4.2.1 Distribution and Density Functions

New slide

As with random variables, a random vector is completely characterised by its cumulative distribution function (cdf) and probability density function (pdf). These are direct generalisations of the case for a RV, and most of the time involve converting a single integral or summation to a multiple integral or summation.

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

$$F_{\mathbf{X}}(\mathbf{x}) \triangleq \Pr\left(\{X_n(\zeta) \le x_n,\, n \in \mathcal{N}\}\right) = \Pr\left(\mathbf{X}(\zeta) \le \mathbf{x}\right) \tag{M:3.2.2}$$

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

$$f_{\mathbf{X}}(\mathbf{x}) = \lim_{\Delta\mathbf{x} \to \mathbf{0}} \frac{\Pr\left(\{x_n < X_n(\zeta) \le x_n + \Delta x_n,\, n \in \mathcal{N}\}\right)}{\Delta x_1 \cdots \Delta x_N} \tag{M:3.2.4}$$

$$= \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} \cdots \frac{\partial}{\partial x_N} F_{\mathbf{X}}(\mathbf{x}) \tag{4.3}$$

where $\Delta\mathbf{x} = \Delta x_1 \Delta x_2 \cdots \Delta x_N$, and $\Delta\mathbf{x} \to \mathbf{0} \triangleq \{\Delta_n \to 0,\, n \in \mathcal{N}\}$. The joint pdf must be multiplied by a certain N-dimensional region $\Delta\mathbf{x}$ to obtain a probability.

Hence, it follows:

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_N} f_{\mathbf{X}}(\mathbf{v})\, dv_N \cdots dv_1 = \int_{-\infty}^{\mathbf{x}} f_{\mathbf{X}}(\mathbf{v})\, d\mathbf{v} \tag{M:3.2.6}$$

As with scalar RVs, the distribution and density functions satisfy the following conditions:

- Properties of **joint-cdf**:

$$0 \le F_{\mathbf{X}}(\mathbf{x}) \le 1, \qquad \lim_{\mathbf{x} \to -\infty} F_{\mathbf{X}}(\mathbf{x}) = 0, \qquad \lim_{\mathbf{x} \to \infty} F_{\mathbf{X}}(\mathbf{x}) = 1 \tag{4.4}$$

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

$$F_{\mathbf{X}}(\mathbf{a}) \le F_{\mathbf{X}}(\mathbf{b}) \quad \text{if} \quad \mathbf{a} \le \mathbf{b} \tag{4.5}$$

  Finally, a valid joint-cdf must have a valid corresponding joint-pdf; it is possible to find a function of multiple parameters which satisfies the properties required of a joint-cdf, but the partial differentials of the cdf do not form a valid joint-pdf. An example is given in the tutorial questions.

- Properties of **joint-pdfs**:

$$f_{\mathbf{X}}(\mathbf{x}) \ge 0, \qquad \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} = 1 \tag{4.6}$$

  Similarly, a valid pdf must have a corresponding valid cdf – although this is virtually always the case for functions that satisfy the properties in Equation 4.6.

- Probability of arbitrary events; note that in general the following relationship is not true!

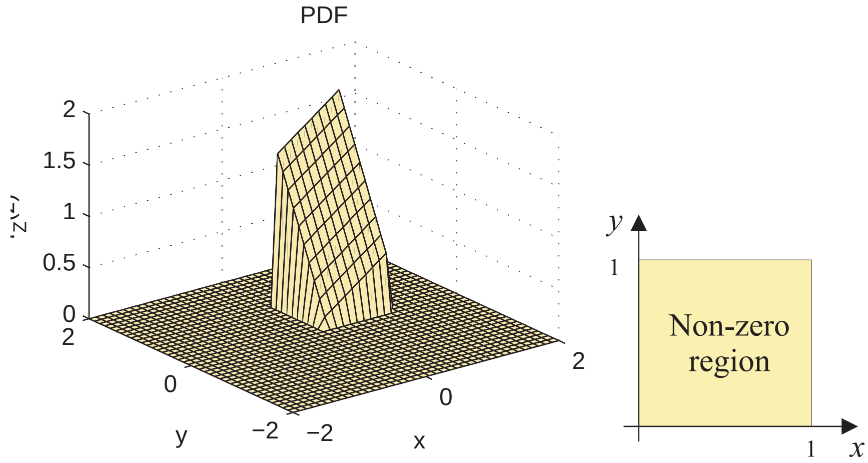
$$\Pr(\mathbf{x}_1 < \mathbf{X}(\zeta) \le \mathbf{x}_2) \ne F_{\mathbf{X}}(\mathbf{x}_2) - F_{\mathbf{X}}(\mathbf{x}_1) = \int_{\mathbf{x}_1}^{\mathbf{x}_2} f_{\mathbf{X}}(\mathbf{v})\, d\mathbf{v} \tag{4.7}$$

  There is an exercise in the tutorial questions that will show you the true relationship for two RVs.

Example 4.1 ([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:

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x + 3y) & 0 \le x,\, y \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

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



(a) A plot of the pdf.

(b) Region of integration.

Figure 4.1: A plot of the probability density function, $f_{\mathbf{Z}}(\mathbf{z})$, for the problem in [Therrien:1992, Example 2.1, Page 20], and a figure showing the region over which the pdf is non-zero, which is the region of integration for calculating the cdf.

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

$$\int_{-\infty}^{\infty} f_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z} = \int_0^1 \int_0^1 \frac{1}{2}(x+3y)\, dx\, dy = \int_0^1 \frac{1}{2}\left[\frac{1}{2}x^2 + 3xy\right]_0^1 dy \tag{4.9}$$

$$= \int_0^1 \frac{1}{4} + \frac{3}{2}y\, dy = \left[\frac{y}{4} + \frac{3y^2}{4}\right]_0^1 = \frac{1}{4} + \frac{3}{4} = 1 \tag{4.10}$$

The pdf and the region over which it is non-zero is shown in Figure 4.1.

The cumulative distribution function is obtained by integrating over both x and y, observing the limits of integration.

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

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

$$F_{\mathbf{Z}}(\mathbf{z}) = \int_{-\infty}^{\mathbf{z}} f_{\mathbf{Z}}(\bar{\mathbf{z}})\, d\bar{\mathbf{z}} = \int_0^y \int_0^x \frac{1}{2}(\bar{x} + 3\bar{y})\, d\bar{x}\, d\bar{y} \tag{4.11}$$

$$= \int_0^y \frac{1}{2}\left(\frac{x^2}{2} + 3x\bar{y}\right) d\bar{y} = \frac{1}{2}\left(\frac{x^2}{2}y + \frac{3xy^2}{2}\right) = \frac{xy}{4}(x+3y) \tag{4.12}$$

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

Hence, in summary, it follows:

$$F_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} 0 & x \leq 0 \quad \text{or} \quad y \leq 0 \\ \frac{xy}{4}(x+3y) & 0 < x,\, y \leq 1 \\ \frac{x}{4}(x+3) & 0 < x \leq 1,\, 1 < y \\ \frac{y}{4}(1+3y) & 0 < y \leq 1,\, 1 < x \\ 1 & 1 < x,\, y < \infty \end{cases} \tag{4.13}$$

□

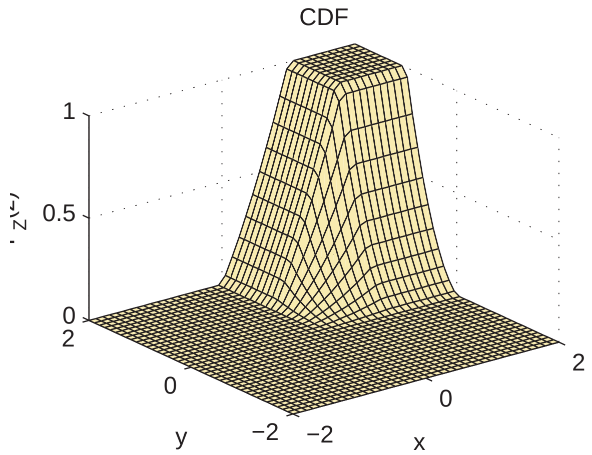
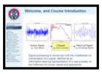


Figure 4.2: A plot of the cumulative distribution function, $F_{\mathbf{Z}}(\mathbf{z})$, for the problem in [Therrien:1992, Example 2.1, Page 20].

The cdf is plotted in Figure 4.2.

4.2.2 Marginal Density Function



New slide

Random vectors lead to the notion of dependence between their components. This notion will be discussed in abstract here, although such dependence between random variables will be emphasised more vividly when the notion of stochastic processes are introduced later in the course.

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 $\mathbf{X}(\zeta)$, such that, for example, if $N = 20$ and $M = 3$,

$$\mathbf{k} = \begin{bmatrix}1 & 5 & 12\end{bmatrix}^T \tag{4.14}$$

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

$$\mathbf{k} = \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_M \end{bmatrix} \quad \text{then} \quad \mathbf{X}_{\mathbf{k}}(\zeta) = \begin{bmatrix} X_{k_1}(\zeta) \\ X_{k_2}(\zeta) \\ \vdots \\ X_{k_M}(\zeta) \end{bmatrix} \tag{4.15}$$

Hence, for example, using the vector $\mathbf{k}$ above, then:

$$\mathbf{X}_{[1,5,12]}(\zeta) = \begin{bmatrix} X_1(\zeta) \\ X_5(\zeta) \\ X_{12}(\zeta) \end{bmatrix} \tag{4.16}$$

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

$$f_{\mathbf{X_k}}(\mathbf{x_k}) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{N-M \text{ integrals}} f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}_{-\mathbf{k}} \tag{4.17}$$

where $\mathbf{x}_{-\mathbf{k}}$ is the vector $\mathbf{x}$ with the elements indexed by the vector $\mathbf{k}$ **removed**.

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

$$f_{X_j}(x_j) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{N-1 \text{ integrals}} f_{\mathbf{X}}(\mathbf{x})\, dx_1 \cdots dx_{j-1} dx_{j+1} \cdots dx_N \tag{M:3.2.5}$$

In the case of a scalar RV, since it is not characterised by a joint pdf, then its pdf might be called a marginal pdf. This technical detail, which seems somewhat unnecessary, is ignored here.

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

Example 4.2 (Marginalisation). This example is again based on [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:

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x+3y) & 0 \le x,\, y \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.18}$$

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:

$$f_X(x) = \int_{\mathbb{R}} f_{\mathbf{Z}}(\mathbf{z})\; dy \tag{4.19}$$

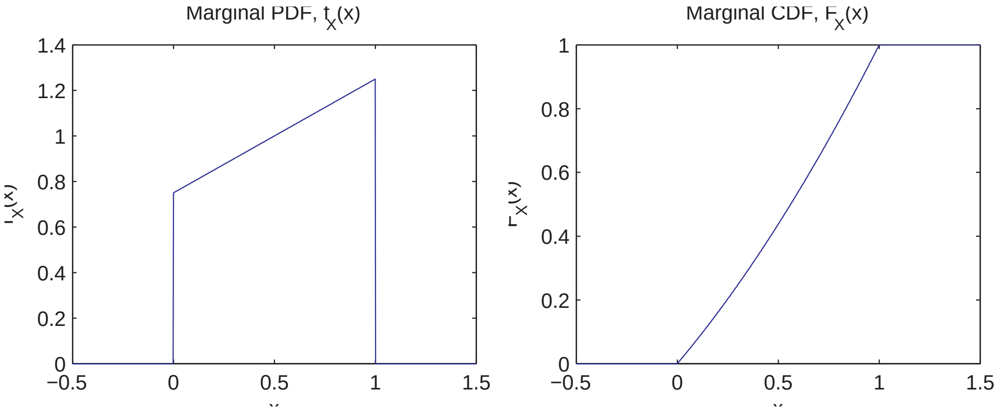
$$f_Y(y) = \int_{\mathbb{R}} f_{\mathbf{Z}}(\mathbf{z})\; dx \tag{4.20}$$

Taking $f_X(x)$, then:

$$f_X(x) = \begin{cases} \frac{1}{2}\int_0^1 (x+3y)\, dy & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.21}$$

which after a simple integration gives:

$$f_X(x) = \begin{cases} \frac{1}{2}\left(x+\frac{3}{2}\right) & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.22}$$



(a) A plot of the marginal-pdf, $X(\zeta)$. (b) A plot of the marginal-cdf, $X(\zeta)$.

Figure 4.3: The marginal-pdf, $f_X(x)$, and cdf, $F_X(x)$, for the RV, $X(\zeta)$.

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

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du = \begin{cases} 0 & x \le 0 \\ \frac{1}{2}\int_0^x \left(u + \frac{3}{2}\right) du & 0 \le x \le 1 \\ \frac{1}{2}\int_0^1 \left(u + \frac{3}{2}\right) du & x > 1 \end{cases} \tag{4.23}$$

Which after, again, a straightforward integration gives:

$$F_X(x) = \begin{cases} 0 & x \le 0 \\ \frac{x}{4}(x+3) & 0 \le x \le 1 \\ 1 & x > 1 \end{cases} \tag{4.24}$$

Note that $\lim_{x\to\infty} F_X(x) = 1$, as expected.

Similarly, it can be shown that:

$$f_Y(y) = \begin{cases} \frac{1}{2}\left(\frac{1}{2} + 3y\right) & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.25}$$

and

$$F_Y(y) = \begin{cases} 0 & y \le 0 \\ \frac{y}{4}(1+3y) & 0 \le y \le 1 \\ 1 & y > 1 \end{cases} \tag{4.26}$$

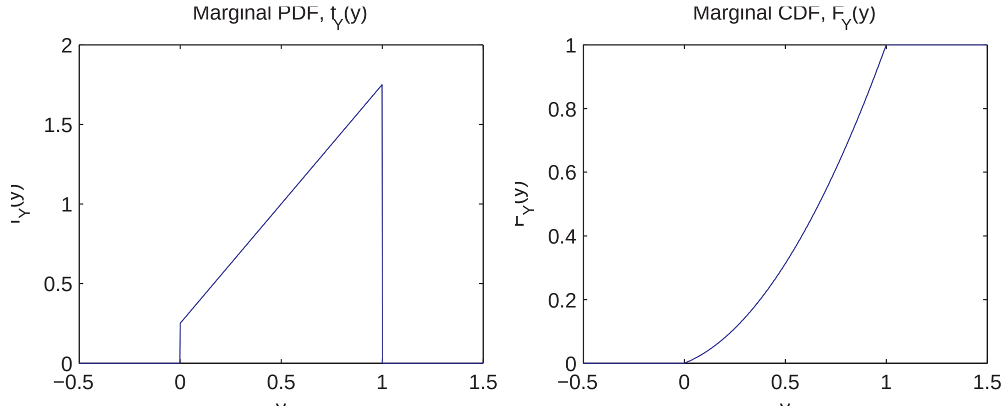
□

The marginal-pdfs and cdfs are shown in Figure 4.3 and Figure 4.4 respectively.

4.2.3 Independence

New slide

The notion of joint RVs leads to the idea of how they relate to one another. Two random variables, $X_1(\zeta)$ and $X_2(\zeta)$ are **independent** if the events $\{X_1(\zeta) \le x_1\}$ and



(a) A plot of the marginal-pdf for $Y(\zeta)$.

(b) A plot of the marginal-cdf for $Y(\zeta)$.

Figure 4.4: The marginal-pdf, $f_Y(y)$, and cdf, $F_Y(y)$, for the RV, $Y(\zeta)$.

$\{X_2(\zeta) \leq x_2\}$ are jointly independent; that is, the events do not influence one another, and

$$\Pr\left(X_1(\zeta) \leq x_1,\, X_2(\zeta) \leq x_2\right) = \Pr\left(X_1(\zeta) \leq x_1\right) \Pr\left(X_2(\zeta) \leq x_2\right) \tag{4.27}$$

This then implies that

$$\begin{aligned} F_{X_1,X_2}(x_1,\, x_2) &= F_{X_1}(x_1)\, F_{X_2}(x_2) \\ f_{X_1,X_2}(x_1,\, x_2) &= f_{X_1}(x_1)\, f_{X_2}(x_2) \end{aligned} \tag{M:3.2.7}$$

Independence will be discussed again later when stochastic processes are introduced.

4.2.4 Complex-valued RVs and vectors

Please note that this section on complex-valued random variables and vectors will not be examined. It is purely for completeness of the notes.

In applications such as (radio) channel equalisation, array processing, and so on, complex signal and noise models are encountered. To help formulate these models, it is necessary to extend the results introduced above to describe complex-valued random variables and vectors. A complex random variable is defined as $X(\zeta) = X_R(\zeta) + jX_I(\zeta)$, where $X_R(\zeta)$ and $X_I(\zeta)$ are both real-valued RVs. Thus, either $X(\zeta)$ can be considered as a mapping from an abstract probability space $\mathcal{S}$ to a complex space $\mathbb{C}$, or perhaps more simply, as a real-valued random vector, $[X_R(\zeta),\, X_I(\zeta)]^T$, with a joint cdf, $F_{X_R,X_I}(x_r,\, x_i)$, and joint pdf, $f_{X_R,X_I}(x_r,\, x_i)$, that can thus lead to a full statistical description.

Thus, the mean of $X(\zeta)$ is defined as:

$$\mathbb{E}\left[X(\zeta)\right] = \mu_X = \mathbb{E}\left[X_R(\zeta) + jX_I(\zeta)\right] = \mu_{X_R} + j\mu_{X_I} \tag{M:3.2.8}$$

and the variance is defined as:

$$\operatorname{var}\left[X\left(\zeta\right)\right] = \sigma_X^2 = \mathbb{E}\left[\left|X\left(\zeta\right) - \mu_X\right|^2\right] \tag{M:3.2.9}$$

which can be shown to equal

$$\operatorname{var}\left[X\left(\zeta\right)\right] = \mathbb{E}\left[\left|X\left(\zeta\right)\right|^2\right] - \left|\mu_X\right|^2 \tag{M:3.2.10}$$

PROOF (EQUIVALENCE OF VARIANCE EXPRESSIONS FOR A COMPLEX-VALUED RV). Beginning with the natural definition of the variance, then:

$$\begin{aligned}
\sigma_X^2 &= \mathbb{E}\left[\left|X\left(\zeta\right) - \mu_X\right|^2\right] && \text{(M:3.2.9)}\\
&= \mathbb{E}\left[\left(X\left(\zeta\right) - \mu_X\right)^*\left(X\left(\zeta\right) - \mu_X\right)\right] && (4.28)\\
&= \mathbb{E}\left[\left|X\left(\zeta\right)\right|^2 - \mu_X^* X\left(\zeta\right) - X^*(\zeta)\mu_X + \left|\mu_X\right|^2\right] && (4.29)\\
&= \mathbb{E}\left[\left|X\left(\zeta\right)\right|^2\right] - \underbrace{\mu_X^*\mathbb{E}\left[X\left(\zeta\right)\right]}_{\mathbb{E}[|\mu_X|^2]} - \underbrace{\mathbb{E}\left[X^*(\zeta)\right]\mu_X}_{\mathbb{E}[|\mu_X|^2]} + \left|\mu_X\right|^2 && (4.30)
\end{aligned}$$

□

giving the desired result.

Similarly, a complex-valued random vector is given by:

$$\mathbf{X}\left(\zeta\right) = \mathbf{X}_R(\zeta) + j\mathbf{X}_I(\zeta) = \begin{bmatrix} X_{R1}(\zeta) \\ \vdots \\ X_{RN}(\zeta)\end{bmatrix} + j\begin{bmatrix} X_{I1}(\zeta) \\ \vdots \\ X_{IN}(\zeta)\end{bmatrix} \tag{M:3.2.11}$$

Again, a complex-valued vector can be considered as a mapping from an abstract probability space to a vector-valued complex space $\mathbb{C}^N$. However, some prefer to consider it a mapping to $\mathbb{R}^{2N}$, although this viewpoint does not always provide an elegant derivation of many results. The joint cdf for $X\left(\zeta\right)$ is defined as:

$$F_{\mathbf{X}}\left(\mathbf{x}\right) \triangleq \Pr\left(\mathbf{X}\left(\zeta\right) \le \mathbf{x}\right) \triangleq \Pr\left(\mathbf{X}_R(\zeta) \le \mathbf{x}_r,\, \mathbf{X}_I(\zeta) \le \mathbf{x}_i\right) \tag{M:3.2.12}$$

while its **joint pdf**, is defined by

$$\begin{aligned}
f_{\mathbf{X}}\left(\mathbf{x}\right) &= \lim_{\Delta\mathbf{x}\to\mathbf{0}} \frac{\Pr\left(\mathbf{x}_r < \mathbf{X}_R(\zeta) \le \mathbf{x}_r + \Delta\mathbf{x}_r,\, \mathbf{x}_i < \mathbf{X}_I(\zeta) \le \mathbf{x}_i + \Delta\mathbf{x}_i\right)}{\Delta x_{r1}\cdots\Delta x_{rN}\Delta x_{i1}\cdots\Delta x_{iN}}\\
&= \frac{\partial}{\partial x_{r1}}\frac{\partial}{\partial x_{i1}}\cdots\frac{\partial}{\partial x_{rN}}\frac{\partial}{\partial x_{iN}} F_{\mathbf{X}}\left(\mathbf{x}\right)
\end{aligned} \tag{M:3.2.13}$$

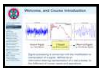
where $\Delta\mathbf{x} = \Delta x_{r1}\Delta x_{i1}\cdots\Delta x_{rN}\Delta x_{iN}$. Moreover, it follows:

$$F_{\mathbf{X}}\left(\mathbf{x}\right) = \int_{-\infty}^{x_{r1}}\int_{-\infty}^{x_{i1}}\cdots\int_{-\infty}^{x_{rN}}\int_{-\infty}^{x_{iN}} f_{\mathbf{X}}\left(\mathbf{v}\right) dv_{r1}dv_{i1}\cdots dv_{rN}dv_{iN} = \int_{-\infty}^{\mathbf{x}} f_{\mathbf{X}}\left(\mathbf{v}\right) d\mathbf{v} \tag{M:3.2.14}$$

Note that the single integral in the last expression is used as a compact notation for a multidimensional integral over all real and imaginary parts, and should not be confused with a complex-contour integral.

These probability functions for a complex-valued random vector or variable possess properties similar to those for real-valued random vectors, and will not be reproduced here. Note, in particular, however, that:

$$\int_{-\infty}^{\infty} f_{\mathbf{X}}\left(\mathbf{v}\right) d\mathbf{v} = 1 \tag{M:3.2.14}$$



4.2.5 Conditional Densities and Bayes's Theorem

New slide

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 probability of two *events* Y given X is defined as

$$\Pr\left(Y \,\middle|\, X\right) = \frac{\Pr(X, Y)}{\Pr(X)} \tag{T:2.35}$$

Defining the event X as:

$$X : \mathbf{x} \le \mathbf{X}(\zeta) \le \mathbf{x} + d\mathbf{x} \tag{T:2.36}$$

and the event Y as:

$$Y : \mathbf{y} \le \mathbf{Y}(\zeta) \le \mathbf{y} + d\mathbf{y} \tag{T:2.37}$$

then

$$\begin{aligned}
\Pr\left(Y \,\middle|\, X\right) &= \frac{\Pr(\mathbf{x} \le \mathbf{X}(\zeta) \le \mathbf{x} + d\mathbf{x},\, \mathbf{y} \le \mathbf{Y}(\zeta) \le \mathbf{y} + d\mathbf{y})}{\Pr(\mathbf{x} \le \mathbf{X}(\zeta) \le \mathbf{x} + d\mathbf{x})} && (4.31)\\
&= \frac{f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) \prod d\mathbf{x}\, d\mathbf{y}}{f_{\mathbf{X}}(\mathbf{x}) \prod d\mathbf{x}} = \left\{ \frac{f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})} \right\} \prod d\mathbf{y} && (4.32)\\
&\triangleq f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) \prod d\mathbf{y} && (4.33)
\end{aligned}$$

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

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) = \frac{f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})} \tag{T:2.39}$$

Note that

$$\int_{\mathbb{R}} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x})\, d\mathbf{y} = \int_{\mathbb{R}} \frac{f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})} d\mathbf{y} = \frac{f_{\mathbf{X}}(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} = 1 \tag{T:2.40}$$

This emphasises that $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x})$ is the density for $\mathbf{Y}(\zeta)$ that depends on $\mathbf{X}(\zeta)$ almost as if it were a parameter. Note that the integral of $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x})$ with respect to (w. r. t.) $\mathbf{x}$ is meaningless.

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:

$$f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})\, f_{\mathbf{Y}}(\mathbf{y}) \tag{T:2.41}$$

as previously defined.

Bayes's rule or Bayes's theorem is based on the fact that the joint pdf of two events can be expressed in terms of either the conditional probability for the first event, or the conditional probability for the second event. Hence, Bayes's theorem for events follows by noting:

$$\Pr(X, Y) = \Pr\left(X \,\middle|\, Y\right) \Pr(Y) = \Pr\left(Y \,\middle|\, X\right) \Pr(X) = \Pr(Y, X) \tag{4.34}$$

and therefore

$$\Pr\left(X \mid Y\right) = \frac{\Pr\left(Y \mid X\right)\Pr\left(X\right)}{\Pr\left(Y\right)} \tag{T:2.42}$$

An analogous expression can be written for density functions. Since

$$f_{\mathbf{XY}}\left(\mathbf{x},\,\mathbf{y}\right) = f_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{y} \mid \mathbf{x}\right) f_{\mathbf{X}}\left(\mathbf{x}\right) = f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x} \mid \mathbf{y}\right) f_{\mathbf{Y}}\left(\mathbf{y}\right) = f_{\mathbf{YX}}\left(\mathbf{y},\,\mathbf{x}\right) \tag{T:2.43}$$

it follows

$$f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x} \mid \mathbf{y}\right) = \frac{f_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{y} \mid \mathbf{x}\right) f_{\mathbf{X}}\left(\mathbf{x}\right)}{f_{\mathbf{Y}}\left(\mathbf{y}\right)} \tag{T:2.44}$$

This result can also be derived by considering an *events* based approach as used above in the derivation of conditional probabilities.

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

$$f_{\mathbf{Y}}\left(\mathbf{y}\right) = \int_{\mathbb{R}} f_{\mathbf{XY}}\left(\mathbf{x},\,\mathbf{y}\right) d\mathbf{x} = \int_{\mathbb{R}} f_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{y} \mid \mathbf{x}\right) f_{\mathbf{X}}\left(\mathbf{x}\right) d\mathbf{x} \tag{4.35}$$

then it follows

$$f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x} \mid \mathbf{y}\right) = \frac{f_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{y} \mid \mathbf{x}\right) f_{\mathbf{X}}\left(\mathbf{x}\right)}{\int_{\mathbb{R}} f_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{y} \mid \mathbf{x}\right) f_{\mathbf{X}}\left(\mathbf{x}\right) d\mathbf{x}} \tag{T:2.45}$$

Bayes's Theorem arises frequently in problems of statistical decision and estimation, the latter of which will be considered later in the course. Suppose that $\mathbf{Y}\left(\zeta\right)$ is an observation of an experiment which depends on some unknown random vector $\mathbf{X}\left(\zeta\right)$; for example, $\mathbf{Y}\left(\zeta\right)$ is $\mathbf{X}\left(\zeta\right)$ observed in additive noise. Then given $\mathbf{X}\left(\zeta\right)$, it is easy to find the *likelihood* of $\mathbf{Y}\left(\zeta\right)$, which is represented by the density $f_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{y} \mid \mathbf{x}\right)$; this is the **likelihood function**, and will again be introduced later in this course. The **prior density**, $f_{\mathbf{X}}\left(\mathbf{x}\right)$, represents the density of the unknown random vector before it is observed. Hence, given the likelihood and the prior, it is possible to calculate the **posterior density**, $f_{\mathbf{X}|\mathbf{Y}}\left(\mathbf{x} \mid \mathbf{y}\right)$, which is the density of the unseen random vector $\mathbf{X}\left(\zeta\right)$ given the observations $\mathbf{Y}\left(\zeta\right)$.

Example 4.3 (The lighthouse problem). A lighthouse is somewhere off a piece of straight coastline at a position α along the shore and a distance β out at sea. It emits a series of short highly collimated flashes (i.e. essentially a single ray of light) at random intervals and hence at random azimuths (i.e. the angle at which the light ray is emitted). These pulses are intercepted on the coast by photo-detectors that record only the fact that a flash has occurred, but not the angle from which it came. N flashes have so far been recorded at positions $\{x_k\}$. Where is the lighthouse?

SOLUTION. The aim of the problem is to estimate the values of α and β from the observations. Estimating both of these parameters from the data is somewhat complicated for this example, and so it will be assumed that the distance out-to-sea, β, is known. The geometry of the lighthouse problem is shown in Figure 4.5.

Given the characteristics of the lighthouse emissions, it seems reasonable to assign a uniform pdf to the azimuth of the observation, or if referring to a single observation, the **datum**, which is given by θ. Hence,

$$f_\Theta\left(\theta\right) = \begin{cases} \frac{1}{\pi} & -\frac{\pi}{2} < \theta < \frac{\pi}{2} \\ 0 & \text{otherwise} \end{cases} \tag{4.36}$$

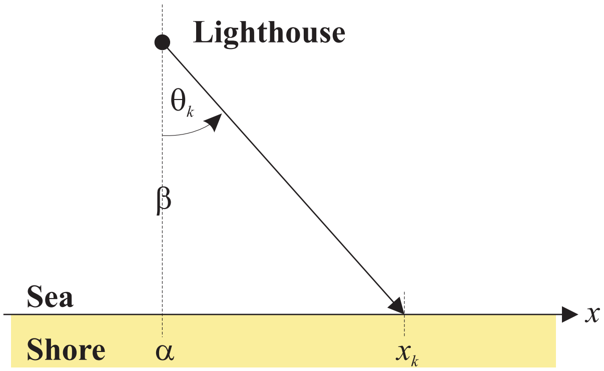


Figure 4.5: The geometry of the lighthouse problem.

The angle must lie between $\pm\frac{\pi}{2}$ radians to have been detected. Since the photo-detectors are only sensitive to position along the coast rather than direction, it is necessary to relate θ to x. An inspection of Figure 4.5 shows that

$$\beta \tan\theta = x - \alpha \tag{4.37}$$

Using the probability transformation rule, it is possible to show that

$$f_X\left(x \mid \alpha\right) = \frac{\beta}{\pi\left[\beta^2 + (x-\alpha)^2\right]} \tag{4.38}$$

where, as a reminder, it is assumed that β is known. This transformation is left as an exercise to the reader. Assuming that the observations are independent, then the joint-pdf of all the data points is given by:

$$\begin{aligned} f_{\mathbf{X}}\left(\mathbf{x} \mid \alpha\right) = f_{\mathbf{X}}\left(x_1, \dots, x_N \mid \alpha\right) &= \prod_{k=1}^{N} f_X\left(x_k \mid \alpha\right) \\ &= \prod_{k=1}^{N} \frac{\beta}{\pi\left[\beta^2 + (x_k-\alpha)^2\right]} \end{aligned} \tag{4.39}$$

The position of the lighthouse is then expressed by:

$$f_A\left(\alpha \mid \mathbf{x}\right) = \frac{f_{\mathbf{X}}\left(\mathbf{x} \mid \alpha\right)\, f_A\left(\alpha\right)}{f_{\mathbf{X}}\left(\mathbf{x}\right)} \tag{4.40}$$

It is reasonable, also, to assign a simple uniform pdf for the *prior density* for the distance along the shore:

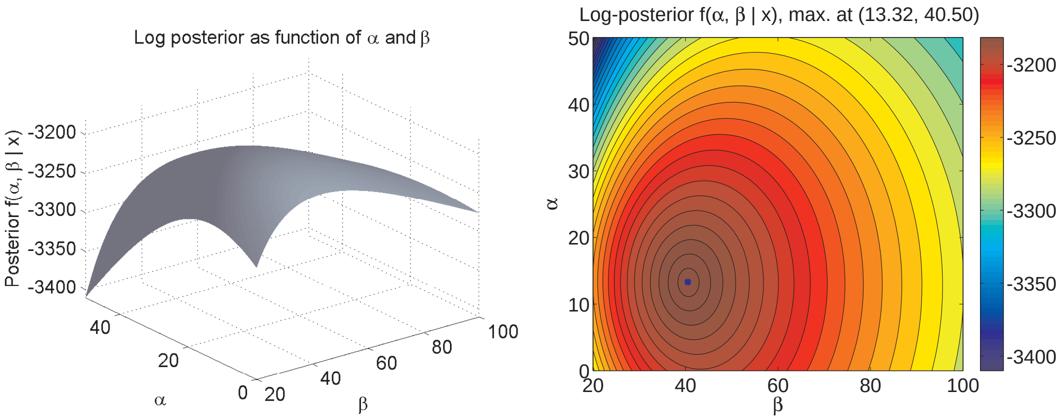
$$f_A\left(\alpha\right) = \begin{cases} \frac{1}{\alpha_{\text{max}}-\alpha_{\text{min}}} & \alpha_{\text{min}} \le \alpha \le \alpha_{\text{max}} \\ 0 & \text{otherwise} \end{cases} \tag{4.41}$$

Hence, it follows that

$$f_A\left(\alpha \mid \mathbf{x}\right) = \frac{f_{\mathbf{X}}\left(\mathbf{x} \mid \alpha\right)\, f_A\left(\alpha\right)}{f_{\mathbf{X}}\left(\mathbf{x}\right)} \propto f_{\mathbf{X}}\left(\mathbf{x} \mid \alpha\right)\, f_A\left(\alpha\right) \tag{4.42}$$

$$\propto \frac{1}{\alpha_{\text{max}}-\alpha_{\text{min}}} \prod_{k=1}^{N} \frac{\beta}{\pi\left[\beta^2 + (x_k-\alpha)^2\right]}, \quad \text{for } \alpha_{\text{min}} \le \alpha \le \alpha_{\text{max}} \tag{4.43}$$

□



(a) Surface plot of the log-posterior.

(b) Contour plot of the log-posterior.

Figure 4.6: Visualising the log-posterior function described in Equation 4.43 when both α and β are unknown. In this case, the number of data-points used is $N = 500$. The actual lighthouse location is at $(\alpha, \beta) = (15, 45)$. Note the error in the estitmae of the maximum value.

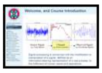
and zero otherwise. Hence, this **posterior density** can be maximised to find the best estimate of the distance along the shore, α. Unfortunately, in this case, this maximisation is not easy.

The result in Equation 4.43 can easily be generalised when both α and β are unknown, and the logarithm of the posterior can be plotted as a function of α and β. The resulting two-dimensional (2-D) function is shown in Figure 4.6 and Figure 4.7 for when the lightouse is actually at $(\alpha, \beta) = (15, 45)$. Note that for $N = 500$ data-points, there is a relatively large error in the estimate, especially when compared with $N = 50000$. This will be discussed in later handouts. Moreover, note that when you run the corresponding MATLAB code, in which the data is generated synthetically, a new estimate is obtained each time. Can you explain why? Finally, if N is small, a typical estimate might be far from the true solution.

A MATLAB script is available on LEARN which plots these functions.

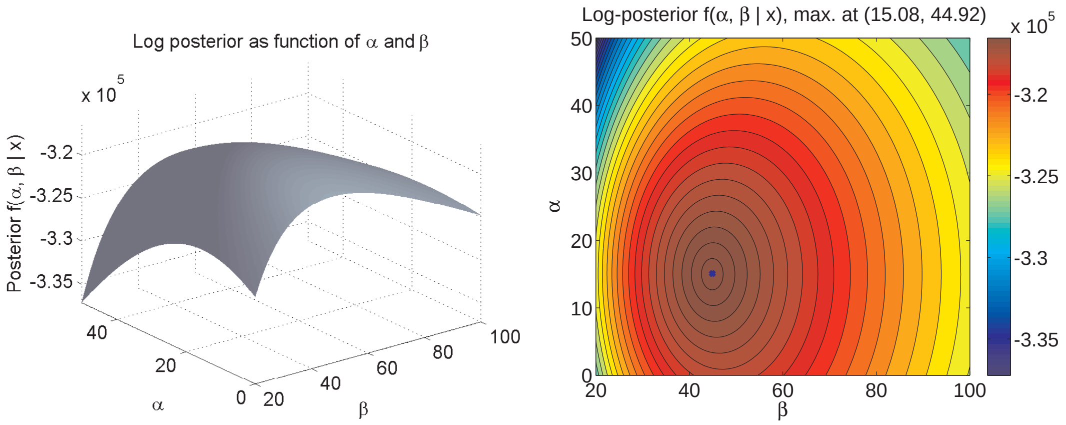
```
thisData = LighthouseProblem(N)
```

4.3 Statistical Description



New slide

As with scalar RVs, the probabilistic descriptions require an enormous amount of information that is not always easy to obtain, or is too complex mathematically for practical use.



(a) Surface plot of the log-posterior.

(b) Contour plot of the log-posterior.

Figure 4.7: Visualising the log-posterior function described in Equation 4.43 when both α and β are unknown. In this case, the number of data-points used is $N = 50000$. The actual lighthouse location is at $(\alpha,\, \beta) = (15, 45)$. Note the error in the estimate of the maximum value is much less than for $N = 500$.

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. Rather than list them all here, they will be introduced where necessary. 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.

The most important statistical descriptors discussed in this section are the **mean vector**, the **correlation matrix** and the **covariance matrix**.

Mean vector The most important statistical operation is the expectation operator. The **mean vector** is the first-moment of the random vector, and is given by:

$$\boldsymbol{\mu}_{\mathbf{X}} = \mathbb{E}\left[\mathbf{X}\left(\zeta\right)\right] = \begin{bmatrix} \mathbb{E}\left[X_1(\zeta)\right] \\ \vdots \\ \mathbb{E}\left[X_N(\zeta)\right] \end{bmatrix} = \begin{bmatrix} \mu_{X_1} \\ \vdots \\ \mu_{X_N} \end{bmatrix} \tag{M:3.2.16}$$

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

$$\mathbf{R}_{\mathbf{X}} \triangleq \begin{bmatrix} \mathbb{E}\left[X_1(\zeta)X_1^*(\zeta)\right] & \cdots & \mathbb{E}\left[X_1(\zeta)X_N^*(\zeta)\right] \\ \vdots & \ddots & \cdots \\ \mathbb{E}\left[X_N(\zeta)X_1^*(\zeta)\right] & \cdots & \mathbb{E}\left[X_N(\zeta)X_N^*(\zeta)\right] \end{bmatrix} \tag{4.44}$$

or, more succinctly,

$$\mathbf{R_X} \triangleq \mathbb{E}\left[\mathbf{X}(\zeta)\,\mathbf{X}^H(\zeta)\right] = \begin{bmatrix} r_{X_1X_1} & \cdots & r_{X_1X_N} \\ \vdots & \ddots & \vdots \\ r_{X_NX_1} & \cdots & r_{X_NX_N} \end{bmatrix} \tag{M:3.2.17}$$

where the superscript H denotes the conjugate transpose operation; in otherwords, for a general $N \times M$ matrix $\mathbf{A} \in \mathbb{C}^{N\times M}$ with complex elements $a_{ij} \in \mathbb{C}$, then

$$\mathbf{A}^H = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{2M} \\ \vdots & \vdots & \cdots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NM} \end{bmatrix}^H = \begin{bmatrix} a_{11}^* & a_{21}^* & \cdots & a_{N1}^* \\ a_{12}^* & a_{22}^* & \cdots & a_{N2}^* \\ \vdots & \vdots & \ddots & \vdots \\ a_{1M}^* & a_{2M}^* & \cdots & a_{NM}^* \end{bmatrix} \in \mathbb{C}^{M\times N} \tag{4.45}$$

The diagonal terms

$$r_{X_iX_i} \triangleq \mathbb{E}\left[|X_i(\zeta)|^2\right], \quad i \in \{1, \dots, N\} \tag{M:3.2.18}$$

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

The off-diagonal terms

$$r_{X_iX_j} \triangleq \mathbb{E}\left[X_i(\zeta)X_j^*(\zeta)\right] = r_{X_jX_i}^*, \quad i \neq j \tag{M:3.2.19}$$

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:

$$r_{X_iX_j} = \mathbb{E}\left[X_i(\zeta)X_j^*(\zeta)\right] = 0, \quad i \neq j \tag{M:3.2.26}$$

Hence, if all the RVs are mutually orthogonal, then the $\mathbf{R_X}$ will be diagonal.

Note that the correlation matrix $\mathbf{R_X}$ is conjugate symmetric, which is also known as **Hermitian**; that is, $\mathbf{R_X} = \mathbf{R_X}^H$.

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

$$\mathbf{\Gamma_X} \triangleq \mathbb{E}\left[\left(\mathbf{X}(\zeta) - \boldsymbol{\mu}_\mathbf{X}\right)\left(\mathbf{X}(\zeta) - \boldsymbol{\mu}_\mathbf{X}\right)^H\right] = \begin{bmatrix} \gamma_{X_1X_1} & \cdots & \gamma_{X_1X_N} \\ \vdots & \ddots & \cdots \\ \gamma_{X_NX_1} & \cdots & \gamma_{X_NX_N} \end{bmatrix} \tag{M:3.2.20}$$

The diagonal terms

$$\gamma_{X_iX_i} \triangleq \sigma_{X_i}^2 = \mathbb{E}\left[|X_i(\zeta) - \mu_{X_i}|^2\right], \quad i \in \{1, \dots, N\} \tag{M:3.2.21}$$

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

The off-diagonal terms

$$\begin{aligned} \gamma_{X_iX_j} &\triangleq \mathbb{E}\left[\left(X_i(\zeta) - \mu_{X_i}\right)\left(X_j(\zeta) - \mu_{X_j}\right)^*\right] \\ &= r_{X_iX_j} - \mu_{X_i}\mu_{X_j}^* = \gamma_{X_jX_i}^*, \quad i \neq j \end{aligned} \tag{M:3.2.22}$$

Sidebar 3 Positive semi-definiteness of Real Matrices

If a matrix $\mathbf{R}$ is real, then the calculation $\mathbf{a}^H\, \boldsymbol{\Gamma}_{\mathbf{X}}\mathbf{a}$ simplifies to only needing to consider any real vector $\mathbf{a}$. This can be shown by writing:

$$\mathbf{a} = \mathbf{a}_R + j\mathbf{a}_I \tag{4.47}$$

where $\mathbf{a}_R$ and $\mathbf{a}_I$ are real column vectors. Hence, assuming that $\boldsymbol{\Gamma}$ is real, it follows that:

$$\begin{aligned} \mathcal{I} = \mathbf{a}^H\boldsymbol{\Gamma}\mathbf{a} &= \left(\mathbf{a}_R + j\mathbf{a}_I\right)^H \left(\boldsymbol{\Gamma}\mathbf{a}_R + j\boldsymbol{\Gamma}\mathbf{a}_I\right) && (4.48)\\ &= \mathbf{a}_R^T\left(\boldsymbol{\Gamma}\mathbf{a}_R + j\boldsymbol{\Gamma}\mathbf{a}_R\right) - j\mathbf{a}_I^T\left(\boldsymbol{\Gamma}\mathbf{a}_R + j\boldsymbol{\Gamma}\mathbf{a}_I\right) && (4.49)\\ &= \mathbf{a}_R^T\boldsymbol{\Gamma}\mathbf{a}_R + j\mathbf{a}_R^T\boldsymbol{\Gamma}\mathbf{a}_R - j\mathbf{a}_I^T\boldsymbol{\Gamma}\mathbf{a}_R + \mathbf{a}_I^T\boldsymbol{\Gamma}\mathbf{a}_I && (4.50) \end{aligned}$$

Now, noting that $\mathcal{I}$ is a scalar quantity, and with $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^T$, $\mathcal{I}$ is also a real scalar quantity. Hence, it can be seen that $\mathbf{a}_R^H\boldsymbol{\Gamma}\mathbf{a}_R = \mathbf{a}_I^H\boldsymbol{\Gamma}\mathbf{a}_R$, therefore giving

$$\mathcal{I} = \mathbf{a}^T\boldsymbol{\Gamma}\mathbf{a} = \mathbf{a}_R^T\boldsymbol{\Gamma}\mathbf{a}_R + \mathbf{a}_I^T\boldsymbol{\Gamma}\mathbf{a}_I \tag{4.51}$$

Since both of these terms are real, then there is no need for both the real and imaginary components of the vector $\mathbf{a}$, and therefore it makes sense to set $\mathbf{a}_I = \mathbf{0}$.

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:

$$\begin{aligned} \mathbf{a}^H\, \mathbf{R}_{\mathbf{X}}\mathbf{a} &\geq 0\\ \mathbf{a}^H\, \boldsymbol{\Gamma}_{\mathbf{X}}\mathbf{a} &\geq 0 \end{aligned} \qquad \text{(T:2.65)}$$

for any complex vector $\mathbf{a}$. This follows since:

$$\mathbf{a}^H\, \mathbf{R}_{\mathbf{X}}\mathbf{a} = \mathbf{a}^H\, \mathbb{E}\left[\mathbf{x}\mathbf{x}^H\right]\mathbf{a} = \mathbb{E}\left[\left|\mathbf{x}^H\mathbf{a}\right|^2\right] \tag{4.46}$$

The covariance matrix $\boldsymbol{\Gamma}_{\mathbf{X}}$ is also a Hermitian matrix. Note that a Hermitian matrix is semi-positive definite if all its eigenvalues are greater than or equal to zero.

Moreover, as for scalar RVs, the covariance, $\gamma_{X_iX_j}$ can also be expressed in terms of the standard deviations of $X_i(\zeta)$ and $X_j(\zeta)$:

$$\rho_{X_iX_j} \triangleq \frac{\gamma_{X_iX_j}}{\sigma_{X_i}\sigma_{X_j}} = \rho^*_{X_jX_i} \qquad \text{(M:3.2.23)}$$

Again, the correlation coefficient measures the degree of statistical similarity between two random variables. Note that:

$$\left|\rho_{X_iX_j}\right| \leq 1, \quad i \neq j, \quad \text{and} \quad \rho_{X_iX_i} = 1 \qquad \text{(M:3.2.24)}$$

If $\left|\rho_{X_iX_j}\right| = 1$, $i \neq j$, then the RVs are said to be *perfectly correlated*. However, if $\rho_{X_iX_j} = 0$, which occurs when the covariance $\gamma_{X_iX_j} = 0$, then the RVs are said to be *uncorrelated*.

Example 4.4 ([Manolakis:2001, Exercise 3.14, Page 145]). Determine whether the following matrices are valid correlation matrices:

$$\mathbf{R}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \qquad \mathbf{R}_2 = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & 1 \end{bmatrix} \tag{4.52}$$

$$\mathbf{R}_3 = \begin{bmatrix} 1 & 1-j \\ 1+j & 1 \end{bmatrix} \qquad \mathbf{R}_4 = \begin{bmatrix} 1 & \frac{1}{2} & 1 \\ \frac{1}{2} & 2 & \frac{1}{2} \\ 1 & 1 & 1 \end{bmatrix} \tag{4.53}$$

SOLUTION. Correlation (and covariance) matrices are Hermitian and positive semidefinite. The first three correlation matrices are Hermitian, and are therefore valid. $\mathbf{R}_4$ is not, and so therefore is not a valid correlation matrix. Next, it is necessary to test whether these matrices are positive semi-definite, and this test is performed below:

1. Setting $\mathbf{a} = [a_1,\, a_2]^T$, then

$$\mathbf{a}^T\mathbf{R}_1\mathbf{a} = \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} a_1 + a_2 \\ a_1 + a_2 \end{bmatrix} = a_1^2 + 2a_1a_2 + a_2^2 = (a_1+a_2)^2 \geq 0 \tag{4.54}$$

for all $a_1,\, a_2$. Thus, this is a valid correlation matrix.

2. Setting $\mathbf{a} = [a_1,\, a_2,\, a_3]^T$, then

$$\mathbf{a}^T\mathbf{R}_2\mathbf{a} = \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} \begin{bmatrix} a_1 + \frac{a_2}{2} + \frac{a_3}{4} \\ \frac{a_1}{2} + a_2 + \frac{a_3}{2} \\ \frac{a_1}{4} + \frac{a_2}{2} + a_3 \end{bmatrix} \tag{4.55}$$

$$= a_1^2 + a_1a_2 + \frac{1}{2}a_1a_3 + a_2^2 + a_2a_3 + a_3^2 \tag{4.56}$$

$$= \frac{1}{2}(a_1+a_2+a_3)^2 + \frac{1}{2}(a_1 - \frac{1}{2}a_3)^2 + \frac{1}{2}a_2^2 + \frac{3}{8}a_3^2 \geq 0 \tag{4.57}$$

for all $a_1,\, a_2$. Thus, this is a valid correlation matrix.

3. Finally, for this complex case, $\mathbf{a} = [a_1,\, a_2]^T$, then

$$\mathbf{a}^H\mathbf{R}_3\mathbf{a} = \begin{bmatrix} a_1^* & a_2^* \end{bmatrix} \begin{bmatrix} a_1 + (1-j)a_2 \\ (1+j)a_1 + a_2 \end{bmatrix} \tag{4.58}$$

$$= |a_1|^2 + (1-j)a_1^*a_2 + (1+j)a_2^*a_1 + |a_2|^2 \tag{4.59}$$

$$= |a_1 + (1-j)a_2|^2 - |a_2|^2 \tag{4.60}$$

□

for all $a_1,\, a_2$. To see that this is not always positive, choose the counter-example: $a_1 = -1+j$ and $a_2 = 1$; then clearly $\mathbf{a}^H\mathbf{R}_3\mathbf{a} = -1 < 0$. Therefore, this is not a valid correlation matrix.

4. As mentioned above, but repeated here for completeness, $\mathbf{R}_4$ is not Hermitian, and is therefore not a valid correlation matrix.

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

$$\mathbf{\Gamma_X} \triangleq \mathbb{E}\left[\left[\mathbf{X}(\zeta) - \boldsymbol{\mu}_\mathbf{X}\right]\left[\mathbf{X}(\zeta) - \boldsymbol{\mu}_\mathbf{X}\right]^H\right] = \mathbf{R_X} - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^H \tag{M:3.3.25}$$

which shows that the two moments have essentially the same amount of information. 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:

$$\begin{aligned} r_{X_iX_j} &= \mathbb{E}\left[X_i(\zeta)\, X_j(\zeta)^*\right] = \mathbb{E}\left[X_i(\zeta)\right]\mathbb{E}\left[X_j^*(\zeta)\right] \\ &= \mu_{X_i}\mu_{X_j}^* \quad \Rightarrow \quad \gamma_{X_iX_j} = 0 \end{aligned} \tag{M:3.3.36}$$

Note, however, that uncorrelatedness does not imply independence, unless the RVs are jointly-Gaussian. If one or both RVs have zero means, then uncorrelatedness also implies orthogonality.

Naturally, the correlation and covariance between two random vectors can also be defined. Let $X(\zeta)$ and $Y(\zeta)$ be random N- and M- vectors.

Cross-correlation is defined as

$$\mathbf{R_{XY}} \triangleq \mathbb{E}\left[\mathbf{X}(\zeta)\,\mathbf{Y}^H(\zeta)\right] = \begin{bmatrix} \mathbb{E}\left[X_1(\zeta)Y_1^*(\zeta)\right] & \cdots & \mathbb{E}\left[X_1(\zeta)Y_M^*(\zeta)\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}\left[X_N(\zeta)Y_1^*(\zeta)\right] & \cdots & \mathbb{E}\left[X_N(\zeta)Y_M^*(\zeta)\right] \end{bmatrix} \tag{M:3.2.28}$$

which is a $N \times M$ matrix. The elements $r_{X_iY_j} = \mathbb{E}\left[X_i(\zeta)Y_j^*(\zeta)\right]$ are the correlations between the RVs $X(\zeta)$ and $Y(\zeta)$.

Cross-covariance is defined as

$$\begin{aligned} \mathbf{\Gamma_{XY}} &\triangleq \mathbb{E}\left[\left\{\mathbf{X}(\zeta) - \boldsymbol{\mu}_\mathbf{X}\right\}\left\{\mathbf{Y}(\zeta) - \boldsymbol{\mu}_\mathbf{Y}\right\}^H\right] \\ &= \mathbf{R_{XY}} - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{Y}^H \end{aligned} \tag{M:3.2.29}$$

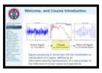
which too is a $N \times M$ matrix. The elements $\gamma_{X_iY_j} = \mathbb{E}\left[\left(X_i(\zeta) - \mu_{X_i}\right)\left(Y_j(\zeta) - \mu_{Y_j}\right)^*\right]$ are the covariances between $X(\zeta)$ and $Y(\zeta)$.

In general, cross-matrices are not square, and even if $N = M$, they are not necessarily symmetric.

Two random-vectors $X(\zeta)$ and $Y(\zeta)$ are said to be:

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

Again, if $\boldsymbol{\mu}_\mathbf{X}$ or $\boldsymbol{\mu}_\mathbf{Y}$ or both are zero vectors, then uncorrelatedness implies orthogonality.



New slide

4.4 Probability Transformation Rule

The probability transformation rule for scalar RVs can be extended to multiple RVs using a similar derivation.

Theorem 4.1 (Probability Transformation Rule). The set of random variables $\mathbf{X}(\zeta) = \{X_n(\zeta),\, n \in \mathcal{N}\}$ where $\mathcal{N} = \{1, \dots, N\}$ are transformed to a new set of RVs, $\mathbf{Y}(\zeta) = \{Y_n(\zeta),\, n \in \mathcal{N}\}$, using the transformations:

$$Y_n(\zeta) = g_n(\mathbf{X}(\zeta)), \quad n \in \mathcal{N} \tag{4.61}$$

or, using an alternative notation,

$$\mathbf{Y}(\zeta) = \mathbf{g}(\mathbf{X}(\zeta)) \tag{4.62}$$

where $\mathbf{g}(\cdot)$ denotes a vector of functions such that $Y_n(\zeta) = g_n(\mathbf{X}(\zeta))$ as above.

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

$$\mathbf{y} = \mathbf{g}(\mathbf{x}_1) = \cdots = \mathbf{g}(\mathbf{x}_M) \tag{4.63}$$

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

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{m=1}^{M} \frac{f_{\mathbf{X}}(\mathbf{x}_m)}{|J(\mathbf{x}_m)|} \tag{4.64}$$

where the **Jacobian** of the transformation, $J_{\mathbf{g}}(\mathbf{x})$, is given by:

$$\begin{aligned} J_{\mathbf{g}}(\mathbf{x}) &\triangleq \frac{\partial(y_1, \dots, y_N)}{\partial(x_1, \dots, x_N)} \\ &= \begin{vmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} & \frac{\partial g_2(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial g_N(\mathbf{x})}{\partial x_1} \\ \frac{\partial g_1(\mathbf{x})}{\partial x_2} & \frac{\partial g_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_N(\mathbf{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1(\mathbf{x})}{\partial x_N} & \frac{\partial g_2(\mathbf{x})}{\partial x_N} & \cdots & \frac{\partial g_N(\mathbf{x})}{\partial x_N} \end{vmatrix} \end{aligned} \tag{T:2.123}$$

It should also be noted, from vector calculus results, that the Jacobian can also be expressed as:

$$\begin{aligned} \frac{1}{J_{\mathbf{g}}(\mathbf{x})} &\triangleq \frac{\partial(x_1, \dots, x_N)}{\partial(y_1, \dots, y_N)} \\ &= \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} & \cdots & \frac{\partial x_N}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_N}{\partial y_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_1}{\partial y_N} & \frac{\partial x_2}{\partial y_N} & \cdots & \frac{\partial x_N}{\partial y_N} \end{vmatrix} \end{aligned} \tag{T:2.123}$$

PROOF. The proof follows a very similar line to that for the scalar RVs case. The definition of the joint-**pdf** is:

$$f_{\mathbf{Y}}(\mathbf{y}) \prod d\mathbf{y} = \Pr(\mathbf{y} < \mathbf{Y}(\zeta) \leq \mathbf{y} + d\mathbf{y}) \tag{4.65}$$

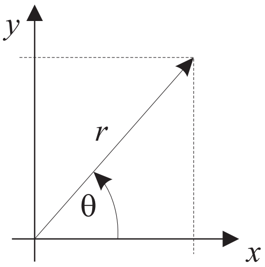


Figure 4.8: The Cartesian and polar coordinate systems.

where $\prod d\mathbf{y} = dy_1\, dy_2 \ldots\, dy_N$. The set of values $\mathbf{x}$ such that $\mathbf{y} < \mathbf{g}(\mathbf{x}) \leq \mathbf{y} + d\mathbf{y}$, consists of the intervals:

$$\mathbf{x}_n < \mathbf{x} \leq \mathbf{x}_n + d\mathbf{x}_n \tag{4.66}$$

The probability that $\mathbf{x}$ lies in this set is, of course,

$$f_{\mathbf{X}}(\mathbf{x}_n) \prod d\mathbf{x}_n = \Pr\left(\mathbf{x}_n < \mathbf{X}(\zeta) \leq \mathbf{x}_n + d\mathbf{x}_n\right) \tag{4.67}$$

Moreover, the transformation from $\mathbf{x}$ to $\mathbf{y}$ is given by the Jacobian:

$$\prod d\mathbf{y} = J_{\mathbf{g}}(\mathbf{x}) \prod d\mathbf{x} \tag{4.68}$$

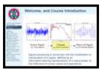
Since these are mutually exclusive sets, then

$$\Pr\left(\mathbf{y} < \mathbf{Y}(\zeta) \leq \mathbf{y} + d\mathbf{y}\right) = \sum_{m=1}^{M} \Pr\left(\mathbf{x}_n < \mathbf{X}(\zeta) \leq \mathbf{x}_n + d\mathbf{x}_n\right) \tag{4.69}$$

$$= \sum_{m=1}^{M} f_{\mathbf{X}}(\mathbf{x}_n)\, \frac{\prod d\mathbf{y}}{J_{\mathbf{g}}(\mathbf{x}_n)} \tag{4.70}$$

□

and thus the desired result is obtained after minor rearrangement.



New slide

4.4.1 Polar Transformation

An important transformation example is the mapping from Cartesian to polar coordinates. Each of these coordinates are shown in Figure 4.8.

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

$$\begin{aligned} r(\zeta) &= \sqrt{X^2(\zeta) + Y^2(\zeta)} \\ \theta(\zeta) &= \arctan \frac{Y(\zeta)}{X(\zeta)} \end{aligned} \tag{4.71}$$

where it is assumed that $r(\zeta) \geq 0$, and $|\theta(\zeta)| \leq \pi$. With this assumption, the transformation $r = \sqrt{x^2 + y^2}$, $\theta = \arctan \frac{y}{x}$ has a single solution:

$$\left.\begin{aligned} x &= r\, \cos\theta \\ y &= r\, \sin\theta \end{aligned}\right\} \quad \text{for } r > 0 \tag{4.72}$$

The Jacobian is given by:

$$J_{\mathbf{g}}(\mathbf{c}) = \frac{\partial(r,\theta)}{\partial(x,y)} = \begin{vmatrix} \frac{\partial\theta}{\partial x} & \frac{\partial r}{\partial x} \\ \frac{\partial\theta}{\partial y} & \frac{\partial r}{\partial y} \end{vmatrix} = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial\theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial\theta} \end{vmatrix}^{-1} \tag{4.73}$$

In the case of polar transformations, $J_{\mathbf{g}}(\mathbf{c})$ simplifies to:

$$J_{\mathbf{g}}(\mathbf{c}) = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix}^{-1} = \frac{1}{r} \tag{4.74}$$

Thus, it follows that:

$$f_{R,\Theta}\left(r,\theta\right) = r f_{XY}\left(r\, \cos\theta,\, r\, \sin\theta\right) \tag{4.75}$$

Example 4.5 (Cartesian to polar transformation of RVs). If $X\left(\zeta\right)$ and $Y\left(\zeta\right)$ are independent and identically distributed (i. i. d.) Gaussian distributed coordinates in Cartesian space, such that $X\left(\zeta\right), Y\left(\zeta\right) \sim \mathcal{N}\left(0,\, \sigma^2\right)$, find the distribution when these are transformed into polar coordinates.

SOLUTION. First, note:

$$f_{XY}\left(x,y\right) = f_X\left(x\right)\, f_Y\left(y\right) = \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{x^2+y^2}{2\sigma^2}\right\} \tag{4.76}$$

Hence, applying the transformation $r = \sqrt{x^2+y^2}$, $\theta = \arctan\frac{y}{x}$, it directly follows that

$$f_{R\Theta}\left(r,\, \theta\right) = \frac{r}{2\pi\sigma^2} \exp\left\{-\frac{r^2}{2\sigma^2}\right\} \mathbb{I}_{[-\pi,\pi]}\left(\theta\right) \mathbb{I}_{\mathbb{R}^+}\left(r\right) \tag{4.77}$$

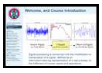
where, as a reminder, $\mathbb{I}_{\mathcal{A}}\left(a\right) = 1$ if $a \in \mathcal{A}$ and zero otherwise. This density is a product of a function of r times a function of θ. Hence, the RVs r and θ are independent with:

$$f_R\left(r\right) = \frac{r}{\sigma^2} \exp\left\{-\frac{r^2}{2\sigma^2}\right\} \mathbb{I}_{\mathbb{R}^+}\left(r\right) \quad \text{and} \quad f_\Theta\left(\theta\right) = \frac{1}{2\pi} \mathbb{I}_{[-\pi,\pi]}\left(\theta\right) \tag{4.78}$$

□

where the scaling factors have been apportioned such that these are proper densities, in the sense that $\int_{\mathbb{R}} f_R\left(r\right)\, dr = \int_{\mathbb{R}} f_\Theta\left(\theta\right)\, d\theta = 1$. Note that θ is uniformally distributed, while r has a **Rayleigh distribution**.

4.4.2 Linear Transformations



New slide

Since linear systems represent such an important class if signal processing systems, it is important to consider **linear transformations** of random vectors. Thus, consider a random vector $\mathbf{Y}(\zeta)$ defined by a linear transformation of the random vector $\mathbf{X}(\zeta)$ through the matrix $\mathbf{A}$:

$$\mathbf{Y}(\zeta) = \mathbf{A}\,\mathbf{X}(\zeta) \tag{M:3.2.32}$$

The matrix $\mathbf{A}$ is not necessarily square and, in particular, if $\mathbf{X}(\zeta)$ is of dimension M, and $\mathbf{Y}(\zeta)$ of dimension N, then $\mathbf{A}$ is of size $N \times M$ (rows by columns).

If $N > M$, then only M $Y_m(\zeta)$ RVs can be independently determined from $\mathbf{X}(\zeta)$. The remaining $N - M$ $Y_m(\zeta)$ RVs can then be obtained from the first M $Y_m(\zeta)$ RVs. If, however, $M > N$, then the random vector $\mathbf{Y}(\zeta)$ can be augmented into an M-vector by introducing the auxiliary RVs,

$$Y_n(\zeta) = X_n(\zeta), \quad \text{for } n > m \tag{M:3.2.33}$$

These additional auxiliary variables must then be marginalised out to obtain the joint-pdf for the original N-vector, $\mathbf{Y}(\zeta)$. The approach of using auxiliary variables is discussed further below in Section 4.4.3.

Both of these cases, for $M \neq N$, lead to less elegant expressions for $f_{\mathbf{Y}}(\mathbf{y})$, and therefore it will be assumed that $M = N$, and that $\mathbf{A}$ is nonsingular.

The Jacobian of a nonsingular linear transformation defined by a matrix $\mathbf{A}$ is simply the absolute value of the determinant of $\mathbf{A}$ as shown in Sidebar 4. Thus, assuming $\mathbf{X}(\zeta)$, $\mathbf{Y}(\zeta)$, and $\mathbf{A}$ are all real, then:

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{f_{\mathbf{X}}\left(\mathbf{A}^{-1}\mathbf{y}\right)}{|\det \mathbf{A}|} \tag{M:3.2.34}$$

In general, determining $f_{\mathbf{Y}}(\mathbf{y})$ is a laborious exercise, except in the case of Gaussian random vectors. In practice, however, the knowledge of $\boldsymbol{\mu}_{\mathbf{Y}}$, $\boldsymbol{\Gamma}_{\mathbf{Y}}$, $\boldsymbol{\Gamma}_{\mathbf{XY}}$ or $\boldsymbol{\Gamma}_{\mathbf{YX}}$ is sufficient information for many algorithms.

Taking expectations of both sides of Equation M:3.2.32, the following relations are found:

Mean vector:

$$\boldsymbol{\mu}_{\mathbf{Y}} = \mathbb{E}\left[\mathbf{A}\,\mathbf{X}(\zeta)\right] = \mathbf{A}\,\boldsymbol{\mu}_{\mathbf{X}} \tag{M:3.2.38}$$

Autocorrelation matrix:

$$\begin{aligned}\mathbf{R}_{\mathbf{Y}} &= \mathbb{E}\left[\mathbf{Y}(\zeta)\,\mathbf{Y}^H(\zeta)\right] = \mathbb{E}\left[\mathbf{A}\mathbf{X}(\zeta)\,\mathbf{X}^H(\zeta)\mathbf{A}^H\right] \\ &= \mathbf{A}\mathbb{E}\left[\mathbf{X}(\zeta)\,\mathbf{X}^H(\zeta)\right]\mathbf{A}^H = \mathbf{A}\mathbf{R}_{\mathbf{X}}\mathbf{A}^H\end{aligned} \tag{M:3.2.39}$$

Autocovariance matrix:

$$\boldsymbol{\Gamma}_{\mathbf{Y}} = \mathbf{A}\boldsymbol{\Gamma}_{\mathbf{X}}\mathbf{A}^H \tag{M:3.2.40}$$

Sidebar 4 Jacobian of a Linear Transformation

A linear transformation of N variables, $\{x_i\}_1^N$, to N variables, $\{y_i\}_1^N$, can either be written in matrix-vector form as shown in Equation M:3.2.33, or equivalently:

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{\mathbf{Y}(\zeta)} = \underbrace{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \dots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}}_{\mathbf{X}(\zeta)} \tag{4.79}$$

or in the scalar form by the linear equation:

$$y_i = \sum_{k=1}^{N} a_{ik}\, x_k \tag{4.80}$$

where a_{ij} is the ith row and jth column of the matrix $\mathbf{A}$. The Jacobian is obtained by calculating:

$$\frac{\partial y_i}{\partial x_j} = \sum_{k=1}^{N} a_{ik}\, \frac{\partial x_k}{\partial x_j} = a_{ij} \tag{4.81}$$

using the fact that

$$\frac{\partial x_k}{\partial x_j} = \begin{cases} 1 & \text{if } k = j \\ 0 & \text{if } k \neq j \end{cases} \tag{4.82}$$

Hence, constructing the Jacobian matrix using Equation 4.81 gives the matrix $\mathbf{A}$.

Cross-correlation matrix:

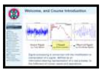
$$\begin{aligned}\mathbf{R_{XY}} &= \mathbb{E}\left[\mathbf{X}(\zeta)\,\mathbf{Y}^H(\zeta)\right] = \mathbb{E}\left[\mathbf{X}(\zeta)\,\mathbf{X}^H(\zeta)\mathbf{A}^H\right] \\ &= \mathbb{E}\left[\mathbf{X}(\zeta)\,\mathbf{X}^H(\zeta)\right]\mathbf{A}^H = \mathbf{R_X}\,\mathbf{A}^H\end{aligned} \tag{M:3.2.42}$$

and hence $\mathbf{R_{YX}} = \mathbf{AR_X}$.

Cross-covariance matrices:

$$\mathbf{\Gamma_{XY}} = \mathbf{\Gamma_X}\,\mathbf{A}^H \quad \text{and} \quad \mathbf{\Gamma_{YX}} = \mathbf{A}\,\mathbf{\Gamma_X} \tag{M:3.2.43}$$

These results will be used to show what happens to a Gaussian random vector under a linear transformation in Section 4.5.



New slide

4.4.3 Auxiliary Variables

The density of a RV that is *one* function $Z(\zeta) = g(X(\zeta), Y(\zeta))$ of two RVs can be determined from the results above, by choosing a convenient **auxiliary variable**. The choice of this auxiliary variable comes with experience, but usually the simpler the better.

Examples might be $W(\zeta) = X(\zeta)$ or $W(\zeta) = Y(\zeta)$. The density of the function $Z(\zeta)$ can then be found by **marginalisation**:

$$f_Z(z) = \int_{\mathbb{R}} f_{WZ}(w, z)\, dw = \sum_{m=1}^{M} \int_{\mathbb{R}} \frac{f_{\mathbf{XY}}(x_m, y_m)}{|J(x_m, y_m)|}\, dw \tag{4.83}$$

Example 4.6 (Sum of two RVs). If $X(\zeta)$ and $Y(\zeta)$ have joint-pdf $f_{XY}(x, y)$, find the pdf of the RV $Z(\zeta) = aX(\zeta) + bY(\zeta)$ for constants a and b.

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

$$J(x, y) = \begin{vmatrix} \frac{\partial w}{\partial x} & \frac{\partial z}{\partial x} \\ \frac{\partial w}{\partial y} & \frac{\partial z}{\partial y} \end{vmatrix} = \begin{vmatrix} 0 & a \\ 1 & b \end{vmatrix} = -a \tag{4.84}$$

Hence, it follows that:

$$f_{WZ}(w, z) = \frac{1}{|a|} f_{XY}\left(\frac{z-bw}{a}, w\right) \tag{4.85}$$

Thus, it follows that:

$$f_Z(z) = \frac{1}{|a|} \int_{\mathbb{R}} f_{XY}\left(\frac{z-bw}{a}, w\right) dw \tag{4.86}$$

□

KEYPOINT! (Choosing the auxiliary variable). Note that you might be concerned about the choice of the auxiliary variable, and what happens if you chose something different to that used here. The answer is that, as long as the auxliary variable is a function of at least one of the RVs, then it doesn't really matter, as the **marginalisation** stage will usually yield the same answer. An example is discussed in Sidebar 5 on page 90. Nevertheless, it usally pays to chose the auxiliary variable carefully to minimise any difficulties in evaluating the marginal-pdf.

Example 4.7 ([Papoulis:1991, Page 149, Problem 6-8]). The RVs $X(\zeta)$ and $Y(\zeta)$ are independent with Rayleigh densities:

$$f_X(x) = \frac{x}{\alpha^2} \exp\left\{-\frac{x^2}{2\alpha^2}\right\} \mathbb{I}_{\mathbb{R}^+}(x) \tag{4.96}$$

$$f_Y(y) = \frac{y}{\beta^2} \exp\left\{-\frac{y^2}{2\beta^2}\right\} \mathbb{I}_{\mathbb{R}^+}(y) \tag{4.97}$$

1. Show that if $Z(\zeta) = \frac{X(\zeta)}{Y(\zeta)}$, then:

$$f_Z(z) = \frac{2\alpha^2}{\beta^2} \frac{z}{\left(z^2 + \frac{\alpha^2}{\beta^2}\right)^2} \mathbb{I}_{\mathbb{R}^+}(z) \tag{4.98}$$

2. Using this result, show that for any $k > 0$,

$$\Pr\left(X(\zeta) \le k\,Y(\zeta)\right) = \frac{k^2}{k^2 + \frac{\alpha^2}{\beta^2}} \tag{4.99}$$

SOLUTION. Considering the first part of the question, then choose the auxiliary variable as $W(\zeta) = X(\zeta)$, then the system $z = \frac{x}{y}$, $w = x$ has the single solution $x = w$, $y = \frac{w}{z}$. The Jacobian is given by:

$$J(x, y) = \text{abs}\begin{vmatrix} \frac{\partial w}{\partial x} & \frac{\partial z}{\partial x} \\ \frac{\partial w}{\partial y} & \frac{\partial z}{\partial y} \end{vmatrix} = \text{abs}\begin{vmatrix} 1 & \frac{1}{y} \\ 0 & -\frac{x}{y^2} \end{vmatrix} = \text{abs}\left|-\frac{x}{y^2}\right| = \left|\frac{z^2}{w}\right| \tag{4.100}$$

The RVs $X(\zeta)$ and $Y(\zeta)$ only take on positive values, since they are Rayleigh distribution, and therefore in this case the Jacobian *can* be simplified to

$$J(x, y) = \frac{z^2}{w} \tag{4.101}$$

Hence, since $X(\zeta)$ and $Y(\zeta)$ are independent,

$$f_{WZ}(w, z) = \frac{w}{z^2} f_X(w)\, f_Y\left(\frac{w}{z}\right) \tag{4.102}$$

$$= \frac{1}{\alpha^2\beta^2} \frac{w^3}{z^3} \exp\left\{-\frac{w^2}{2}\left(\frac{1}{\alpha^2} + \frac{1}{z^2\beta^2}\right)\right\} \mathbb{I}_{\mathbb{R}^+\times\mathbb{R}^+}(w, z) \tag{4.103}$$

$$= \frac{\hat{\alpha}^2}{z^3\alpha^2\beta^2}\left[w^2\, \frac{w}{\hat{\alpha}^2} \exp\left\{-\frac{w^2}{2\hat{\alpha}^2}\right\}\right] \mathbb{I}_{\mathbb{R}^+\times\mathbb{R}^+}(w, z) \tag{4.104}$$

Sidebar 5 What if you chose a complicated auxiliary variable?

Consider Example 4.6 and suppose that rather than chosing $W(\zeta) = Y(\zeta)$, you accidentally chose something more complicated such as:

$$W(\zeta) = \frac{X(\zeta)}{Y(\zeta)} \tag{4.87}$$

Will the resulting expression for $f_Z(z)$ be the same as Equation 4.86? The answer can be seen through an example, or a more detailed generic analysis. Here, we show an example. While the joint-pdf $f_{WZ}(w,\, z)$ will be different from Equation 4.85, it is the **marginalisation** stage that ensures the expressions for $f_Z(z)$ are the same. For the auxiliary variable shown in Equation 4.87, noting that $Z(\zeta) = aX(\zeta) + bY(\zeta)$, then

$$x = w\,y \quad \Rightarrow \quad z = awy + by = y(aw+b) \tag{4.88}$$

$$y = \frac{z}{aw+b}, \quad x = \frac{wz}{aw+b} \tag{4.89}$$

The Jacobian is given by:

$$J = \operatorname{abs} \begin{bmatrix} \frac{\partial z}{\partial x} & \frac{\partial z}{\partial y} \\ \frac{\partial w}{\partial x} & \frac{\partial w}{\partial y} \end{bmatrix} = \operatorname{abs} \begin{bmatrix} a & b \\ \frac{1}{y} & -\frac{x}{y^2} \end{bmatrix} \tag{4.90}$$

$$= \operatorname{abs} \frac{ax+by}{y^2} = \operatorname{abs} \frac{z}{y^2} = \operatorname{abs} \frac{(aw+b)^2}{z} \tag{4.91}$$

For simplicity, assume that $(x,\, y) > 0$.[a] Then, the joint-pdf is given by:

$$f_{WZ}(w,\, z) = \frac{z}{(b+aw)^2} f_{XY}\left(\frac{wz}{aw+b}, \frac{z}{aw+b}\right) \tag{4.92}$$

This is clearly different to that in Equation 4.85. However, the marginal for $Z(\zeta)$ is:

$$f_Z(z) = \int \frac{z}{(b+aw)^2} f_{XY}\left(\frac{wz}{aw+b}, \frac{z}{aw+b}\right) dw \tag{4.93}$$

Let $\theta = \frac{z}{aw+b}$, such that $d\theta = -\frac{az}{(aw+b)^2}\, dw$, and also note that

$$\frac{wz}{aw+b} = \theta\, w = \theta\left(\frac{z-b\theta}{\theta\, a}\right) = \frac{z-b\theta}{a} \tag{4.94}$$

Substituting into Equation 4.93, and noting that the minus sign in the differential term will get absorbed into the limits of the integral, then Equation 4.93 becomes:

$$f_Z(z) = \frac{1}{a} \int f_{XY}\left(\frac{z-b\theta}{a},\, \theta\right) d\theta \tag{4.95}$$

which is indeed equivalent to Equation 4.86.

[a] This ensures that it is not necessary to worry about the absolute value of the Jacobian. Depending on the range of values that $X(\zeta)$ and $Y(\zeta)$ take on, this proof will need to be tightened up to take account of the absolute value of the Jacobian.

where $\hat{\alpha}^2 = \alpha^2 \frac{z^2}{z^2+\frac{\alpha^2}{\beta^2}}$. Integrating over all values of w gives:

$$f_Z(z) = \int_{\mathbb{R}^+} f_{XZ}(w,\, z)\, dw = \frac{\hat{\alpha}^2}{z^3 \alpha^2 \beta^2} \int_0^\infty w^2 \, \frac{w}{\hat{\alpha}^2} \exp\left\{ -\frac{w^2}{2\hat{\alpha}^2} \right\} dw \tag{4.105}$$

The integral is the **second moment** of a Rayleigh distribution. It can be shown that

$$\int_0^\infty w^2 \, \frac{w}{\hat{\alpha}^2} \exp\left\{ -\frac{w^2}{2\hat{\alpha}^2} \right\} dw = 2\hat{\alpha}^2 \tag{4.106}$$

Finally, therefore,

$$f_Z(z) = \frac{2\hat{\alpha}^4}{z^3 \alpha^2 \beta^2} \mathbb{I}_{\mathbb{R}^+}(z) = \frac{2\alpha^2}{\beta^2} \frac{z}{\left(z^2 + \frac{\alpha^2}{\beta^2}\right)^2} \mathbb{I}_{\mathbb{R}^+}(z) \tag{4.107}$$

For the second part of the question, notice that:

$$\Pr\left(X(\zeta) \le kY(\zeta)\right) = \Pr\left(Z(\zeta) \le k\right) = \int_0^k f_Z(z)\, dz \tag{4.108}$$

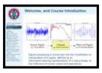
$$= \frac{\alpha^2}{\beta^2} \int_0^k \frac{2z}{\left(z^2 + \frac{\alpha^2}{\beta^2}\right)^2} dz = -\frac{\alpha^2}{\beta^2} \left[\frac{1}{z^2 + \frac{\alpha^2}{\beta^2}} \right]_0^k \tag{4.109}$$

$$= \frac{\alpha^2}{\beta^2} \left[\frac{1}{\frac{\alpha^2}{\beta^2}} - \frac{1}{k^2 + \frac{\alpha^2}{\beta^2}} \right] = 1 - \frac{\frac{\alpha^2}{\beta^2}}{k^2 + \frac{\alpha^2}{\beta^2}} \tag{4.110}$$

□

which gives the desired result when these fractions are combined.

4.5 Multivariate Gaussian Density Function



New slide

Gaussian random vectors and Gaussian random sequences, as will be seen in the following handouts, 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:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{N}{2}} \left|\mathbf{\Gamma}_{\mathbf{X}}\right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} \left(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}\right)^T \mathbf{\Gamma}_{\mathbf{X}}^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}\right) \right] \tag{M:3.2.44}$$

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

$$f_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_{\mathbf{X}},\, \mathbf{\Gamma}_{\mathbf{X}}\right) \tag{4.111}$$

Note the difference between the notation used here, and the notation used to indicate when a random vector is distributed, or drawn, from a normal distribution:

$$\mathbf{X}(\zeta) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Gamma}_{\mathbf{X}}) \tag{4.112}$$

The term in the exponent of Equation M:3.2.44 is a positive definite quadratic function of x_n, and can be written as:

$$(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})^T \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \langle \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} \rangle_{ij} (x_i - \mu_i)(x_j - \mu_j) \tag{M:3.2.45}$$

where $\langle \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} \rangle_{ij}$ denotes the (i, j)th element of $\boldsymbol{\Gamma}_{\mathbf{X}}^{-1}$. It is therefore straightforward to calculate the marginal distribution for the RV $X_n(\zeta)$ by marginalising over all the other RVs. The details are left as an exercise for the reader.

The complex-valued normal random vector has pdf:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\pi^N |\boldsymbol{\Gamma}_{\mathbf{X}}|} \exp\left[-(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})^H \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})\right] \tag{M:3.2.47}$$

again with mean $\boldsymbol{\mu}_{\mathbf{X}}$ and covariance $\boldsymbol{\Gamma}_{\mathbf{X}}$. For a more detail discussion of complex random variables, see [Therrien:1991].

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 $\boldsymbol{\Gamma}_{\mathbf{X}}$. All other higher-order moments can be obtained from these parameters.

   Theorem 4.2 (Moments of a Gaussian RV). The moments of a Gaussian RV $X(\zeta) \sim \mathcal{N}(0, \sigma_x^2)$, are given by:

   $$\mathbb{E}\left[X^k(\zeta)\right] = \begin{cases} 1 \cdot 3 \cdots (k-1)\sigma_x^k & k \text{ even} \\ 0 & k \text{ odd} \end{cases} \tag{4.113}$$

   PROOF. Since $f_X(x)$ is an even function, then it follows that the odd moments are zero. The proof for the even moments then follows by using integration by parts to obtain a recursive relationship between $\mathbb{E}\left[X^k(\zeta)\right]$ and $\mathbb{E}\left[X^{k+2}(\zeta)\right]$. This is left as an exercise for the reader.

   This theorem can be extended to the multivariate case.

2. If the components of $\mathbf{X}(\zeta)$ are mutually uncorrelated, then they are also independent. This property has an important consequence in **blind signal separation** or **independent component analysis (ICA)**.

3. A linear transformation of a normal random vector is also normal. This can readily be seen as follows, where the proof assumes a real normal random vector; the proof for a complex normal random vector follows a similar line. Noting that for a linear transformation,

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{f_{\mathbf{X}}\left(\mathbf{A}^{-1}\mathbf{y}\right)}{|\det \mathbf{A}|} \tag{M:3.2.34}$$

then if $f_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \,\middle|\, \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Gamma}_{\mathbf{X}}\right)$, it follows:

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|\det \mathbf{A}|} \frac{1}{(2\pi)^{\frac{N}{2}} |\boldsymbol{\Gamma}_{\mathbf{X}}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\mathbf{A}^{-1}\mathbf{y} - \boldsymbol{\mu}_{\mathbf{X}}\right)^T \boldsymbol{\Gamma}_{\mathbf{X}}^{-1}\left(\mathbf{A}^{-1}\mathbf{y} - \boldsymbol{\mu}_{\mathbf{X}}\right)\right] \tag{4.114}$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}} \left|\mathbf{A}\boldsymbol{\Gamma}_{\mathbf{X}}\mathbf{A}^T\right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\mathbf{y} - \mathbf{A}\boldsymbol{\mu}_{\mathbf{X}}\right)^T \mathbf{A}^{-T}\boldsymbol{\Gamma}_{\mathbf{X}}^{-1}\mathbf{A}^{-1}\left(\mathbf{y} - \mathbf{A}\boldsymbol{\mu}_{\mathbf{X}}\right)\right] \tag{4.115}$$

where it has been noted that $\left|\mathbf{A}\boldsymbol{\Gamma}_{\mathbf{X}}\mathbf{A}^T\right|^{\frac{1}{2}} = |\mathbf{A}||\boldsymbol{\Gamma}_{\mathbf{X}}|^{\frac{1}{2}}$. Thus, using the expressions for $\boldsymbol{\mu}_{\mathbf{Y}}$ and $\boldsymbol{\Gamma}_{\mathbf{Y}}$ above, it directly follows that

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\boldsymbol{\Gamma}_{\mathbf{Y}}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}\right)^T \boldsymbol{\Gamma}_{\mathbf{Y}}^{-1}\left(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}\right)\right] \tag{4.116}$$

$$= \mathcal{N}\left(\mathbf{y} \,\middle|\, \boldsymbol{\mu}_{\mathbf{Y}}, \boldsymbol{\Gamma}_{\mathbf{Y}}\right) \tag{4.117}$$

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. This can be shown as follows, assuming real random vectors and that $\mathbf{X}(\zeta) \in \mathbb{R}^N$, $\mathbf{Y}(\zeta) \in \mathbb{R}^M$; as usual, a similar derivation follows for the complex case. Defining the joint random vector:

$$\mathbf{Z}(\zeta) = \begin{bmatrix} \mathbf{X}(\zeta) \\ \mathbf{Y}(\zeta) \end{bmatrix} \tag{T:2.101}$$

then the corresponding mean vector and covariance matrix is given by:

$$\boldsymbol{\mu}_{\mathbf{Z}} = \mathbb{E}\left[\begin{bmatrix} \mathbf{X}(\zeta) \\ \mathbf{Y}(\zeta) \end{bmatrix}\right] = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{X}} \\ \boldsymbol{\mu}_{\mathbf{Y}} \end{bmatrix} \tag{T:2.102}$$

$$\boldsymbol{\Gamma}_{\mathbf{Z}} = \mathbb{E}\left[\begin{bmatrix} \mathbf{X}(\zeta) - \boldsymbol{\mu}_{\mathbf{X}} \\ \mathbf{Y}(\zeta) - \boldsymbol{\mu}_{\mathbf{Y}} \end{bmatrix} \begin{bmatrix} \mathbf{X}(\zeta) - \boldsymbol{\mu}_{\mathbf{X}} & \mathbf{Y}(\zeta) - \boldsymbol{\mu}_{\mathbf{Y}} \end{bmatrix}^H\right] = \begin{bmatrix} \boldsymbol{\Gamma}_{\mathbf{X}} & \boldsymbol{\Gamma}_{\mathbf{XY}} \\ \boldsymbol{\Gamma}_{\mathbf{XY}}^H & \boldsymbol{\Gamma}_{\mathbf{Y}} \end{bmatrix} \tag{T:2.103}$$

Hence, the **joint-pdf** is given by:

$$f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{Z}}(\mathbf{z}) = \mathcal{N}\left(\mathbf{z} \,\middle|\, \boldsymbol{\mu}_{\mathbf{Z}}, \boldsymbol{\Gamma}_{\mathbf{Z}}\right) \tag{4.118}$$

$$= \frac{1}{(2\pi)^{\frac{N+M}{2}} |\boldsymbol{\Gamma}_{\mathbf{Z}}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{Z}}\right)^T \boldsymbol{\Gamma}_{\mathbf{Z}}^{-1}\left(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{Z}}\right)\right] \tag{4.119}$$

But by substituting for $\mathbf{z}$, $\boldsymbol{\mu}_{\mathbf{z}}$ and $\boldsymbol{\Gamma}_{\mathbf{z}}$ in terms of the $\mathbf{x}$ and $\mathbf{y}$ components and their respective means and covariances, it can be shown that the marginal densities are also Gaussian, where:

$$f_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \,\middle|\, \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Gamma}_{\mathbf{X}}\right) \tag{4.120}$$

$$f_{\mathbf{Y}}(\mathbf{y}) = \mathcal{N}\left(\mathbf{y} \,\middle|\, \boldsymbol{\mu}_{\mathbf{Y}}, \boldsymbol{\Gamma}_{\mathbf{Y}}\right) \tag{4.121}$$

Moreover, since

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) = \frac{f_{\mathbf{X}\mathbf{Y}}(\mathbf{x}, \mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})} \tag{T:2.39}$$

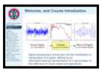
then the conditional density is also Gaussian, given by:

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) = \frac{1}{(2\pi)^{\frac{M}{2}} \left|\boldsymbol{\Gamma}_{\mathbf{Y}|\mathbf{X}}\right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}}\right)^T \boldsymbol{\Gamma}_{\mathbf{Y}|\mathbf{X}}^{-1} \left(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}}\right)\right] \tag{T:2.106}$$

where

$$\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Gamma}_{\mathbf{X}\mathbf{Y}}^{H} \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}) \tag{T:2.108}$$

$$\boldsymbol{\Gamma}_{\mathbf{Y}|\mathbf{X}} = \boldsymbol{\Gamma}_{\mathbf{Y}} - \boldsymbol{\Gamma}_{\mathbf{X}\mathbf{Y}}^{H} \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} \boldsymbol{\Gamma}_{\mathbf{X}\mathbf{Y}} \tag{T:2.109}$$



New slide

4.6 Characteristic Functions

The **characteristic function** and **moment generating function** for a scalar random variable can be extended to deal with random vectors. Essentially, these are defined as the multi-dimensional Fourier transform of the joint-pdf. Hence, the characteristic function is:

$$\Phi_{\mathbf{X}}(\boldsymbol{\xi}) \triangleq \mathbb{E}\left[e^{j\boldsymbol{\xi}^T \mathbf{X}(\zeta)}\right] = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})\, e^{j\boldsymbol{\xi}^T \mathbf{x}}\, d\mathbf{x} \tag{4.122}$$

Similarly, the moment generating function is given by:

$$\bar{\Phi}_{\mathbf{X}}(\mathbf{s}) \triangleq \mathbb{E}\left[e^{\mathbf{s}^T \mathbf{X}(\zeta)}\right] = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})\, e^{\mathbf{s}^T \mathbf{x}}\, d\mathbf{x} \tag{4.123}$$

The characteristic function for a real-valued Gaussian random vector is given by:

$$\Phi_{\mathbf{X}}(\boldsymbol{\xi}) = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})\, e^{j\boldsymbol{\xi}^T \mathbf{x}}\, d\mathbf{x} \tag{4.124}$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}} \left|\boldsymbol{\Gamma}_{\mathbf{X}}\right|^{\frac{1}{2}}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})^T \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})\right] e^{j\boldsymbol{\xi}^T \mathbf{x}}\, d\mathbf{x} \tag{4.125}$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}} \left|\boldsymbol{\Gamma}_{\mathbf{X}}\right|^{\frac{1}{2}}} \int_{-\infty}^{\infty} \exp\left[-\frac{\mathbf{x}^T \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} \mathbf{x} - 2\left(\boldsymbol{\mu}_{\mathbf{X}}^T \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} + j\boldsymbol{\xi}^T\right)\mathbf{x} + \boldsymbol{\mu}_{\mathbf{X}}^T \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} \boldsymbol{\mu}_{\mathbf{X}}}{2}\right] d\mathbf{x} \tag{4.126}$$

Using the integral identity:

$$\int_{\mathbb{R}^P} \exp\left\{-\frac{1}{2}\left[\alpha + 2\mathbf{y}^T\boldsymbol{\beta} + \mathbf{y}^T\boldsymbol{\Gamma}\mathbf{y}\right]\right\} d\mathbf{y} = \frac{(2\pi)^{\frac{P}{2}}}{|\boldsymbol{\Gamma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left[\alpha - \boldsymbol{\beta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\beta}\right]\right\} \tag{4.127}$$

where $\mathbf{y} \in \mathbb{R}^P$ is a P-dimensional column vector, then it follows by setting $\alpha = \boldsymbol{\mu}_{\mathbf{X}}^T\boldsymbol{\Gamma}_{\mathbf{X}}^{-1}\boldsymbol{\mu}_{\mathbf{X}}$, $\boldsymbol{\beta} = -\left(\boldsymbol{\Gamma}_{\mathbf{X}}^{-1}\boldsymbol{\mu}_{\mathbf{X}} + j\boldsymbol{\xi}\right)^T$, $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_{\mathbf{X}}^{-1}$, $\mathbf{y} = \mathbf{x}$ and $P = N$, that:

$$\Phi_{\mathbf{X}}(\boldsymbol{\xi}) = \exp\left[-\frac{1}{2}\left\{\boldsymbol{\mu}_{\mathbf{X}}^T\boldsymbol{\Gamma}_{\mathbf{X}}^{-1}\boldsymbol{\mu}_{\mathbf{X}} - \left(\boldsymbol{\mu}_{\mathbf{X}}^T\boldsymbol{\Gamma}_{\mathbf{X}}^{-1} + j\boldsymbol{\xi}^T\right)\boldsymbol{\Gamma}_{\mathbf{X}}\left(\boldsymbol{\Gamma}_{\mathbf{X}}^{-1}\boldsymbol{\mu}_{\mathbf{X}} + j\boldsymbol{\xi}\right)\right\}\right] \tag{4.128}$$

which after multiplying out gives:

$$\Phi_{\mathbf{X}}(\boldsymbol{\xi}) = \exp\left[j\boldsymbol{\xi}^T\boldsymbol{\mu}_{\mathbf{X}} - \frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\Gamma}_{\mathbf{X}}\boldsymbol{\xi}\right] \tag{M:3.2.46}$$

where, of course, $\boldsymbol{\xi}^T = [\xi_1, \dots, \xi_N]$. It can be shown that the characteristic function for the complex-valued normal random vector is given by

$$\Phi_{\mathbf{X}}(\boldsymbol{\xi}) = \exp\left[j\Re\{\boldsymbol{\xi}^H\boldsymbol{\mu}_{\mathbf{X}}\} - \frac{1}{4}\boldsymbol{\xi}^H\boldsymbol{\Gamma}_{\mathbf{X}}\boldsymbol{\xi}\right] \tag{M:3.2.50}$$

4.7 Higher-Order Statistics

Random vectors, and random processes as introduced in the forthcoming lectures, can also be characterised by higher-order moments. These, again, are a generalisation of the equivalent definitions for scalar-random variables. However, they become significantly more complicated for random vectors since the various products of the random variables creates a very large set of combinations. These will not be discussed in this course, although an introduction can be found in [Therrien:1992, Section 4.10.1]. As an example, taken from [Manolakis:2000, Page 89], it is noted that the fourth-order moment of a normal random vector

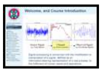
$$\mathbf{X}(\zeta) = \begin{bmatrix} X_1(\zeta) & X_2(\zeta) & X_3(\zeta) & X_4(\zeta) \end{bmatrix}^T \tag{4.129}$$

can be expressed in terms of its second order moments. For the real case when $\mathbf{X}(\zeta) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_{\mathbf{X}})$, then:

$$\begin{aligned}\mathbb{E}\left[X_1(\zeta)X_2(\zeta)X_3(\zeta)X_4(\zeta)\right] =& \mathbb{E}\left[X_1(\zeta)X_2(\zeta)\right]\mathbb{E}\left[X_3(\zeta)X_4(\zeta)\right] \\ &+ \mathbb{E}\left[X_1(\zeta)X_3(\zeta)\right]\mathbb{E}\left[X_2(\zeta)X_4(\zeta)\right] \\ &+ \mathbb{E}\left[X_1(\zeta)X_4(\zeta)\right]\mathbb{E}\left[X_2(\zeta)X_3(\zeta)\right]\end{aligned} \tag{M:3.2.53}$$

Note that each RV appears only once in each term. It is also possible to define **higher-order cumulants** which can be extremely useful; for example, they are identically zero for Gaussian random processes, which can help identify whether a process is Gaussian or not.

4.8 Sum of Independent Random Variables



New slide

Theorem 4.3 (Sum of Random Variables and Vectors). If $\mathbf{X}(\zeta)$ and $\mathbf{Y}(\zeta)$ have joint-pdf, $f_{\mathbf{XY}}(\mathbf{x},\, \mathbf{y})$, then $\mathbf{Z}(\zeta) = \mathbf{X}(\zeta) + \mathbf{Y}(\zeta)$ has density function:

$$f_{\mathbf{Z}}(\mathbf{z}) \triangleq f_{\mathbf{X}+\mathbf{Y}}(\mathbf{z}) = \int_{\mathbb{R}} f_{\mathbf{XY}}(\mathbf{x},\, \mathbf{z}-\mathbf{x})\, d\mathbf{x} \tag{4.130}$$

PROOF. Define the event $Z = \{(x,y) : x + y \leq z\}$. Then:

$$\Pr(\mathbf{X}+\mathbf{Y} \leq \mathbf{x}) = \iint_{Z} f_{\mathbf{XY}}(\mathbf{u},\, \mathbf{v})\, d\mathbf{u}\, d\mathbf{v} = \int_{\mathbf{v}\in\mathbb{R}} \int_{\mathbf{u}=-\infty}^{\mathbf{z}-\mathbf{v}} f_{\mathbf{XY}}(\mathbf{u},\, \mathbf{v})\, d\mathbf{u}\, d\mathbf{v} \tag{4.131}$$

and by making the substitution $\mathbf{w} = \mathbf{u} + \mathbf{v}$.

$$= \int_{\mathbf{v}\in\mathbb{R}} \int_{\mathbf{w}=-\infty}^{\mathbf{z}} f_{\mathbf{XY}}(\mathbf{w}-\mathbf{u},\, \mathbf{v})\, d\mathbf{w}\, d\mathbf{v} \tag{4.132}$$

$$= \int_{\mathbf{w}=-\infty}^{\mathbf{z}} \int_{\mathbf{v}\in\mathbb{R}} f_{\mathbf{XY}}(\mathbf{u},\, \mathbf{w}-\mathbf{v})\, d\mathbf{u}\, d\mathbf{v} \triangleq \int_{\mathbf{w}=-\infty}^{\mathbf{z}} f_{\mathbf{X}}(\mathbf{v})\, d\mathbf{v} \tag{4.133}$$

□

giving the result as required. This result can also be obtained using the probability transformation rule and an auxiliary variable. This is left as an exercise for the reader.

Theorem 4.4 (Sum of Independent Random Variables and Vectors). If $\mathbf{X}(\zeta)$ and $\mathbf{Y}(\zeta)$ are independent, this result becomes

$$f_{\mathbf{Z}}(\mathbf{z}) \triangleq f_{\mathbf{X}+\mathbf{Y}}(\mathbf{z}) = \int_{\mathbb{R}} f_{\mathbf{X}}(\mathbf{x})\, f_{\mathbf{Y}}(\mathbf{z}-\mathbf{x})\, d\mathbf{x} \tag{4.134}$$

$$= \int_{\mathbb{R}} f_{\mathbf{X}}(\mathbf{z}-\mathbf{y})\, f_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y} = f_{\mathbf{X}}(\mathbf{z}) * f_{\mathbf{Y}}(\mathbf{y}) \tag{4.135}$$

PROOF. Follows trivially by writing $f_{\mathbf{XY}}(\mathbf{x},\, \mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})\, f_{\mathbf{Y}}(\mathbf{y})$

Independent RVs can also be dealt with using **characteristic functions** as introduced in the lecture on scalar random variables.

If $Z(\zeta) = X(\zeta) + Y(\zeta)$, then its characteristic function is given by:

$$\Phi_Z(\xi) \triangleq \mathbb{E}\left[e^{j\xi\, Z(\zeta)}\right] = \mathbb{E}\left[e^{j\xi[X(\zeta)+Y(\zeta)]}\right] = \mathbb{E}\left[e^{j\xi\, X(\zeta)}\right] \mathbb{E}\left[e^{j\xi\, Y(\zeta)}\right] \tag{M:3.2.59}$$

where the last inequality follows from independence. More explicitly, observe that:

$$\Phi_Z(\xi) = \mathbb{E}\left[e^{j\xi[X(\zeta)+Y(\zeta)]}\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,\, y)\, e^{j\xi[x+y]}\, dx\, dy \tag{4.136}$$

and noting that due to independence $f_{XY}(x, y) = f_X(x)\, f_Y(y)$, then

$$\Phi_Z(\xi) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_X(x)\, f_Y(y)\, e^{j\xi\, x} e^{j\xi\, y}\, dx\, dy \tag{4.137}$$

$$= \left\{\int_{-\infty}^{\infty} f_X(x)\, e^{j\xi\, x}\, dx\right\}\left\{\int_{-\infty}^{\infty} f_Y(y)\, e^{j\xi\, y}\, dy\right\} \tag{4.138}$$

giving the desired result.

Hence, from the convolution property of the Fourier transform, it follows directly from this result that

$$f_Z(z) = f_X(x) * f_Y(y) \tag{M:3.2.61}$$

This result can be generalised to the summation of M independent RVs:

$$Y(\zeta) = \sum_{k=1}^{M} c_k\, X_k(\zeta) \tag{M:3.2.55}$$

where $\{c_k\}_1^M$ is a set of fixed (deterministic) coefficients.

It follows straightforwardly that:

$$\Phi_Y(\xi) \triangleq \mathbb{E}\left[e^{j\xi\, Y(\zeta)}\right] = \prod_{k=1}^{M} \mathbb{E}\left[e^{j\xi\, c_k X_k(\zeta)}\right] = \prod_{k=1}^{M} \Phi_{X_k}(c_k \xi) \tag{M:3.2.72}$$

Hence, the pdf of $Y(\zeta)$ is given by:

$$f_Y(y) = \frac{1}{|c_1|} f_{X_1}\left(\frac{y}{c_1}\right) * \frac{1}{|c_2|} f_{X_2}\left(\frac{y}{c_2}\right) * \cdots * \frac{1}{|c_M|} f_{X_M}\left(\frac{y}{c_M}\right) \tag{M:3.2.73}$$

where, implicitly, the Fourier transform of a frequency scaled signal has been used, which is equivalent to using the probability transformation rule for a scalar random variable.

Theorem 4.5 (Mean and variance of sum of independent RVs). Using the linearity of the expectation operator, and taking expectations of both sides of Equation M:3.2.55, then:

$$\mu_Y = \sum_{k=1}^{M} c_k\, \mu_{X_k} \tag{M:3.2.56}$$

Moreover, assuming independence, then the variance of $Y(\zeta)$ is given by:

$$\sigma_Y^2 = \mathbb{E}\left[\left|\sum_{k=1}^{M} c_k\, \mu_{X_k} - \mu_{X_k}\right|^2\right] = \sum_{k=1}^{M} |c_k|^2\, \sigma_{X_k}^2 \tag{M:3.2.57}$$

PROOF. These results follow from the linearity of the expectation operator, and the independence property of the random variables. The proof is left as an exercise for the reader.

Finally, the cumulant generating, or second characteristic, function can be used to determine the nth-order cumulants for $Y(\zeta)$.

Recall that

$$\Psi_X(\xi) \triangleq \ln \Phi_X(\xi) = \ln \mathbb{E}\left[e^{k\,\xi\,X(\zeta)}\right] \tag{4.139}$$

Then, from Equation M:3.2.72,

$$\Psi_Y(\xi) \triangleq \ln \mathbb{E}\left[e^{j\xi\,Y(\zeta)}\right] = \sum_{k=1}^{M} \ln \mathbb{E}\left[e^{j\xi\,c_k X_k(\zeta)}\right] = \sum_{k=1}^{M} \Psi_{X_k}(c_k\xi) \tag{M:3.2.74}$$

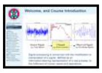
Therefore, it can readily be shown that the cumulants of $Y(\zeta)$ are given by:

$$\kappa_Y^{(n)} = \sum_{k=1}^{M} c_k^n\, \kappa_{X_k}^{(n)} \tag{M:3.2.75}$$

It is left as an exercise for the reader to demonstrate this.

When these results are extended to the sum of an infinite number of statistically independent random variables, a powerful theorem known as the central limit theorem (CLT) is obtained.

Another interesting concept develops when the sum of i. i. d. random variables preserve their distribution, which results in so-called **stable distributions**. Examples are the Gaussian and Cauchy distributions.



4.8.1 Central limit theorem

New slide

Consider the random variable $Y(\zeta)$ given by:

$$Y_M(\zeta) = \sum_{k=1}^{M} c_k\, X_k(\zeta) \tag{M:3.2.55}$$

What is the distribution of $Y_M(\zeta)$ as $M \to \infty$?

If $Y_M(\zeta)$ is a sum of i. i. d. RVs with a stable distribution, the distribution of $Y_M(\zeta)$ also converges to a stable distribution. If the distributions are not stable and, in particular, have finite variance, then the CLT reveals the distribution for $\lim_{M\to\infty} Y_M(\zeta)$.

Theorem 4.6 (Central limit theorem). Let $\{X_k(\zeta)\}_{k=1}^M$ be a collection of RVs that are independent and identically distributed and for which the mean and variance of each RV exists and is finite, such that $\mu_X = \mu_{X_k} < \infty$ and $\sigma_X = \sigma_{X_k}^2 < \infty$ for all $k = \{1, \dots, M\}$. Then the distribution of the normalised random variable

$$\hat{Y}_M(\zeta) = \frac{Y_M(\zeta) - \mu_{Y_M}}{\sigma_{Y_M}} \quad \text{where} \quad Y_M(\zeta) = \sum_{k=1}^{M} X_k(\zeta) \tag{M:3.2.55}$$

approaches that of a normal random variable with zero mean and unit standard deviation as $M \to \infty$; in other words,

$$\lim_{M\to\infty} f_{\hat{Y}_M}(y) = \mathcal{N}\left(y \,\middle|\, 0,\, 1\right) \tag{4.140}$$

PROOF. Since the $X_k(\zeta)$'s are i. i. d., then $\mu_{Y_M} = M\mu_X$ and $\sigma^2_{Y_M} = M\sigma^2_X$. Let

$$Z_k(\zeta) = \frac{X_k(\zeta) - \mu_X}{\sigma_X} \tag{4.141}$$

such that $\mu_{Z_k} = \mu_Z = 0$, $\sigma^2_{Z_k} = \sigma^2_Z = 1$ and:

$$\hat{Y}_M(\zeta) = \frac{1}{\sqrt{M}} \sum_{k=1}^{M} Z_k(\zeta) \tag{4.142}$$

Noting that if $V(\zeta) = a\,U(\zeta)$ for some real-scalar a then

$$\Phi_V(\xi) = \mathbb{E}\left[e^{j\xi\, aU(\zeta)}\right] = \Phi_U(a\xi) \tag{4.143}$$

Hence, from Equation M:3.2.72, the characteristic function for $\hat{Y}_M(\zeta)$ is given by:

$$\Phi_{\hat{Y}_M}(\xi) = \prod_{k=1}^{M} \Phi_{Z_k}\left(\frac{\xi}{\sqrt{M}}\right) \tag{4.144}$$

Since the $X_k(\zeta)$'s and therefore the $Z_k(\zeta)$'s are i. i. d., then $\Phi_{Z_k}(\xi) = \Phi_Z(\xi)$, or:

$$\Phi_{\hat{Y}_M}(\xi) = \Phi_Z^M\left(\frac{\xi}{\sqrt{M}}\right) \tag{4.145}$$

From the previous chapter on scalar random variables,

$$\Phi_Z(\xi) = \mathbb{E}\left[e^{j\xi\, Z(\zeta)}\right] = \sum_{n=0}^{\infty} \frac{(j\xi)^n}{n!} \mathbb{E}\left[Z^n(\zeta)\right] \tag{4.146}$$

and therefore

$$\Phi_{\hat{Y}_M}(\xi) = \left\{\sum_{n=0}^{\infty} \frac{1}{n!}\left(\frac{j\xi}{\sqrt{M}}\right)^n \mathbb{E}\left[Z^n(\zeta)\right]\right\}^M = \left\{1 + \frac{j\xi\mu_Z}{\sqrt{M}} - \frac{\xi^2\sigma_Z^2}{2M} + \mathcal{O}\left(\left\{\frac{\xi}{\sqrt{M}}\right\}^3\right)\right\}^M \tag{4.147}$$

$$= \left\{1 - \frac{\xi^2}{2M} + \mathcal{O}\left(\left\{\frac{\xi}{\sqrt{M}}\right\}^3\right)\right\}^M \quad \to e^{-\frac{1}{2}\xi^2} \quad \text{as } M \to \infty \tag{4.148}$$

using the limit that:

$$\lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n = e^x \tag{4.149}$$

□

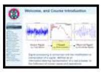
This last term is the characteristic function of the $\mathcal{N}\left(y \mid 0,\, 1\right)$ distribution.

5

Principles of Estimation Theory

This handout presents an introduction to estimation theory, including the notion of an estimator, measures of performance of the estimator (bias, variance, mean-squared error (MSE), the Cramér-Rao lower-bound (CRLB), and consistency). Discusses various estimators such as maximum-likelihood estimate (MLE), least-squares, and Bayesian estimators.

5.1 Introduction



New slide

- Thus far, the theory and material presented in this lecture course have assumed that either the probability density function (pdf) or statistical values, such as mean, covariance, or higher order statistics, associated with a problem are fully known. As a result, all required probabilities, and statistical functions could either be derived from a set of assumptions about a particular problem, or were given *a priori*.

- In most practical applications, this is the exception rather than the rule. In fact, unless the process by which observations, such as random values or vectors, are generated is known exactly, such that desired pdf or statistical properties could be theoretically calculated, there is absolutely no reason why they should be known *a priori*.

- The properties and parameters of random events must be obtained by collecting and analysing finite set of measurements. Again, it would be impossible or very rare indeed to known the ensemble of realisations of a sample space, and it will always be the case in practical applications that only a few realisations will ever be observed.

- This handout will consider the problem of **Parameter Estimation**. This refers to the estimation of a parameter that is fixed, but is unknown. For example, given a collection of observations that are known to be from a Gaussian distribution with unknown mean, estimate the mean from the observations.

5.1.1 A (Confusing) Note on Notation

Note that, unfortunately, from this point onwards, a slightly different (and abusive use of) notation for random quantities is used than what was presented in the first set of handouts. In the literature, the nth-order particular **observation** of a random variable are written as lower-case letters, possibly using subscripts such as x_n, but also often using square brackets, such as $x[n]$. This is all fine; except that for convenience, lower-case letters are often also used to mean the **random variable** itself meaning that, in different contexts, $x[n]$ can mean both a particular observation, as well as a potentially random value ($x[n] = X(\zeta)$. Where possible, upper-case letters are used to denote random elements, but this isn't always true.

The reason for this sloppiness is due to the notation used to describe **random processes** in the next lecture course, where the representation of a random process in the frequency domain is discussed, and upper-case letters are exclusively reserved to denote spectral representations. Moreover, lower-case letters for time-series are generally more recognisable and readable, and helps with the clarity of the presentation (where, as will be seen, $x[n]$ is short-hand notation for $x[n, \zeta]$).

Since this handout leads onto the notation of stochastic processes in the next course, this sloppy notation will be introduced now, but note that where the existing notation can be used without ambiguity in exam questions, it will be.

5.1.2 Examples of parameter estimation

To motivate this handout, this section lists a number of potential problems in which parameters might wish to be estimated.

Frequency Estimation Consider estimating the spectral content of a harmonic process $x[n]$ consisting of a single-tone, given by

$$x[n] = A_0 \cos(\omega_0 n + \phi_0) + w[n] \tag{5.1}$$

where A_0, ϕ_0, and ω_0 are *unknown* constants, and where $w[n]$ is an additive white Gaussian noise (AWGN) process with zero-mean and variance σ^2. It is desired to estimate the unknown constants, namely the amplitude A_0, phase ϕ_0, and frequency ω_0 from a realisation of the random process $x[n]$.

Sampling Distribution Parameters It is known that a set of observations, $\{x[n]\}_0^{N-1}$, are drawn from a sampling distribution with unknown parameters $\boldsymbol{\theta}$, such that:

$$x[n] \sim f_X(x \mid \boldsymbol{\theta}) \tag{5.2}$$

Sidebar 6 The taxi-cab problem

The following **taxicab problem** has been part of the orally transmitted folklore in the area of elementary parameter estimation for several decades [Jaynes:2003, Page 190], and is essentially an application of estimating the parameters of a sampling distribution from a small sample size.

It goes as follows: you are travelling on a night train; on awakening from sleep, you notice that the train has stopped at some unknown town, and all you can see is a taxicab with the number 27 on it. What, then, is your guess as to the number N of taxicabs in the town, which would in turn give a clue as to the size of the town?

Many people intuitively answer that there seems to be something about the choice $N_{est} = 2 \times 27 = 54$ that recommends itself; but few can offer a convincing rationale for this. The obvious *model* that seems to apply is that there will be N taxicabs numbered 1 through N, and, given N, the taxicab observed is equally likely to be any of them. Given that model, it is deductively known that $N \geq 27$, but from that point on, the reasoning depends on what metric is being used for deciding what a good estimator is.

If the problem seems to abstract by virtue of just one observation, consider observing a number of taxi's, say 2 or 3 taxi's with numbers 27, 13, and 28. Now what would your estimate be, and how many taxi's would you prefer to see before estimating the value of N?

This problem might seem rather academic, but has actually in the past been far from it.

For example, if it is known that $x[n] \sim \mathcal{U}_{[a,\,b]}$, then it might be of interest to estimate the parameters a and b.

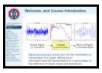
Estimate of Moments It might be of interest to estimate the moments of a set of observations, $\{x[n]\}_0^{N-1}$, for example $\mu_X = \mathbb{E}[x[n]]$ and $\sigma_X^2 = \text{var}[x[n]]$.

Constant value in noise An example which covers the various cases above is estimating a "direct current" (DC) constant in noise:

$$x[n] = A + w[n]\,, \quad n \in \{0, \dots, N-1\} \tag{5.3}$$

This list isn't exhaustiive, but gives an example of the type of **parameter estimation** problems that need to be addressed.

5.2 Properties of Estimators



New slide

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:

$$\hat{\theta} = \hat{\theta}[\mathcal{X}] = \hat{\theta}\left[\{x[n]\}_0^{N-1}\right] \tag{5.4}$$

Sidebar 7 German Tank Problem

In the statistical theory of estimation, the problem of estimating the maximum of a discrete uniform distribution from sampling without replacement is known in English as the **German tank problem**, due to its application in World War II to the estimation of the number of German tanks.

In this scenario, an *intelligence officer* has spotted a number of enermy tanks, with serial numbers that were assumed to be sequentially numbered from 1 to N. Given these observations, what is the prediction of the number of tanks produced? `http://en.wikipedia.org/wiki/German_tank_problem`

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 random variable (RV), and thus has a mean and variance. As an example of an estimator, consider estimating the mean μ_X of a random variate, $X(\zeta)$, from N observations $\mathcal{X} = \{x[n]\}_0^{N-1}$. The most natural estimator is a simple arithmetic average of these observations, given by the **sample mean**:

$$\hat{\mu}_X = \hat{\theta}[\mathcal{X}] = \frac{1}{N}\sum_{n=0}^{N-1} x[n] \tag{M:3.6.1}$$

Similarly, a natural estimator of the variance, σ_X^2, of the random variable $X(\zeta)$, $x[n]$, would be:

$$\hat{\sigma}_X^2 = \hat{\theta}'[\mathcal{X}] = \frac{1}{N}\sum_{n=0}^{N-1} \left(x[n] - \hat{\mu}_X\right)^2 \tag{M:3.6.2}$$

Thus, to demonstrate that these estimates are RVs, consider repeating the procedure for calculating the sample mean and sample variance from a large number of difference sets of realisations. Then a large number of estimates of μ_X and σ_X^2, denoted by the set $\{\hat{\mu}_X\}$ and $\{\hat{\sigma}_X^2\}$ respectively, is obtained, and these can be used to generate a histogram showing the distribution of the estimates.

The set of N observations, $\{x[n]\}_{n=0}^{N-1}$ can be regarded as one realisation of the random process $\{x[n,\zeta]\}_{n=0}^{N-1}$ which, technically, is defined on an N-dimensional sample space. Hence, the estimator $\hat{\theta}\left[\{x[n,\zeta]\}_0^{N-1}\right]$ becomes a RV whose probability density function can be obtained from the joint-pdf of the random variables $\{x[n,\zeta]\}_0^{N-1}$ using the probability transformation rule. This distribution is called the **sampling distribution** of the estimator, and is a fundamental concept in estimation theory because it provides all the information needed to evaluate the quality of an estimator.

Now, the sampling distribution of a *good* estimator should be concentrated as closely as possible around the parameter that it estimates. To determine how *good* an estimator is, and how different estimators of the same parameter compare with one another,

Sidebar 8 Expectation w. r. t. what?

Note that the expectation is taken with respect to the pdf of the data $\mathcal{X}$, denoted by $p\left(\mathcal{X} \mid \theta\right)$. Thus, more precisely one would write:

$$B(\hat{\theta}) \triangleq \mathbb{E}_{p(\mathcal{X} \mid \theta)}\left[\hat{\theta}\right] - \theta \tag{5.5}$$

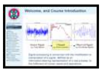
where

$$\mathbb{E}_{p(\mathcal{X} \mid \theta)}\left[\hat{\theta}\right] \triangleq \int_{\Theta} \hat{\theta}\left(\mathcal{X}\right) p\left(\mathcal{X} \mid \theta\right)\, d\mathcal{X} \tag{5.6}$$

However, often in textbooks and the literature, the pdf with which the expecation is taken against is omitted.

it is necessary to determine their sampling distributions. Of course, in practice, the joint-pdf for the random process $x[n,\zeta]$ is rarely known, so frequently it is not possible to obtain the sampling distribution. However, it is possible to estimate the statistical properties of the sampling distribution, such as lower-order moments (mean, variance, mean-squared error, and so forth), and that is the subject of this handout.

5.2.1 Bias of estimator



New slide

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

$$B(\hat{\theta}) \triangleq \mathbb{E}\left[\hat{\theta}\right] - \theta \tag{M:3.6.3}$$

It is important to appreciate that the expectation is taken with respect to (w. r. t.) the observed data *given* the true parameter θ.

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:

$$\epsilon_b(\hat{\theta}) \triangleq \frac{B(\hat{\theta})}{\theta} = \frac{\mathbb{E}\left[\hat{\theta}\right]}{\theta} - 1, \quad \theta \neq 0 \tag{M:3.6.4}$$

Example 5.1 (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 $\mathbb{E}\left[\hat{\mu}_x\right] = \mathbb{E}\left[\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] = \frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left[x[n]\right] = \frac{N\mu_X}{N} = \mu_X$.

When $B(\hat{\theta}) = 0$, the estimator is said to be **unbiased** and the pdf of the estimator is centered exactly at the true value of θ. Generally, estimators that are unbiased should be selected, such as the sample mean above, or very nearly unbiased. However, as will be seen later, it is not always wise to select an unbiased estimator. That an estimator is unbiased does not necessarily mean that it is a good estimator, only that it guarantees *on* *average* that it will attain the true value. It might have a higher variance, as discussed below, than a biased estimator. On the other hand, biased estimators are ones that are characterised by a systematic error, which presumably should not be present, and a persistent bias will always result in a poor estimator.

[Therrien:1992, Section 6.1.3, Page 290] gives a more formal definition of unbiasedness, and this is as follows:

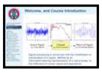
Definition 5.1 (Bias of an estimator). An estimate $\hat{\theta}_N$, based on N data observations, of a parameter θ is **unbiased** if

$$\mathbb{E}\left[\hat{\theta}_N\right] = \theta \tag{5.7}$$

Otherwise, the estimate is **biased** with bias $B(\hat{\theta}_N) = \mathbb{E}\left[\hat{\theta}_N\right] - \theta$. An estimate is **asymptotically unbiased** if

$$\lim_{N\to\infty} \mathbb{E}\left[\hat{\theta}_N\right] = \theta \tag{5.8}$$

♢



New slide

5.2.2 Variance of estimator

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

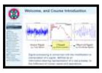
$$\operatorname{var}\left[\hat{\theta}\right] = \sigma_{\hat{\theta}}^2 \triangleq \mathbb{E}\left[\left|\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right|^2\right] \tag{M:3.6.5}$$

This, as with any variance value, measures the *spread* of the pdf of $\hat{\theta}$ around the mean. Therefore, it would, at first sight, seem sensible to select an estimate with the smallest variance. 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:

$$\epsilon_r \triangleq \frac{\sigma_{\hat{\theta}}}{\theta}, \quad \theta \neq 0 \tag{M:3.6.6}$$



New slide

5.2.3 Mean square error

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

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[\left|\hat{\theta} - \theta\right|^2\right] = \sigma_{\hat{\theta}}^2 + |B(\hat{\theta})|^2 \tag{M:3.6.7}$$

Again, it is important to remember that the expectation in the MSE term is w. r. t. the data, x, as discussed in Sidebar 8 page 105.

PROOF (RELATIONSHIP BETWEEN MSE, VARIANCE AND BIAS OF AN ESTIMATOR.). Rewriting Equation M:3.6.7 by substracting and adding the mean of the estimator gives:

$$\begin{aligned}\text{MSE}(\hat{\theta}) &= \mathbb{E}\left[|\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right] - (\theta - \mathbb{E}\left[\hat{\theta}\right])|^2\right] && (5.9)\\ &= \mathbb{E}\left[|\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]|^2\right] - \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right])^*(\theta - \mathbb{E}\left[\hat{\theta}\right])\right] && (5.10)\\ &\quad - \mathbb{E}\left[(\theta - \mathbb{E}\left[\hat{\theta}\right])(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right])^*\right] + \mathbb{E}\left[|\theta - \mathbb{E}\left[\hat{\theta}\right]|^2\right] && (5.11)\end{aligned}$$

Now, note that $\mathbb{E}\left[|\theta - \mathbb{E}\left[\hat{\theta}\right]|^2\right] = |\theta - \mathbb{E}\left[\hat{\theta}\right])|^2$, since both θ and $\mathbb{E}\left[\hat{\theta}\right]$ are deterministic values. Moreover,

$$\begin{aligned}\mathbb{E}\left[(\theta - \mathbb{E}\left[\hat{\theta}\right])^*(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right])\right] &= (\theta - \mathbb{E}\left[\hat{\theta}\right])^*\mathbb{E}\left[\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right] && (5.12)\\ &= (\theta - \mathbb{E}\left[\hat{\theta}\right])^*\left\{\mathbb{E}\left[\hat{\theta}\right] - \mathbb{E}\left[\hat{\theta}\right]\right\} = 0 && (5.13)\end{aligned}$$

giving:

$$\text{MSE}(\hat{\theta}) = \underbrace{\mathbb{E}\left[|\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]|^2\right]}_{\sigma^2_{\hat{\theta}}} + \underbrace{|\theta - \mathbb{E}\left[\hat{\theta}\right]|^2}_{B(\hat{\theta})} \qquad \text{(M:3.6.9)}$$

□

as required.

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

$$\hat{\theta}_{MSE} = \arg_{\hat{\theta}} \min \text{MSE}(\hat{\theta}) \qquad (5.14)$$

This measures the average mean squared deviation of the estimator from its true value. Unfortunately, the last expression in the right hand side (RHS) of Equation M:3.6.7 indicates that adoption of this natural criterion leads to unrealisable estimators; ones which cannot be written solely as a function of the data.

To see how this problem arises, note from Equation M:3.6.7 that the MSE is composed of errors due to the variance of the estimator, as well as the bias. This inevitable leads to an optimal estimator that is a function of the true parameter value.

Note that when finding the minimum MSE through application of Equation 5.14, the argument (or parameter) that is minimised is usually a parameter that defines the structure of the **estimator** and is not necessarily the unknown parameter of interest. Thus, a parameter α might affect the functional form of the estimator such that $\hat{\theta} = \hat{\theta}[\mathcal{X}, \alpha]$, and it is actually α that is used as the variable parameter in the optimisation. The following example demonstrates these issues.

Example 5.2 ([Kay:1993, Example 2.1, Page 16]). Consider the observations

$$x[n] = A + w[n], \quad n \in \{0, \dots, N-1\} \qquad \text{(K:2.2)}$$

where A is the parameter to be estimated, and $w[n]$ is white Gaussian noise (WGN). The parameter A can take on any value in the interval $-\infty < A < \infty$. A reasonable estimator for the average value of $x[n]$ is:

$$\hat{A}_a = a\,\frac{1}{N}\sum_{n=0}^{N-1} x[n] \tag{5.15}$$

If $a = 1$, then this is just the sample mean. Due to the linearity properties of the expectation operator, then it can be seen, as in the previous example, that:

$$\mathbb{E}\left[\hat{A}_1\right] = \mathbb{E}\left[\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] = \frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left[x[n]\right] = A \tag{5.16}$$

for all A. Hence, the sample mean is unbiased. However, consider finding the optimal (modified) estimator $\hat{A}_a$ by finding the value of a that minimises the MSE. Then noting that:

$$\mathbb{E}\left[\hat{A}_a\right] = \mathbb{E}\left[a\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] = aA \tag{5.17}$$

which, incidentally is a **biased estimate**, and also noting that the samples $x[n]$ are independent and identically distributed (i. i. d.) such that $\mathbb{E}\left[x[n]\,x[m]\right] = \sigma^2\,\delta(n-m) + A^2$ since $\sigma_x^2 = \sigma^2$, then:

$$\operatorname{var}\left[\hat{A}_a\right] = \mathbb{E}\left[\left|\left\{a\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right\} - aA\right|^2\right] \tag{5.18}$$

$$= \frac{a^2}{N^2}\mathbb{E}\left[\sum_{n=0}^{N-1}\sum_{m=0}^{N-1} x[n]\,x[m] - 2AN\sum_{n=0}^{N-1} x[n] + N^2A^2\right] \tag{5.19}$$

$$= \frac{a^2}{N^2}\left\{N\left[\sigma^2 + NA^2\right] - 2N^2A^2 + N^2A^2\right\} = \frac{a^2\sigma^2}{N} \tag{5.20}$$

Hence, the MSE is given by:

$$\text{MSE}(\hat{A}_a) = \operatorname{var}\left[\hat{A}_a\right] + |B(\hat{A}_a)|^2 = \frac{a^2\sigma^2}{N} + (a-1)^2A^2 \tag{5.21}$$

In order to find the minimum mean-square error (MMSE), then differentiate this and set to zero:

$$\frac{d\text{MSE}(\hat{A}_a)}{da} = \frac{2a\sigma^2}{N} + 2(a-1)A^2 \tag{5.22}$$

which is equal to zero when

$$a_{\text{opt}} = \frac{A^2}{A^2 + \frac{\sigma^2}{N}} \tag{5.23}$$

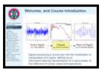
⋈

Thus, unfortunately, the optimal value of a depends upon the unknown parameter A. The estimator is therefore not realisable, and this is since the bias term is a function of A. It would therefore seem that any criterion which depends on the bias of the estimator will lead to an unrealisable estimator. Although this is generally true, on occasion realisable MMSE estimators can be found.

From a practical viewpoint, therefore, the MMSE estimator needs to be abandoned. An alternative approach is to constrain the bias to be zero, and find the estimator that minimises the variance. Such an estimator is termed the minimum variance unbiased estimator (MVUE). Note that the MSE of an unbiased estimator is just the variance.

It should be noted, however, that the MMSE criterion is the basis of most least-squares algorithms as will be seen later in the course, and is also intimately connected with Gaussian processes. However, in those contexts, the meaning and application is somewhat different, as will be seen.

5.2.4 Cramer-Rao Lower Bound



New slide

Being able to place a lower bound on the variance of any unbiased estimator process to be an extremely useful tool in practice. At best, it allows the identification of a minimum variance unbiased (MVU) estimator. This will be the case if the estimator attains the bound for all values of the unknown parameter. At worst, it provides a benchmark against which the performance of any unbiased estimator can be compared. Moreover, it highlights the physical impossibility of finding an unbiased estimator whose variance is less than the bound, and this can be useful in signal processing feasibility studies. Although many such bounds on the variance of an estimator exists, the Cramér-Rao lower-bound (CRLB) is by far the easiest to determine. Additionally, the theory of the CRLB provides a condition for which it is possible to determine whether an estimator exists that attains the 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).

Theorem 5.1 (CRLB - scalar parameter). Recalling $\{x[n]\}_0^{N-1}$ is just one realisation of the RVs $\{x[n,\zeta]\}_0^{N-1}$, defined on an N-dimensional space, then 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:

$$\operatorname{var}\left[\hat{\theta}\right] \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta}\right)^2\right]} \tag{M:3.6.17}$$

Alternatively, it may also be expressed as:

$$\operatorname{var}\left[\hat{\theta}\right] \geq -\frac{1}{\mathbb{E}\left[\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta^2}\right]} \tag{M:3.6.18}$$

The function $\ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)$ is called the **log-likelihood** function of θ. A discussion about the likelihood-function is given in Sidebar 9.

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

$$\hat{\theta} - \theta = K(\theta) \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta} \tag{K:3.7}$$

Sidebar 9 The likelihood function

The likelihood function is discussed in detail in Section 5.3. As has been noted throughout this course, given a physical model of a problem, it is possible to write down the joint density of the RVs $\mathbf{X}(\zeta) = \{x[n,\zeta]\}_{n=0}^{N-1}$, which depends on a fixed but unknown parameter vector $\boldsymbol{\theta}$: it is given by $f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})$, and can be viewed as a function of $\mathbf{x}$.

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

$$\mathcal{L}(\boldsymbol{\theta};\, \mathbf{x}) \equiv f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})\big|_{\text{fixed } \mathbf{x}, \text{ variable } \boldsymbol{\theta}} \tag{5.24}$$

Thus, the likelihood function $\mathcal{L}(\boldsymbol{\theta};\, \mathbf{x})$ should be intepreted as a function of $\boldsymbol{\theta}$ *given* $\mathbf{x}$. However, it is important to note that $\mathcal{L}(\boldsymbol{\theta};\, \mathbf{x}) \equiv f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ is not necessarily a pdf since, in general, it does not integrate to one over $\boldsymbol{\theta}$:

$$\int \mathcal{L}(\boldsymbol{\theta};\, \mathbf{x})\, d\boldsymbol{\theta} = \int f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})\, d\boldsymbol{\theta} \neq 1 \tag{5.25}$$

Note, however, that according to Bayes's theorem:

$$\int f_{\Theta}(\boldsymbol{\theta} \mid \mathbf{x})\, d\boldsymbol{\theta} = \int \frac{f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})\; f_{\Theta}(\boldsymbol{\theta})}{f_{\mathbf{X}}(\mathbf{x})}\, d\boldsymbol{\theta} = 1 \tag{5.26}$$

or alternatively, a weighted version of the likelihood gives rise to the probability of the observations:

$$\int \mathcal{L}(\boldsymbol{\theta};\, \mathbf{x})\, f_{\Theta}(\boldsymbol{\theta})\, d\boldsymbol{\theta} = f_{\mathbf{X}}(\mathbf{x}) \tag{5.27}$$

In otherwords, it is simply important to not intepret the likelihood function as a pdf, and simply to be carefull with the manipulations.

for some function $K(\theta)$, and where $\hat{\theta} = \hat{\theta}(\mathbf{x})$ is a function of the data only and, importantly, not a function of the true value of θ. Alternatively, a more useful way of writing Equation K:3.7 is to determine whether the log-likelihood function can be written in the form:

$$\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta} = I(\theta)\left(\hat{\theta} - \theta\right), \qquad \text{where } I(\theta) = K^{-1}(\theta). \tag{5.28}$$

The estimator $\hat{\theta}$ which attains this bound is the MVUE, and the minimum variance is given by $K(\theta)$. Note that an estimator which is unbiased and attains the CRLB is also said to be an **efficient estimator** in that it efficiently used the data.

PROOF. If $\hat{\theta}$ is unbiased, then $\mathbb{E}\left[\hat{\theta} - \theta\right] = 0$, which may be expressed as:

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (\hat{\theta} - \theta) f_{\mathbf{X}}(\mathbf{x} \mid \theta)\, d\mathbf{x} = 0 \tag{M:3.6.11}$$

Differentiating w. r. t. the true parameter θ, and assuming a real-value $\hat{\theta}$, then:

$$0 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta}\left[(\hat{\theta} - \theta) f_{\mathbf{X}}(\mathbf{x} \mid \theta)\right] d\mathbf{x} \tag{5.29}$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (\hat{\theta} - \theta) \frac{\partial f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta} d\mathbf{x} - \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x} \mid \theta)\, d\mathbf{x}}_{=1} \tag{M:3.6.12}$$

Note that here it has been assumed differentiation and integration may be interchanged. This is generally true except when the domain of the pdf for which it is nonzero depends on the known parameter. Using the fact that

$$\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta} = \frac{1}{f_{\mathbf{X}}(\mathbf{x} \mid \theta)} \frac{\partial f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta} \tag{5.30}$$

or,

$$\frac{\partial f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta} = \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta} f_{\mathbf{X}}(\mathbf{x} \mid \theta) \tag{M:3.6.13}$$

then substituting into Equation M:3.6.12 gives:

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left\{(\hat{\theta} - \theta) \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta}\right\} f_{\mathbf{X}}(\mathbf{x} \mid \theta)\, d\mathbf{x} = 1 \tag{M:3.6.14}$$

which can be written using the expectation operator as:

$$\mathbb{E}\left[(\hat{\theta} - \theta) \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta}\right] = 1 \tag{M:3.6.15}$$

Now, using the **Cauchy-Schwartz inequality** (see [Papoulis:1991]), which states that:

$$\left|\mathbb{E}\left[\mathbf{X}(\zeta)\mathbf{Y}(\zeta)\right]\right|^2 \leq \mathbb{E}\left[|\mathbf{X}(\zeta)|^2\right] \mathbb{E}\left[|\mathbf{Y}(\zeta)|^2\right] \tag{5.31}$$

then squaring both sides of Equation M:3.6.15 gives

$$1 = \mathbb{E}^2 \left[(\hat{\theta} - \theta) \frac{\partial \ln f_{\mathbf{X}} \left(\mathbf{x} \mid \theta \right)}{\partial \theta} \right] \leq \mathbb{E} \left[(\hat{\theta} - \theta)^2 \right] \mathbb{E} \left[\left(\frac{\partial \ln f_{\mathbf{X}} \left(\mathbf{x} \mid \theta \right)}{\partial \theta} \right)^2 \right] \tag{M:3.6.16}$$

Note that the Cauchy-Schwartz inequality becomes and equality iff the two integrands that are implicit in the expectation operator are related by a constant multiplier, independent of $\mathbf{x}$. That is, when:

$$(\hat{\theta} - \theta)^2 f_{\mathbf{X}} \left(\mathbf{x} \mid \theta \right) = K(\theta) \left(\frac{\partial \ln f_{\mathbf{X}} \left(\mathbf{x} \mid \theta \right)}{\partial \theta} \right)^2 f_{\mathbf{X}} \left(\mathbf{x} \mid \theta \right) \tag{5.32}$$

or, alternatively,

$$\hat{\theta} - \theta = K(\theta) \frac{\partial \ln f_{\mathbf{X}} \left(\mathbf{x} \mid \theta \right)}{\partial \theta} \tag{K:3.7}$$

This is the minimum variance unbiased estimator. Since the estimator is unbiased, then $\operatorname{var}\left[\hat{\theta}\right] = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$, and therefore:

$$\operatorname{var}\left[\hat{\theta}\right] \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial \ln f_{\mathbf{X}} (\mathbf{x} \mid \theta)}{\partial \theta} \right)^2 \right]} \tag{M:3.6.17}$$

To derive the second form by starting with the simple condition that:

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}} \left(\mathbf{x} \mid \theta \right) \, d\mathbf{x} = 1 \tag{5.33}$$

Differentiating once w. r. t. to $\theta$and using Equation M:3.6.13 gives

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial f_{\mathbf{X}} \left(\mathbf{x} \mid \theta \right)}{\partial \theta} d\mathbf{x} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial \ln f_{\mathbf{X}} \left(\mathbf{x} \mid \theta \right)}{\partial \theta} f_{\mathbf{X}} \left(\mathbf{x} \mid \theta \right) \, d\mathbf{x} = 0 \tag{5.34}$$

and differentiating again gives:

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{\partial^2 \ln f_{\mathbf{X}} \left(\mathbf{x} \mid \theta \right)}{\partial \theta^2} f_{\mathbf{X}} \left(\mathbf{x} \mid \theta \right) + \left\{ \frac{\partial \ln f_{\mathbf{X}} \left(\mathbf{x} \mid \theta \right)}{\partial \theta} \right\}^2 f_{\mathbf{X}} \left(\mathbf{x} \mid \theta \right) \right) \, d\mathbf{x} = 0 \tag{5.35}$$

which gives the desired result

$$\mathbb{E} \left[\frac{\partial^2 \ln f_{\mathbf{X}} \left(\mathbf{x} \mid \theta \right)}{\partial \theta^2} \right] = -\mathbb{E} \left[\left\{ \frac{\partial \ln f_{\mathbf{X}} \left(\mathbf{x} \mid \theta \right)}{\partial \theta} \right\}^2 \right] \tag{5.36}$$

□

This can then be substituted into Equation M:3.6.17.

Note that a generalisation of the CRLB for biased estimates is given by:

$$\operatorname{var}\left[\hat{\theta}\right] \geq \frac{\left(1+\frac{\partial B(\hat{\theta})}{\partial \theta}\right)^2}{\mathbb{E}\left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta}\right)^2\right]} \tag{5.37}$$

where $B(\hat{\theta})$ is the bias as previously defined. The proof follows a very similar line as given above, and is left as an exercise for the reader.

Example 5.3 ([Kay:1993, Example 3.3, Page 31]). Consider again the observations

$$x[n] = A + w[n], \quad n \in \{0, \dots, N-1\} \tag{K:2.2}$$

where A is the parameter to be estimated, and $w[n]$ is WGN. The parameter A can take on any value in the interval $-\infty < A < \infty$. Determine the CRLB for an estimator, $\hat{A}$, of the parameter A.

SOLUTION. Since the transformation between $x[n]$ and $x[n]$ is linear, with a multiplication factor of 1, the *likelihood function* can be written down as:

$$f_{\mathbf{X}}(\mathbf{x} \mid A) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[n]-A)^2\right] \tag{5.38}$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] \tag{5.39}$$

Note, a more detailed derivation of this likelihood is given in Sidebar 10 on page 114. Taking the first derivative of the **log-likelihood** gives:

$$\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid A)}{\partial A} = \frac{\partial}{\partial A}\left[-\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] \tag{5.40}$$

$$= \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A) = \frac{N}{\sigma^2}(\hat{\mu}_X - A) \tag{K:3.8}$$

where $\hat{\mu}_X$ is the sample mean. Differentiating again, then:

$$\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} \mid A)}{\partial A^2} = -\frac{N}{\sigma^2} \tag{5.41}$$

and noting that this second derivative is constant, then the CRLB is given by:

$$\operatorname{var}\left[\hat{A}\right] \geq \frac{\sigma^2}{N} \tag{K:3.9}$$

□

Comparing Equation K:3.7 and Equation K:3.8, then it is clear that the sample mean attains the bound, such that $\hat{A} = \mu_X$, and must therefore be the MVUE. Hence, the minimum variance will also be given by $\operatorname{var}\left[\hat{A}\right] = \frac{\sigma^2}{N}$.

Sidebar 10 Likelihood Derivation for Signal in Noise

A common model for a set of observations $\mathcal{X} = \{x[n]\}_0^{N-1}$ is the signal in noise:

$$x[n] = s[n;\, \boldsymbol{\theta}] + w[n]\,, \quad w[n] \sim \mathcal{N}\left(0,\, \sigma_w^2\right) \tag{5.42}$$

where $s[n;\, \boldsymbol{\theta}]$ denotes a parametric model for the underlying signal, and is dependent on a parameter (vector) θ. The noise process $w[n]$ is assumed to be i. i. d.; therefore, since $x[n]$ does not depend on previous values of either the input, $w[n]$, or the observed process, $x[n]$, it follows that $x[n]$ is also i. i. d..

Conditional on θ and a particular time index n, the pdf for the observed sample $x[n]$ can be obtained using the probability transformation rule. Hence, noting that there is one unique solution $w[n] = x[n] - s[n;\, \boldsymbol{\theta}]$, and that the Jacobian of the transformation is given by:

$$J_{w[n]\to x[n]} = \frac{\partial x[n]}{\partial w[n]} = 1 \tag{5.43}$$

it follows that

$$f_X\left(x[n] \mid \boldsymbol{\theta}\right) = \frac{f_W\left(x[n] - s[n;\, \boldsymbol{\theta}]\right)}{J_{w[n]\to x[n]}} = \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left\{-\frac{\left(x[n] - s[n;\, \boldsymbol{\theta}]\right)^2}{2\sigma_w^2}\right\} \tag{5.44}$$

where it is implicitly understood that $f_X\left(x[n] \mid \boldsymbol{\theta}\right) = f_X\left(x[n] \mid \boldsymbol{\theta},\, \sigma_w^2\right)$ also depends on the noise variance σ_w^2 although this isn't always explicitly written. Since the $x[n]$'s are i. i. d., then it follows that:

$$f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right) = f_{\mathbf{X}}\left(x[0]\,,\, \dots,\, x[N-1] \mid \boldsymbol{\theta}\right) \tag{5.45}$$

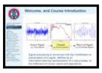
$$= \prod_{n=0}^{N-1} f_X\left(x[n] \mid \boldsymbol{\theta}\right) \tag{5.46}$$

$$= \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left\{-\frac{\left(x[n] - s[n;\, \boldsymbol{\theta}]\right)^2}{2\sigma_w^2}\right\} \tag{5.47}$$

$$= \frac{1}{\left(2\pi\sigma_w^2\right)^{\frac{N}{2}}} \exp\left\{-\frac{\sum_{n=0}^{N-1}\left(x[n] - s[n;\, \boldsymbol{\theta}]\right)^2}{2\sigma_w^2}\right\} \tag{5.48}$$

Note, therefore, that many of the examples in this handout have a likelihood function that take this form. Nevertheless, it is important to derive these results carefully each time you attempt to solve a problem, as a different model might give a different result. Moreover, this derivation should be included in any example questions that you tackle.

5.2.5 Consistency of an Estimator



New slide

If the MSE of the estimator,

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[|\hat{\theta} - \theta|^2\right] = \sigma_{\hat{\theta}}^2 + |B(\hat{\theta})|^2 \tag{M:3.6.7}$$

can be made to approach 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**.

Note that [Therrien:1992, Section 6.1.3, Page 290] gives a slightly more formal definition of a **consistent estimator:**

Definition 5.2 (Consistent Estimator). An estimate $\hat{\theta}_N$, based on N data observations, is **consistent** if

$$\lim_{N\to\infty} \Pr\left(\left|\hat{\theta}_N - \theta\right| < \epsilon\right) = 1 \tag{5.49}$$

♢

for any arbitrarily small number ϵ. The sequence of estimates $\{\hat{\theta}_N\}_0^\infty$ is said to **converge in probability** to the true value of the parameter θ.

Example 5.4 ([Manolakis:2001, Exercise 3.32, Page 147]). The Cauchy distribution with mean μ is given by:

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+(x-\mu)^2}, \quad x \in \mathbb{R} \tag{5.50}$$

Let $\{x_k\}_{k=0}^{N-1}$ be N i. i. d. RVs with this distribution. Consider the mean estimator based on these samples:

$$\hat{\mu} = \frac{1}{N}\sum_{k=0}^{N-1} x_k \tag{5.51}$$

Determine whether $\hat{\mu}$ is a consistent estimator of μ.

SOLUTION. It is simplest to use the definition that an estimator is consistent if $\lim_{N\to\infty} \text{MSE}(\theta) = 0$, where

$$\text{MSE}(\theta) = \mathbb{E}\left[|\hat{\theta} - \theta|^2\right] = \sigma_{\hat{\theta}}^2 + |B(\hat{\theta})|^2 \tag{M:3.6.7}$$

and

$$\sigma_{\hat{\theta}}^2 \triangleq \mathbb{E}\left[\left|\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right|^2\right] \equiv \mathbb{E}\left[\left|\hat{\theta}\right|^2\right] - \mathbb{E}^2\left[\hat{\theta}\right] \tag{M:3.6.5}$$

Hence, by noting that $\mathbb{E}[\hat{\mu}] = \mu$, such that $|B(\hat{\theta})|^2 = 0$, then the MSE is given by:

$$\text{MSE}(\theta) = \sigma^2_{\hat{\theta}} = \mathbb{E}\left[|\hat{\mu}|^2\right] - \mathbb{E}^2[\hat{\mu}] \tag{5.52}$$

$$\equiv \mathbb{E}\left[|\hat{\mu} - \mathbb{E}[\hat{\mu}]|^2\right] = \mathbb{E}\left[\left|\frac{1}{N}\sum_{k=0}^{N-1} x_k - \mu\right|^2\right] \tag{5.53}$$

$$\equiv \frac{1}{N^2}\sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\mathbb{E}[x_k\,x_l] - \mu^2 \tag{5.54}$$

Since the samples are independent and identically distributed (i. i. d.), then the autocorrelation function is given by:

$$\mathbb{E}[x_k\,x_l] = \begin{cases}\mathbb{E}[x_k]\,\mathbb{E}[x_l] & k \neq l \\ \mathbb{E}[x_k^2] & k = l\end{cases} \tag{5.55}$$

$$= \begin{cases}\mu^2 & k \neq l \\ \mu^2 + \sigma^2 & k = l\end{cases} \tag{5.56}$$

$$= \sigma^2\,\delta(k-l) + \mu^2 \tag{5.57}$$

Hence,

$$\text{MSE}(\theta) = \frac{1}{N^2}\sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\left(\sigma^2\,\delta(k-l) + \mu^2\right) - \mu^2 \tag{5.58}$$

$$= \frac{1}{N^2}\sum_{k=0}^{N-1}\left(\sigma^2 + N\mu^2\right) - \mu^2 \tag{5.59}$$

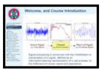
$$= \frac{1}{N}\left(\sigma^2 + N\mu^2\right) - \mu^2 = \frac{\sigma^2}{N} \tag{5.60}$$

□

Since the variance for a Cauchy distribution is unbounded, such that $\sigma^2 \to \infty$, then $\lim_{N\to\infty}\text{MSE}(\theta)$ does not converge to zero, and is therefore **not consistent**.

5.2.6 Efficiency of an estimator

Definition 5.3 (Efficiency of an estimator). An estimate is said to be **efficient** w. r. t. another estimate if it has a lower variance. Thus, if $\hat{\theta}_N$ is an estimator that depends on N observations and is both **unbiased** and **efficient** with respect to $\hat{\theta}_{N-1}$ for all N, then $\hat{\theta}_N$ is a **consistent** estimate.



5.2.7 Estimating Multiple Parameters

New slide

Multiple parameters occur in, for example, estimating the statistical properties of a random time-series, estimating the parameters of a curve fitted to a set of data, estimating any model described by a set of parameters. To deal with these vectors of parameters, the previous results can be extended and defined in an analogous way.

A vector of parameters, $\boldsymbol{\theta}$, of a random event $X(\zeta)$ can be estimated from a set of observations, $\mathcal{X} = \{x[n]\}_0^{N-1}$, using some function:

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}[\mathcal{X}] = \hat{\boldsymbol{\theta}}\left[\{x[n]\}_0^{N-1}\right] \tag{5.61}$$

The definitions of **unbiasedness**, **consistency**, **efficiency**, and the **CRLB** are all straightforward extensions of the definitions and results for scalar parameter estimates.

Assuming $\boldsymbol{\theta}$ is a $P \times 1$ parameter vector, these properties are:

Unbiased Estimator An estimate $\hat{\boldsymbol{\theta}}_N$ is **unbiased** if

$$\mathbb{E}\left[\hat{\boldsymbol{\theta}}_N\right] = \boldsymbol{\theta} \tag{5.62}$$

Otherwise, the estimate is **biased** with bias $\mathbf{b}(\hat{\boldsymbol{\theta}}_N) = \mathbb{E}\left[\hat{\boldsymbol{\theta}}_N\right] - \boldsymbol{\theta}$. An estimate is **asymptotically unbiased** if

$$\lim_{N\to\infty} \mathbb{E}\left[\hat{\boldsymbol{\theta}}_N\right] = \boldsymbol{\theta} \tag{5.63}$$

Consistent Estimator An estimate $\hat{\boldsymbol{\theta}}_N$, based on N data observations, is **consistent** if

$$\lim_{N\to\infty} \Pr\left(\left|\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}\right| < \epsilon\right) = 1 \tag{5.64}$$

for any arbitrarily small number ϵ. The sequence of estimates $\{\hat{\boldsymbol{\theta}}_N\}_0^\infty$ is said to **converge in probability** to the true value of the parameter θ.

Efficient Estimator An estimate $\hat{\boldsymbol{\theta}}$ is said to be **efficient** w. r. t. another estimate $\hat{\boldsymbol{\theta}}'$ if the difference of their covariance matrices $\boldsymbol{\Gamma}_{\hat{\boldsymbol{\theta}}'} - \boldsymbol{\Gamma}_{\hat{\boldsymbol{\theta}}}$ is positive definite. This implies that the variance of every component of $\hat{\boldsymbol{\theta}}$ must be smaller than the variance of the corresponding component of $\hat{\boldsymbol{\theta}}'$. If $\hat{\boldsymbol{\theta}}_N$ is unbiased and efficient with respect to $\hat{\boldsymbol{\theta}}_{N-1}$ for all N, then $\hat{\boldsymbol{\theta}}_N$ is a **consistent estimate**.

Theorem 5.2 (CRLB - real parameter vectors). This theorem is only for real parameter vectors. Complex-parameter vectors are slightly more detailed, but the principle no different, as highlighted by the note following this theorem. Assuming that the estimator $\hat{\boldsymbol{\theta}}$ is unbiased, then the vector parameter CRLB will place a bound on the variance of each element. This CRLB for a vector parameter is similar in concept to the scalar form, but requires a little more slickness in mathematical presentation. Define the gradient of the log-likelihood function to be:

$$\mathbf{s} \equiv \mathbf{s}(\mathbf{x}; \boldsymbol{\theta}) \triangleq \nabla_{\boldsymbol{\theta}} \ln f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta}) \tag{T:6.43}$$

The vector s is called the **score** for $\boldsymbol{\theta}$ based on x and can be shown to have zero mean (see [Therrien:1993, Problem 6.8, Page 331]). If $\hat{\boldsymbol{\theta}}$ is substituted for $\boldsymbol{\theta}$, the score is a measure of the optimality of the estimate, which scores near $\mathbf{0}_{P\times 1}$ being more desirable (albeit, not necessarily revealing the optimum solution); it follows that the maximum-likelihood estimate (MLE) introduced in the next section has a score of exactly $\mathbf{0}_{P\times 1}$. The covariance of the score vector is known as the **Fisher information matrix**, and is assumed to be nonsingular:

$$\mathbf{J}(\boldsymbol{\theta}) = \mathbb{E}\left[\mathrm{s}(\mathrm{x};\boldsymbol{\theta})\,\mathrm{s}^T(\mathrm{x};\boldsymbol{\theta})\right] \tag{T:6.42}$$

The Fisher information matrix can also be written in the following equivalent form:

$$[\mathbf{J}(\boldsymbol{\theta})]_{ij} = -\mathbb{E}\left[\frac{\partial^2 \ln f_{\mathbf{X}}\left(\mathrm{x} \mid \boldsymbol{\theta}\right)}{\partial\theta_i\partial\theta_j}\right] \tag{K:3.21}$$

If $\hat{\boldsymbol{\theta}}$ is any unbiased estimate, and $\boldsymbol{\Gamma}_{\hat{\boldsymbol{\theta}}}$ is its covariance matrix, then the CRLB can be stated as:

$$\boldsymbol{\Gamma}_{\hat{\boldsymbol{\theta}}} \geq \mathbf{J}^{-1}(\boldsymbol{\theta}) \tag{5.65}$$

where the notation $\geq$ means that the difference matrix $\boldsymbol{\Gamma}_{\hat{\boldsymbol{\theta}}} - \mathbf{J}^{-1}(\boldsymbol{\theta})$ is positive definite. This bound is satisfied with equality iff the estimate satisfies an equation of the form:

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \mathbf{J}^{-1}(\boldsymbol{\theta})\mathrm{s}(\mathrm{x};\boldsymbol{\theta}) \tag{T:6.47}$$

where $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathrm{x})$ is a function of the data only (and, importantly, not a function of the true value of $\boldsymbol{\theta}$. Note that an estimator which is unbiased and attains the CRLB is also said to be an **efficient estimator** in that it efficiently used the data.

PROOF. For a full proof, see [Therrien:1992, Page 298], or [Kay:1993]. However, the proof is relatively straightforward and is analogous to the proof for the case of the scalar real parameter. It is currently omitted from this document.

The CRLB derived here can, of course, be applied to complex parameters by separating the parameter into real and imaginary parts, and including those parts separately into the real vector $\boldsymbol{\theta}$. It is possible to develop a direct complex version of this bound, and this is discussed in [Therrien:1992, Page 298].

Example 5.5 ([Kay:1993, Example 3.7, Page 41] - Line fitting). Consider the problem of fitting a line to a set of observations, that is dependent on the observation index n. This, given a random process $X(\zeta, n) = x[n]$, and the model

$$x[n] = A + Bn + w[n]\,, \quad n \in \{0,\, 1,\, \dots,\, N-1\} \tag{5.66}$$

where $w[n]$ is WGN. Determine the CRLB for the slope B and the intercept A.

SOLUTION. The 2×2 Fisher information matrix is given by:

$$\begin{aligned}
\mathbf{J}(\boldsymbol{\theta}) &= \mathbb{E}\left[\mathrm{s}(\mathrm{x};\boldsymbol{\theta})\,\mathrm{s}^T(\mathrm{x};\boldsymbol{\theta})\right] && \text{(T:6.42)}\\
&= \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \ln f_{\mathbf{X}}\left(\mathrm{x} \mid \boldsymbol{\theta}\right)\, \nabla_{\boldsymbol{\theta}}^T \ln f_{\mathbf{X}}\left(\mathrm{x} \mid \boldsymbol{\theta}\right)\right] && \text{(5.67)}\\
&= \begin{bmatrix} \mathbb{E}\left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathrm{x}\mid\boldsymbol{\theta})}{\partial A}\right)^2\right] & \mathbb{E}\left[\frac{\partial \ln f_{\mathbf{X}}(\mathrm{x}\mid\boldsymbol{\theta})}{\partial A}\frac{\partial \ln f_{\mathbf{X}}(\mathrm{x}\mid\boldsymbol{\theta})}{\partial B}\right] \\ \mathbb{E}\left[\frac{\partial \ln f_{\mathbf{X}}(\mathrm{x}\mid\boldsymbol{\theta})}{\partial B}\frac{\partial \ln f_{\mathbf{X}}(\mathrm{x}\mid\boldsymbol{\theta})}{\partial A}\right] & \mathbb{E}\left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathrm{x}\mid\boldsymbol{\theta})}{\partial B}\right)^2\right] \end{bmatrix} && \text{(5.68)}
\end{aligned}$$

where the notation $\boldsymbol{\theta} = [A,\, B]^T$ is used as a shorthand. Alternatively, the elements of the Fisher information matrix can be found using:

$$[\mathbf{J}(\boldsymbol{\theta})]_{ij} = -\mathbb{E}\left[\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right] \tag{K:3.21}$$

This alternative expression is often a more straightforward method for evaluating the Fisher information matrix, and will be used here. As in the case of a DC signal in WGN, the likelihood function can be written as

$$f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - A - Bn)^2\right] \tag{5.69}$$

from which the following derivatives follow:

$$\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})}{\partial A} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n] - A - Bn) \tag{5.70}$$

$$\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})}{\partial B} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n] - A - Bn)\, n \tag{5.71}$$

and

$$\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})}{\partial A^2} = -\frac{N}{\sigma^2} \tag{5.72}$$

$$\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})}{\partial A \partial B} = -\frac{1}{\sigma^2}\sum_{n=0}^{N-1} n \tag{5.73}$$

$$\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})}{\partial B^2} = -\frac{1}{\sigma^2}\sum_{n=0}^{N-1} n^2 \tag{5.74}$$

Using the identities that

$$\sum_{n=1}^{N} n = \frac{1}{2}N(N+1) \quad \text{and} \quad \sum_{n=1}^{N} n^2 = \frac{1}{6}N(N+1)(2N+1) \tag{5.75}$$

and noting that the second-order derivatives do not depend on $\mathbf{x}$ and therefore equal their expected values, then the Fisher information can be written as follows:

$$\mathbf{J}(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\begin{bmatrix} N & \frac{N(N-1)}{2} \\ \frac{N(N-1)}{2} & \frac{N(N-1)(2N-1)}{6} \end{bmatrix} \tag{5.76}$$

Inverting this matrix yields:

$$\mathbf{J}^{-1}(\boldsymbol{\theta}) = \sigma^2\begin{bmatrix} \frac{2(2N-1)}{N(N+1)} & -\frac{6}{N(N+1)} \\ -\frac{6}{N(N+1)} & \frac{12}{N(N^2-1)} \end{bmatrix} \tag{5.77}$$

Hence, it can be deduced that the variances for the individual parameters is given by the CRLB or:

$$\operatorname{var}\left[\hat{A}\right] \geq \frac{2(2N-1)\sigma^2}{N(N+1)} \tag{5.78}$$

$$\operatorname{var}\left[\hat{B}\right] \geq \frac{12\sigma^2}{N(N^2-1)} \tag{5.79}$$

Finally, note that a MVUE, if it exists, satisfies the relationship:

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \mathbf{J}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}} \ln f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right) \tag{T:6.47}$$

where the estimator $\hat{\boldsymbol{\theta}}$ depends on the observations only, and not the true parameter $\boldsymbol{\theta}$; if this were not the case, then the MVUE cannot exist physically. Hence, using the expressions for the terms in the RHS

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \mathbf{J}(\boldsymbol{\theta})^{-1}\begin{bmatrix} \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})}{\partial A} \\ \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})}{\partial B}\end{bmatrix} \tag{5.80}$$

$$\begin{bmatrix} \hat{A} - A \\ \hat{B} - B \end{bmatrix} = \sigma^2 \begin{bmatrix} \frac{2(2N-1)}{N(N+1)} & -\frac{6}{N(N+1)} \\ -\frac{6}{N(N+1)} & \frac{12}{N(N^2-1)} \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left(x[n] - A - Bn\right) \\ \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left(x[n] - A - Bn\right)n \end{bmatrix} \tag{5.81}$$

$$= \begin{bmatrix} \frac{2(2N-1)}{N(N+1)}\sum_{n=0}^{N-1}\left(x[n] - A - Bn\right) - \frac{6}{N(N+1)}\sum_{n=0}^{N-1}\left(x[n] - A - Bn\right)n \\ -\frac{6}{N(N+1)}\sum_{n=0}^{N-1}\left(x[n] - A - Bn\right) + \frac{12}{N(N^2-1)}\sum_{n=0}^{N-1}\left(x[n] - A - Bn\right)n \end{bmatrix} \tag{5.82}$$

$$= \frac{2}{N(N+1)}\begin{bmatrix} (2N-1)\sum_{n=0}^{N-1} x[n] - 3\sum_{n=0}^{N-1} n\, x[n] \\ -3\sum_{n=0}^{N-1} x[n] + \frac{6}{(N-1)}\sum_{n=0}^{N-1} n\, x[n] \end{bmatrix} - \begin{bmatrix} A \\ B \end{bmatrix} \tag{5.83}$$

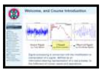
where again the identities for $\sum_{n=0}^{N-1} n$ and $\sum_{n=0}^{N-1} n^2$ have been used, and the terms not involving the data have been grouped, simplified, and ultimately either cancelled or rearranged into the second column vector on the RHS. Hence, it follows that:

$$\begin{bmatrix} \hat{A} \\ \hat{B} \end{bmatrix} = \frac{2}{N(N+1)}\begin{bmatrix} (2N-1)\sum_{n=0}^{N-1} x[n] - 3\sum_{n=0}^{N-1} n\, x[n] \\ -3\sum_{n=0}^{N-1} x[n] + \frac{6}{(N-1)}\sum_{n=0}^{N-1} n\, x[n] \end{bmatrix} \tag{5.84}$$

□

Since the estimator is not dependent on the true value of the parameters, then this is indeed the MVUE for the line fitting problem. It would not be straightforward to have intuitively determined what this estimator should have been without using the CRLB.

This previous example leads to an interesting observation. Note first that the CRLB for $\hat{A}$ has increased over that obtained when B is known, for in the latter case, it can be determined that $\operatorname{var}\left[\hat{A}\right] \geq \frac{\sigma^2}{N}$, which for $N \geq 2$, is less than $\frac{2(2N-1)\sigma^2}{N(N+1)}$. This relates to quite a general result that asserts that *the CRLB always increases as more parameters are estimated.*



5.3 Maximum Likelihood Estimation

New slide

This section now investigates an alternative to the MVUE, which is desirable in situations where the MVUE does not exist, or cannot be found even if it does exist. This estimator, which is based on the **maximum likelihood principle**, is overwhelmingly the most popular approach to *practical* estimators. It has the advantage of being a *recipe procedure*, allowing it to be implemented for complicated problems. Additionally, for most cases of practical interest, its performance is optimal for large enough data records. Specifically, it is approximately the MVUE estimator due to its approximate efficiency. For these reasons, almost all practical estimators are based on the maximum likelihood principle.

The joint density of the RVs $\mathbf{X}(\zeta) = \{x[n,\zeta]\}_0^{N-1}$, which depends on fixed but unknown parameter vector $\boldsymbol{\theta}$, is given by $f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$. 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 $f_{\mathbf{X}}\left(\hat{\mathbf{x}} \mid \boldsymbol{\theta}\right)$. In other-words, the MLE for a parameter $\boldsymbol{\theta}$ is that estimate that makes the *given* value of the observation vector the *most likely value*.

This point cannot be over-emphasised; it is common to think of $f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$ as a function of $\mathbf{x}$; now it is necessary to turn this thinking around, and view $f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$ as a function of $\boldsymbol{\theta}$, for a given $\mathbf{x}$.

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

$$\hat{\boldsymbol{\theta}}_{ml}(\mathbf{x}) = \arg_{\boldsymbol{\theta}} \max f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right) \tag{T:6.40}$$

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

Assuming a differentiable likelihood function, and that $\boldsymbol{\theta} \in \mathbb{R}^P$, the MLE is found from

$$\begin{bmatrix} \frac{\partial f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta})}{\partial \theta_P} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{5.85}$$

or, more simply,

$$\nabla_{\boldsymbol{\theta}} f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right) \triangleq \frac{\partial f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} = \mathbf{0}_{P\times 1} \tag{K:7.35}$$

where $\mathbf{0}_{P\times 1}$ denotes the $P \times 1$ vector of zero elements. This property is listed in the next section for further information. If multiple solutions to this exist, then the one that maximises the likelihood function is the MLE.

There is a slight abuse of notation here, in that $\mathbf{x}$ is used to denote both the argument in $f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$, as well as the *given parameter* in the likelihood function. However, this strict distinction is not important here, although it can be useful to be more careful in advanced work of this nature.

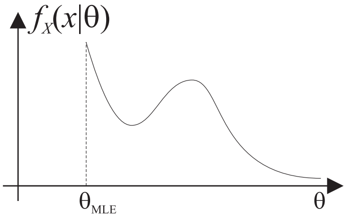
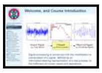


Figure 5.1: A single parameter MLE that occurs at a boundary, and therefore for which $\left.\frac{\partial f_{\mathbf{X}}(\mathbf{x}\,|\,\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}_{ml}} \neq 0$. Hence, in this case, a MLE and the MVUE are not necessarily equal.



5.3.1 Properties of the MLE

New slide

1. The MLE satisfies

$$\left.\nabla_{\boldsymbol{\theta}} f_{\mathbf{X}}\left(\mathbf{x}\,|\,\boldsymbol{\theta}\right)\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ml}} = \mathbf{0}_{P\times 1} \tag{T:6.41a}$$

$$\left.\nabla_{\boldsymbol{\theta}} \ln f_{\mathbf{X}}\left(\mathbf{x}\,|\,\boldsymbol{\theta}\right)\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ml}} = \mathbf{0}_{P\times 1} \tag{T:6.41b}$$

where $\boldsymbol{\theta} \in \mathbb{R}^{P\times 1}$. Note that in the case of a scalar parameter, θ, then these expressions reduce to the simpler form:

$$\left.\frac{\partial f_{\mathbf{X}}\left(\mathbf{x}\,|\,\theta\right)}{\partial\theta}\right|_{\theta=\hat{\theta}_{ml}} = 0 \tag{T:6.10a}$$

$$\left.\frac{\partial \ln f_{\mathbf{X}}\left(\mathbf{x}\,|\,\theta\right)}{\partial\theta}\right|_{\theta=\hat{\theta}_{ml}} = 0 \tag{T:6.10b}$$

2. If an MVUE exists and the MLE does not occur at a boundary, then the MLE *is* the MVUE. If the MLE occurs at the boundary, then the derivative of the likelihood function is not necessarily equal to zero, as shown, for example, in Figure 5.1. An example of such a case is discussed in the tutorial exercises.

PROOF (EQUIVALENCE OF MVUE AND MLE). For clarity and simplicity, only the proof for the scalar case is given. The extension to parameter vectors is straightforward. As shown in the derivation of the CRLB, the MVUE satisfies:

$$\hat{\theta} - \theta = K(\theta)\frac{\partial \ln f_{\mathbf{X}}\left(\mathbf{x}\,|\,\theta\right)}{\partial\theta} \tag{5.86}$$

The MLE satisfies

$$\left.\frac{\partial f_{\mathbf{X}}\left(\mathbf{x}\,|\,\theta\right)}{\partial\theta}\right|_{\theta=\hat{\theta}_{ml}} = 0 \tag{5.87}$$

$$\left.\frac{\partial \ln f_{\mathbf{X}}\left(\mathbf{x}\,|\,\theta\right)}{\partial\theta}\right|_{\theta=\hat{\theta}_{ml}} = 0 \tag{5.88}$$

Hence, setting $\theta = \hat{\theta}_{ml}$ and substituting these into one another, gives:

$$\hat{\theta} - \hat{\theta}_{ml} = K(\hat{\theta}_{ml})\left.\frac{\partial \ln f_{\mathbf{X}}\left(\mathbf{x}\,|\,\theta\right)}{\partial\theta}\right|_{\theta=\hat{\theta}_{ml}} = 0 \tag{5.89}$$

Hence,

$$\hat{\theta} = \hat{\theta}_{ml} \tag{5.90}$$

□

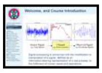
3. If the pdf, $f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$, of the data $\mathbf{x}$ satisfies certain *regularity* conditions, then the MLE of the unknown parameter $\boldsymbol{\theta}$ is asymptotically distributed (for large data records) according to a Gaussian distribution:

$$\hat{\boldsymbol{\theta}}_{ml} \sim \mathcal{N}\left(\boldsymbol{\theta},\, \mathbf{J}^{-1}(\boldsymbol{\theta})\right) \tag{5.91}$$

where $\mathbf{J}(\boldsymbol{\theta})$ is known as the **Fisher information** evaluated at the true value of the unknown parameter.

From the asymptotic distribution, the MLE is seen to be asymptotically unbiased and asymptotically attains the CRLB. It is therefore *asymptotically efficient*, and hence *asymptotically optimal*. This, of course, leads to the key question of how large does the data set need to be for these asymptotic properties to apply.

5.3.2 DC Level in white Gaussian noise



New slide

An example of the maximum likelihood principle begins with the scalar case, and again deals with a DC level in WGN.

Example 5.6 ([Therrien:1991, Example 6.1, Page 282]). A constant but unknown signal is observed in additive WGN. That is,

$$x[n] = A + w[n] \quad \text{where} \quad w[n] \sim \mathcal{N}\left(0,\, \sigma_w^2\right) \tag{5.92}$$

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

SOLUTION. Since $x[n] = A + w[n]$, then consider the probability transformation from $w[n]$ to $x[n]$. Then it is clear that

$$f_X\left(x[n] \mid A\right) = f_W\left(w[n] \mid A\right) = f_W\left(x[n] - A\right) \tag{5.93}$$

Moreover, since this is a memoryless system, and $w(n)$ are i. i. d., then so is $x[n]$, and therefore:

$$f_{\mathbf{X}}\left(\mathbf{x} \mid A\right) = \prod_{n \in \mathcal{N}} f_W\left(x[n] - A\right) = \frac{1}{\left(2\pi\sigma_w^2\right)^{\frac{N}{2}}} \exp\left\{-\frac{\sum_{n \in \mathcal{N}} \left(x[n] - A\right)^2}{2\sigma_w^2}\right\} \tag{5.94}$$

The **log-likelihood** is given by the logarithm of the likelihood function, and is usually a simpler function to minimise, at least for distributions which involve exponential functions. Hence, for this case, the log-likelihood is given by:

$$\ln f_{\mathbf{X}}\left(\mathbf{x} \mid A\right) = -\frac{N}{2}\ln(2\pi\sigma_w^2) - \frac{\sum_{n \in \mathcal{N}} \left(x[n] - A\right)^2}{2\sigma_w^2} \tag{5.95}$$

Differentiating this expression w. r. t. A gives

$$\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid A)}{\partial A} = \frac{\sum_{n\in\mathcal{N}}(x[n]-A)}{\sigma_w^2} \tag{5.96}$$

and setting this to zero yields the MLE:

$$\hat{A}_{ml} = \frac{1}{N}\sum_{n\in\mathcal{N}} x[n] \tag{5.97}$$

□

This is the **sample mean**, and it has already been seen that this is an efficient estimator. Hence, the MLE is efficient. This result is true in general; if an **efficient estimator** exists, the *maximum likelihood procedure* will produce it.

Example 5.7 ([Kay:1993, Example 7.3, Page 162]). The previous example of a DC level in WGN is considered again, except that in this case, the DC level is assumed to be positive $(A > 0)$, and the variance of $w[n]$ is now proportional to A. Thus, for a large value of A, a higher noise power is expected. Thus, the observations may be modelled as:

$$x[n] = A + w[n] \quad \text{where} \quad w[n] \sim \mathcal{N}\left(0,\, A\sigma_w^2\right) \tag{5.98}$$

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

SOLUTION. Following the development of the previous example, the pdf for the observed data and, equivalently, the likelihood function is given by:

$$f_{\mathbf{X}}(\mathbf{x} \mid A) = \frac{1}{(2\pi A\sigma_w^2)^{\frac{N}{2}}} \exp\left\{-\frac{\sum_{n\in\mathcal{N}}(x[n]-A)^2}{2A\sigma_w^2}\right\} \tag{5.99}$$

and thus the log-likelihood function is given by:

$$\ln f_{\mathbf{X}}(\mathbf{x} \mid A) = -\frac{N}{2}\ln(2\pi A\sigma_w^2) - \frac{\sum_{n\in\mathcal{N}}(x[n]-A)^2}{2A\sigma_w^2} \tag{5.100}$$

Differentiating the log-likelihood function w. r. t. A gives:

$$\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} \mid A)}{\partial A} = -\frac{N}{2A} + \frac{4A\sigma_w^2 \sum_{n\in\mathcal{N}}(x[n]-A) + 2\sigma_w^2 \sum_{n\in\mathcal{N}}(x[n]-A)^2}{4A^2\sigma_w^4} \tag{5.101}$$

$$= -\frac{N}{2A} + \frac{\sum_{n\in\mathcal{N}}(x[n]-A)}{A\sigma_w^2} + \frac{\sum_{n\in\mathcal{N}}(x[n]-A)^2}{2A^2\sigma_w^2} \tag{5.102}$$

and setting this equal to zero produces:

$$AN\sigma_w^2 = \sum_{n\in\mathcal{N}}\left\{(x[n]-A)^2 + 2A(x[n]-A)\right\} \tag{5.103}$$

$$A^2 + A\,\sigma_w^2 = \frac{1}{N}\sum_{n\in\mathcal{N}} x^2[n] \tag{5.104}$$

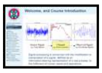
Solving for $\hat{A} > 0$ gives:

$$\hat{A} = -\frac{\sigma_w^2}{2} + \sqrt{\frac{\sigma_w^4}{4} + \frac{1}{N}\sum_{n \in \mathcal{N}} x^2[n]} \tag{5.105}$$

□

Finally, that $\hat{A}$ indeed maximises the log-likelihood function can be verified by examining the second derivative.

5.3.3 MLE for Transformed Parameter



New slide

Theorem 5.3 (Invariance Property of the MLE). The invariance property is discussed further in [Kay:1993, Theorem 7.2, Page 176] and [Kay:1993, Theorem 7.4, Page 185], for scalar and vector parameters respectively. The following theorem is presented for vector parameters, and can be simplified accordingly for scalar parameters. The MLE of the parameter $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$, where 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

$$\hat{\boldsymbol{\alpha}}_{ml} = \mathbf{g}(\hat{\boldsymbol{\theta}}_{ml}) \tag{5.106}$$

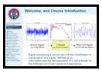
where $\hat{\boldsymbol{\theta}}_{ml}$ is the MLE of $\boldsymbol{\theta}$.

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 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:

$$\bar{p}_T\left(\mathbf{x} \mid \boldsymbol{\alpha}\right) = \max_{\boldsymbol{\theta}:\boldsymbol{\alpha}=\mathbf{g}(\boldsymbol{\theta})} f_{\mathbf{X}}\left(\mathbf{x} \mid \boldsymbol{\theta}\right) \tag{5.107}$$

♢

5.4 Least Squares

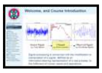


New slide

The estimators discussed so far have attempted to find an optimal or nearly optimal (for large data records) estimator by considering the class of unbiased estimators and determining the one exhibiting minimum variance, 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**.

The principle or method of least squares dates back to 1821 when Carl Friedrich Gauss used the method to determine the orbit of the asteroid Ceres by formulating the estimation problem as an optimisation problem.

A salient feature of the method is that *no probabilistic assumptions* are made about the data; only a *signal model* is assumed. The advantage is that it is a simpler procedure to find a parameter estimate since, for the MVUE and MLE, the pdf must either be known, or computable from the information in the problem, which makes these estimates difficult to compute and implement. As will be seen, it turns out that the least-squares estimate (LSE) can be calculated when just the first and second moments are known, and through the solution of *linear* equations. Hence, the method has a broader range of possible applications. On the negative side, no claims about optimality can be made, and furthermore, the statistical performance cannot be assessed without some specific assumptions about the probabilistic structure of the data.



5.4.1 The Least Squares Approach

New slide

Thus far, in determining a good estimator, the focus has been on finding one that is unbiased and has minimum variance. Hence, it is sought to minimise the average discrepancy between the estimate and the true parameter value. For unbiased estimates, this corresponds to minimising the variance of the estimator.

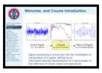
In the least-squares (LS) approach, it is sought to minimise the squared difference between the given, or observed, data $x[n]$ and the assumed, or hidden, signal or noiseless data. Here it is assumed that the hidden or unobserved signal is generated by some model which, in turn, depends on some unknown parameter $\boldsymbol{\theta}$. Due to observation noise or model inaccuracies, the observation $x[n]$, is a perturbed version of $s[n]$. 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:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} (x[n] - s[n])^2 \tag{K:8.1}$$

where $s[n] = s[n; \boldsymbol{\theta}]$ is a function of $\boldsymbol{\theta}$. The LSE is given by:

$$\hat{\boldsymbol{\theta}}_{LSE} = \arg_{\boldsymbol{\theta}} \min J(\boldsymbol{\theta}) \tag{5.108}$$

Note that no probabilistic assumptions have been made about the data $x[n]$ and that the method is equally valid for Gaussian as well as non-Gaussian noise. Of course, the performance of the LSE will depend on the properties of the corrupting noise, as well as any modelling errors. LSEs are usually applied in situations where a precise statistical characterisation of the data or noise process is unknown. They are also applied when an optimal estimator cannot be found, or may be too complicated to apply in practice.



5.4.2 DC Level

New slide

Again, start by considering an example with a scalar parameter. The case with vector parameters follows a similar line.

Example 5.8 ([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 $n \in \mathcal{N} = \{0, \dots, N-1\}$. Calculate the LSE of the unknown signal A.

SOLUTION. According to the LS approach, then:

$$\hat{A}_{LSE} = \arg_A \min J(A) \quad \text{where} \quad J(A) = \sum_{n=0}^{N-1} (x[n] - A)^2 \tag{5.109}$$

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

$$\hat{A}_{LSE} = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \tag{5.110}$$

□

which is the sample mean estimator.

This LSE cannot, however, be claimed to be optimal in the MVU sense, but only in that it minimises the LS error. If it is known that $x[n] = A + w[n]$, where $w[n]$ is zero-mean WGN, then the LSE will also be the MVUE, but otherwise not.

5.4.3 Nonlinear Least Squares

Example 5.9 (Sinusoidal Frequency Estimation). Again, it is assumed that an observed signal, $x[n]$, is a perturbed version of an unknown signal, $s[n]$, which is modelled as

$$s[n] = \cos 2\pi f_0 n \tag{5.111}$$

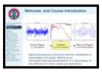
in which the frequency f_0 is to be estimated. The LSE can be found by minimising

$$J(f_0) = \sum_{n=0}^{N-1} (x[n] - \cos 2\pi f_0 n)^2 \tag{5.112}$$

⋈

In contrast to the DC level signal for which the minimum is easily found, here the LS error function is highly nonlinear in the parameter f_0. The minimisation cannot be done in closed form. Since the error criterion is a quadratic function of the signal, a signal that is *linear* in the unknown parameter yields a quadratic function for J, as in the previous example. The minimisation is then easily carried out. A signal model that is *linear in the unknown parameter* is said to generate a **linear least squares** problem. **Nonlinear least squares** problems are solved via grid searches or iterative minimisation methods.

5.4.4 Linear Least Squares



New slide

Again, assume that an observed signal, $\{x[n]\}_0^{N-1}$, is a perturbed version of an unknown signal, $\{s[n]\}_0^{N-1}$, where each of these processes can be written by the random vectors:

$$\mathbf{s} = \begin{bmatrix} s[0] & s[1] & \cdots & s[N-1] \end{bmatrix}^T \text{ and } \mathbf{x} = \begin{bmatrix} x[0] & x[1] & \cdots & x[N-1] \end{bmatrix}^T \tag{5.113}$$

The signal, $s[n]$, can be written as a linear combination of P known functions, $\{h_k[n]\}_{k=1}^P$, with weighting parameters $\{\theta_k\}_{k=1}^P$; thus:

$$s[n] = \sum_{k=1}^{P} \theta_k \, h_k[n] \tag{5.114}$$

Writing this in matrix-vector notation, it follows that:

$$\underbrace{\begin{bmatrix} s[0] \\ s[1] \\ \vdots \\ s[N-1] \end{bmatrix}}_{\mathbf{s}} = \underbrace{\begin{bmatrix} h_1[0] & h_2[0] & \cdots & h_P[0] \\ h_1[1] & h_2[1] & \cdots & h_P[1] \\ \vdots & \vdots & \ddots & \vdots \\ h_1[N-1] & h_2[N-1] & \cdots & h_P[N-1] \end{bmatrix}}_{\mathbf{H}} \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_P \end{bmatrix}}_{\boldsymbol{\theta}} \tag{5.115}$$

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

$$\mathbf{s} = \mathbf{H}\,\boldsymbol{\theta} \tag{K:8.8}$$

As shown above, $\mathbf{H}$ is a known $N \times P$ matrix, where $N > P$, and must be of full rank. It is referred to as the **observation matrix**. The LSE is found by minimising:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} |x[n] - s[n]|^2 = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \tag{K:8.9}$$

This can be written as:

$$J(\boldsymbol{\theta}) = \mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{H}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} \tag{5.116}$$

and using the two identities that:

$$\frac{\partial \mathbf{b}^T \mathbf{a}}{\partial \mathbf{a}} = \mathbf{b} \quad \text{and} \quad \frac{\partial \mathbf{a}^T\mathbf{B}\mathbf{a}}{\partial \mathbf{a}} = \left(\mathbf{B} + \mathbf{B}^T\right)\mathbf{a} \tag{5.117}$$

then observing in this case $\mathbf{B} = \mathbf{H}^T\mathbf{H} = \mathbf{B}^T$ it follows that

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \theta} = -2\mathbf{H}^T\mathbf{x} + 2\mathbf{H}^T\mathbf{H}\theta \tag{5.118}$$

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

$$\hat{\boldsymbol{\theta}}_{LSE} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x} \tag{K:8.10}$$

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**.

Requiring $\mathbf{H}$ to be full rank guarantees the invertibility of $\mathbf{H}^T\mathbf{H}$. The minimum LS error is found from Equation K:8.9 and Equation K:8.10:

$$\begin{aligned} J_{\min} = J(\hat{\boldsymbol{\theta}}) &= \left(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}\right)^T \left(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}\right) && (5.119) \\ &= \left(\mathbf{x} - \mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}\right)^T \left(\mathbf{x} - \mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}\right) && (5.120) \\ &= \mathbf{x}^T \left(\mathbf{I}_N - \mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\right) \left(\mathbf{I}_N - \mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\right)\mathbf{x} && (5.121) \end{aligned}$$

Now, the matrix $\mathbf{A} = \mathbf{I}_N - \mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T$ is an **idempotent matrix** in that it has the property $\mathbf{A}^2 = \mathbf{A}$. This follows from noting that:

$$\mathbf{A}^2 = \mathbf{I}_N - 2\mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T + \underbrace{\mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\,\mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T}_{=\mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T} = \mathbf{A} \tag{5.122}$$

Hence,

$$J_{\min} = \mathbf{x}^T\left(\mathbf{I}_N - \mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\right)\mathbf{x} \tag{K:8.11}$$

Other forms for $J_{\min}$ are:

$$\begin{aligned} J_{\min} &= \mathbf{x}^T\mathbf{x} - \mathbf{x}^T\mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x} && \text{(K:8.12)} \\ &= \mathbf{x}^T\left(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right) && \text{(K:8.13)} \end{aligned}$$

5.4.5 Weighted Linear Least Squares

An extension of the linear LS problem is **weighted linear least squares**. Instead of minimising Equation K:8.9, an $N \times N$ positive definite, and by definition, therefore symmetric, weighting matrix $\mathbf{W}$, so that

$$J(\boldsymbol{\theta}) = \left(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right)^T\mathbf{W}\left(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right) \tag{K:8.14}$$

If, for instance, $\mathbf{W}$ is diagonal with diagonal elements $[\mathbf{W}]_{ii} = w_i > 0$, then the LS error of Equation K:8.1 reduces to:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} w_n\left(x[n] - s[n]\right)^2 \tag{5.123}$$

The rationale for introducing weighting factors into the error criterion is to emphasise the contributions of those data samples that are deemed to be more reliable. Hence, consider again Example 5.8 on page 126, and assume that $x[n] = A + w[n]$, where $w[n]$ is a zero-mean uncorrelated noise signal with variance σ_n^2; if σ_n^2 is large compared with A, then the estimate of the underlying signal $s[n] = A$ from $x[n]$ will be unreliable. Thus, it would seem reasonable to choose a weighting factor of $w_n = \frac{1}{\sigma_n^2}$.

Example 5.10 ([Kay:1993, Problem 8.8, Page 276]). Find the weighted least squares estimate of an unknown signal, $s[n] = A$, from an observed signal $x[n]$, where the known weighting factors are given by $w_n = \frac{1}{\sigma_n^2}$.

SOLUTION. The weighted LS error is given by:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}\left(x[n] - A\right)^2 \tag{5.124}$$

Differentiating w. r. t. A, and setting to zero gives:

$$0 = \sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}\left(x[n] - A\right) \tag{5.125}$$

Rearranging gives straightforwardly:

$$\hat{A}_{LSE} = \frac{\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2} x[n]}{\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}} \tag{5.126}$$

□

The general form of the weighted LSE is readily shown to be:

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T \mathbf{W} \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{W} \mathbf{x} \tag{K:8.16}$$

and its minimum LS error is

$$J_{\min} = \mathbf{x}^T \left(\mathbf{W} - \mathbf{W}\mathbf{H} \left(\mathbf{H}^T \mathbf{W} \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{W}\right) \mathbf{x} \tag{K:8.17}$$

5.5 Bayesian Parameter Estimation

Using the method of maximum likelihood (or least squares) to infer the values of a parameter has several significant limitations:

1. First, the likelihood function does not use information other than the data itself to *infer* the values of the parameters. No prior knowledge, stated before the data is observed, is utilised regarding the possible or probable values that the parameters might take. In many applications, a physical understanding of the problem at hand, or of the circumstances surrounding how an experiment is conducted, can suggest that some values of the parameters are impossible, and that some are more likely to occur than others.

   There are cases where the maximum-likelihood estimate (MLE) can return parameter estimates outside the sensible range of the parameters, or outside the physical constraints of the system under consideration.

2. The likelihood function on its own does not limit the number of parameters in a model used to fit the data. The number of parameters is chosen in advance, by the Signal Processing Engineer, but the likelihood function does not indicate whether the number of parameters chosen is more than necessary to model the data, or less than needed.

   In general, the more parameters used to model the data, the better the model will fit the data. For example, a data set consisting of N observations can always be described exactly by a model with N parameters. However, suppose that a model is used to describe a particular realisation of a stochastic process with no error by using N parameters to model N observations. If another realisation of that random process is generated, then a new model is required to describe the new data with no error. Often the new parameter estimates can be vastly different to the old parameter set.

   This problem arises from the tendency to attempt to over-parameterize the data; there is clearly a tradeoff between modelling a signal with no error and having a more complicated or sophisticated model. With this in mind, model simplicity is the key to maximising the *degree of consistency* between parameter estimates computed from independent realisations of a process.

There are methods to this **model order selection** problem: these include final prediction error (FPE), Akaike's information criterion (AIC), minimum description length (MDL), Parzen's criterion autoregressive transfer function (CAT) and B-Information criterion (BIC). However, it would be preferable to have a parameter estimation method that explicitly takes into account the fact that the model order is unknown. Although **model selection** will not be discussed in detail in this course, **Bayesian parameter estimation** is a framework in which it is consistent and straightforward to consider the **model order** as simply another unknown parameter.

5.5.1 Bayes's Theorem (Revisited)

Suppose N observations, $\mathbf{x} = \{x[n]\}_0^{N-1}$, of a random process, $x[n,\zeta]$, is denoted by $\mathbf{X}(\zeta) = \{x[n,\zeta]\}_0^{N-1}$. It is assumed that this process can be assigned a signal model, $\mathcal{I}_k$, such that it is possible to write down a **likelihood function**:

$$\mathcal{L}_k\left(\boldsymbol{\theta}_k;\, \mathbf{x}\right) = p_{\mathbf{X}|\boldsymbol{\Theta}_k}\left(\mathbf{x} \mid \boldsymbol{\theta}_k,\, \mathcal{I}_k\right) \tag{5.127}$$

where $\boldsymbol{\theta}_k$ is an unknown parameter vector which characterises the k-th signal model, $\mathcal{I}_k$. Suppose all the knowledge *prior* to observing the data regarding the probability of the values of the parameters of model $\mathcal{I}_k$ is summarised by the probability density function, $p_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k \mid \mathcal{I}_k\right)$. Then **Bayes's theorem** gives:

$$p_{\boldsymbol{\Theta}_k|\mathbf{X}}\left(\boldsymbol{\theta}_k \mid \mathbf{x},\, \mathcal{I}_k\right) = \frac{p_{\mathbf{X}|\boldsymbol{\Theta}_k}\left(\mathbf{x} \mid \boldsymbol{\theta}_k,\, \mathcal{I}_k\right)\, p_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k \mid \mathcal{I}_k\right)}{p_{\mathbf{X}}\left(\mathbf{x} \mid \mathcal{I}_k\right)} \tag{5.128}$$

Equation 5.128 is composed of the following terms:

Prior: $p_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k \mid \mathcal{I}_k\right)$ summarises all the knowledge of the values of the parameters $\boldsymbol{\theta}_k$ *prior* to observing the data;

Likelihood: $p_{\mathbf{X}|\boldsymbol{\Theta}_k}\left(\mathbf{x} \mid \boldsymbol{\theta}_k,\, \mathcal{I}_k\right)$, is determined by the signal model $\mathcal{I}_k$;

Evidence: $p_{\mathbf{X}}\left(\mathbf{x} \mid \mathcal{I}_k\right)$, which is the normalising expression in Equation 5.128, is known as the **Bayesian evidence**. Since the left hand side (LHS) must integrate to unity to be a valid pdf, then it follows:

$$p_{\mathbf{X}}\left(\mathbf{x} \mid \mathcal{I}_k\right) = \int_{\boldsymbol{\Theta}_k} p_{\mathbf{X}|\boldsymbol{\Theta}_k}\left(\mathbf{x} \mid \boldsymbol{\theta}_k,\, \mathcal{I}_k\right)\, p_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k \mid \mathcal{I}_k\right)\, d\boldsymbol{\theta}_k \tag{5.129}$$

This term is of interest in model selection; in cases where only one model is under consideration, this term may be considered as a constant, since it is not a function of the unknown parameters $\boldsymbol{\theta}_k$.

Posterior: $p_{\boldsymbol{\Theta}_k|\mathbf{X}}\left(\boldsymbol{\theta}_k \mid \mathbf{x},\, \mathcal{I}_k\right)$ is the joint **posterior pdf** for the unknown parameters $\boldsymbol{\theta}_k$ given the observations $\mathbf{x}$. It summarises the state of knowledge about the parameters *after* the data is observed.

The posterior density may be used for parameter estimation, and various estimators exist. One common estimator is the value of $\boldsymbol{\theta}_k$ that maximises the posterior pdf:

$$\hat{\boldsymbol{\theta}}_k = \arg_{\boldsymbol{\theta}_k} \max p_{\boldsymbol{\Theta}_k|\mathbf{X}}\left(\boldsymbol{\theta}_k \mid \mathbf{x}, \mathcal{I}_k\right) \tag{5.130}$$

This is known as the maximum *a posteriori* (MAP) estimate.

Note that in order to simplify the notation, Bayes's theorem is frequently written as:

$$p\left(\boldsymbol{\theta}_k \mid \mathbf{x}, \mathcal{I}_k\right) = \frac{p\left(\mathbf{x} \mid \boldsymbol{\theta}_k, \mathcal{I}_k\right)\, p\left(\boldsymbol{\theta}_k \mid \mathcal{I}_k\right)}{p\left(\mathbf{x} \mid \mathcal{I}_k\right)} \tag{5.131}$$

It is understood in Equation 5.131 that the probability density functions, $p\left(\cdot \mid \cdot\right)$, are identified based on its context. In other-words, it is important to realise that each term in Equation 5.131 represents a different functional form for the pdfs.

In cases where there is only one model in consideration, Equation 5.131 simplifies further to:

$$p\left(\boldsymbol{\theta} \mid \mathbf{x}, \mathcal{I}\right) = \frac{p\left(\mathbf{x} \mid \boldsymbol{\theta}, \mathcal{I}\right)\, p\left(\boldsymbol{\theta} \mid \mathcal{I}\right)}{p\left(\mathbf{x} \mid \mathcal{I}\right)} \tag{5.132}$$

5.5.2 The Removal of Nuisance Parameters

One of the more interesting features of the Bayesian paradigm is the ability to remove *nuisance parameters*: these are parameters that are of no interest in the analysis. Consider a signal model, $\mathcal{I}$, that involves two parameters, α and β. In this case, Bayes's theorem may be written as:

$$p\left(\alpha, \beta \mid \mathbf{x}, \mathcal{I}\right) = \frac{p\left(\mathbf{x} \mid \alpha, \beta, \mathcal{I}\right)\, p\left(\alpha, \beta \mid \mathcal{I}\right)}{p\left(\mathbf{x} \mid \mathcal{I}\right)} \tag{5.133}$$

It might be that it is only of interest to estimate α, and that an estimate of β is unnecessary. The **marginal a posteriori pdf** for α can be obtained by **marginalising** over the random variable β:

$$\begin{aligned} p\left(\alpha \mid \mathbf{x}, \mathcal{I}\right) &= \int p\left(\alpha, \beta \mid \mathbf{x}, \mathcal{I}\right) d\beta \\ &= \frac{1}{p\left(\mathbf{x} \mid \mathcal{I}\right)} \int p\left(\mathbf{x} \mid \alpha, \beta, \mathcal{I}\right)\, p\left(\alpha, \beta \mid \mathcal{I}\right) d\beta \end{aligned} \tag{5.134}$$

Marginalisation, also known as **marginal inference**, is an appealing procedure when the integral in Equation 5.134 can be calculated in closed form. In such cases, the **marginal posterior density** is reduced in dimensionality since the parameter β is no longer present in the term $p\left(\alpha \mid \mathbf{x}, \mathcal{I}\right)$. Note that marginalisation necessitates a loss of information; the integration in Equation 5.134 means that all the information about the value of β is lost.

If the marginal posterior density is used for parameter estimation, then the value of α that maximises the marginal posterior pdf:

$$\hat{\alpha} = \arg_{\alpha} \max p\left(\alpha \mid \mathbf{x}, \mathcal{I}\right) = \arg_{\alpha} \max \int p\left(\alpha, \beta \mid \mathbf{x}, \mathcal{I}\right) d\beta \tag{5.135}$$

is known as the maximum marginal *a posteriori* (MMAP) estimate.

5.5.3 Prior Probabilities

The selection of **prior densities** is a highly involved topic for discussion, and is only briefly mentioned here. A prior density is selected to describe ones state of knowledge, or lack of it, about the value of a parameter before it is observed.

One can claim to have no knowledge whatsoever about the value of a parameter prior to observing the data. This state of ignorance may be described by using a prior pdf that is very broad and flat relative to the likelihood function. The most intuitively obvious non-informative prior is a **uniform density**. This prior is typically used for discrete distributions, or for unbounded real value parameters:

$$p\left(\boldsymbol{\theta}_k \mid \mathcal{I}_k\right) = k \tag{5.136}$$

where k is a constant. In the case of an uniform prior, parameter estimates obtained from a MAP estimate are identical to those obtained using maximum likelihood. The problem with the uniform prior in Equation 5.136 is that is is not normalisable, and is therefore not a valid pdf.

Prior probabilities are non-informative if they convey ignorance of the parameter values before observing the data *compared* with the state of knowledge afterwards. Therefore, the prior pdf need only be diffuse in relation to the likelihood function. Thus, to avoid the normalisation problem with the uniform prior, frequently a Gaussian prior is adopted:

$$p\left(\boldsymbol{\theta}_k \mid \mathcal{I}_k\right) = \frac{1}{\left(2\pi\delta^2\right)^{\frac{P}{2}}} \exp\left[-\frac{\boldsymbol{\theta}_k^T\boldsymbol{\theta}_k}{2\delta^2}\right] \tag{5.137}$$

where P is the number of parameters inside the vector $\boldsymbol{\theta}_k$. The parameter δ is known as a **hyper-parameter**, and needs to be chosen somehow. To indicate ignorance of the value of a parameter, δ should be set to a large value. Alternatively, it is possible to assign another prior to the hyper-parameter δ itself. This hyper-prior will be characterised by hyper-hyper-parameters.

Often a prior is chosen for mathematical convenience. In many situations, the likelihood function has an exponential form. For the ease of analysis, the prior density can be chosen to be **conjugate** to the likelihood function so that the **posterior density** is of the same functional form as the likelihood. In general, however, it is desirable to convey all prior knowledge in a prior density function; this is problem specific, and is discussed in many many research texts.

5.5.4 General Linear Model

The general linear model has previously been introduced in the discussion on the method of least squares. Any data that may be described in terms of a linear combination of basis functions with an additive Gaussian noise component satisfies the general linear model. Suppose that the observed data may be described by a signal model of the form:

$$x[n] = \sum_{p=1}^{P} a_p\, g_p[n] + e[n]\,, \qquad \text{where} \qquad 0 \le n \le N-1 \tag{5.138}$$

and $g_p(n)$ is the value of a time-dependent model or basis function evaluated at time index n, and $e[n]$ is WGN with variance σ_e^2: thus, $e[n] \sim \mathcal{N}(0, \sigma_e^2)$. Consider writing Equation 5.138 for all values of n:

$$\underbrace{\begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} g_1[0] & g_2[0] & \cdots & g_P[0] \\ g_1[1] & g_2[1] & \cdots & g_P[1] \\ \vdots & \vdots & \ddots & \vdots \\ g_1[N-1] & g_2[N-1] & \cdots & g_P[N-1] \end{bmatrix}}_{\mathbf{G}} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_P \end{bmatrix}}_{\mathbf{a}} + \underbrace{\begin{bmatrix} e[0] \\ e[1] \\ \vdots \\ e[N-1] \end{bmatrix}}_{\mathbf{e}} \tag{5.139}$$

In other-words, Equation 5.138 may be written as:

$$\mathbf{x} = \mathbf{G}\,\mathbf{a} + \mathbf{e} \tag{5.140}$$

where $\mathbf{x}$ is an $N \times 1$ vector of observations, $\mathbf{e}$ is an $N \times 1$ vector of i. i. d. Gaussian noise samples, $\mathbf{G}$ is a $N \times P$ matrix, and $\mathbf{a}$ is a $P \times 1$ vector of parameters. The columns of matrix $\mathbf{G}$ are the basis functions evaluated at each time index, and the basis functions themselves are a function of some unknown parameters $\boldsymbol{\theta}$. For example, the basis functions might be sinusoids, and $\boldsymbol{\theta}$ denotes the frequencies of these sinusoids.

The vector-matrix equation in Equation 5.140 is linear in the parameter vector $\mathbf{a}$; hence, the model in Equation 5.140 is often called the **LITP** model. Now, consider finding the likelihood function $p(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{a}, \sigma_e^2, \mathcal{I})$, where $\boldsymbol{\theta}$ is the unknown parameter vector of the basis functions that form the matrix $\mathbf{G}$. The probability density function for the noise vector is given by:

$$p\left(\mathbf{e} \mid \sigma_e^2\right) = \frac{1}{(2\pi\sigma_e^2)^{\frac{N}{2}}} \exp\left[-\frac{\mathbf{e}^T\mathbf{e}}{2\sigma_e^2}\right] \tag{5.141}$$

Now, suppose that $\mathbf{G}$ is not a function of the observations $\mathbf{x}$; the probability transformation from the random vector $\mathbf{e}$ to the random vector $\mathbf{x}$ is linear, and has unity Jacobian. Hence, the likelihood function for the observations is given by:

$$p\left(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{a}, \sigma_e^2, \mathcal{I}\right) = \frac{1}{(2\pi\sigma_e^2)^{\frac{N}{2}}} \exp\left[-\frac{(\mathbf{x} - \mathbf{G}\,\mathbf{a})^T(\mathbf{x} - \mathbf{G}\,\mathbf{a})}{2\sigma_e^2}\right] \tag{5.142}$$

where $\mathcal{I}$ indicates all the known information in the chosen signal model. Now, suppose that the aim is to infer the values of the parameters of the basis functions, $\boldsymbol{\theta}$, without inferring the values of the nuisance parameters, namely the linear parameters, $\mathbf{a}$, and the variance of the white noise, σ_e^2. The Bayesian methodology is thus applied. First some priors are required for the variance and the linear parameters.

The variance term is known as a **scale parameter** and is a measure of scale or magnitude. A vague non-informative prior that is usually assigned to scale parameters is the **inverse-Gamma density**; the reason for this is not discussed here. Therefore:

$$p\left(\sigma_e^2 \mid \alpha_e, \beta_e\right) = \mathcal{IG}\left(\sigma_e^2 \mid \alpha_e, \beta_e\right) = \begin{cases} 0 & \text{if } \sigma_e^2 < 0, \\ \frac{\alpha_e^{\beta_e}}{\Gamma(\beta_e)} \left(\sigma_e^2\right)^{-(\beta_e+1)} e^{-\frac{\alpha_e}{\sigma_e^2}} & \text{if } \sigma_e^2 \geq 0, \end{cases} \tag{5.143}$$

Note that α_e and β_e are **hyper-parameters**. Further, for linear parameters, it is usual to apply a vague Gaussian prior similar to that in Equation 5.137:

$$p\left(\mathbf{a}\,|\,\sigma_e^2,\,\mathcal{I}\right)=\mathcal{N}\left(\mathbf{a}\,\middle|\,0,\,\delta^2\sigma_e^2\mathbf{I}_P\right)=\frac{1}{\left(2\pi\delta^2\sigma_e^2\right)^{\frac{P}{2}}}\exp\left[-\frac{\mathbf{a}^T\mathbf{a}}{2\delta^2\sigma_e^2}\right] \tag{5.144}$$

where $\mathbf{I}_P$ is the $P\times P$ identity matrix. Note that the prior $p\left(\mathbf{a}\,|\,\sigma_e^2,\,\delta,\,\mathcal{I}\right)$ is conditional on σ_e^2; the choice of this prior allows both σ_e^2 and $\mathbf{a}$ to be marginalised analytically. The hyper-parameters $\delta,\,\alpha_e,\,\beta_e$ are all assumed to be known.

Using Bayes's theorem, the posterior density for all the parameters $\boldsymbol{\theta},\,\mathbf{a},\,\sigma_e^2$ is given by:

$$p\left(\boldsymbol{\theta},\,\mathbf{a},\,\sigma_e^2\,\middle|\,\mathbf{x},\,\mathcal{I}\right)\propto p\left(\mathbf{x}\,|\,\boldsymbol{\theta},\,\mathbf{a},\,\sigma_e^2,\,\mathcal{I}\right)p\left(\boldsymbol{\theta},\,\mathbf{a},\,\sigma_e^2\,\middle|\,\mathcal{I}\right) \tag{5.145}$$

where the evidence term is considered as a constant and therefore omitted, and $\propto$ indicates proportionality. The prior term factorises as:

$$p\left(\boldsymbol{\theta},\,\mathbf{a},\,\sigma_e^2\right)=p\left(\boldsymbol{\theta}\right)p\left(\mathbf{a}\,|\,\sigma_e^2\right)p\left(\sigma_e^2\right) \tag{5.146}$$

where the dependence on the model $\mathcal{I}$ has been dropped for convenience. Thus, the joint posterior density is given by:

$$\begin{aligned}p\left(\boldsymbol{\theta},\,\mathbf{a},\,\sigma_e^2\,\middle|\,\mathbf{x},\,\mathcal{I}\right)\;&\propto p\left(\boldsymbol{\theta}\right)\,\frac{1}{\left(2\pi\sigma_e^2\right)^{\frac{N}{2}}}\exp\left[-\frac{\left(\mathbf{x}-\mathbf{G}\,\mathbf{a}\right)^T\left(\mathbf{x}-\mathbf{G}\,\mathbf{a}\right)}{2\sigma_e^2}\right]\\&\quad\times\frac{1}{\left(2\pi\delta^2\sigma_e^2\right)^{\frac{P}{2}}}\exp\left[-\frac{\mathbf{a}^T\mathbf{a}}{2\delta^2\sigma_e^2}\right]\frac{\alpha_e^{\beta_e}}{\Gamma(\beta_e)}\,\left(\sigma_e^2\right)^{-(\beta_e+1)}\,e^{-\frac{\alpha_e}{\sigma_e^2}}\end{aligned} \tag{5.147}$$

Since the observations and hyper-parameters are known, and therefore constant from the perspective of the posterior density, then after some manipulation, this may be written as

$$p\left(\boldsymbol{\theta},\,\mathbf{a},\,\sigma_e^2\,\middle|\,\mathbf{x},\,\mathcal{I}\right)\propto\frac{p\left(\boldsymbol{\theta}\right)}{\left(\sigma_e^2\right)^{\frac{N+P}{2}+\beta_e+1}}\exp\left[-\frac{\mathbf{a}^T\left(\mathbf{G}^T\mathbf{G}+\delta^{-2}\mathbf{I}_P\right)\mathbf{a}-2\mathbf{x}^T\mathbf{G}\mathbf{a}+\mathbf{x}^T\mathbf{x}+2\alpha_e}{2\sigma_e^2}\right] \tag{5.148}$$

The linear parameters $\mathbf{a}$ can be marginalised out using the identity:

$$\int_{\mathbb{R}^P}\exp\left\{-\frac{1}{2}\left[\alpha+2\mathbf{y}^T\boldsymbol{\beta}+\mathbf{y}^T\boldsymbol{\Gamma}\mathbf{y}\right]\right\}d\mathbf{y}=\frac{\left(2\pi\right)^{\frac{P}{2}}}{\left|\boldsymbol{\Gamma}\right|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left[\alpha-\boldsymbol{\beta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\beta}\right]\right\} \tag{5.149}$$

To perform this, set $\mathbf{y}=\mathbf{a}$, $\boldsymbol{\Gamma}=\frac{1}{\sigma_e^2}\left(\mathbf{G}^T\mathbf{G}+\delta^{-2}\mathbf{I}_P\right)$, $\alpha=\frac{\mathbf{x}^T\mathbf{x}+2\alpha_e}{\sigma_e^2}$, and $\boldsymbol{\beta}=-\frac{1}{\sigma_e^2}\mathbf{G}^T\mathbf{x}$, so that

$$p\left(\boldsymbol{\theta},\,\sigma_e^2\,\middle|\,\mathbf{x},\,\mathcal{I}\right)=\int_{-\infty}^{\infty}p\left(\boldsymbol{\theta},\,\mathbf{a},\,\sigma_e^2\,\middle|\,\mathbf{x},\,\mathcal{I}\right)\,d\mathbf{a} \tag{5.150}$$

$$\propto\frac{p\left(\boldsymbol{\theta}\right)}{\sqrt{\det\left|\mathbf{G}^T\mathbf{G}+\delta^{-2}\mathbf{I}_P\right|\left(\sigma_e^2\right)^{R+1}}}\exp\left[-\frac{\mathbf{x}^T\mathbf{x}+2\alpha_e-\mathbf{x}^T\mathbf{G}\left(\mathbf{G}^T\mathbf{G}+\delta^{-2}\mathbf{I}_P\right)^{-1}\mathbf{G}^T\mathbf{x}}{2\sigma_e^2}\right] \tag{5.151}$$

where $R = \frac{N+2\beta_e}{2}$. Finally, the variance can be marginalised using the fact that the inverse-Gamma pdf implies:

$$1 = \int_0^\infty \mathcal{IG}\left(\sigma^2 \,|\, \alpha,\, \beta,\ \right) d\sigma^2 = \int_0^\infty \frac{\alpha^\beta}{\Gamma(\beta)} \left(\sigma^2\right)^{-(\beta+1)} e^{-\frac{\alpha}{\sigma^2}}\, d\sigma^2 \tag{5.152}$$

and therefore:

$$\int_0^\infty \left(\sigma^2\right)^{-(\beta+1)} e^{-\frac{\alpha}{\sigma^2}}\, d\sigma^2 = \frac{\Gamma(\beta)}{\alpha^\beta} \tag{5.153}$$

Hence, this gives the **marginal a posterior pdf** for the parameters $\boldsymbol{\theta}$ as

$$\begin{aligned} p\left(\boldsymbol{\theta} \,|\, \mathbf{x}, \mathcal{I}\right) &= \int_0^\infty p\left(\boldsymbol{\theta},\, \sigma_e^2 \,|\, \mathbf{x}, \mathcal{I}\right) d\sigma_e^2 \\ &\propto p\left(\boldsymbol{\theta}\right) \frac{\left[\mathbf{x}^T\mathbf{x} + 2\alpha_e - \mathbf{x}^T\mathbf{G}\left(\mathbf{G}^T\mathbf{G} + \delta^{-2}\mathbf{I}_P\right)^{-1}\mathbf{G}^T\mathbf{x}\right]^{-\left(\frac{N}{2}+\beta_e\right)}}{\sqrt{\det\left|\mathbf{G}^T\mathbf{G} + \delta^{-2}\mathbf{I}_P\right|}} \end{aligned} \tag{5.154}$$

The MMAP estimate can be found by maximising this expression with respect to the parameters $\boldsymbol{\theta}$ which are *implicitly* incorporated in the basis matrix $\mathbf{G}$.

It is important to realise that the expression in Equation 5.154 is a function of the basis parameters $\boldsymbol{\theta}$ only. This means that there is no need to know about the standard deviation, σ_e^2, nor the values of the linear parameters to infer the values of $\boldsymbol{\theta}$. Moreover, since the integrals in the marginalisation process have been performed analytically, the dimensionality of the parameter space has been reduced for each parameter integrated out. This reduction of the dimensionality is a property of Bayesian marginal estimates and is a major advantage in many applications.

Example 5.11 (Frequency estimation). An application of the general linear model is in frequency estimation. Suppose that a signal, $s[n]$, is modelled as the sum of sinusoids:

$$s[n] = \sum_{p=1}^P \left(a_p \sin\omega_p n + b_p \cos\omega_p n\right) \tag{5.155}$$

where the coefficients $\{a_p,\, b_p\}_1^P$ are the amplitudes, $\{\omega_p\}_1^P$ are the frequencies, and P is the model order. As usual, it is implicitly assumed that the sampling period $T = 1$ and that the frequencies $\{\omega_p\}_1^P$ are normalised to between 0 and π. The signal, $s[n]$, is observed in white Gaussian noise (WGN) with unknown variance σ_e^2:

$$x[n] = s[n] + e[n] = \sum_{p=1}^P \left(a_p \sin\omega_p n + b_p \cos\omega_p n\right) + e[n] \tag{5.156}$$

This model can be written in the linear in the parameters (LITP) form by defining the matrix:

$$\mathbf{G} = \begin{bmatrix} 0 & 1 & 0 & 1 & \cdots & 0 & 1 \\ \sin\omega_1 & \cos\omega_1 & \sin\omega_2 & \cos\omega_2 & \cdots & \sin\omega_P & \cos\omega_P \\ \sin 2\omega_1 & \cos 2\omega_1 & \sin 2\omega_2 & \cos 2\omega_2 & \cdots & \sin 2\omega_P & \cos 2\omega_P \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \sin \ell\omega_1 & \cos \ell\omega_1 & \sin \ell\omega_2 & \cos \ell\omega_2 & \cdots & \sin \ell\omega_P & \cos \ell\omega_P \end{bmatrix} \tag{5.157}$$

where $\ell = N - 1$. Hence, with the parameter vector defined as:

$$\mathbf{a} = \begin{bmatrix} a_1 & b_1 & a_2 & b_2 & \cdots & a_P & b_P \end{bmatrix}^T \tag{5.158}$$

the **marginal a posterior pdf** for the unknown frequencies $\{\omega_p\}_1^P$ is given by:

$$p\left(\{\omega_p\}_1^P \mid \mathbf{x}\right) \propto p\left(\{\omega_p\}_1^P\right) \frac{\left[\mathbf{x}^T\mathbf{x} + 2\alpha_e - \mathbf{x}^T\mathbf{G}\left(\mathbf{G}^T\mathbf{G} + \delta^{-2}\mathbf{I}_{2P}\right)^{-1}\mathbf{G}^T\mathbf{x}\right]^{-\left(\frac{N}{2}+\beta_e\right)}}{\sqrt{\det\left|\mathbf{G}^T\mathbf{G} + \delta^{-2}\mathbf{I}_{2P}\right|}} \tag{5.159}$$

⋈

where the parameter vector, $\mathbf{a}$, is of dimension $2P$, and therefore the size of $\mathbf{G}$ is $N \times 2P$.

The MMAP estimate can be found by maximising this w. r. t. the frequencies $\{\omega_p\}_1^P$. Note that the hyper-parameters and a prior for $\{\omega_p\}_1^P$ must also be chosen; typically, a uniform prior on ω_p between 0 and π will be sufficient.

5.5.4.1 Model Selection using Bayesian Evidence

Next, the Bayesian evidence term is considered:

$$p_{\mathbf{X}}\left(\mathbf{x} \mid \mathcal{I}_k\right) = \int_{\boldsymbol{\Theta}_k} p_{\mathbf{X}|\boldsymbol{\Theta}_k}\left(\mathbf{x} \mid \boldsymbol{\theta}_k, \mathcal{I}_k\right) \, p_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k \mid \mathcal{I}_k\right) \, d\boldsymbol{\theta}_k \tag{5.160}$$

This term can be used to select signal models and noise statistics appropriate to the observed data. To clarify, in this equation, $\boldsymbol{\Theta}_k$ is the parameter space, and $\mathcal{I}_k$ denotes the structure of the k-th model. The term $\mathcal{I}_k$ represents the joint assumption of *both* the noise statistics and the signal model; together, this is called the **data model**. It is important to note that the integral in Equation 5.160 is the likelihood multiplied by the prior *integrated* over *all* the parameters in that data model. In the case of discrete distributions, the integration simplifies to a summation.

Consider a set of competing possible data models labelled $\{\mathcal{I}_k\}_1^M$ proposed to describe a given set of observations. Bayes's theorem can be used to find the posterior density of each model given the data:

$$p_{\mathcal{I}|\mathbf{X}}\left(\mathcal{I}_k \mid \mathbf{x}\right) = \frac{p_{\mathbf{X}|\mathcal{I}}\left(\mathbf{x} \mid \mathcal{I}_k\right) \, p_{\mathcal{I}}\left(\mathcal{I}_k\right)}{p_{\mathbf{X}}\left(\mathbf{x}\right)} \tag{5.161}$$

where the probability of the observations is given by:

$$p_{\mathbf{X}}\left(\mathbf{x}\right) = \sum_{k=1}^{M} p_{\mathbf{X}|\mathcal{I}}\left(\mathbf{x} \mid \mathcal{I}_k\right) \, p_{\mathcal{I}}\left(\mathcal{I}_k\right) \tag{5.162}$$

If all the models are equally likely *a priori*, then

$$p_{\mathcal{I}}\left(\mathcal{I}_k\right) = \frac{1}{M} \tag{5.163}$$

Therefore, the posterior probability of a model is given by the **relative evidence**:

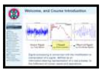
$$p_{\mathcal{I}|\mathbf{x}}\left(\mathcal{I}_k \mid \mathbf{x}\right) = \frac{p_{\mathbf{x}|\mathcal{I}}\left(\mathbf{x} \mid \mathcal{I}_k\right)}{\sum\limits_{k=1}^{M} p_{\mathbf{x}|\mathcal{I}}\left(\mathbf{x} \mid \mathcal{I}_k\right)} \tag{5.164}$$

This expression constitutes the evidence framework for the selection of signal models. It is important to realise that in terms of real data, the correct data model may not be in the set chosen. It is only possible to compare the candidate models that have been considered to determine which models are more plausible.

6 Monte Carlo Methods

This handout discusses the problem of generating sequences of random numbers or variates, for use in numerical simulations, including Monte Carlo integration and optimisation.

6.1 Introduction



New slide

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

Optimisation: involves finding the solution to

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} h(\boldsymbol{\theta}) \tag{6.1}$$

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. An example of a complicated optimisation problem might be finding the maximum of the equation:

$$h(x) = (\cos 50x + \sin 20x)^2\,, \quad 0 \leq x \leq 1 \tag{6.2}$$

This function is plotted in Figure 6.1.

Integration: involves evaluating an integral,

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{6.3}$$

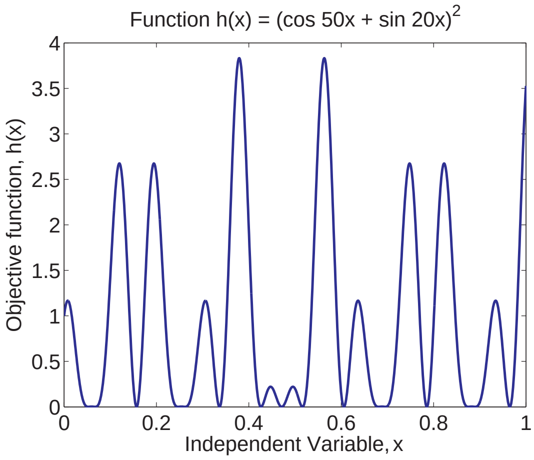


Figure 6.1: Plot of the function in Equation 6.2.

that cannot explicitly be calculated in *closed form*. For example, the Gaussian-error function:

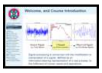
$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}} \, d\theta \tag{6.4}$$

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

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

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \int_{\Omega} f(\boldsymbol{\theta}, \boldsymbol{\omega}) \, d\boldsymbol{\omega} \tag{6.5}$$

The reader is encouraged to honestly consider how many problems they solve reduce to either an integration or an optimisation problem.



New slide

6.1.1 Deterministic Numerical Methods

There are various deterministic solutions to the optimisation and integration problems. A browse through [Press:1992, Chapters 4 and 10], for example, reveals a variety of well-known approaches:

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.

Integration: Most deterministic integration is only feasible in one-dimension, and many methods rely on classic formulas for equally spaced abscissas:

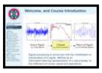
1. simple Riemann integration;
2. standard and extended Simpson's and Trapezoidal rules;
3. refinements such as Romberg Integration.

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.

Some examples of deterministic numerical solutions to these problems are considered in Section 6.1.1.1 and Section 6.1.1.2.

6.1.1.1 Deterministic Optimisation



New slide

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 in some cases, it can be robust.

Gradient methods are typically based on the Newton-Raphson algorithm which solves the equation $\nabla h(\boldsymbol{\theta}) = \mathbf{0}$. For a scalar function, $h(\boldsymbol{\theta})$, of a vector of independent variables $\boldsymbol{\theta}$, a sequence $\boldsymbol{\theta}_n$ is produced such that:

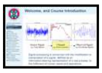
$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \left(\nabla\, \nabla^T h\left(\boldsymbol{\theta}_n\right)\right)^{-1} \nabla h\left(\boldsymbol{\theta}_n\right) \tag{6.6}$$

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

The primary difficulty in evaluating Equation 6.6 is the computation of the Hessian term $\nabla\, \nabla^T h\left(\boldsymbol{\theta}_n\right)$. However, it is not crucial to obtain an exact estimate of the Hessian in order to reduce the cost function at each iteration. In fact, any *positive definite* matrix will suffice, and often a matrix proportional to the identity matrix is used.

The Broyden-Fletcher-Goldfarb-Shannon (BFGS) algorithm, for example, constructs an approximate Hessian matrix by analyzing successive gradient vectors, and by assuming that the function can be locally approximated as a quadratic function in the region around the optimum.

6.1.1.2 Deterministic Integration



New slide

Numerical computation of the scalar case of the integral in Equation 6.7 can be done using simple **Riemann integration**, or by improved methods such as the **trapezoidal rule**. For example, the

$$\mathcal{I} = \int_a^b f(\theta)\, d\theta, \tag{6.7}$$

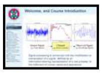
where θ is a scalar, and $b > a$, can be solved with the trapezoidal rule using

$$\hat{I} = \frac{1}{2}\sum_{k=0}^{N-1} (\theta_{k+1} - \theta_k)\,(f(\theta_k) + f(\theta_{k+1})) \tag{6.8}$$

where the θ_k's constitute an ordered partition of $[a,\, b]$. Another formula is **Simpson's rule**:

$$\hat{I} = \frac{\delta}{3}\left\{ f(a) + 4\sum_{k=1}^{N} f(\theta_{2k-1}) + 2\sum_{k=1}^{N} h(\theta_{2k}) + f(b) \right\} \tag{6.9}$$

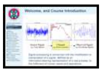
in the case of equally spaced samples with $\delta = \theta_{k+1} - \theta_k$.



6.1.2 Monte Carlo Numerical Methods

New slide

Monte Carlo methods are stochastic techniques, in which random numbers are generated and use to examine some problem.



6.1.2.1 Monte Carlo Integration

New slide

Consider the integral,

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{6.10}$$

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:

$$\mathcal{I} = \int_{\Theta} \frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{6.11}$$

where the function $\pi(\boldsymbol{\theta}) > 0$, $\boldsymbol{\theta} \in \Theta$ is a probability density function (pdf) which satisfies the normalised expression:

$$\int_{\Theta} \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} = 1 \tag{6.12}$$

It can now be seen that Equation 6.57 can be viewed as an expectation of the function $h(\boldsymbol{\theta}) = f(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})^{-1}$ over the pdf of $\pi(\boldsymbol{\theta})$. In other-words, Equation 6.57 becomes

This may be written as an expectation:

$$\mathcal{I} = \mathbb{E}_{\pi}\left[\frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right] \tag{6.13}$$

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 $\pi(\boldsymbol{\theta})$,

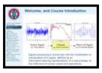
$$\boldsymbol{\theta}^{(k)} \sim \pi(\boldsymbol{\theta}), \quad k \in \mathcal{N} = \{0,\, \ldots,\, N-1\} \tag{6.14}$$

2. Calculate the sample average of the expectation in Equation 6.13 using

$$\hat{\mathcal{I}} = \frac{1}{N}\sum_{k=0}^{N-1} \frac{f(\boldsymbol{\theta}^{(k)})}{\pi(\boldsymbol{\theta}^{(k)})} \approx \mathbb{E}_\pi\left[\frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right] \tag{6.15}$$

This technique is known as **importance sampling** because the function $f(\boldsymbol{\theta})$ is sampled with the density $\pi(\boldsymbol{\theta})$, thereby giving more *importance* to some values of $f(\boldsymbol{\theta})$ than others.

6.1.2.2 Stochastic Optimisation



New slide

There are two distinct approaches to the Monte Carlo optimisation (here, maximisation) of the objective function $h(\boldsymbol{\theta})$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} h(\boldsymbol{\theta}) \tag{6.16}$$

The first method is broadly known as an **exploratory approach**, while the second approach is based on a **probabilistic approximation** of the objective function.

Exploratory approach This approach is an exploratory method in that it is concerned with fast *explorations* of the sample space rather than working with the objective function directly.

For example, Equation 6.16 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:

$$\hat{\boldsymbol{\theta}} \approx \arg\max_{\{\boldsymbol{\theta}^{(k)}\}} h\left(\boldsymbol{\theta}^{(k)}\right) \tag{6.17}$$

Typically, when no specific features regarding the function $h\left(\boldsymbol{\theta}\right)$, are taken into account, $\pi(\boldsymbol{\theta})$ will take on a uniform distribution over Θ. Although this method converges as $N \to \infty$, the method is very slow: one can usually do better by finding a density $\pi(\boldsymbol{\theta})$ that is related to $h\left(\boldsymbol{\theta}\right)$, but this requires some additional insight into the function $h\left(\boldsymbol{\theta}\right)$.

Stochastic Approximation

- The Monte Carlo EM algorithm

A more sophisticated approach to **stochastic exploration** is based on the deterministic gradient-based methods. A modified form of Equation 6.6 is:

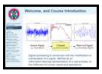
$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + \mathbf{G}_n \nabla h\left(\boldsymbol{\theta}_n\right) \tag{6.18}$$

where $\mathbf{G}_n$ is a sequence which may approximate the Hessian of $h\left(\boldsymbol{\theta}_n\right)$ in order to ensure the algorithm converges.

6.1.2.3 Implementation issues

Monte Carlo methods rely on the assumption that is is possible to simulate **samples** or **variates** $\{\boldsymbol{\theta}^{(k)}\}$ from the density $\pi(\theta)$.

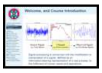
The next sections address how such samples can be obtained.



New slide

6.2 Generating Random Variables

This section discusses a variety of techniques for generating random variables from a different distributions.



New slide

6.2.1 Uniform Variates

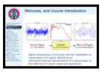
The foundation underpinning all stochastic simulations is the ability to generate a sequence of independent and identically distributed (i. i. d.) uniform random variates over the range $(0, 1]$. All random variates are generated using techniques that assume **uniform random variates** are available.

Random variates are *pseudo* or *synthetic* and not truly random since they are usually generated using a recurrence of the form:

$$x_{n+1} = (a\, x_n + b) \mod m \tag{6.19}$$

This is known as the linear congruential generator. For the purposes of generating random variates, it is importance that knowledge of a particular set of variates gives no discernible knowledge of the next variate drawn *provided* that the transformation in Equation 6.19 is unknown. Of course, given the sample x_0, and the parameters $\{a,\, b,\, m\}$, the samples $\{x_1,\, \dots,\, x_n\}$ are always the same.

However, suitable values of a, b and m can be chosen such that the random variates pass all statistical tests of randomness.



New slide

6.2.2 Transformation Methods

It is possible to sample from a number of extremely important probability distributions by being able to sample from the simplest of distribution functions, namely the uniform density, and then applying various probability transformation methods. *Assuming* that it is possible to sample from the uniform distribution, this section gives an overview of the methods for obtaining **variates** from other well-known distributions.

Beyond the basic definitions of random variables (RVs), the fundamental probability transformation rule forms the basis of most of the methods described in this section.

Theorem 6.1 (Probability transformation rule). Denote the real roots of $y = g(x)$ by $\{x_n,\, n \in \mathcal{N}\}$, such that

$$y = g(x_1) = \cdots = g(x_N) \tag{6.20}$$

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

$$f_Y\,(y) = \sum_{n=1}^{N} \frac{f_X\,(x_n)}{|g'(x_n)|} \tag{6.21}$$

where $g'(x)$ is the derivative with respect to (w. r. t.) x of $g(x)$.

PROOF. The proof is given in the handout on scalar random variables.

6.2.3 Generating white Gaussian noise (WGN) samples

Recall that the **probability transformation rule** takes random variables from one distribution as inputs and outputs random variables in a new distribution function:

Theorem 6.2 (Probability transformation rule (revised)). If $\{x_1, \dots x_n\}$ are random variables with a joint-pdf $f_X\,(x_1, \dots, x_n)$, and if $\{y_1, \dots y_n\}$ are random variables obtained from functions of $\{x_k\}$, such that $y_k = g_k(x_1,\, x_2 \dots x_n)$, then the joint-pdf, $f_Y\,(y_1, \dots, y_n)$, is given by:

$$f_Y\,(y_1, \dots, y_n) = \frac{1}{|J(x_1, \dots, x_n)|} f_X\,(x_1, \dots, x_n) \tag{6.22}$$

where $J(x_1,\, \dots,\, x_n)$ is the **Jacobian** of the transformation given by:

$$J(x_1, \dots, x_n) = \frac{\partial(y_1, \dots y_n)}{\partial(x_1, \dots x_n)} \tag{6.23}$$

♢

One particular well-known example is the *Box-Muller* (1958) transformation that takes two uniformly distributed random variables, and transforms them to a bivariate Gaussian distribution. Consider the transformation between two uniform random variables given by,

$$f_{X_k}\,(x_k) = \mathbb{I}_{0,1}\,(x_k)\,, \quad k = 1, 2 \tag{6.24}$$

where $\mathbb{I}_{\mathcal{A}}\,(x) = 1$ if $x \in \mathcal{A}$, and zero otherwise, and the two random variables y_1, y_2 given by:

$$y_1 = \sqrt{-2 \ln x_1}\, \cos 2\pi x_2 \tag{6.25}$$

$$y_2 = \sqrt{-2 \ln x_1}\, \sin 2\pi x_2 \tag{6.26}$$

It follows, by rearranging these equations, that:

$$x_1 = \exp\left[-\frac{1}{2}(y_1^2 + y_2^2)\right] \tag{6.27}$$

$$x_2 = \frac{1}{2\pi} \arctan \frac{y_2}{y_1} \tag{6.28}$$

The Jacobian determinant can be calculated as:

$$J(x_1, x_2) = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{vmatrix} = \begin{vmatrix} \frac{-1}{x_1\sqrt{-2\ln x_1}} \cos 2\pi x_2 & -2\pi\sqrt{-2\ln x_1}\,\sin 2\pi x_2 \\ \frac{-1}{x_1\sqrt{-2\ln x_1}} \sin 2\pi x_2 & 2\pi\sqrt{-2\ln x_1}\,\cos 2\pi x_2 \end{vmatrix} = \frac{2\pi}{x_1} \tag{6.29}$$

Hence, it follows:

$$f_Y\left(y_1, y_2\right) = \frac{x_1}{2\pi} = \left[\frac{1}{\sqrt{2\pi}} e^{-y_1^2/2}\right]\left[\frac{1}{\sqrt{2\pi}} e^{-y_2^2/2}\right] \tag{6.30}$$

since the domain $[0, 1]^2$ is mapped to the range $(-\infty, \infty)^2$, thus covering the range of real numbers. This is the product of y_1 alone and y_2 alone, and therefore each y is i. i. d. according to the normal distribution, as required.

Consequently, this transformation allows one to sample from a uniform distribution in order to obtain samples that have the same pdf as a Gaussian random variable.

Example 6.1 (MSc. Exam Question, 2005).

1. Let U be a random variable generated from a uniform pdf on the interval $[0,\ 1]$, such that

$$f_U\left(u\right) = \begin{cases} 1, & \text{if } 0 \leq u \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

Show the random variable $X = -\frac{1}{\lambda}\log U$ has an exponential distribution with parameter λ, where $\log U$ is the natural logarithm of U.

2. Let Y be a Beta random variable with parameters α and $1-\alpha$, where $0 \leq \alpha < 1$, such that it has pdf:

$$f_Y\left(y\right) = \begin{cases} \frac{1}{B(\alpha, 1-\alpha)} y^{\alpha-1}\left(1-y\right)^{-\alpha}, & 0 \leq y \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

where $B(a, b)$ is the Beta function.

The independent random variables X, from part 1, and Y are transformed to give two new random variables $W = X$ and $Z = XY$.

Show that the joint-pdf of W and Z is given by:

$$f_{WZ}\left(w,\ z\right) = \begin{cases} \frac{\lambda}{B(\alpha,1-\alpha)}\, e^{-\lambda\, w}\, z^{\alpha-1}\, \left(w-z\right)^{-\alpha}, & \text{if } (w,\ z) \in \mathcal{R} \\ 0, & \text{otherwise} \end{cases}$$

and write down the region $\mathcal{R}$ over which the density is non-zero.

3. Hence, show that the marginal-pdf of the random variable Z is Gamma distributed. Use the substitution $g = \lambda(w - z)$ where appropriate.

You may assume that the Beta function may be written as:

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \qquad \text{where} \qquad \Gamma(p) = \int_0^\infty x^{p-1}\, e^{-x}\, dx \qquad \bowtie$$

is the Gamma function with $\Gamma(1) = 1$.

4. Suppose two random number generators are available, one which generates samples from a uniform distribution, and the other from a beta distribution.

   Describe an algorithm that generates random samples from a Gamma distribution.

SOLUTION. 1. The transformation $X = g(U) = -\frac{1}{\lambda}\log U$ for $0 \le u \le 1$ has a single root:

$$u = \begin{cases} e^{-\lambda x} & \text{if } x \ge 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.31}$$

The derivative of the function $X = g(U)$ for $0 \le u \le 1$ is given by:

$$g'(u) = \frac{dg(u)}{du} = -\frac{1}{\lambda u} \tag{6.32}$$

Hence, noting that the pdf for the RV U is uniform, then the pdf for X is:

$$f_X(x) = \sum_{n=1}^{N} \frac{f_U(u_n)}{|g'(u_n)|} = \begin{cases} \frac{1}{\frac{1}{\lambda u}} = \lambda u & \text{if } 0 \le u \le 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.33}$$

which gives the desired exponential distribution with pdf:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \ge 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.34}$$

2. Consider the transformation from the two RVs X and Y to the two new random variables $W = X$ and $Z = XY$. In this case, the probability transformation rule for two random variables is required. This is a straightforward extension of the scalar case, but the Jacobian needs to be evaluated:

$$J = \frac{\partial(w,\, z)}{\partial(x, y)} = \begin{vmatrix} \frac{\partial w}{\partial x} & \frac{\partial w}{\partial y} \\ \frac{\partial z}{\partial x} & \frac{\partial z}{\partial y} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ y & x \end{vmatrix} = x = w \tag{6.35}$$

Moreover, note that there is one root of the transformation, and this is given by:

$$x = w \qquad \text{and} \qquad y = \frac{z}{w} \tag{6.36}$$

Since X and Y are independent RVs, the joint-pdf of W and Z is therefore:

$$f_{WZ}(w,\, z) = \frac{1}{J} f_{XY}(x,\, y) = \frac{1}{w} f_X(w)\, f_Y\left(\frac{z}{w}\right) \tag{6.37}$$

Note that if $x < 0$, then $f_X(x) = 0$. Moreover, if $y < 0$ or $y > 1$, then $f_Y(y) = 0$. Thus, z varies between $0 \times w$ and $1 \times w$. Thus, the regions of non-zero probability density is shown in Figure 6.2

Figure 6.2: Region of non-zero probability density

Substituting for $f_X(x)$ and $f_Y(y)$ in the non-zero region gives:

$$f_{WZ}(w,\, z) = \frac{1}{w}\,\lambda e^{-\lambda\, w}\,\frac{1}{B(\alpha, 1-\alpha)}\left(\frac{z}{w}\right)^{\alpha-1}\left(1-\frac{z}{w}\right)^{-\alpha} \tag{6.38}$$

$$= \frac{\lambda}{B(\alpha, 1-\alpha)} e^{-\lambda\, w}\, z^{\alpha-1} w^{-\alpha}\left(\frac{w-z}{w}\right)^{-\alpha} \tag{6.39}$$

which gives the desired result:

$$f_{WZ}(w,\, z) = \begin{cases} \frac{\lambda}{B(\alpha,1-\alpha)} e^{-\lambda\, w}\, z^{\alpha-1}\,(w-z)^{-\alpha} & w \geq 0 \quad \text{and} \quad 0 \leq z \leq w \\ 0 & \text{otherwise} \end{cases} \tag{6.40}$$

3. The marginal-pdf of Z is given by integrating over w:

$$f_Z(w) = \int_z^\infty f_{WZ}(w,\, z)\; dw \tag{6.41}$$

The limits of this integration are obtained by looking back at Figure 6.2, and considering the values of w for a fixed value of z. Hence, for $z > 0$,

$$f_Z(z) = \int_z^\infty \frac{\lambda}{B(\alpha, 1-\alpha)} e^{-\lambda\, w}\, z^{\alpha-1}\,(w-z)^{-\alpha}\; dw \tag{6.42}$$

$$= \frac{\lambda}{B(\alpha, 1-\alpha)} z^{\alpha-1} \int_z^\infty e^{-\lambda\, w}\;(w-z)^{-\alpha}\; dw \tag{6.43}$$

Making the substitution $g = \lambda(w-z)$, such that when $w = z$, $g = 0$, and when $w \to \infty$, $g \to \infty$. Further, $dg = \lambda\, dw$. Therefore,

$$f_Z(z) = \frac{\lambda}{B(\alpha, 1-\alpha)} z^{\alpha-1} \int_0^\infty e^{-(g+\lambda\, z)}\,\left(\frac{g}{\lambda}\right)^{-\alpha}\frac{dg}{\lambda} \tag{6.44}$$

$$= \frac{\lambda^\alpha}{B(\alpha, 1-\alpha)} z^{\alpha-1}\, e^{-\lambda\, z} \int_0^\infty e^{-g}\, g^{-\alpha}\, dg \tag{6.45}$$

Finally, using the identities given in the question:

$$B(\alpha, 1-\alpha) = \frac{\Gamma(\alpha)\Gamma(1-\alpha)}{\Gamma(1)} \qquad \text{where} \qquad \Gamma(1-\alpha) = \int_0^\infty x^{1-\alpha-1}\, e^{-x}\, dx \tag{6.46}$$

where $\Gamma(1) = 1$, then it follows that:

$$f_Z(z) = \frac{\lambda^\alpha}{\Gamma(\alpha)\Gamma(1-\alpha)} z^{\alpha-1}\, e^{-\lambda\, z}\,\Gamma(1-\alpha) = \frac{\lambda^\alpha}{\Gamma(\alpha)} z^{\alpha-1}\, e^{-\lambda\, z}, \quad z \geq 0 \tag{6.47}$$

and zero otherwise, which, using the definition given in the notes, is a Gamma distribution with parameters λ and α: $f_Z(z) = \Gamma(z \mid \lambda,\, \alpha)$.

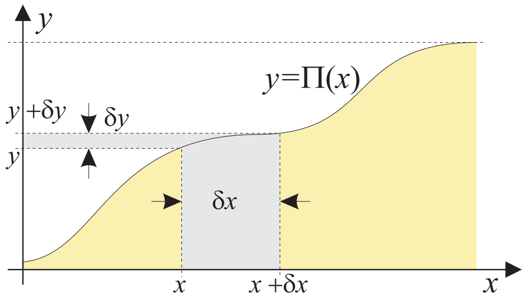
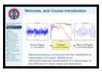


Figure 6.3: A simple derivation of the inverse transform method

4. To generate a Gamma random variable, assuming that a uniform and beta random number generators are available, the algorithm is thus:
   (a) Generate random variate, u, between 0 and 1 from uniform generator.
   (b) Generate variate, y, from the beta generator with parameters $\alpha,\ 1-\alpha$.
   (c) Calculate $x = -\frac{1}{\lambda}\log u$.
   (d) Calculate product $z = xy$; z is a variate from a Gamma distribution with parameters λ and α. □

Note, in the above example, a Beta generator is required. It is possible to generate Beta random variates when the distribution has integer parameters using *order statistics*.

6.2.4 Inverse Transform Method


New slide

There are various ways of deriving the inverse transform method, but a straightforward approach follows a similar line to the derivation of the probability transformation rule.

Referring to Figure 6.3, suppose that $X(\zeta)$ and $Y(\zeta)$ are RVs related by the function $Y(\zeta) = \Pi(X(\zeta))$. The function $\Pi(\zeta)$ is monotonically increasing so that there is only one solution to the equation $y = \Pi(x)$, and this solution is denoted by$x = \Pi^{-1}(y)$.

Writing the probability transformation rule in an inverted form:

$$f_X(x) = \frac{d\Pi(x)}{dx} f_Y(y) \tag{6.48}$$

Now, suppose $\Pi(x)$ only takes on values in the range $[0,1]$, and that $Y(\zeta) \sim \mathcal{U}_{[0,\,1]}$ is a uniform random variable. If the function $\Pi(x)$ is the cumulative distribution function (cdf) corresponding to a desired pdf $\pi(x)$, then since $\pi(x)$ and $\Pi(x)$ are related by the equation

$$\pi(x) = \frac{d\Pi(x)}{dx} \tag{6.49}$$

it follows that

$$f_X(x) = \pi(x), \quad \text{where} \quad x = \Pi^{-1}(y) \tag{6.50}$$

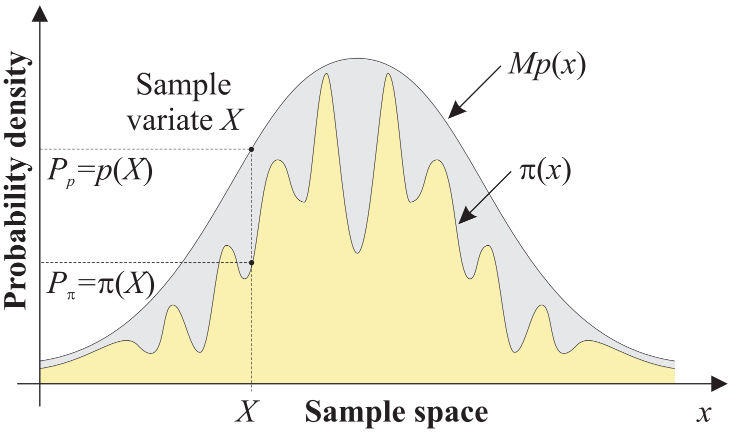


Figure 6.4: Rejection sampling

In otherwords, if

$$U(\zeta) \sim \mathcal{U}_{[0,\,1]},\; X(\zeta) = \Pi^{-1}U(\zeta) \sim \pi\left(x\right) \tag{6.51}$$

Example 6.2 (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).

6.2.5 Acceptance-Rejection Sampling

New slide

For most distributions, it is often difficult or even impossible to directly simulate using either the inverse transform or probability transformations. If if the distribution could be represented in an usable form, such as a transformation or as mixture, it would in principle be possible to exploit directly the probabilistic properties to derive a simulation method; unfortunately, it is not usually possible to make such representations.

Thus, **acceptance-rejection sampling** is a flexible class of methods that relies on the simpler requirement of finding a density $p\left(x\right)$ from which it is *easy* to sample from, where $Mp\left(x\right) > \pi\left(x\right)$.

The basic idea of acceptance-rejection sampling is shown in Figure 6.4. It is desired to sample from the distribution $\pi\left(x\right)$ which cannot be sampled from using the transform methods above. However, assume it has been possible to find a proper density $p\left(x\right)$ and a constant M such that $Mp\left(x\right) > \pi\left(x\right)$. This is shown in Figure 6.4 as a *generous envelope* around the desired function. For simplicity of explanation, assume that $M = 1$.

Imagine now that a sample variate X has been drawn from the density $p\left(x\right)$. This sample has been drawn with probability $P_g\,\delta x$ where $P_g = p\left(X\right)$. However, if the sample were really to have been drawn from the desired distribution, it should have probability $P_\pi\,\delta x$ where $P_\pi = \pi\left(X\right)$. Hence, on average, you would expect to have too many variates that take on the value X by a factor of

$$u(X) = \frac{P_p}{P_\pi} = \frac{p\left(X\right)}{\pi\left(X\right)} \tag{6.52}$$

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*. This throwing away of samples is also called *discarding samples*, or *rejecting samples*.

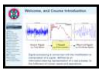
Rather than drawing a large number of samples and discarding a certain proportion, the accept-reject method will accept a sample with a certain probability given by:

$$P_a = \Pr\left(\text{accept variate } X\right) = \frac{\pi\left(X\right)}{Mp\left(x\right)} \tag{6.53}$$

This leads to the full accept-reject algorithm which takes the form:

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.

6.2.5.1 Envelope and Squeeze Methods



New slide

A problem with many sampling methods, which can make the density $\pi\left(x\right)$ difficult to simulate, is down to the complexity of the function $\pi\left(x\right)$ itself; the function may require substantial computing time at each evaluation.

It is possible to reduce the algorithmic complexity of the accept-reject algorithm by looking for another computationally simple function, $q\left(x\right)$ which *bounds* $\pi\left(x\right)$ *from below*.

In the case that the proposed variate X satisfies $q\left(X\right) \leq \pi\left(X\right)$, then considering the probability of acceptance in the accept-reject algorithm the proposed variate X should be accepted when $U \leq \frac{q(X)}{Mp(x)}$, since this also satisfies $U \leq \frac{\pi(X)}{Mp(x)}$. This is shown graphically in Figure 6.6.

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

$$P_{\bar{\pi}} = \frac{1}{M}\int q\left(x\right)\,dx \tag{6.54}$$

which is the probability that $\pi\left(x\right)$ is not evaluated.

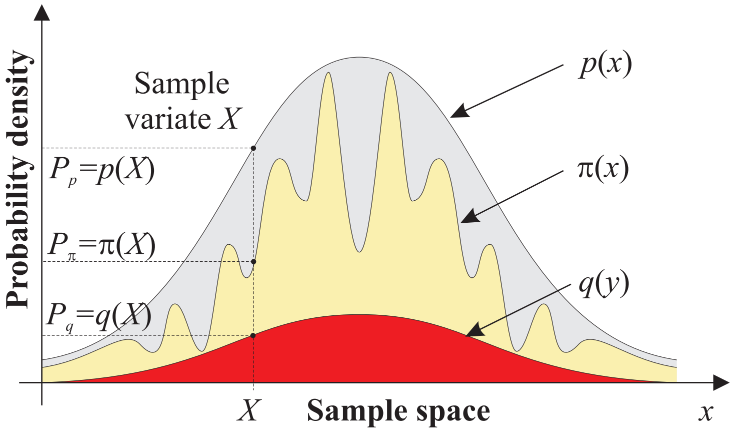
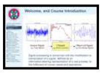


Figure 6.5: Envelope Rejection sampling

6.2.6 Importance Sampling



New slide

The problem with accept-reject sampling methods is finding the envelope functions and the constant M. This difficulty can easily be resolved if the eventual application of the samples is considered, rather than considering the sampling process as an end to-itself.

The simplest application of **importance sampling** is in Monte Carlo integration. Suppose that is is desired to evaluate the function:

$$\mathcal{I} = \int_\Theta f(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{6.55}$$

In principle, this integral can be solved by drawing samples from the density $f(\boldsymbol{\theta})$ and finding those values of $\boldsymbol{\theta}$ that lie in the region of integration: $\boldsymbol{\theta} \in \Theta$. In other words, an empirical average of $\mathcal{I}$ is:

$$\hat{\mathcal{I}} = \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{I}_\Theta \left(\boldsymbol{\theta}^{(k)}\right), \quad \text{where} \quad \boldsymbol{\theta}^{(k)} \sim f(\boldsymbol{\theta}) \tag{6.56}$$

where $\mathbb{I}_\mathcal{A}(a)$ is the indicator function, and is equal to one if $a \in \mathcal{A}$ and zero otherwise.

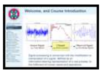
It is often difficult to sample directly from $f(\boldsymbol{\theta})$, and in any case, there are other problems with the estimator in Equation 6.56. A best estimate is as follows:

Defining an *easy-to-sample-from* density $\pi(\boldsymbol{\theta}) > 0$, $\forall \boldsymbol{\theta} \in \Theta$:

$$\mathcal{I} = \int_\Theta \frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \mathbb{E}_\pi \left[\frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right], \tag{6.57}$$

leads to an estimator based on the **sample expectation**;

$$\hat{\mathcal{I}} = \frac{1}{N} \sum_{k=0}^{N-1} \frac{f(\boldsymbol{\theta}^{(k)})}{\pi(\boldsymbol{\theta}^{(k)})} \tag{6.58}$$



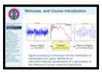
New slide

6.2.7 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).

6.3 Markov chain Monte Carlo Methods

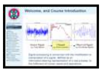


New slide

In the previous chapter on sampling random variables, the variates are drawn from an independent process.

A **Markov chain** is the first generalisation of an independent process, where each *state* of a Markov chain depends on the previous state only.

6.3.1 The Metropolis-Hastings algorithm



New slide

The **Metropolis-Hastings algorithm** is an extremely flexible method for producing a random sequence of samples from a given density.

Metropolis-Hastings explores the parameter space of the density $\pi\left(x\right)$ by means of a random walk. Unlike the accept-reject algorithm, each new sample is proposed as a random perturbation of a previously accepted variate. The **Metropolis-Hastings** algorithm is as follows, given a previously drawn sample $X_{(k)}$:

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:

$$X^{(k+1)} = \begin{cases} Y & \text{with probability } \rho(X^{(k)},\, Y) \\ X^{(k)} & \text{with probability } 1 - \rho(X^{(k)},\, Y) \end{cases} \tag{6.59}$$

   where the acceptance ratio function $\rho(x,\, y)$ is given by:

$$\rho(x,\, y) = \min\left\{ \frac{\pi\left(y\right)}{g\left(y \mid x\right)} \left(\frac{\pi\left(x\right)}{g\left(x \mid y\right)} \right)^{-1}, 1 \right\} \equiv \min\left\{ \frac{\pi\left(y\right)}{\pi\left(x\right)} \frac{g\left(x \mid y\right)}{g\left(y \mid x\right)}, 1 \right\} \tag{6.60}$$

This calculation is represented graphically in Figure 6.6.

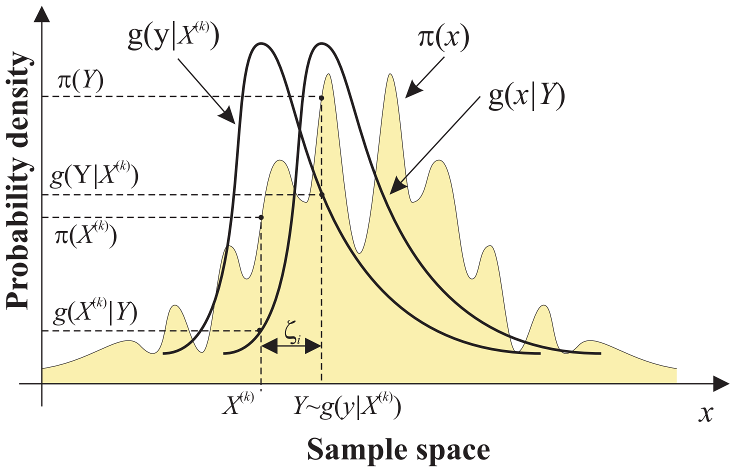
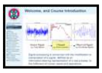


Figure 6.6: Graphical representation of the Metropolis-Hastings algorithm.

6.3.1.1 Gibbs Sampling


New slide

Gibbs sampling is a Monte Carlo method that facilitates sampling from a multivariate density function, $\pi\left(\theta_0,\,\theta_1,\,\dots,\,\theta_M\right)$ by drawing successive samples from marginal densities of smaller dimensions.

Using the probability chain rule,

$$\pi\left(\{\theta_m\}_{m=1}^{M}\right) = \pi\left(\theta_\ell \mid \{\theta_m\}_{m=1,m\neq\ell}^{M}\right)\,\pi\left(\{\theta_m\}_{m=1,m\neq\ell}^{M}\right) \tag{6.61}$$

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.

This proceeds as follows assuming the components are initialised with values $\theta_0^{(0)},\,\theta_1^{(0)},\,\dots,\,\theta_M^{(0)}$

First iteration:

$$\begin{aligned}
\theta_1^{(1)} &\sim \pi\left(\theta_1 \mid \theta_2^{(0)},\,\theta_3^{(0)},\,\theta_4^{(0)},\dots,\,\theta_M^{(0)}\right)\\
\theta_2^{(1)} &\sim \pi\left(\theta_2 \mid \theta_1^{(1)},\,\theta_3^{(0)},\,\theta_4^{(0)},\dots,\,\theta_M^{(0)}\right)\\
\theta_3^{(1)} &\sim \pi\left(\theta_3 \mid \theta_1^{(1)},\,\theta_2^{(1)},\,\theta_4^{(0)},\dots,\,\theta_M^{(0)}\right)\\
\vdots & \qquad\quad \vdots\\
\theta_M^{(1)} &\sim \pi\left(\theta_M \mid \theta_1^{(1)},\,\theta_2^{(1)},\,\theta_4^{(1)},\dots,\,\theta_{M-1}^{(1)}\right)
\end{aligned} \tag{6.62}$$

Second iteration:

$$\begin{aligned}
\theta_1^{(2)} &\sim \pi\left(\theta_1 \mid \theta_2^{(1)},\, \theta_3^{(1)},\, \theta_4^{(1)}, \ldots,\, \theta_M^{(1)}\right)\\
\theta_2^{(2)} &\sim \pi\left(\theta_2 \mid \theta_1^{(2)},\, \theta_3^{(1)},\, \theta_4^{(1)}, \ldots,\, \theta_M^{(1)}\right)\\
\theta_3^{(2)} &\sim \pi\left(\theta_3 \mid \theta_1^{(2)},\, \theta_2^{(2)},\, \theta_4^{(1)}, \ldots,\, \theta_M^{(1)}\right)\\
\vdots &\qquad \vdots\\
\theta_M^{(2)} &\sim \pi\left(\theta_M \mid \theta_1^{(2)},\, \theta_2^{(2)},\, \theta_4^{(2)}, \ldots,\, \theta_{M-1}^{(2)}\right)
\end{aligned} \tag{6.63}$$

$k+1$**-th iteration:**

$$\begin{aligned}
\theta_1^{(k+1)} &\sim \pi\left(\theta_1 \mid \theta_2^{(k)},\, \theta_3^{(k)},\, \theta_4^{(k)}, \ldots,\, \theta_M^{(k)}\right)\\
\theta_2^{(k+1)} &\sim \pi\left(\theta_2 \mid \theta_1^{(k+1)},\, \theta_3^{(k)},\, \theta_4^{(k)}, \ldots,\, \theta_M^{(k)}\right)\\
\theta_3^{(k+1)} &\sim \pi\left(\theta_3 \mid \theta_1^{(k+1)},\, \theta_2^{(k+1)},\, \theta_4^{(k)}, \ldots,\, \theta_M^{(k)}\right)\\
\vdots &\qquad \vdots\\
\theta_M^{(k+1)} &\sim \pi\left(\theta_M \mid \theta_1^{(k)},\, \theta_2^{(k)},\, \theta_4^{(k)}, \ldots,\, \theta_{M-1}^{(k)}\right)
\end{aligned} \tag{6.64}$$

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 $\pi\left(\theta_0,\, \theta_1,\, \ldots,\, \theta_M\right)$.