

**Probability, Random Variables and Signals, and
Classical Estimation Theory
UDRC-EURASIP Summer School, 28th June 2021**



Accompanying Notes

Dr James R. Hopgood

Copyright © 2021 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, Monday 14th June, 2021.
Last printed revision with minor corrections, 28 June, 2021.

Typeset by the author with the L^AT_EX 2_ε Documentation System, with A_MS-L^AT_EX 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 principal author of this document and development of this course is Dr James R. Hopgood, with material written from 2004 onwards. These documents are continually being revised, updated, and expanded. The author:

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

- Mulgrew 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.homepages.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 Recipes 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

I	Course overview and exemplar applications	3
1	Aims and Objectives	4
1.1	Obtaining the Latest Version of these Handouts	4
1.2	Welcome	6
1.3	Introduction and Overview	8
1.3.1	Module Abstract	10
1.3.2	Description and Learning Outcomes	10
1.3.3	Prerequisites	13
1.4	Recommended Texts and Learning Resources	14
1.4.1	Recommended Texts: Prerequisite Material	17
1.4.2	Further Recommended Reading	18
1.4.3	Additional Resources	19
1.4.4	Convention for Equation Numbering	20
2	Signal Processing	21
2.1	What is Signal Processing?	22
2.1.1	Modern Signal Processing Applications	23
2.1.2	The fields of Signal Processing, Automatic Control, and Communications	28
2.1.3	From Studio to the Ear	29
2.1.4	Case Study: Digital Audio Processing	32
2.1.5	Why Study Signals and Communications?	35
2.2	Fundamental Signal Processing Problems	35
2.2.1	Extracting Signals from Other Signals	36
2.2.2	Correcting Distortions in Measured Signals	37
2.2.3	Indirect Parameter Estimation	37
2.2.4	Tools for solving these problems	40
2.3	What are Signals and Systems?	42
2.3.1	Mathematical Representation of Signals	43
2.3.1.1	Continuous-time and discrete-time signals	43

2.3.1.2	Other types of signals	44
2.3.2	Mathematical Representation of Systems	45
2.3.3	Deterministic Signals	47
2.4	Motivation	48
2.4.1	Speech Modelling and Recognition	49
2.4.2	Blind System Identification	50
2.4.3	Blind Signal Separation	51
2.4.4	Data Compression	52
2.4.5	Enhancement of Signals in Noise	53
2.5	Passive and Active Target Localisation	54
2.6	Passive Target Localisation Methodology	55
2.6.1	Source Localization Strategies	55
2.6.2	Geometric Layout	56
2.6.3	Ideal Free-field Model	56
2.7	Indirect time-difference of arrival (TDOA)-based Methods	57
2.7.1	Hyperbolic Least Squares Error Function	57
2.7.2	TDOA estimation methods	58
2.7.2.1	GCC TDOA estimation	58
2.7.2.2	generalised cross correlation (GCC) Processors	59
2.8	Direct Localisation Methods	60
2.8.1	Steered Response Power Function	60
2.8.2	Conclusions	62

II Probability, Random Variables, and Estimation Theory 63

3	Probability Theory 64
3.1	Introduction 65
3.2	The Notion of Probability 66
3.3	Classical Definition of Probability 70
3.3.1	Using the Classical Definition 71
3.3.2	Difficulties with the Classical Definition 74
3.3.3	Discussion: Bertrand's Paradox 75
3.4	Axiomatic Definition 78
3.4.1	Properties of Axiomatic Probability 79
3.4.2	Set Theory 81
3.4.3	Countable Spaces and Principle of Total Probability 85
3.4.4	The Real Line 88
3.5	Conditional Probability 89
3.6	Bayes's Rule 92

4	Scalar Random Variables	95
4.1	Abstract	96
4.2	Definition	97
4.2.1	Distribution functions	99
4.2.2	Kolmogorov's Axioms	100
4.3	Density functions	101
4.4	Properties: Distributions and Densities	104
4.5	Common Continuous RVs	105
4.6	Probability transformation rule	110
4.7	Expectations	114
4.7.1	Properties of expectation operator	115
4.8	Moments	118
4.8.1	Central Moments	120
4.8.2	Relationship between Moments	120
4.8.3	Higher-Order Statistics	123
4.9	Characteristic Functions	126
4.9.1	The probability generating function	130
4.9.2	Cumulants	133
5	Multiple Random Variables	134
5.1	Abstract	135
5.2	Definition of Random Vectors	136
5.2.1	Distribution and Density Functions	137
5.2.2	Complex-valued RVs and vectors	141
5.2.3	Marginal Density Function	143
5.2.4	Independence	147
5.2.5	Conditionals and Bayes's	148
5.3	Probability Transformation Rule	156
5.3.1	Polar Transformation	159
5.3.2	Generating Gaussian distributed samples	161
5.3.3	Auxiliary Variables	165
5.4	Statistical Description	169
5.4.1	Mean Vectors and Correlation Matrices	170
5.4.2	Properties of Correlation Matrices	176
5.4.3	Further Statistical Descriptions	180
5.5	Linear Transformations	183
5.6	Multivariate Gaussian Density Function	187
5.6.1	Deriving the Multivariate Gaussian	188

5.6.2	Properties of Multivariate Gaussians	190
5.7	Characteristic Functions	192
5.8	Higher-Order Statistics	195
5.9	Sum of Independent Random Variables	196
5.10	Central limit theorem	199
6	Estimation Theory	203
6.1	Introduction	204
6.1.1	A (Confusing) Note on Notation	205
6.1.2	Examples of parameter estimation	205
6.2	Properties of Estimators	206
6.2.1	What makes a good estimator?	210
6.2.2	Bias of estimator	212
6.2.3	Variance of estimator	213
6.2.4	Mean square error	214
6.2.5	Consistency of an Estimator	218
6.2.6	Cramer-Rao Lower Bound	221
6.2.7	Estimating Multiple Parameters	229
6.3	Maximum Likelihood Estimation	236
6.3.1	Properties of the maximum-likelihood estimate (MLE)	237
6.3.2	DC Level in white Gaussian noise	239
6.3.3	MLE for Transformed Parameter	240
6.4	Least Squares	242
6.4.1	The Least Squares Approach	243
6.4.2	DC Level	244
6.4.3	Nonlinear Least Squares	245
6.4.4	Linear Least Squares	246
6.4.5	Weighted Linear Least Squares	249
6.5	Bayesian Parameter Estimation	250
6.5.1	Bayes's Theorem (Revisited)	251
6.5.2	The Removal of Nuisance Parameters	252
6.5.3	Prior Probabilities	252
6.5.4	General Linear Model	253
6.5.4.1	Model Selection using Bayesian Evidence	256

7	MonteCarlo	258
7.1	Introduction	258
7.1.1	Deterministic Numerical Methods	259
7.1.1.1	Deterministic Optimisation	260
7.1.1.2	Deterministic Integration	260
7.1.2	Monte Carlo Numerical Methods	260
7.1.2.1	Monte Carlo Integration	261
7.1.2.2	Stochastic Optimisation	262
7.1.2.3	Implementation issues	262
7.2	Generating Random Variables	262
7.2.1	Uniform Variates	263
7.2.2	Transformation Methods	263
7.2.3	Generating white Gaussian noise (WGN) samples	263
7.2.4	Inverse Transform Method	267
7.2.5	Acceptance-Rejection Sampling	268
7.2.5.1	Envelope and Squeeze Methods	269
7.2.6	Importance Sampling	270
7.2.7	Other Methods	271
7.3	Markov chain Monte Carlo Methods	271
7.3.1	The Metropolis-Hastings algorithm	271
7.3.1.1	Gibbs Sampling	272
III	Stochastic Processes	275
8	Linear Systems Review	276
8.1	Introduction	276
8.2	Signal Classification	277
8.2.1	Types of signal	278
8.2.2	Energy and Power Signals	282
8.2.2.1	Motivation for Energy and Power Expressions	283
8.2.2.2	Formal Definitions for Energy and Power	286
8.2.2.3	Units of Energy and Power	288
8.2.2.4	Power for Periodic Signals	288
8.3	Fourier Series and transforms	289
8.3.1	Complex Fourier series	289
8.3.1.1	Common Fourier Series Expansions	291
8.3.1.2	Dirichlet Conditions	291

8.3.1.3	Parseval's Theorem	292
8.3.2	Fourier transform	294
8.3.2.1	Parseval's Theorem	295
8.3.3	The DTFT	297
8.3.4	Discrete Fourier transform	298
8.3.4.1	Parseval's Theorem for Finite Length Signals	299
8.3.4.2	The DFT as a Linear Transformation	299
8.3.4.3	Properties of the DFT	300
8.4	Discrete-time systems	302
8.4.1	Basic discrete-time signals	302
8.4.2	The z -transform	303
8.4.2.1	Bilateral z -transform	305
8.4.2.2	Properties of the z -transform	308
8.4.2.3	The Unilateral z -transform	310
8.4.3	LTI systems	312
8.4.3.1	Matrix-vector formulation	312
8.4.3.2	Transform-domain analysis	313
8.4.3.3	Frequency response	313
8.4.3.4	Periodic Inputs	313
8.4.4	Rational transfer functions	314
9	Stochastic Processes	315
9.1	A Note on Notation	315
9.2	Definition of a Stochastic Process	316
9.2.1	Interpretation of Sequences	317
9.2.2	Description using probability density functions (pdfs)	319
9.3	Second-order Statistical Description	320
9.3.1	Example of Calculating Autocorrelations	320
9.4	Types of Stochastic Processes	324
9.5	Stationary Processes	328
9.5.1	Order-N and strict-sense stationarity	329
9.5.2	Wide-sense stationarity	329
9.5.3	WSS Properties	335
9.5.4	Wide-sense cyclo-stationarity	340
9.5.5	Quasi-stationarity	343
9.6	Estimating statistical properties	345
9.6.1	Ensemble and Time-Averages	346
9.6.2	Ergodicity	346

9.6.3	More Details on Mean-Ergodicity	347
9.7	Joint Signal Statistics	351
9.7.1	Types of Joint Stochastic Processes	352
9.8	Correlation Matrices	353
9.9	Markov Processes	355
10	Power Spectral Density	357
10.1	Introduction	358
10.2	Motivating the power spectral density	360
10.2.1	Informal Motivation	360
10.2.2	Formal Statistical Derivation	361
10.3	The power spectral density	364
10.3.1	Properties of the power spectral density	365
10.3.2	General form of the PSD	366
10.4	The cross-power spectral density	368
10.5	Complex Spectral Density Functions	369
10.6	Table of bilateral z -transforms	371
11	Linear Systems Theory	374
11.1	Systems with Stochastic Inputs	375
11.2	Methods for Calculating Input-Output Statistics	376
11.3	LTI Systems with Stationary Inputs	378
11.3.1	Input-output Statistics of a linear time-invariant (LTI) System	380
11.3.2	System identification	385
11.4	LTV Systems with Nonstationary Inputs	388
11.4.1	Input-output Statistics of a linear time-varying (LTV) System	388
11.4.2	Effect of Linear Transformations on Cross-correlation	389
11.5	Time-Domain Analysis with Difference Equations	391
11.6	Frequency-Domain Analysis of LTI systems	395
12	Linear Signal Models	401
12.1	Abstract	401
12.2	The Ubiquitous WGN Sequence	402
12.2.1	Generating WGN samples	402
12.2.2	Filtration of WGN	403
12.3	Nonparametric and parametric models	404
12.4	Parametric Pole-Zero Signal Models	405
12.4.1	Types of pole-zero models	405
12.4.2	All-pole Models	407

12.4.2.1	Frequency Response of an All-Pole Filter	407
12.4.2.2	Impulse Response of an All-Pole Filter	408
12.4.2.3	Autocorrelation of the Impulse Response	409
12.4.2.4	All-Pole Modelling and Linear Prediction	410
12.4.2.5	Autoregressive Processes	410
12.4.2.6	Autocorrelation Function from AR parameters	411
12.4.3	All-Zero models	413
12.4.3.1	Frequency Response of an All-Zero Filter	413
12.4.3.2	Impulse Response	414
12.4.3.3	Autocorrelation of the Impulse Response	415
12.4.3.4	Moving-average processes	416
12.4.3.5	Autocorrelation Function for MA Process	416
12.4.4	Pole-Zero Models	417
12.4.4.1	Pole-Zero Frequency Response	417
12.4.4.2	Impulse Response	418
12.4.4.3	Autocorrelation of the Impulse Response	419
12.4.4.4	Autoregressive Moving-Average Processes	420
12.5	Estimation of AR Model Parameters from Data	420
12.5.1	LS AR parameter estimation	421
12.5.2	Autocorrelation Method	422
12.5.3	Covariance Method	423

IV Advanced Topics 427

13 Passive Target Localisation 428

13.1	Introduction	428
13.1.1	Structure of the Tutorial	428
13.2	Recommended Texts	430
13.3	Why Source Localisation?	431
13.4	ASL Methodology	431
13.4.1	Source Localization Strategies	433
13.4.2	Geometric Layout	434
13.4.3	Ideal Free-field Model	435
13.4.4	TDOA and Hyperboloids	435
13.5	Indirect TDOA-based Methods	436
13.5.1	Spherical Least Squares Error Function	439
13.5.1.1	Two-step Spherical LSE Approaches	441

13.5.1.2	Spherical Intersection Estimator	442
13.5.1.3	Spherical Interpolation Estimator	442
13.5.1.4	Other Approaches	443
13.5.2	Hyperbolic Least Squares Error Function	443
13.5.2.1	Linear Intersection Method	444
13.5.3	TDOA estimation methods	445
13.5.3.1	GCC TDOA estimation	445
13.5.3.2	CPSD for Free-Field Model	446
13.5.3.3	GCC Processors	447
13.5.3.4	Adaptive Eigenvalue Decomposition	447
13.6	Direct Localisation Methods	450
13.6.1	Steered Response Power Function	450
13.6.2	Conceptual Intepretation of SRP	451
13.7	DUET Algorithm	452
13.7.1	Effect of Reverberation and Noise	455
13.7.2	Estimating multiple targets	455
13.8	Further Topics	455

Acronyms

2-D	two-dimensional
3-D	three-dimensional
A2DP	Advanced Audio Distribution Profile
AC	autocorrelation
ACF	autocorrelation function
ACS	autocorrelation sequence
ADC	analogue-to-digital converter
AED	adaptive eigenvalue decomposition
AIC	Akaike's information criterion
AIR	acoustic impulse response
AR	autoregressive
ARMA	autoregressive moving average
ASL	acoustic source localisation
AVS	acoustic vector sensor
AWGN	additive white Gaussian noise
BFGS	Broyden-Fletcher-Goldfarb-Shannon
BIBO	bounded-input, bounded-output
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
CMOS	complementary metal-oxide-semiconductor
CPSD	cross-power spectral density
CRLB	Cramér-Rao lower-bound
CTFT	continuous-time Fourier transform
DAB	digital audio broadcasting
DAT	digital audio tape
DAW	digital audio workstation
DC	"direct current"

DFT	discrete Fourier transform
DNA	deoxyribonucleic acid
DSP	digital signal processing
DTFT	discrete-time Fourier transform
DUET	degenerate unmixing estimation technique
DVB	digital video broadcasting
DVD	digital versatile disc
DVD-A	digital versatile disc-audio
ECG	electrocardiogram
EEG	electroencephalogram
FFT	Fast Fourier transform
FIM	Fisher information matrix
FIR	finite impulse response
FLAC	free lossless audio codec
FPE	final prediction error
FS	Fourier series
FT	Fourier transform
GCC	generalised cross correlation
GCC-PHAT	GCC-phase transform (PHAT)
HCI	human-computer interface
ICA	independent component analysis
IDFT	inverse-DFT
KL	Karhunen-Loeve
LHS	left hand side
LI	linear intersection
LITP	linear in the parameters
LS	least-squares
LSE	least-squares estimate
LSE	least squares error
LTI	linear time-invariant
LTV	linear time-varying
MA	moving average
MAP	maximum <i>a posteriori</i>
MDL	minimum description length
MEG	magnetoencephalography
MEMS	micro-electromechanical systems
MGF	moment generating function
ML	maximum-likelihood

MLE	maximum-likelihood estimate
MMAP	maximum marginal <i>a posteriori</i>
MMSE	minimum mean-square error
MP3	MPEG-1 Audio Layer 3
MPEG	Moving Picture Experts Group
MRI	magnetic resonance imaging
MS	mean-square
MSC	magnitude square coherence
MSE	mean-squared error
MVU	minimum variance unbiased
MVUE	minimum variance unbiased estimator
NMRI	nuclear magnetic resonance imaging
PGF	probability generating function
PHAT	phase transform
PHD	Ph.D. thesis
PSD	power spectral density
PTL	passive target localisation
RHS	right hand side
RIR	room impulse response
ROC	region of convergence
Radar	RAdio Detection And Ranging
SACD	super-audio CD
SAR	synthetic aperture RADAR
SBF	steered beamformer
SCOT	Smoothed Coherence Transform
SI	spherical interpolation
SLAM	simultaneous localisation and mapping
SNR	signal-to-noise ratio
SRC	stochastic region contraction
SRP	steered response power
SSP	statistical signal processing
SSS	strict-sense stationary
STFT	short-time Fourier transform
SX	spherical intersection
TDOA	time-difference of arrival
TF	time-frequency
TFR	time-frequency representation
UAV	unmanned aerial vehicle

ULA	uniform linear array
WDO	W-disjoint orthogonality
WGN	white Gaussian noise
WSP	wide-sense periodic
WSS	wide-sense stationary
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

10;

Part I

Course overview and exemplar applications

1

Introduction, 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 Einstein

This handout also provides an introduction to signals and systems, and an overview of statistical signal processing applications. This is relevant to provide context and motivation for studying this branch of signal and information processing.

1.1 Obtaining the Latest Version of these Handouts

- 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. The course is based on MSc level materials.

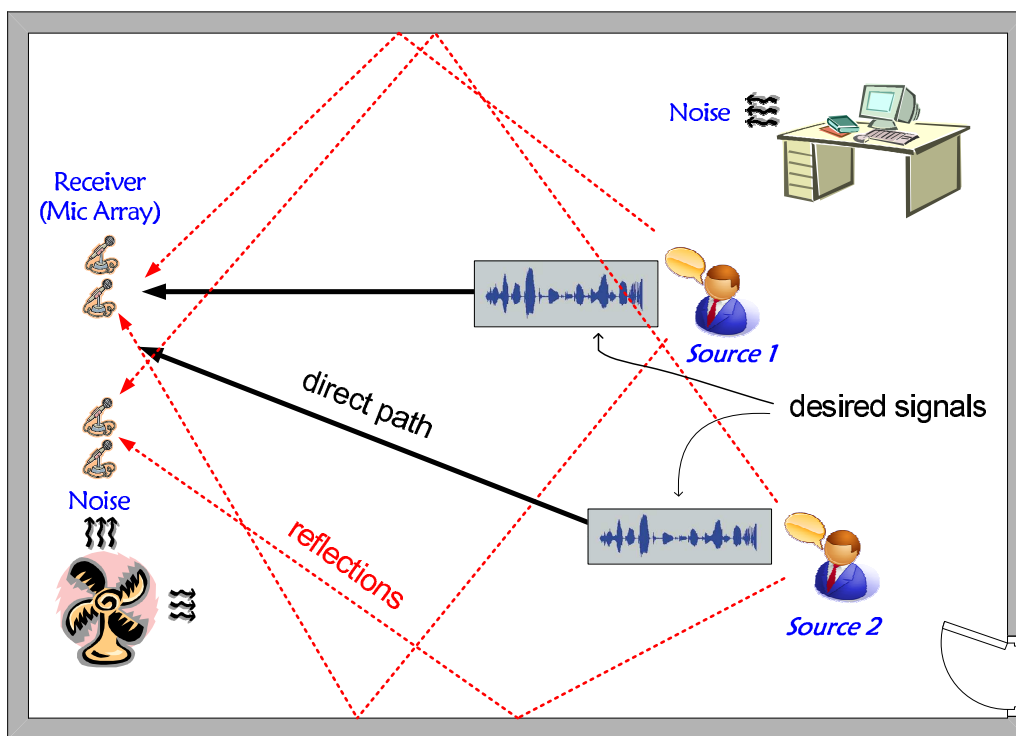


Figure 1.1: Source localisation and blind source separation (BSS). An example of topics using statistical signal processing.

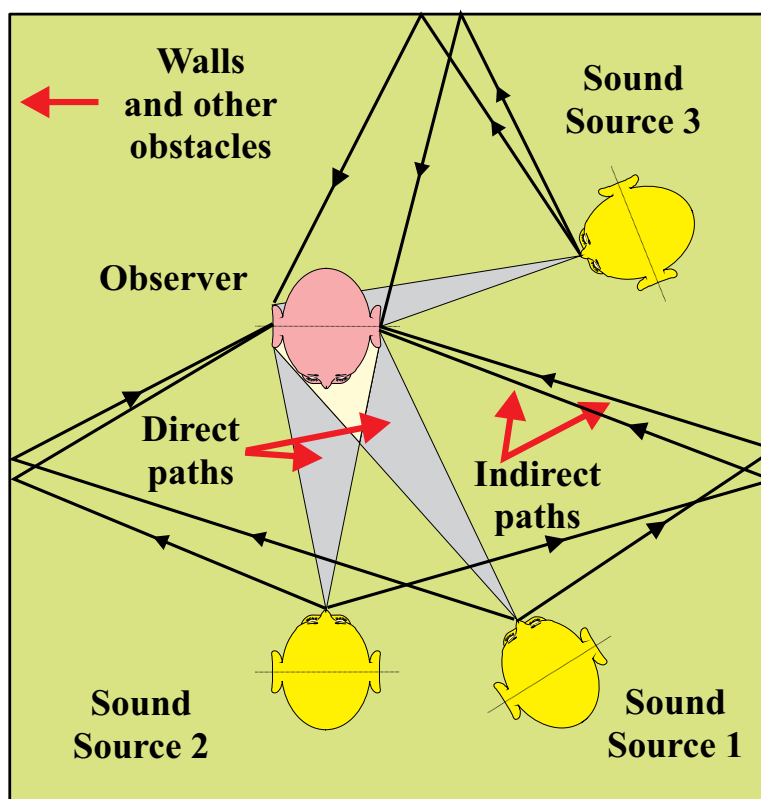


Figure 1.2: Humans turn their head in the direction of interest in order to reduce interference from other directions; *joint detection, localisation, and enhancement*. An application of probability and estimation theory, and statistical signal processing.

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

- This tutorial is being continually updated, and feedback is welcomed. The hardcopy documents published or online may differ slightly 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 obtained from the author, Dr James R. Hopgood, by emailing him at:
mailto:james.hopgood@ed.ac.uk
(Update: The notes are no longer online due to the desire to maintain copyright control on the document.)
- Extended thanks to the many MSc students over the past 16 years who have helped proof-read and improve these documents.

1.2 Welcome

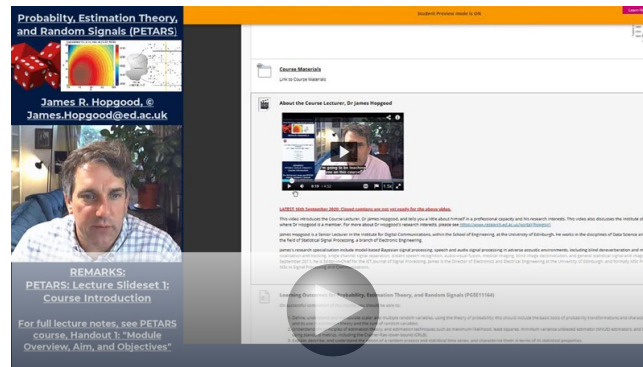
The Probability, Estimation Theory, And Random Signals module introduces the fundamental statistical tools that are required to analyse and describe advanced signal processing algorithms within this MSc programme.

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.



http://media.ed.ac.uk/media/1_6wt1ez10

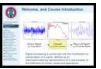
Video Summary: This video introduces the Course Lecturer, Dr James Hopgood, and tells you a little about himself in a professional capacity and his research interests. This video also discusses the Institute of Digital Communications, where Dr Hopgood is a member. For more about Dr Hopgood's research interests, please see <https://www.research.ed.ac.uk/portal/jhopgood1>.



http://media.ed.ac.uk/media/1_1y8dtumu

Video Summary: This video shows you how to navigate the LEARN virtual learning environment. It shows how to navigate course content and the course guide.

1.3 Introduction and Overview



New slide

Topic Summary 1 Course aims and objectives, overview, key themes

Topic Objectives:

- Awareness of the aims and objectives of the course.
- Highlight the learning outcomes of the course.
- List the mathematical prerequisites for the course.
- Lists the main themes of the course.

Topic Activities:

Type	Details	Duration	Progress
Watch video	12 : 12 minute video	3× video length	
Read Handout	Read page 8 to page ??	8 mins/page	

Probability, Estimation Theory, and Random Signals (PETARS)
James R. Hoggood, ©

Introduction and Overview

Source Signal (e.g. Clean Speech) → Channel (e.g. Room Acoustics) → Observed Signal (e.g. Reverberant Speech)

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

It is assumed you have a grounding in DSP. This module will take you to the next level; a tour of the exciting, fascinating, and active research area of *statistical signal processing*.

http://media.ed.ac.uk/media/1_q42rrjjf

Video Summary: This video gives a very brief introduction to signal processing, describes the course aims and objectives from a high-level, the learning outcomes, and prerequisites needed to study the course. The video also discusses the key themes studied in this course.

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). This module will take you to the next level; a tour of the exciting, fascinating, and active research area of *statistical signal processing*.

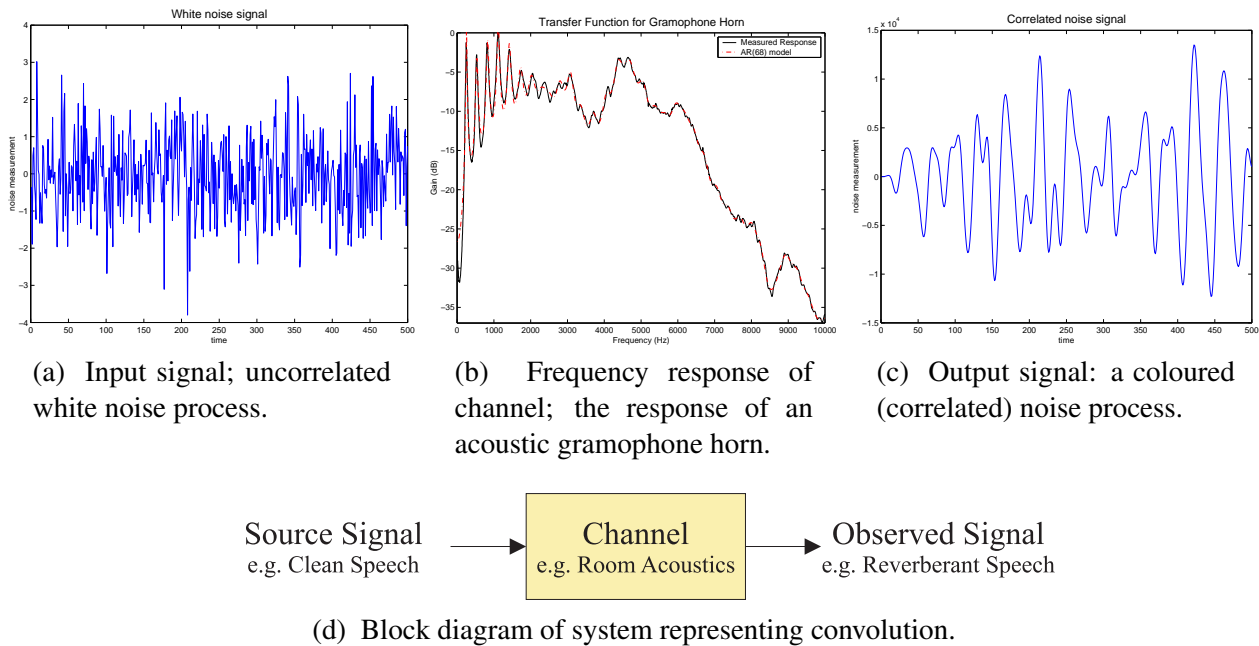


Figure 1.3: Solutions to the so-called *blind deconvolution problem* require statistical signal processing methods.

Sidebar 1 Signal Processing

The IEEE Signal Processing Society makes the following statement regarding signal processing.

The technology we use, and even rely on, in our everyday lives – computers, radios, video, cell phones – is enabled by signal processing, a branch of electrical engineering that models and analyzes data representations of physical events.

Signal processing is at the heart of our modern world, powering today's entertainment and tomorrow's technology. It's at the intersection of biotechnology and social interactions. It enhances our ability to communicate and share information.

Signal processing is the science behind our digital lives.

Recently, machine learning techniques have been applied to aspects of signal processing, blurring the lines between the sciences, and causing many shared applications between the two.

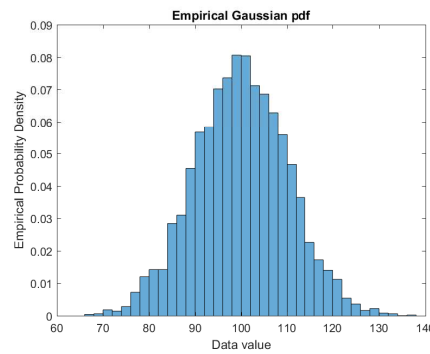


Figure 1.4: Empirical Gaussian probability density function.

1.3.1 Module Abstract

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**.

1.3.2 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 1.3.3 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 powerful 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 have:

1. acquired sufficient expertise in this area to understand and implement **spectral estimation, signal modelling, parameter estimation, and adaptive filtering** techniques;
2. developed an understanding of the basic concepts and methodologies in statistical signal processing that provides the foundation for **further study, research, and application to new problems**.

PETARS Learning Outcomes There are five key learning outcomes for the full PETARS course. On completion of this course, the student will be able to:

- Define, understand and manipulate scalar and multiple random variables, using the theory of probability; this should include the basic tools of probability transformations and characteristic functions, moments, the central limit theorem (CLT) and its use in estimation theory and the sum of random variables.
- Understand the principles of estimation theory, and estimation techniques such as maximum-likelihood, least squares, minimum variance unbiased estimator (MVUE) estimators, and Bayesian estimation; be able to characterise the estimator using standard metrics, including the Cramér-Rao lower-bound (CRLB).

- Explain, describe, and understand the notion of a random process and statistical time series, and characterise them in terms of its statistical properties.
- Define, describe, and understand the notion of the power spectral density of stationary random processes, and be able to analyse and manipulate them; 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.
- 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; apply least squares, maximum-likelihood, and Bayesian estimators to model based signal processing problems.

These are broken down further in the expanded Learning Outcomes below.

Expanded 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 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 CRLB and 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.

1.3.3 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 the section on *Learning Resources* later in this Handout.

The screenshot shows a web page titled 'Probability, Estimation Theory and Random Signals (PETARS) (MSc)'. The course organiser is Dr James Hopgood. The page lists five recommended books with their covers and availability information:

- Probability and random processes for electrical and computer engineers** by Therrien, Charles W., Tummala, Murali, CRC Press, 2012. Available at Murray Library Murray Library, King's Buildings (RESERVE): TK153 The. and more locations.
- Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing** by Dimitris G. Manolakis, Vinay K Ingle; Stephen M Kogon, 2005. Available at Murray Library Murray Library, King's Buildings (RESERVE): Folio TK5102.9 Man. and more locations.
- Fundamentals of statistical signal processing** by Kay, Steven M., Prentice-Hall PTR, Prentice-Hall signal processing series, ©1993-1998. Available at Murray Library Murray Library, King's Buildings (RESERVE): TK5102.5 Kay. and more locations.
- Probability, random variables, and stochastic processes** by Papoulis, Athanasios, Pillai, S. Unnikrishna, McGraw-Hill, 2002. Available at Murray Library Murray Library, King's Buildings (RESERVE): QA273 Pap.
- Digital signal processing: principles, algorithms, and applications** by Proakis, John G., Manolakis, Dimitris G, Second edition., New York: Toronto: New York, Macmillan Maxwell Macmillan Canada Maxwell Macmillan, 2005. Available at Murray Library Murray Library, King's Buildings (STANDARD LOAN): TK5102.5 Pro.

Figure 1.5: The Resource List page, accessible from LEARN, lists the course textbooks, and how to find them in the University.

1.4 Recommended Texts and Learning Resources

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.

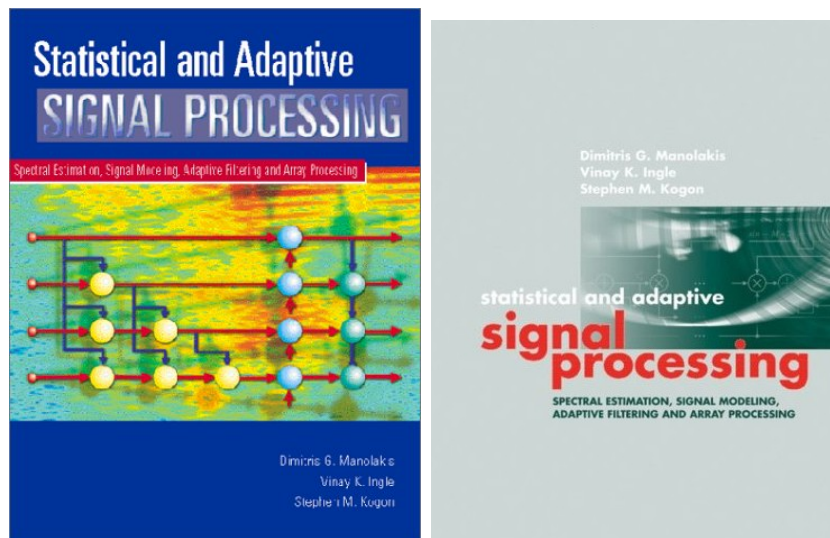
IDENTIFIERS – *Paperback*, ISBN10: 0070400512, ISBN13: 9780070400511

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

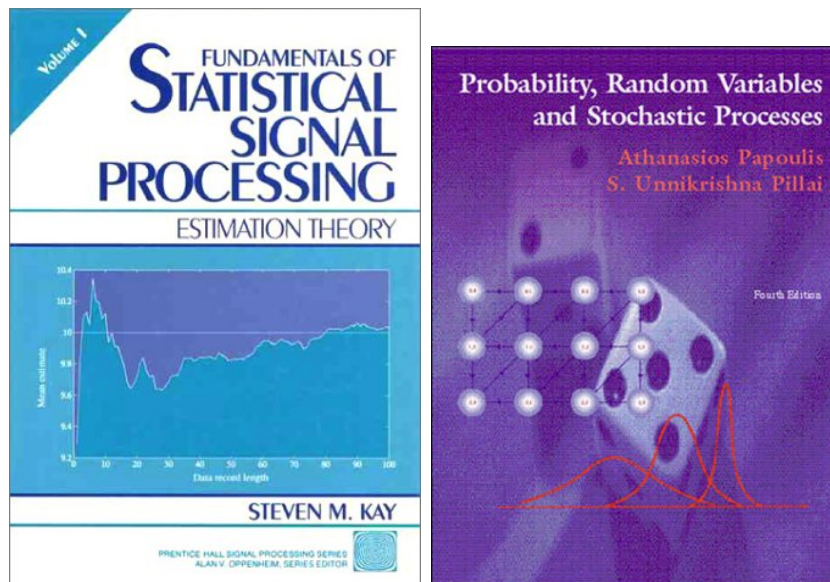
Images of the book covers are shown in Figure 1.6. For further reading, or an alternative perspective on the subject matter, other recommended text books for this module include:



(a) Cover of *paperback* version.

(b) Cover of *hardback* version.

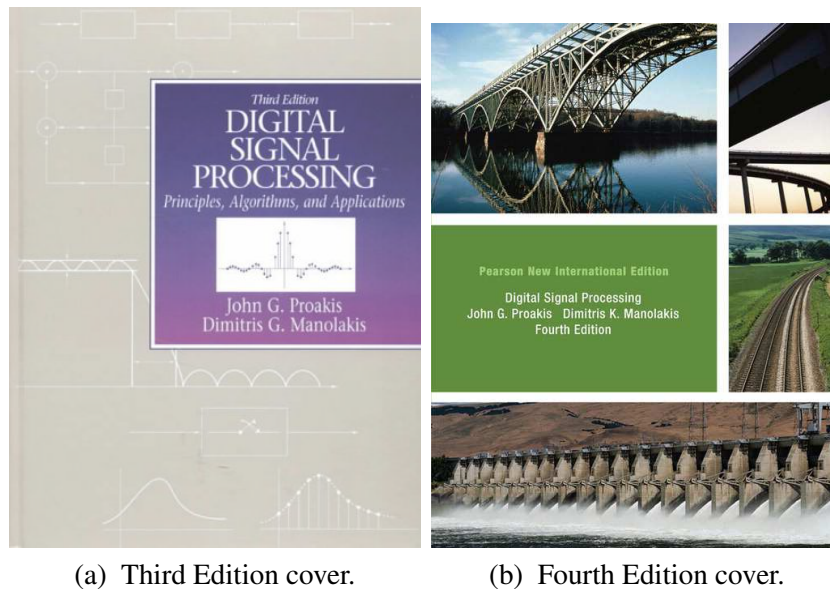
Figure 1.6: The main **course text** for this module: [Manolakis:2000].



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

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

Figure 1.7: Additional recommended texts for the course.



(a) Third Edition cover.

(b) Fourth Edition cover.

Figure 1.8: **Course text:** further reading for digital signal processing and mathematics, [Proakis:1996].

1. 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

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 1.7. The material in [Kay:1993] is mainly covered in Handout 6 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 SSP module.

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

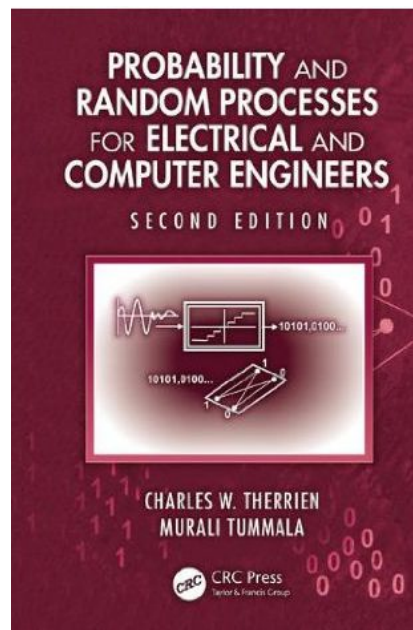


Figure 1.9: Further reading for statistical signal processing, [Therrien:2011].

1.4.1 Recommended Texts: Prerequisite Material

As mentioned in the section on mathematic pre-requisites above, 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*:

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]:

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

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

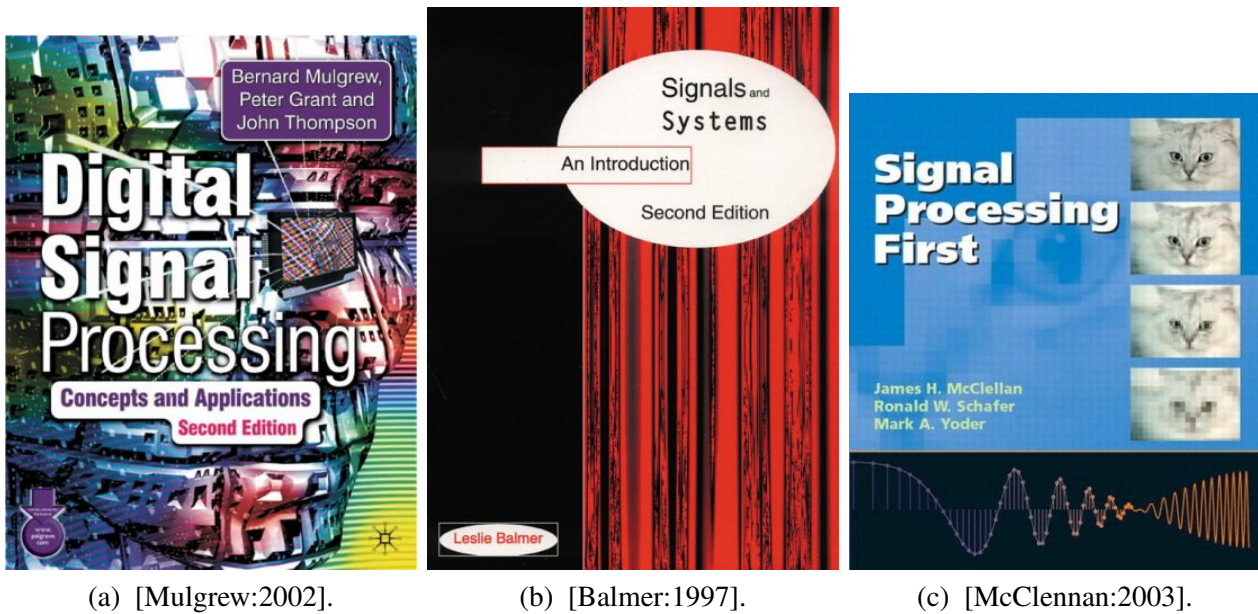


Figure 1.10: Undergraduate texts on Signals and Systems.

See <http://www.homepages.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

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, Inc., 2003.

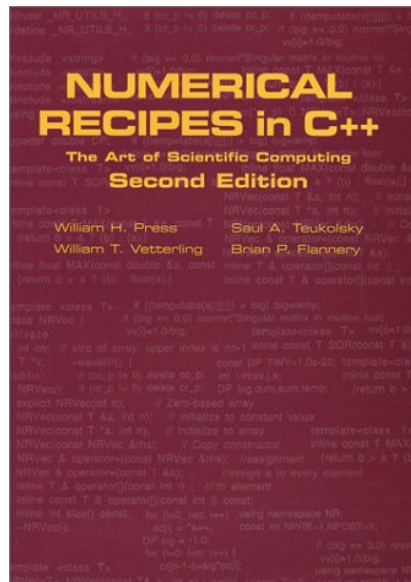
IDENTIFIERS – *Paperback*, ISBN10: 0131202650, ISBN13: 9780131202658

Hardback, ISBN10: 0130909998, ISBN13: 9780130909992

1.4.2 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 Recipes in C: The Art of Scientific Computing*, Second edition, Cambridge University Press, 1992.



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

Figure 1.11: Further reading for numerical methods and mathematics.

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.

1.4.3 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>

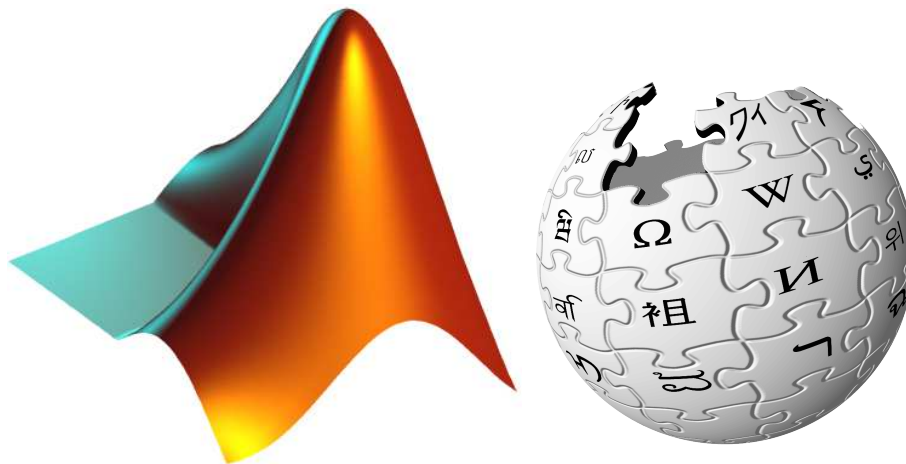
- Connexions is an environment for collaboratively developing, freely sharing, and rapidly publishing scholarly content on the Web. A wide variety of technical lectures can be found at:

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, 2001 – present.



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

(b) Wikipedia, The Free Encyclopedia.

Figure 1.12: Some useful resources.

See <http://en.wikipedia.org/>

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

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., 1998 – present.

See <http://www.google.co.uk>

– End-of-Topic 1: Learning resources –



1.4.4 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.

2

Applications of Signal Processing

We live in a society exquisitely dependent on science and technology, in which hardly anyone knows anything about science and technology.

Carl Sagan

This handout begins by motivating the need for this course material by looking at key application areas and concepts that will be studied in detail during the lectures.

2.1 What is Signal Processing?

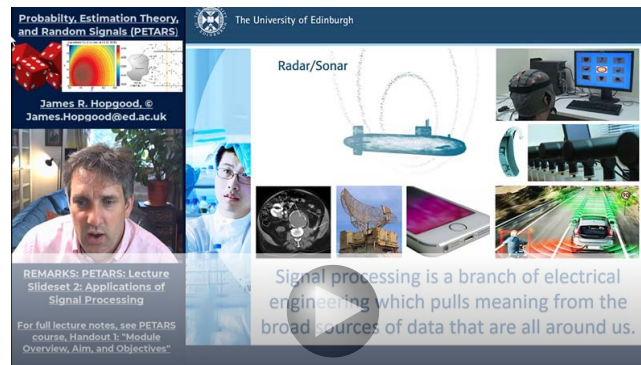
Topic Summary 2 What is Signal Processing?

Topic Objectives:

- Learn a high-level overview of signal processing.
- Identify signal processing in our daily lives.
- Understand why signal processing has become common-place.

Topic Activities:

Type	Details	Duration	Progress
Watch video	13 : 41 minute video	3×video length	
Discussion Board	Your views of signal processing	15 minutes	
Read Handout	Read page 22 to page 27	8 mins/page	



http://media.ed.ac.uk/media/1_t0qrik06

Video Summary: This video explains the role of signal processing in powering modern communications, entertainment, transportation, and healthcare systems, in addition to numerous industrial and defence applications. It explains why signal processing techniques have grown substantially over the past few decades in terms of improvements in signal processing algorithms as well as other key enabling technologies, such as low-power computing platforms, sensor technologies, and advances in battery technology.

Signal processing is a branch of electrical engineering which pulls meaning from the broad sources of data that are all around us.

Signal processing is at the heart of our modern world: signal processing powers modern communications (including voice recognition), modern entertainment (including motion sensing-gaming), tomorrow's transportation (including autonomous vehicles), and healthcare.

A nice introduction for the general public is presented in a YouTube video from the IEE Signal Processing Society, as shown in Figure 2.1.



<http://youtu.be/R90ciUoxcJU>

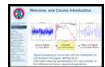
Figure 2.1: A video from the IEEE Signal Processing Society explaining *What is Signal Processing?*

2.1.1 Modern Signal Processing Applications

The last decade has seen a large number of domestic products which are heavily dependent on sophisticated *signal processing algorithms*. Some of these products are actually worth getting excited about in the sense they are extremely clever, and signal processing isn't restricted to simple removal of basic background noise (either in images or in audio). Some examples include:

- Microsoft Kinect, as shown in Figure 2.3, which includes skeletal tracking, depth estimation, acoustic noise cancellation, and speech identification and recognition; a demonstration of this will be given in lectures;
- Low-cost low-flying unmanned aerial vehicles (UAVs), which includes sophisticated algorithms for self-geolocation using on-board cameras and other sensors, and simultaneous localisation and mapping (SLAM), and on-board sensing of objects and targets; see Figure 2.2
- Video calling such as Skype and Facetime, which requires good audio, image, and video compression for network communication and online streaming;
- Computer-based music analysis, especially for game play, such as Guitar Hero and Rocksmith;
- Room acoustic calibration (or *correction*) techniques in audio-visual setups (for example, most major audio-visual AV receivers);
- Far-field speech enhancement for voice assistance (Amazon Echo, Google Home);
- Digital image manipulation and processing using desktop software (*Photoshopping images*).

These are domestic applications which have grown over recent years, and of course are in addition to *medical imaging*, *defence*, *meteorological*, and *geophysical* applications, amongst many others



New slide



<http://youtu.be/Gj-5RNdUz3I>

Figure 2.2: A research UAV from Ascending Technologies: <http://www.asctec.de/en/uav-uas-drone-products/asctec-firefly>

as described below. It is important, however, to appreciate *why* digital techniques have grown substantially over the past few decades. Reasons include:

1. the dramatic improvement in computational power available on low-power devices due to the microelectronics revolution and advances in battery power;
2. the almost universal adoption of digital media, both audio and video, over the past two decades;
3. the vast improvements in sensor modalities including micro-electromechanical systems (MEMS) microphones and complementary metal-oxide-semiconductor (CMOS) cameras, as well as other MEMS devices such as accelerometers on mobile devices;
4. advances in understanding and performance of optimisation algorithms, estimation theory, and signal filters.

Signal processing is the technology that allows the manipulation, efficient storage, and analysis of signals that are recorded using a variety of sensor technologies, on electronic hardware. It is vital to appreciate that many of the electronic products, domestic, civilian, or military, are reliant on the processing of measured signals, from Radio Detection And Ranging (Radar) (see Figure 2.5), to magnetic resonance imaging (MRI), through to cameras and microphones, or temperature sensors. It is vital to appreciate that most electronic products require some form of signal processing.



Figure 2.3: Hands-free human-computer interface (HCI).



Figure 2.4: UAVs used for package deliveries.



Figure 2.5: Radar of the type used for detection of aircraft. It rotates steadily sweeping the airspace with a narrow beam. Air Force Museum, by Bukvoed / CC BY-SA 3.0.

KEYPOINT! (Discussion Topic). Signal Processing as a subject has strong overlaps with other disciplines, such as machine learning in Computer Science, applied statistics in Mathematics and Econometrics, and remote sensing in the Geosciences. Using the discussion boards, think about and try and answer the questions:

1. What is signal processing and communications?
2. What applications have signal processing, communications, and machine learning had an impact on in society?
3. How do sensors play an important role in signal processing?

– End-of-Topic 2: **What is Signal Processing?** –



2.1.2 The fields of Signal Processing, Automatic Control, and Communications

Topic Summary 3 Applications of Signal Processing and Communications

Topic Objectives:

- Examples signal processing applications.
- Privacy aware signal processing.
- Example of a signal processing and communication system.

Topic Activities:

Type	Details	Duration	Progress
Watch video	9 : 25 minute video	3× video length	
Read Handout	Read page 28 to page 31	8 mins/page	

From Studio to the Ear

Live recording

Music Production

Wireless Transmission

Listening on a portable media player

From an instrument being played through to listening on Advanced Audio Distribution Profile (A2DP) Bluetooth headphones via a portable media player.

http://media.ed.ac.uk/media/1_cwkcy5dq

Video Summary: This video considers in more detail some applications of signal processing, including biomedical, surveillance and homeland security, target tracking and navigation, mobile communications, and speech enhancement and recognition. The video then considers the application of delivering live music to a remote listener wearing a wireless headset. The different signal processing and communication systems involved in this application are discussed. This video provides background information for the MSc in Signal Processing and Communications.

Although this course has been written with a bias towards *electronic engineering*, the mathematical tools and techniques introduced are fundamental in many other areas of Engineering. They are not limited to the examples given in this course by any stretch of the imagination. More significantly, this course initially covers continuous-time analogue signals, and then moves onto discrete-time signals. Discrete-time digital signals are the basis of modern digital and statistical signal processing, and is used in a plethora of modern Engineering problems. Modern advances in statistical signal processing, control, and communications include:

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 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 ranges. 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 flight 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.

To consider how signal processing plays a role in modern domestic products, Section 2.1.3 considers how audio is streamed to your phone.

2.1.3 From Studio to the Ear

As an immediate application of **signal and system** theory, consider the Engineering processes that have occurred in delivering down-loadable music to your phone, either high-definition formats such as free lossless audio codec (FLAC) files (much preferred and strongly encouraged) or lossy-compressed files (if you really really must and don't appreciate sonic quality). A very simplified diagram is shown in Figure 2.6.

A sound is generated in a room, which generates a sound pressure wave which propagates throughout the room until reaching a microphone. This electro-mechanical device converts the sound pressure wave into an **analogue continuous-time signal** which appears as a voltage waveform. This signal is



New slide

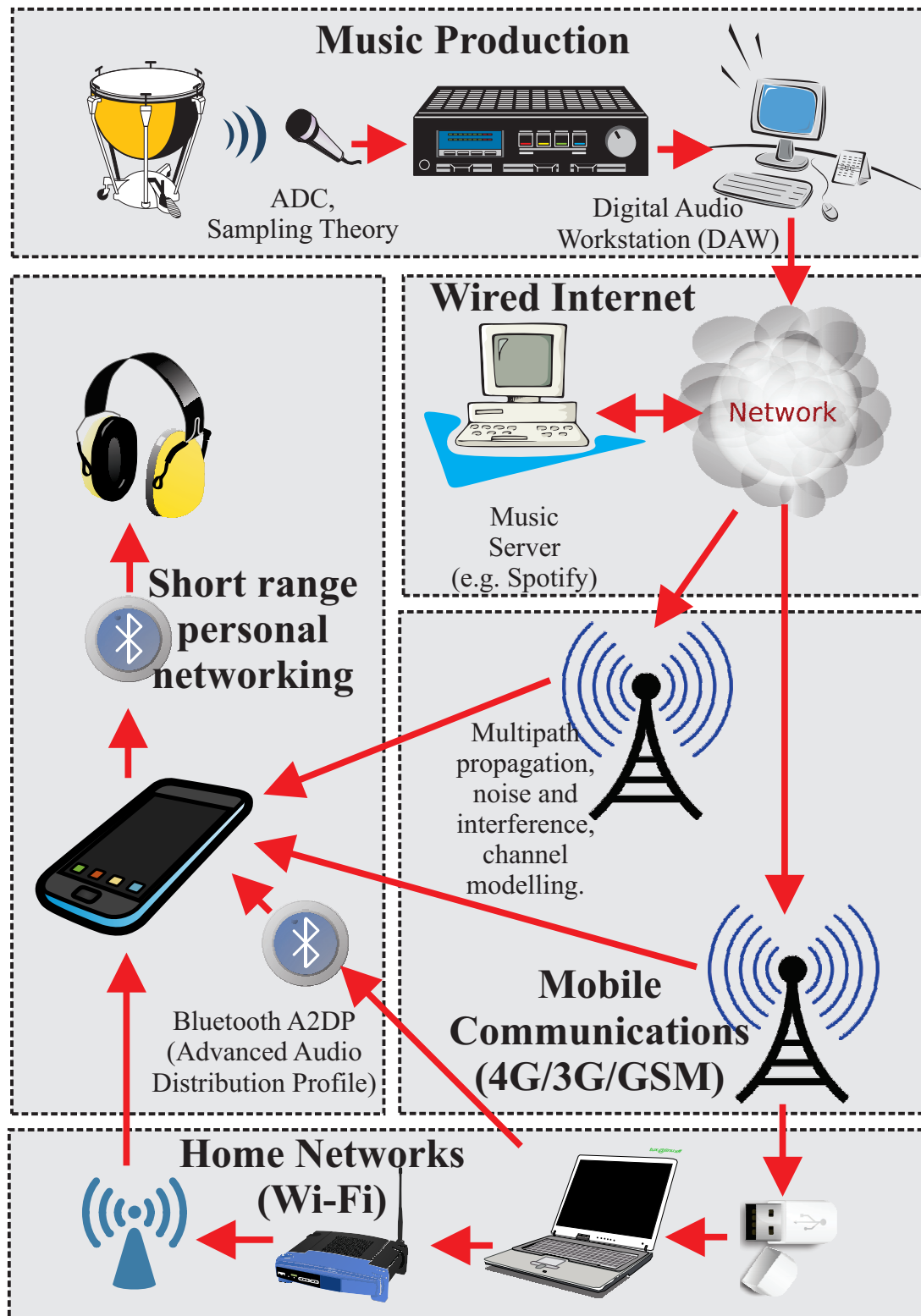


Figure 2.6: From an instrument being played through to listening on Advanced Audio Distribution Profile (A2DP) Bluetooth headphones via a portable media player.

sampled by an analogue-to-digital converter (ADC), which quantises and samples the signal, thereby producing a **discrete-time digital signal** that can be stored in finite-precision memory on a computer or digital recording device. This digital representation can then be processed on a digital audio workstation (DAW) which will compose various audio tracks and add any special-effects. Once the musical track is complete, this can then be delivered via the Internet to an online music server, probably in a compressed format (using perceptual compression). This audio track can then be delivered via a mobile network to a laptop or phone, which can then relay the signal to a set of Bluetooth headphones using the A2DP bluetooth mode.¹ This process involves a number of signal analysis and processing methods, such as sampling the analogue signal to produce a digital signal; it also involves systems, such as the effect of the acoustics on the propagation of sound, or the circuitry within the ADC; it also involves various communication systems, including wired baseband systems, medium-range wireless systems, and short-range personal wireless systems. This course provides an introduction to the understanding and analysis of these systems.

– End-of-Topic 3: **Examples of Signal Processing** –



2.1.4 Case Study: Digital Audio Processing

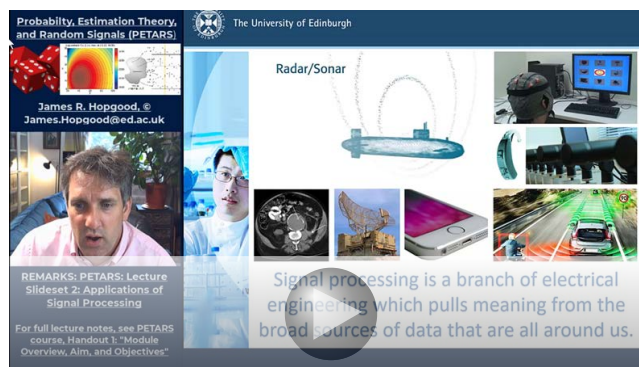
Topic Summary 4 Topic Title TBC

Topic Objectives:

- Objectives TBC.

Topic Activities:

Type	Details	Duration	Progress
Watch video	13 : 41 minute video	3×video length	
Discussion Board	Your views of signal processing	15 minutes	
Read Handout	Read page 32 to page 33	8 mins/page	



http://media.ed.ac.uk/media/1_t0qrik06

Video Summary: To be completed. Video above is a temporary link

From an electronic Engineering perspective, **signals** and **systems** is the foundation for the revolution in digital audio and video processing. Sophisticated digital electronic devices are common-place in modern everyday life; games consoles, mobile telephones, digital audio recording and playback devices, digital audio broadcasting (DAB), digital video broadcasting (DVB), digital versatile disc (DVD) video, and audio and visual streams using Moving Picture Experts Group (MPEG) compression schemes, are all very familiar to us.

These devices are the direct result of over six decades of research and innovation in the areas of **information theory** and **signal processing**.

It is common knowledge that, for example, a MPEG-1 Audio Layer 3 (MP3) player encodes an audio **signal** as a binary sequence of *ones* and *zeros*. However, such a statement isn't saying very much since, for example, word processing documents are also encoded as *ones* and *zeros*. So what makes an audio file different to an arbitrary electronic document?

To understand thoroughly how MP3 works, more pertinent questions are:

- How is a continuous-time analogue signal turned into a discrete sequence of binary numbers, and what are the properties of this binary sequence?

¹See http://en.wikipedia.org/wiki/Bluetooth_profile#Advanced_Audio_Distribution_Profile_.28A2DP.29

- How many *ones* and *zeros* are needed to represent the audio signal? If they are stored as bytes, how many bytes are needed to represent each individual audio sample? How many audio samples must be recorded to faithfully reproduce the real-world analogue signal?
- The MP3 standard uses a compression technique based on the characteristics of the human-hearing mechanism; it incorporates a method known as **perceptual masking** which removes (or masks) signal components that are not perceived by the human brain. What tools are used to characterise the properties of human-hearing, and how are these acoustical properties expressed in terms of an algorithm that runs on a **digital signal processing (DSP)**?
- How is an analogue signal recreated from a sequence of *ones* and *zeros*, and how can the deficiencies of our electronic systems be overcome by clever schemes with how the data is encoded in the first place?

The issue of using **signals** and **systems** theory to overcome the deficiencies of electronics is the basis of two recent data-formats that are available for high-quality audio reproduction. The compact disc (CD) player dominated the digital audio market from the mid-1980's until the early 2000's. Although other web-driven formats now dominate, such as MP3 and other proprietary formats, in the 1990's, the music industry initially pushed two new high-end audio formats: SACD and DVD-A. These formats store more data than the traditional CD, despite the fact that CDs already store just enough data to accurately encode the audio stream. By storing much more information than needed, SACD and DVD-A can use several tricks which mean that cheaper and less accurate electronics are needed in the playback device. How exactly do these tricks work? This will be answered later in the course.

The physical-media based SACD and DVD-A are essentially a failed format, primarily because of their high-prices, the lack of interest in multi-channel audio formats at the time, and the fact that there is sufficient download bandwidth to avoid physical-media for music. Nevertheless, stereo HD audio files such as 24/96 formats are increasingly becoming available in a download format such as FLAC and ALAC, amongst others. The insight gained from the SACD and DVD-A are the same as for downloadable HD audio formats, and Sony is in now pushing the hi-res audio format with considerable drive: <http://www.sony.co.uk/electronics/hi-res-audio>.

– End-of-Topic 4: **case studies of signal processing** –





(a) The Blu-Ray Disc Logo

(b) The digital versatile disc-audio (DVD-A) logo.



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

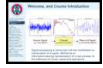


(d) The free lossless audio codec (FLAC) logo.



(e) Sony and Hi-Res Audio.

Figure 2.7: High-quality audio formats. Note that SACD and DVD-A are essentially a failed format, but HD audio files such as 24/96 formats are increasingly becoming available in a download format such as FLAC.



2.1.5 Why Study Signals and Communications?

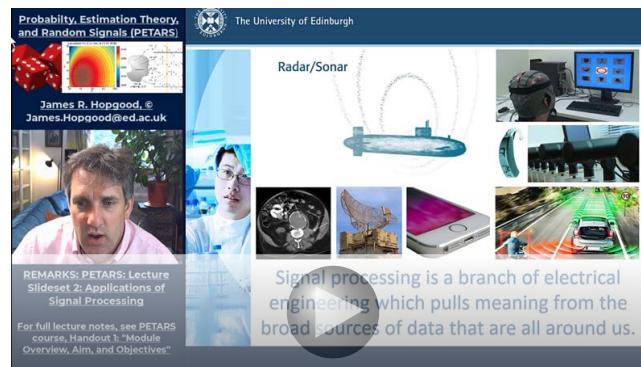
Topic Summary 5 Topic Title TBC

Topic Objectives:

- Objectives TBC.

Topic Activities:

Type	Details	Duration	Progress
Watch video	13 : 41 minute video	3× video length	
Discussion Board	Your views of signal processing	15 minutes	
Read Handout	Read page 35 to page 37	8 mins/page	



http://media.ed.ac.uk/media/1_t0qrik06

Video Summary: To be completed. Video above is a temporary link

The need for formal analysis of signals and systems stems from a number of viewpoints which will become apparent as the course progresses. In the meantime, it perhaps is simplest to begin with, as an example, the circuit shown in Figure 2.8. You might have analysed this **linear system** in other courses in your degree; the most likely analysis you will have tried is evaluating the output of the circuit when a sinusoidal signal is applied to the input. We will cover this again in this course, but could you calculate the output of the system if a microphone were connected to the input of the circuit? In such a scenario, the microphone converts a sound pressure wave into an electrical signal as the result of an instrument being played or some arbitrary spoken speech.

KEYPOINT! (Analysing system output to an arbitrary input). Evaluating the output of a linear system to an arbitrary signal is made possible by using signal analysis techniques such as the **Fourier series** and **Fourier transforms**.



2.2 Fundamental Signal Processing Problems

Consider three fundamental signal processing problems:

1. Extracting *desired* signals from other signals.

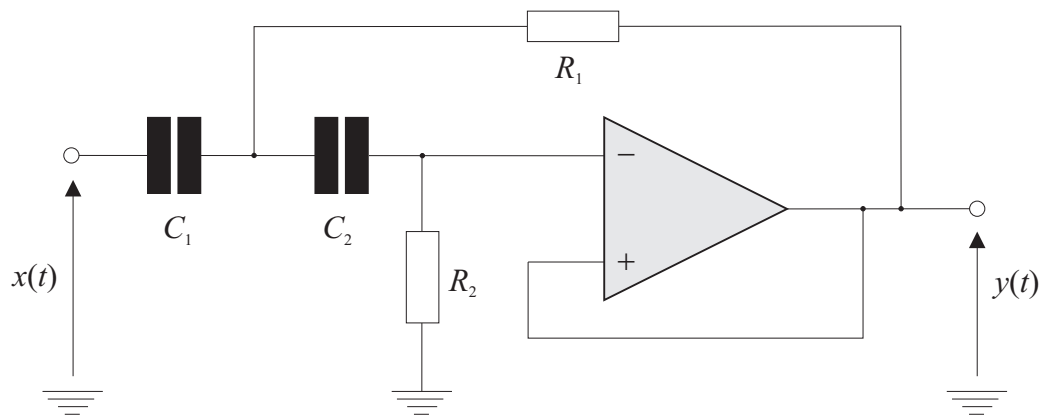
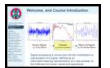
Figure 2.8: Second-order active **high-pass** filter.

Figure 2.9: Person undergoing an magnetoencephalography (MEG). National Institute of Mental Health.

2. Correcting *distortions* in measured signals.
3. Extracting *estimates* of indirect quantities from observed signals.

We shall briefly consider each of these fundamental applications in turn, and then consider what tools we need to solve these problems.

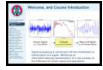
2.2.1 Extracting Signals from Other Signals



The generic problem of extracting signals from a mixture of other signals covers a wide range of applications, from simple noise reduction or removal, through to signal separation problems. As an example application, consider functional neuroimaging technique for mapping brain activity, called MEG, seen in Figure 2.9. This technique records magnetic fields produced by electrical currents naturally occurring in the brain using very sensitive magnetometers. These signals are extremely small; moreover, due to the number of electrodes present, a number of signals are measured, and there is a variety of interferences from other electromagnetic signals in the human body.

In the examples shown in Figure 2.10a, there are 148 signals of length 1695 samples over 10 seconds, or a sampling frequency of 169.55 Hz. In order to extract the brain activity, it is necessary to remove interference resulting from the heart. This interference overlaps with the desired frequencies in the

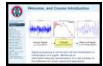
brain activity, and therefore cannot be removed with a basic filter. This requires a technique called blind source separation (BSS), which requires models for the underlying interfering signals, as well as a model for the system which mixes the signals. The extracted signals are shown in Figure 2.10b, which show the signal resulting from the heart (can you calculate the patient's heart-rate?).



2.2.2 Correcting Distortions in Measured Signals

New slide

While visible-spectrum camera images are usually very high quality, remote imaging or sensing technologies are significantly less so. Techniques such as synthetic aperture RADAR (SAR) produce noisy images with much distortion. Signal processing techniques can be used to significantly improve the quality of the image, as shown in Figure 2.11.



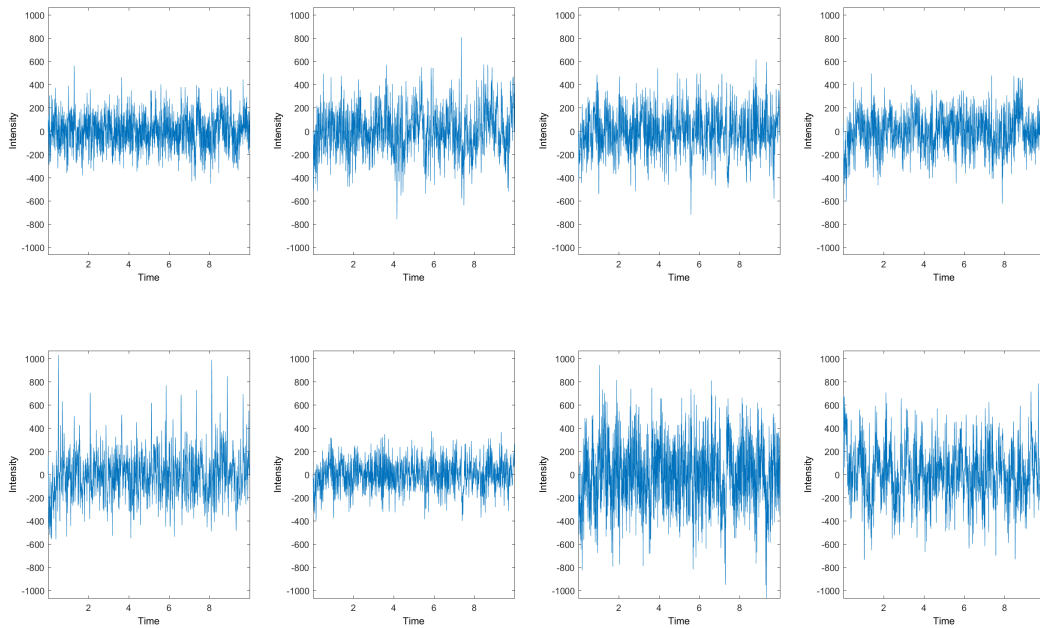
2.2.3 Indirect Parameter Estimation

New slide

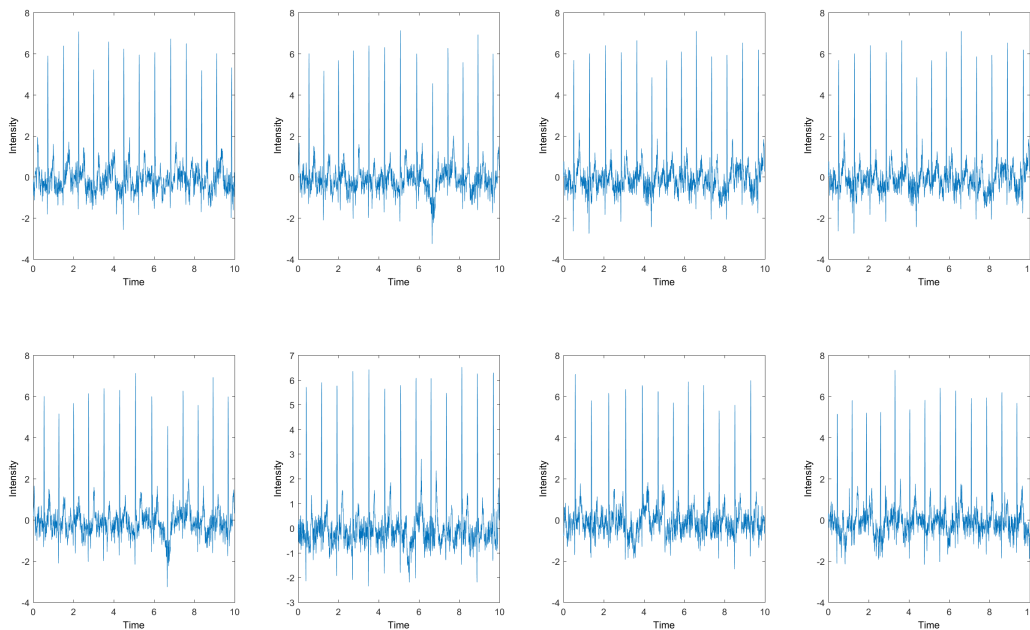
A further application of signal processing is the estimation of a quantity indirectly from measured signals. Figure 2.12 shows a multi-static radar system that uses multiple transmit and receive antenna's to locate an aircraft. The underlying signals are pulse chirps transmitted and received, but the quantity of interest is the actual position of the aircraft.

– End-of-Topic 5: **fundamental signal processing problems** –





(a) Example MEG signals.



(b) Extracted heart interference. Data kindly supplied by Dr Javier Escudero (School of Engineering, University of Edinburgh).

Figure 2.10: Signal processing of MEG signals.

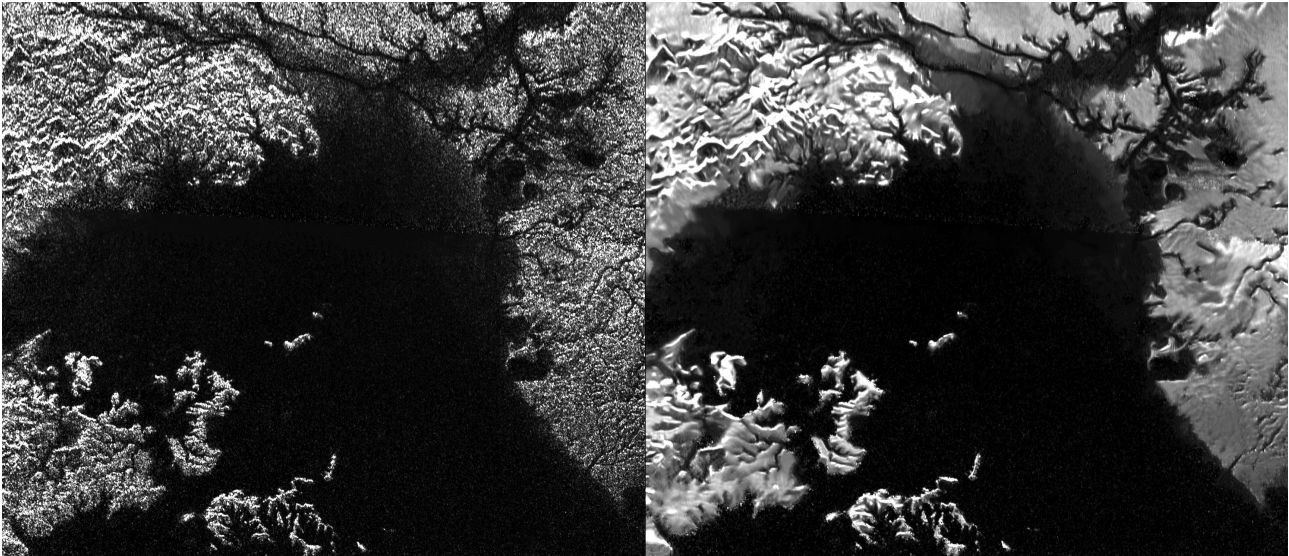


Figure 2.11: SAR and clearer despeckled views of Titan – Ligeia Mare. NASA/JPL-Caltech/ASI. Presented here are side-by-side comparisons of a traditional Cassini SAR view and one made using a new technique for handling electronic noise that results in clearer views of Titan's surface.

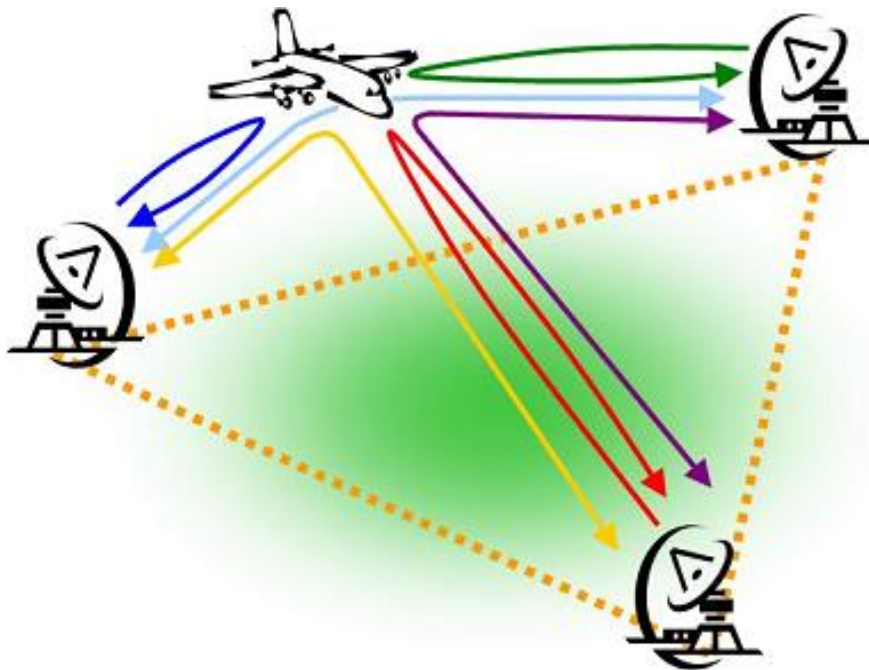


Figure 2.12: A multistatic RADAR Multistatic system, by Srdoughty / CC BY-SA 3.0.

2.2.4 Tools for solving these problems

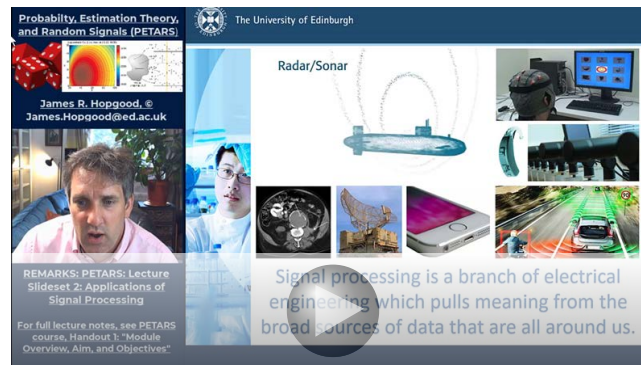
Topic Summary 6 Topic Title TBC

Topic Objectives:

- Objectives TBC.

Topic Activities:

Type	Details	Duration	Progress
Watch video	13 : 41 minute video	3× video length	
Discussion Board	Your views of signal processing	15 minutes	
Read Handout	Read page 40 to page 40	8 mins/page	



http://media.ed.ac.uk/media/1_t0qrik06

Video Summary: To be completed. Video above is a temporary link

In each application scenario considered in this section, it is necessary to:

- Understand the nature and structure of the signal in the real world.
- Understand the nature of how the signal was acquired by our data processing system.
- Understand how the signals are effected by propagation through systems.
- Design systems that can modify or change the signals to our needs.

An example of the different signal processing chains is shown in Figure 2.13, and will be discussed further in lectures (and expanded on here in due course).

– End-of-Topic 6: **The Signal Processing Chain** –



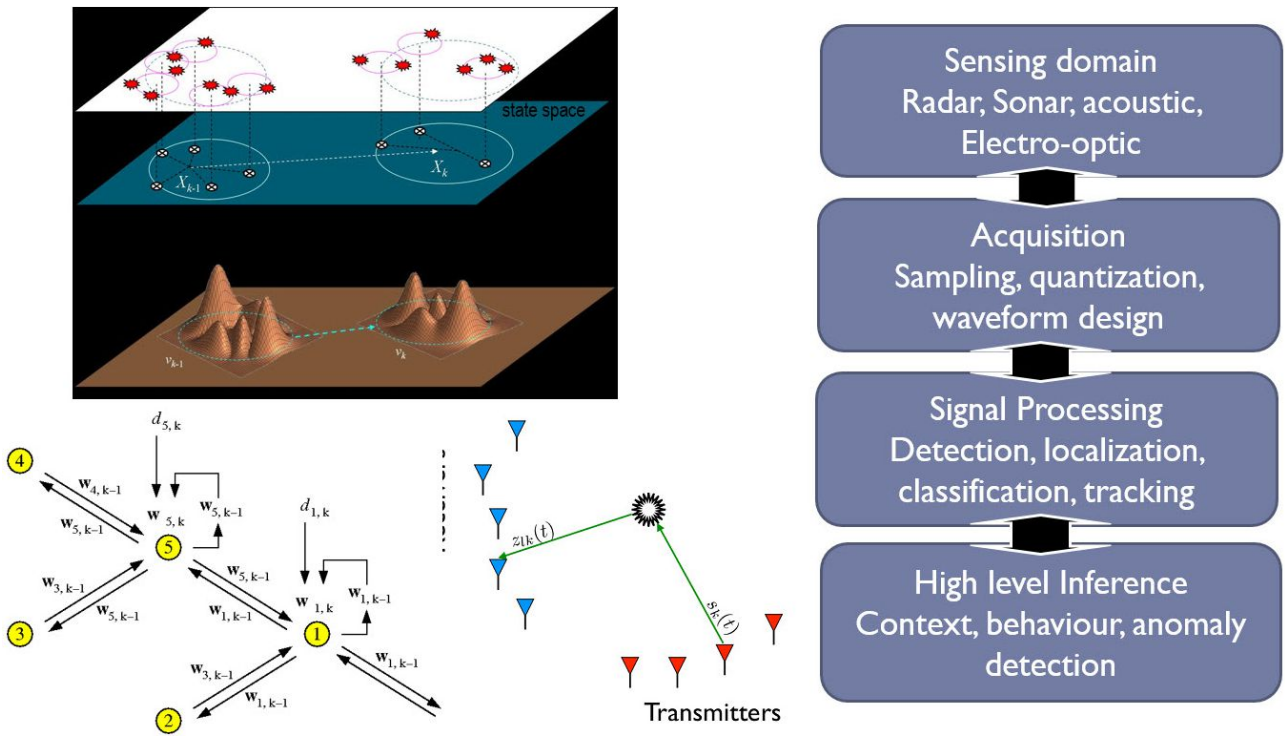
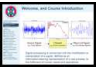


Figure 2.13: The signal processing chain.

2.3 What are Signals and Systems?



Topic Summary 7 Topic Title TBC

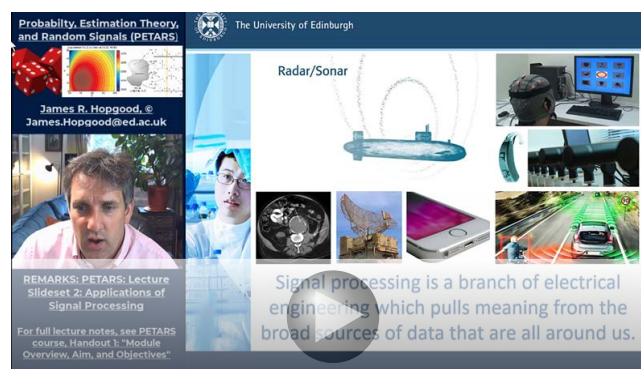
New slide

Topic Objectives:

- Objectives TBC.

Topic Activities:

Type	Details	Duration	Progress
Watch video	13 : 41 minute video	3× video length	
Discussion Board	Your views of signal processing	15 minutes	
Read Handout	Read page 42 to page 47	8 mins/page	



http://media.ed.ac.uk/media/1_t0qrik06

Video Summary: To be completed. Video above is a temporary link

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 **signals**. In general, **systems** operate on **signals** to produce new signals or new signal representations. For example, an audio 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.

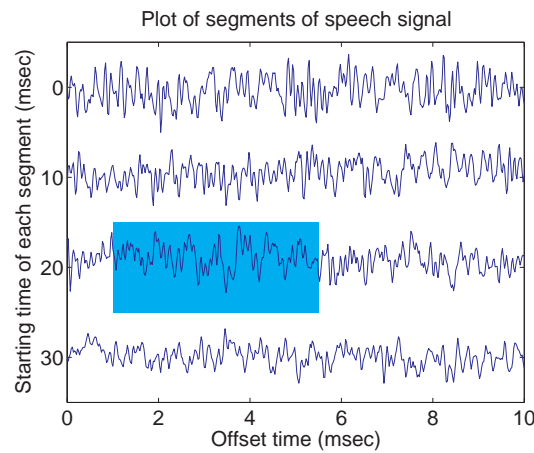


Figure 2.14: Plot of part of a speech signal. This signal can be represented by the function $s(t)$, where t is the independent variable representing time. The shaded region is shown in more detail in Figure 2.15.

2.3.1 Mathematical Representation of Signals

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 2.14. 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.

2.3.1.1 Continuous-time and discrete-time signals

The signal shown in Figure 2.14 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. 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:

$$s[n] = s(nT_s), \quad n \in \{0, 1, 2, \dots\}. \quad (2.1)$$

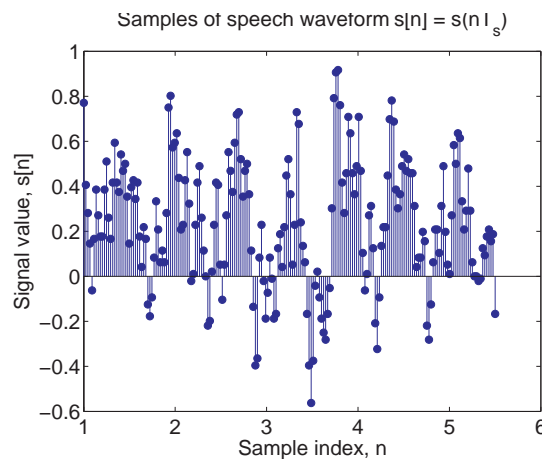


Figure 2.15: Example of a discrete-time signal. This is a sampled version of the shaded region shown in Figure 2.14.

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 to (and is not yet consistent in these notes either).

Figure 2.15 shows an example of a short segment of the speech waveform from Figure 2.14, 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 $s(t)$ 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 largest **sampling period** that can be used to retain all or most of the information about the original signal.

2.3.1.2 Other types of signals

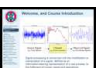
While many signals can be considered as evolving patterns in time, many other signals are not time-varying patterns at all. For example, an image formed by focusing light through a lens onto an imaging array is a spatial pattern. Thus, an image is represented mathematically as a function of two independent 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 2.16; 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 2.16 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:

$$p[m, n] = p(m\Delta_x, n\Delta_y), \quad m, n \in \{0, 1, \dots, N-1\}. \quad (2.2)$$

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



New slide



Figure 2.16: Example of a signal that can be represented by a function of two spatial variables.

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.

2.3.2 Mathematical Representation of Systems

A **system** manipulates, changes, records, or transmits **signals**. 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

$$y(t) = \mathcal{T} \{x(t)\} \quad (2.3)$$

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

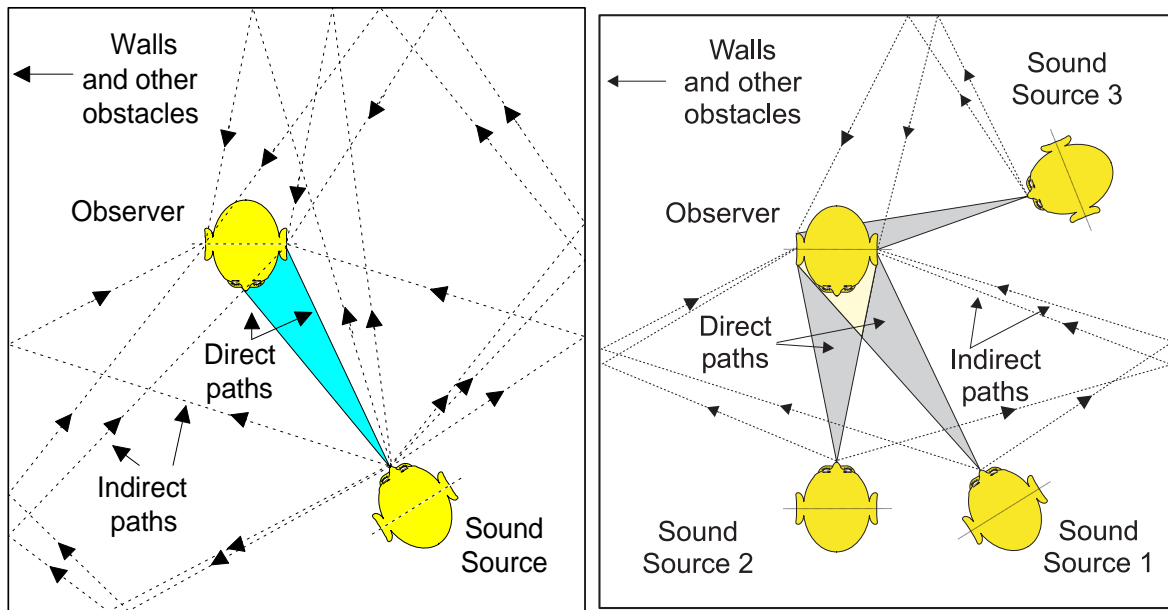
$$y(t) = [x(t)]^2 \quad (2.4)$$

Figure 2.17 and Figure 2.19 show how signals can be generated and observed in a real application. In Figure 2.17, the sound source and the information received by the observer, or microphone, are the **signals**; the room acoustics represent the **system**. Figure 2.18 shows the **input signal** to the system, a *characterisation of the system*, and the resulting **output signal**. In Figure 2.19, 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:



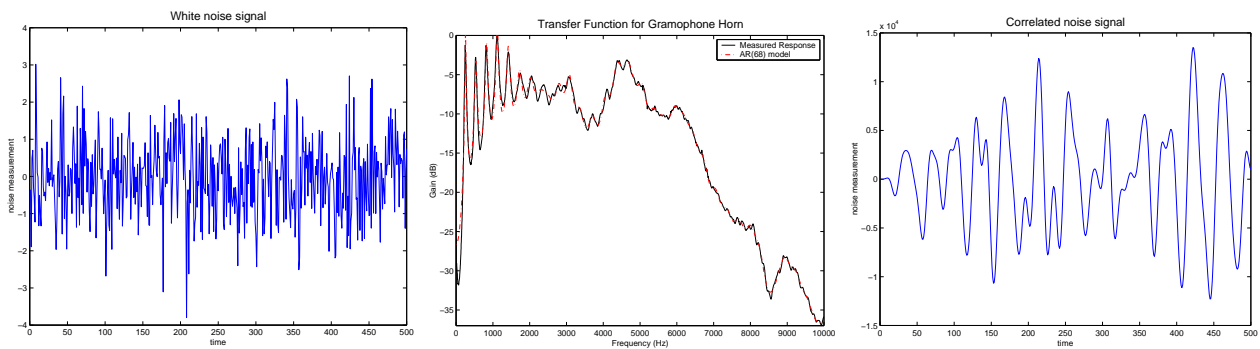
New slide



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

(b) Many sound sources within a room.

Figure 2.17: Observed signals in room acoustics.



(a) Source signal into a system.

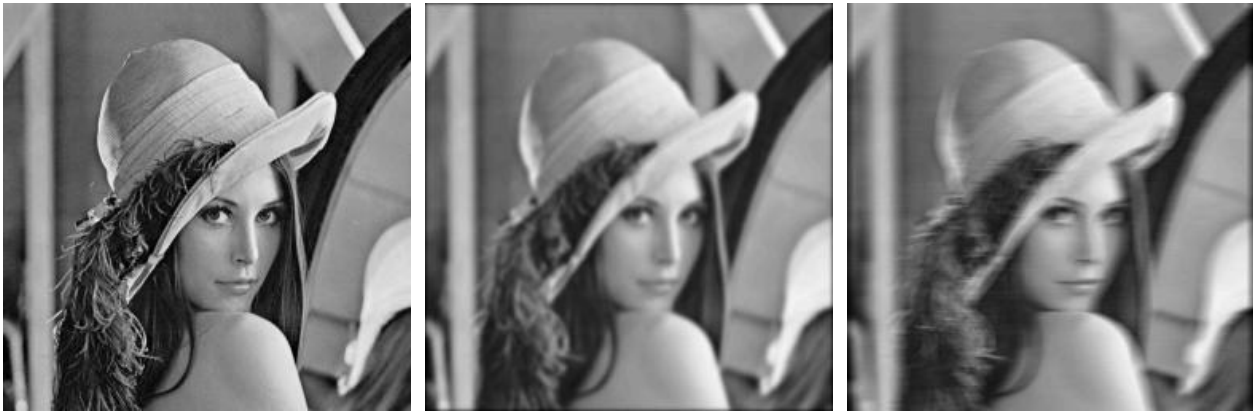
(b) A frequency response representing the characteristics of the system.

(c) The system output.



(d) Block diagram representation of signal paths.

Figure 2.18: 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 2.19: A blind image deconvolution problem; restoration of natural photographic images.

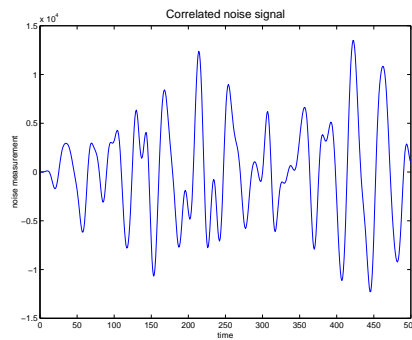


Figure 2.20: Amplitude-verses-time plot.

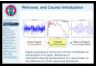
Signal processing is concerned with the modification or manipulation of a signal, defined as an information-bearing representation of a real process, that has been passed through a *system*, to the fulfillment of human needs and aspirations.

2.3.3 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 2.20; 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.



2.4 Motivation for Signal Modelling



Topic Summary 8 Topic Title TBC

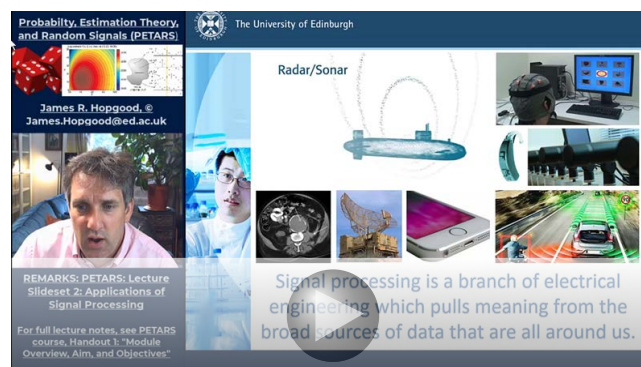
New slide

Topic Objectives:

- Objectives TBC.

Topic Activities:

Type	Details	Duration	Progress
Watch video	13 : 41 minute video	3×video length	
Discussion Board	Your views of signal processing	15 minutes	
Read Handout	Read page 48 to page 52	8 mins/page	



http://media.ed.ac.uk/media/1_t0qrik06

Video Summary: To be completed. Video above is a temporary link

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

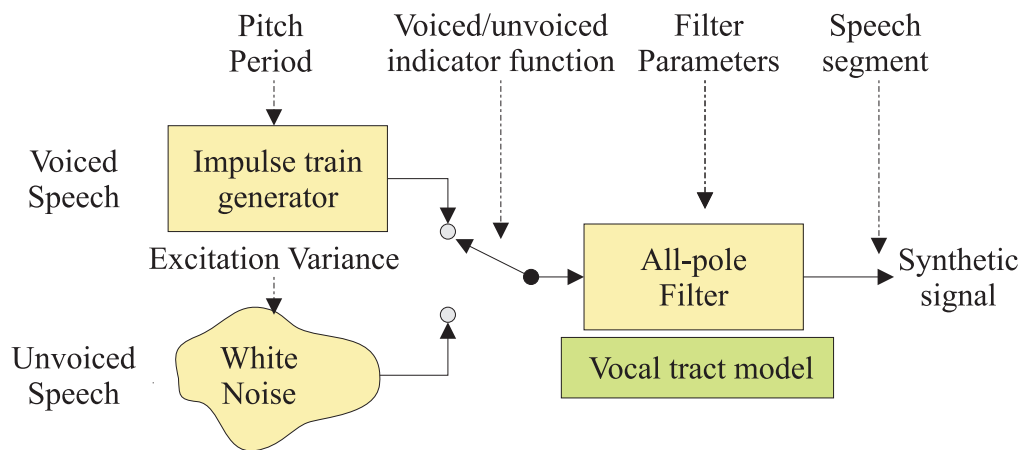


Figure 2.21: The speech synthesis model.

image, as shown in Figure 2.19; 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
- Single Channel Blind System Identification
- Blind Signal Separation
- Data Compression
- Enhancement of Signals in Noise

2.4.1 Speech Modelling and Recognition

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;



New slide



Figure 2.22: Solutions to the blind deconvolution problem requires advanced statistical signal processing.

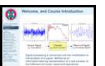
- unvoiced speech is characteristic of consonants at the beginning of syllables, fricatives (*/f/*, */s/* sounds), and a combination of these.

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 possess a form that can be modelled using the statistical models introduced in later lectures.

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 2.21. 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**.

2.4.2 Single Channel Blind System Identification



New slide

Consider the following abstract problem that is shown in Figure 2.22:

- 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, $\mathbf{x} = \{x[n], n \in \mathbb{Z}\}$,² 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:

$$x[n] = h[n] * s[n] = \sum_{k \in \mathbb{Z}} h[n - k] s[k] \quad (2.5)$$

- **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 .

²The 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.

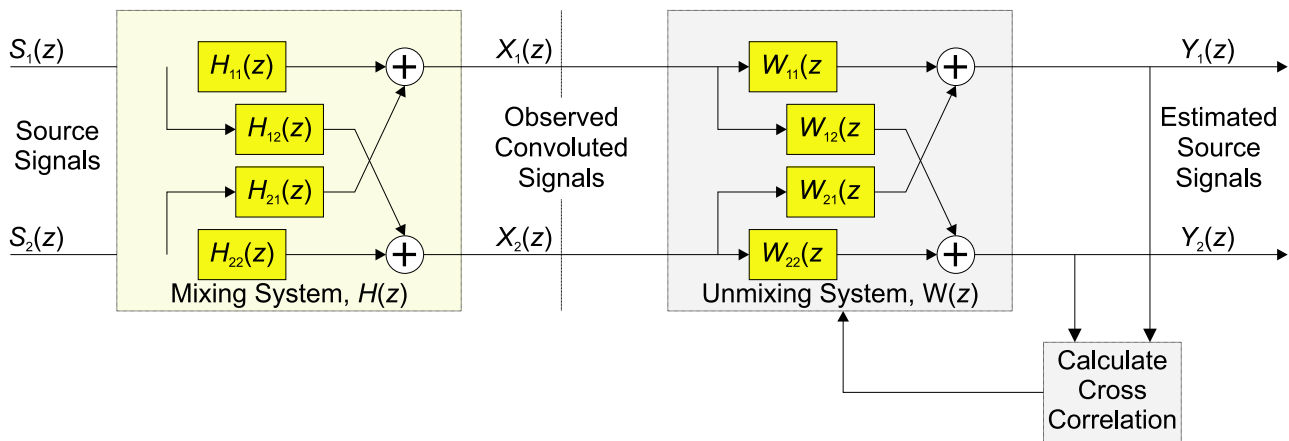


Figure 2.23: Standard signal separation using the independent component assumption.

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.

2.4.3 Blind Signal Separation

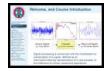
An extremely broad and fundamental problem in signal processing is BSS, and an important special case is the separation of a mixture of audio signals in an acoustic environment. Typical applications include the separation of overlapping speech signals, the separation of musical instruments, enhancement 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 2.17).

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 2.23 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.



New slide



2.4.4 Data Compression

New slide

Three basic 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.

– End-of-Topic 8: Examples of Signal Processing Applications –





Figure 2.24: High-quality audio formats.

2.4.5 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 versatile 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

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.

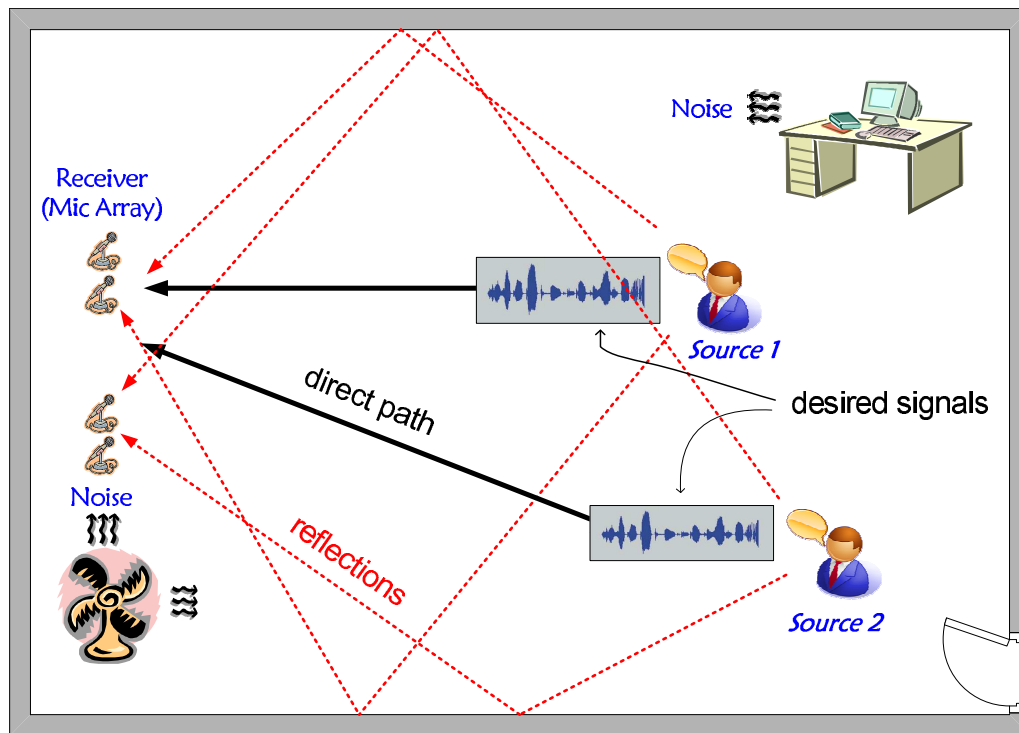


Figure 2.25: Passive source localisation and BSS.

2.5 Passive and Active Target Localisation

This section presents a standard application in signal processing, namely passive target localisation. Active target localisation will be considered during the day as well, but this section will focus on the passive scenario. The aim of this section is to present, briefly, solutions to this problem, without restricting the notation used. If the mathematics is somewhat alien, then great, as the rest of this tutorial will explain the terms and concepts used here. An expanded version of this section, with a focus on acoustic source localisation, is included at the end of this handout.

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

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

- Passive localisation is particularly challenging.

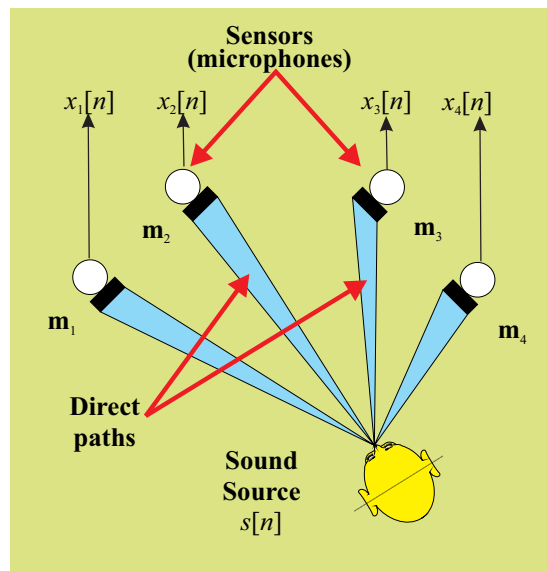
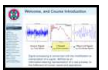


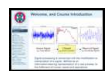
Figure 2.26: Ideal free-field model.

2.6 Passive Target Localisation Methodology



- In general, most passive target localisation (PTL) techniques rely on the fact that an impinging wavefront reaches one acoustic sensor before it reaches another (spatio-temporal diversity). *New slide*
- Many PTL algorithms are designed assuming there is no multipath or reverberation present, the *free-field assumption*.

2.6.1 Source Localization Strategies



New slide

Existing source localisation methods can loosely be divided into three generic strategies:

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

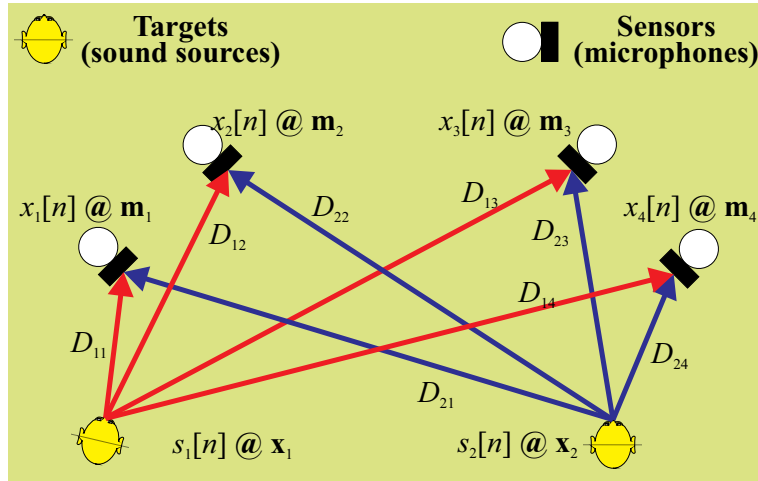


Figure 2.27: Geometry assuming a free-field model.

2.6.2 Geometric Layout

Suppose there is a:

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

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

$$T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k) \triangleq T_{ij}(\mathbf{x}_k) = \frac{|\mathbf{x}_k - \mathbf{m}_i| - |\mathbf{x}_k - \mathbf{m}_j|}{c} \quad (2.6)$$

where c is the speed of the impinging wavefront.

2.6.3 Ideal Free-field Model

- In an anechoic free-field environment, as depicted in Figure 2.26, the signal from source k , denoted $s_k(t)$, propagates to the i -th sensor at time t as:

$$x_{ik}(t) = \alpha_{ik} s_k(t - \tau_{ik}) + b_{ik}(t) \quad (2.7)$$

where $b_{ik}(t)$ denotes additive noise, and α_{ik} is the attenuation.

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

$$X_{ik}(\omega) = \alpha_{ik} S_k(\omega) e^{-j\omega\tau_{ik}} + B_{ik}(\omega) \quad (2.8)$$

On the assumption of **geometrical wave propagation**, which assumes high frequencies, a point source of single frequency ω , at position \mathbf{x}_k in free space, emits a pressure wave $P_{(\mathbf{x}_k, \mathbf{m}_i), t}(\omega)$ at time t and at position \mathbf{m}_i :

$$P_{(\mathbf{x}_k, \mathbf{m}_i)}(\omega, t) = P_0 \frac{\exp[j\omega(r/c - t)]}{r} \quad (2.9)$$

where $t \in \mathbb{R}$ is time, and $r = |\mathbf{x}_k - \mathbf{m}_i|$.

- The additive noise source is assumed to be uncorrelated with the source and noise sources at other sensors.
- The TDOA between the i -th and j -th sensor is given by:

$$\tau_{ijk} = \tau_{ik} - \tau_{jk} = T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k) \quad (2.10)$$

2.7 Indirect TDOA-based Methods

This is typically a two-step procedure in which:

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

2.7.1 Hyperbolic Least Squares Error Function

KEYPOINT! (Underlying Concept). Suppose that for each pair of sensors, i and j , a TDOA corresponding to source k is somehow estimated, and this is denoted by τ_{ijk} . One approach to ASL is to minimise the total error between the measured TDOAs and the TDOAs predicted by the geometry given an assumed target position.

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

$$\epsilon_{ij}(\mathbf{x}_k) = \tau_{ijk} - T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k) \quad (2.11)$$

where the error is considered as a function of the source position \mathbf{x}_k .

- The total error as a function of target position

$$J(\mathbf{x}_k) = \sum_{i=1}^N \sum_{j \neq i=1}^N \epsilon_{ij}(\mathbf{x}_k) = \sum_{i=1}^N \sum_{j \neq i=1}^N (\tau_{ijk} - T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k))^2 \quad (2.12)$$

where

$$T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k) \triangleq T_{ij}(\mathbf{x}_k) = \frac{|\mathbf{x}_k - \mathbf{m}_i| - |\mathbf{x}_k - \mathbf{m}_j|}{c} \quad (2.13)$$

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

2.7.2 TDOA estimation methods

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

GCC algorithm most popular approach assuming an ideal free-field model. It has the advantages that

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

However, GCC-based methods

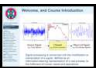
- fail when multipath is high;
- focus of current research is on combating the effect of multipath.

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

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

Note that both methods assume that the signals received at the sensors arise as the result of a single source, and that if there are multiple sources, the signals will first need to be separated into different contributions of the individual sources.

2.7.2.1 GCC TDOA estimation

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

- The TDOA estimate between two microphones i and j is obtained as the time lag that maximises the cross-correlation between the filtered versions of the microphone outputs:

$$\hat{\tau}_{ij} = \arg \max_{\ell} r_{x_i x_j}[\ell] \quad (2.14)$$

where the signal received at microphone i is given by $x_i[n]$, and where x_i should not be confused with the location of the source k , which is denoted by $\mathbf{x}_k = [x_k, y_k, z_k]^T$.

- The cross-correlation function is given by

$$r_{x_i x_j}[\ell] = \mathcal{F}^{-1} \left(\Psi_{x_1 x_2} (e^{j\omega T_s}) \right) \quad (2.15)$$

$$= \mathcal{F}^{-1} \left(\Phi (e^{j\omega T_s}) P_{x_1 x_2} (e^{j\omega T_s}) \right) \quad (2.16)$$

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

$$P_{x_1 x_2} (e^{j\omega T_s}) = \mathbb{E} [X_1 (e^{j\omega T_s}) X_2 (e^{j\omega T_s})] \quad (2.17)$$

The cross-power spectral density (CPSD) can be estimated in a variety of means. The choice of the filtering term or frequency domain weighting function, $\Phi (e^{j\omega T_s})$, leads to a variety of different GCC methods for TDOA estimation.

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

$$\angle P_{x_i x_j}(\omega) = -j\omega T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k) \quad (2.18)$$

In other words, all the TDOA information is conveyed in the phase rather than the amplitude of the CPSD. This therefore suggests that the weighting function can be chosen to remove the amplitude information.

2.7.2.2 GCC Processors

The most common choices for the GCC weighting term are listed in the table below. In particular, the phase transform (PHAT) is considered in detail.

Processor Name	Frequency Function
Cross Correlation	1
PHAT	$\frac{1}{ P_{x_1 x_2}(e^{j\omega T_s}) }$
Roth Impulse Response	$\frac{1}{P_{x_1 x_1}(e^{j\omega T_s})}$ or $\frac{1}{P_{x_2 x_2}(e^{j\omega T_s})}$
SCOT	$\frac{1}{\sqrt{P_{x_1 x_1}(e^{j\omega T_s}) P_{x_2 x_2}(e^{j\omega T_s})}}$
Eckart	$\frac{P_{s_1 s_1}(e^{j\omega T_s})}{P_{n_1 n_1}(e^{j\omega T_s}) P_{n_2 n_2}(e^{j\omega T_s})}$
Hannon-Thomson or ML	$\frac{ \gamma_{x_1 x_2}(e^{j\omega T_s}) ^2}{ P_{x_1 x_2}(e^{j\omega T_s}) (1 - \gamma_{x_1 x_2}(e^{j\omega T_s}) ^2)}$

where $\gamma_{x_1 x_2}(e^{j\omega T_s})$ is the normalised CPSD or **coherence function** is given by

$$\gamma_{x_1 x_2}(e^{j\omega T_s}) = \frac{P_{x_1 x_2}(e^{j\omega T_s})}{\sqrt{P_{x_1 x_1}(e^{j\omega T_s}) P_{x_2 x_2}(e^{j\omega T_s})}} \quad (2.19)$$

The PHAT-GCC approach can be written as:

$$r_{x_i x_j}[\ell] = \int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}} \Phi(e^{j\omega T_s}) P_{x_1 x_2}(e^{j\omega T_s}) e^{j\ell\omega T} d\omega \quad (2.20)$$

$$= \int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}} \frac{1}{|P_{x_1 x_2}(e^{j\omega T_s})|} |P_{x_1 x_2}(e^{j\omega T_s})| e^{j\angle P_{x_1 x_2}(e^{j\omega T_s})} e^{j\ell\omega T} d\omega \quad (2.21)$$

$$= \int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}} e^{j(\ell\omega T + \angle P_{x_1 x_2}(e^{j\omega T_s}))} d\omega \quad (2.22)$$

$$= \delta(\ell T_s + \angle P_{x_1 x_2}(e^{j\omega T_s})) \quad (2.23)$$

$$= \delta(\ell T_s - T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k)) \quad (2.24)$$

- In the absence of reverberation, the GCC-PHAT (GCC-PHAT) algorithm gives an impulse at a lag given by the TDOA divided by the sampling period.

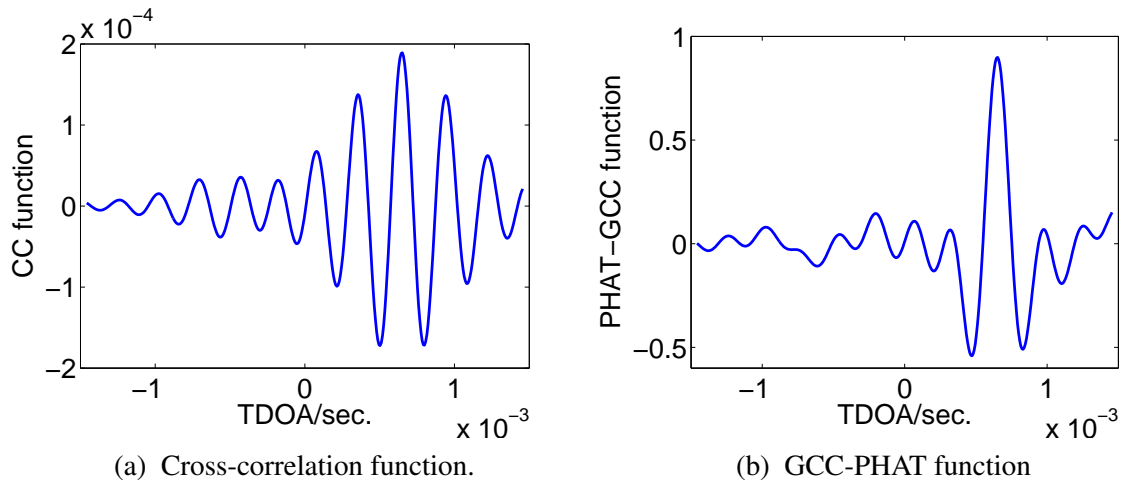


Figure 2.28: Normal cross-correlation and GCC-PHAT functions for a frame of speech.

2.8 Direct Localisation Methods

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

2.8.1 Steered Response Power Function

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

The frequency domain **delay-and-sum beamformer** steered to a spatial position $\hat{\mathbf{x}}_k$ such that $\hat{\tau}_{pk} = |\hat{\mathbf{x}} - \mathbf{m}_p|$, using the notation in Equation 13.8, is given by:

$$S(\hat{\mathbf{x}}) = \int_{\Omega} \left| \sum_{p=1}^N W_p(e^{j\omega T_s}) X_p(e^{j\omega T_s}) e^{j\omega \hat{\tau}_{pk}} \right|^2 d\omega \quad (2.25)$$

Expanding, rearranging the order of integration and summation, taking expectations of both sides and setting $\Phi_{pq}(e^{j\omega T_s}) = W_p(e^{j\omega T_s}) W_q^*(e^{j\omega T_s})$ gives

$$\mathbb{E}[S(\hat{\mathbf{x}})] = \sum_{p=1}^N \sum_{q=1}^N r_{x_i x_j}[\hat{\tau}_{pqk}] \quad (2.26)$$

$$\equiv \sum_{p=1}^N \sum_{q=1}^N r_{x_i x_j} \left[\frac{|\mathbf{x}_k - \mathbf{m}_i| - |\mathbf{x}_k - \mathbf{m}_j|}{c} \right] \quad (2.27)$$

In other words, the SRP is the sum of all possible pairwise GCC functions evaluated at the time delays hypothesised by the target position.

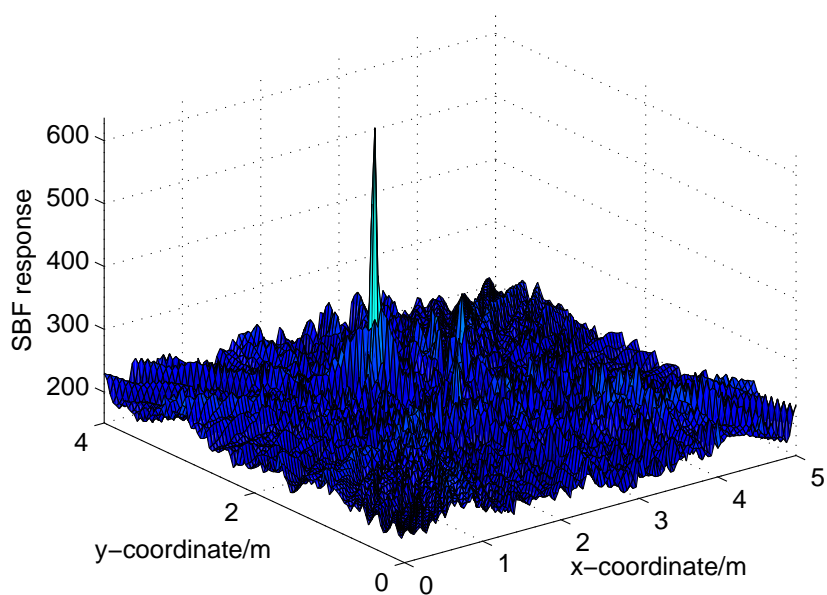


Figure 2.29: SBF response from a frame of speech signal. The integration frequency range is 300 to 3500 Hz (see Equation 13.84). The true source position is at $[2.0, 2.5]m$. The grid density is set to 40 mm.

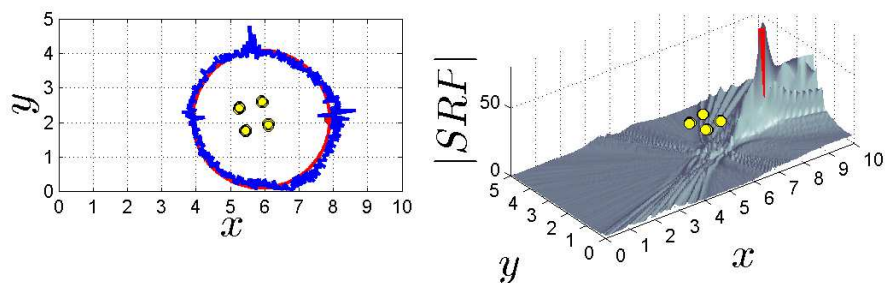
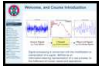


Figure 2.30: An example video showing the SBF changing as the source location moves.

2.8.2 Conclusions



New slide

To fully appreciate the algorithms in PTL, we need:

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

Part II

Probability, Random Variables, and Estimation Theory

3

Review of Basic Probability Theory



All knowledge degenerates into probability.

David Hume

This handout motivates the need for and gives a review of the fundamentals of probability theory. The idea is to motivate the definitions of cumulative distribution functions (cdfs) and probability density functions (pdfs) in the next handout, which form the foundation of statistical estimation theory and signal processing.

3.1 Introduction

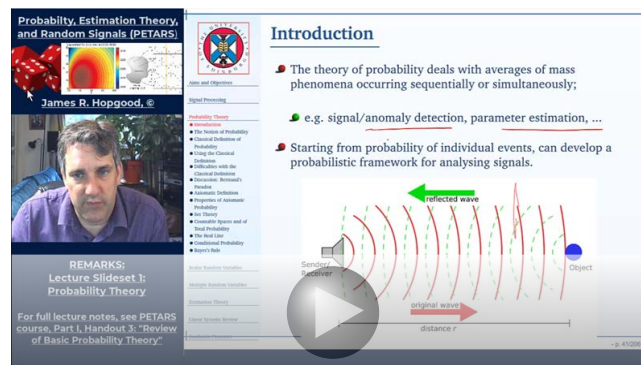
Topic Summary 9 Motivating Empirical Probability

Topic Objectives:

- Introduce uncertainty through a simple example.
- Discuss general applications of probability.
- Use law-of-large numbers to define empirical probability.

Topic Activities:

Type	Details	Duration	Progress
Watch video	10.54 mins video	3× video length	
Read Handout	Read page 65 to page 69	8 mins/page	
Discussion Board	Discuss Taxi-Cab Problem	20 mins	



http://media.ed.ac.uk/media/0_3jxfljjc

Video Summary: This video motivates probability by considering the simplest of problems in the presence of uncertainty. It considers the tools we need to study problems, and the notion of probability. This begins by discussing how the law-of-large numbers leads to the definition of empirical probability through counting successes in a series of Bernoulli trials. The definition of empirical probability, or relative frequency, which will then lead onto classical probability in the next lecture.

To motivate the need for probability theory, consider the simplest of problems in the presence of uncertainty. What tools are needed to study this problem?

- The notion of probability and random variables;
- The notion of probability density functions (pdfs);
- The notion of independence of observations;
- The notion of estimation theory and uncertainty quantification, some of which are highlighted in Figure 3.1 which shows a method called Kernel Density Estimation.

These will be studied in turn throughout this course; we will start off looking at the basics of probability.

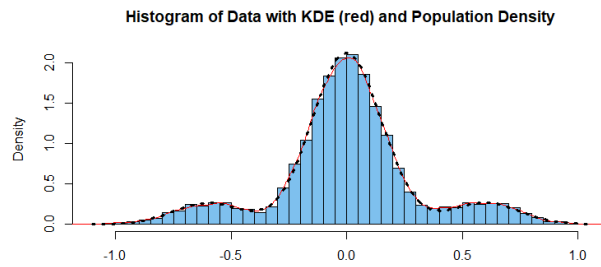


Figure 3.1: Kernel density estimation for modelling observation data.

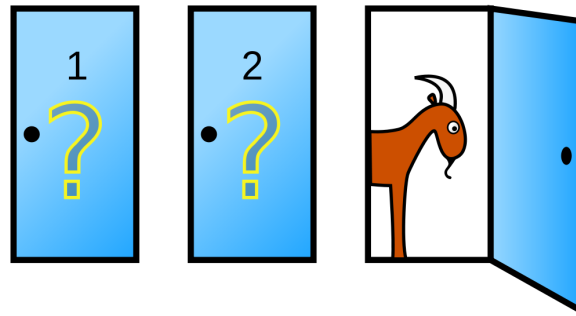


Figure 3.2: Is the infamous Monty-Hall problem counter-intuitive or not?

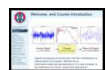
Students are exposed to probability at school from a relatively young age. It is not the intention of this course to go over basic probability again. Instead, the purpose is to:

- enhance a fundamental understanding of probability that enable us develop more complex concepts;
- identify limitations of classical definitions;
- reaffirm that human intuition with regards to probability is often wrong; and that careful and systematic analysis is often needed.

- KEYPOINT! (Probability).**
- The theory of probability deals with averages of mass phenomena occurring sequentially or simultaneously;
 - e.g. signal/anomaly detection, parameter estimation, ...
 - Starting from probability of individual events, can develop a probabilistic framework for analysing signals.

3.2 The Notion of Probability

The theory of probability deals with averages of mass phenomena occurring sequentially or simultaneously. In signal processing and communications, this phenomena might include signal returns in active radar or sonar detection (see Figure 3.3), detection of acoustic events in



New slide

Sidebar 2 The Venice Water-Taxi Problem

Understanding probability and statistics helps understand simple, but important, questions related to estimating the parameters of a sampling distribution from a small sample size.

On a trip to Venice (in July 2016), it was observed that the water taxis appeared to be numbered in sequential order from number 1 up-wards (a water-taxi with the number 1 on the side was observed, and only positive integer valued taxi designations).



Assuming that all taxis are in service, suppose we wanted to guess the number N of water taxis in Venice, based purely on the taxi numbers observed. Let's assume we observed a taxi with the number 304 on the side. What is our best guess of N ?

The solution will be discussed in detail in Chapter 5, but now is a good time to think about it in advance of learning the techniques that will help us answer the question. Moreover, suppose we observe more taxis, perhaps with the numbers 157, 202, 11, 248; how will our estimate change?



This problem might seem rather academic, but has actually in the past been far from it, as discussed in Chapter 5. A well known example is called the **German tank problem**.

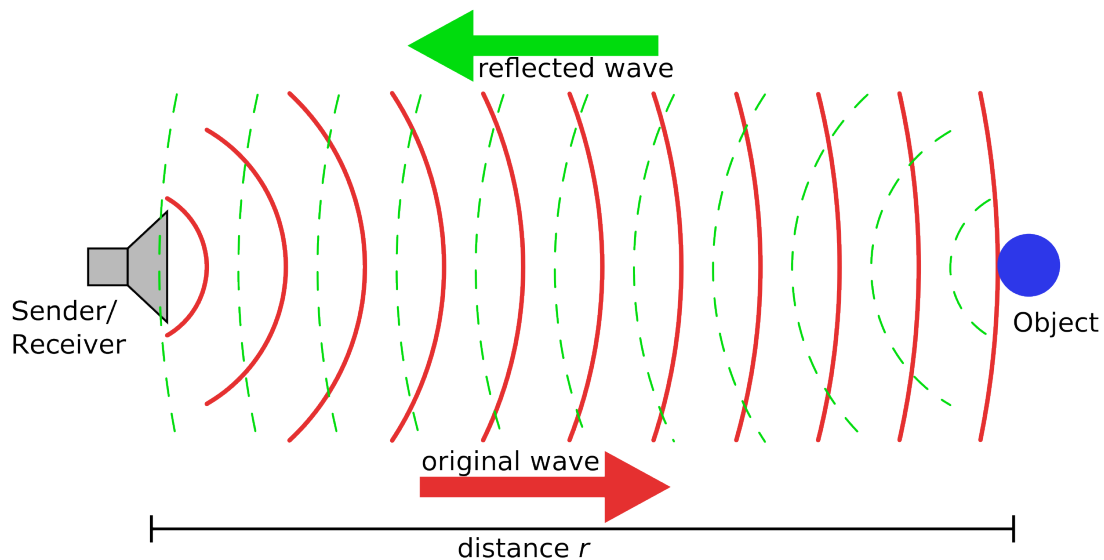


Figure 3.3: Active radar system; Drawing by Georg Wiora (Dr. Schorsch) / CC BY-SA

environmental sound analysis, anomaly detection in communication systems, parameter estimation, and so forth.

How does one start considering the notion and meaning of probability, and how can it be extended to modelling signals and events? To address this, it is first important to consider fundamentals such as the **probability** of individual events, from which a probabilistic framework for analysing signals can be obtained. To motivate the definition of probability, it is first *observed* that in many fields certain **averages** approach a constant value as the number of observations increases. 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.

This is formalised through the principal of the law of large numbers. As an illustration of the law of large numbers, consider a particular sequence of rolls of a single six-sided dice. As the number of rolls in the sequence increases, the average of the values of all the results approaches the theoretical **mean value** of $\frac{1}{6} \sum_{k=1}^6 k = 3.5$, as shown in Figure 3.4. While different sequences (or trials) would show a different *shape* over a small number of throws (at the left of Figure 3.4), over a large number of rolls (to the right of Figure 3.4) they would be extremely similar.

It follows from the law of large numbers that the **empirical probability** of success in a series of Bernoulli trials will converge to the theoretical probability. In the theory of probability and statistics, a Bernoulli trial (or binomial trial) is a random experiment with exactly two possible outcomes, “success” and “failure”, in which the probability of success is the same every time the experiment is conducted. For a Bernoulli random variable, the expected value is the theoretical probability of success, and the average of n such variables (assuming they are independent and identically distributed (i.i.d.)) is precisely the relative frequency. Therefore, the law of large numbers justifies the empirical probability, relative frequency, or experimental probability of an event is the ratio of the number of outcomes in which a specified event occurs to the total number of trials, not in a theoretical sample space but in an actual experiment. In a more general sense, empirical probability estimates probabilities from experience and observation.

Therefore, 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.

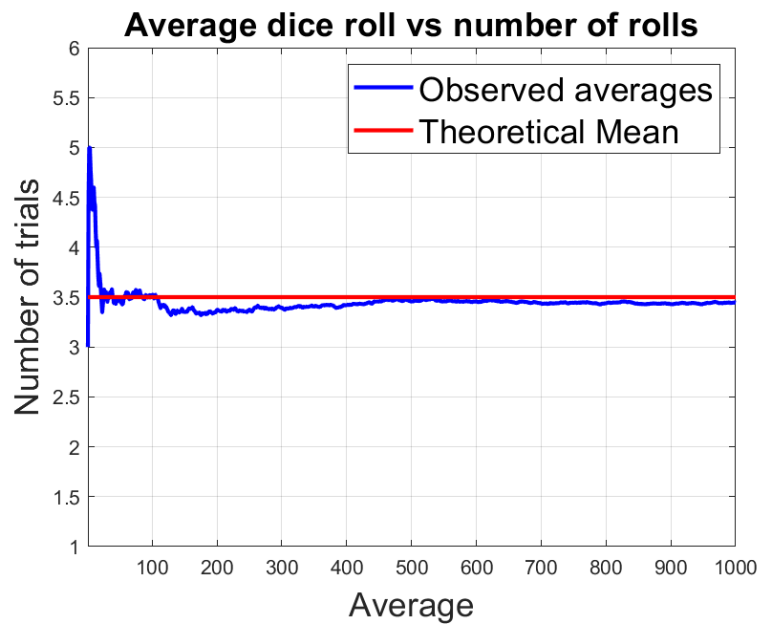


Figure 3.4: Illustrating the law-of-large numbers through throwing of a 6-sided dice.

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:

$$\Pr(A) \approx \frac{n_A}{n} \quad (3.1)$$

provided that n is *sufficiently large*.

This is called the **empirical probability**, **experimental probability**, or **relative frequency**, and is an *estimator of probability*.

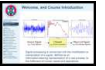
Note that this frequentist interpretation and the language used is all very imprecise, and phrases such as *high degree of certainty*, *close to*, and *sufficiently large* have no clear meaning. These terms will be more precisely defined as concepts are introduced throughout this course.

- Moreover, another problem with this definition is that it implies an experiment needs to be performed in order to define a probability. In the next section, we will move away from this restriction.

– End-of-Topic 9: **Introduction to Probability, The Law-of-Large Numbers, and Empirical Probability** –



3.3 Classical Definition of Probability



Topic Summary 10 Classical Probability

New slide

Topic Objectives:

- Introduce the definition of classical probability.
- Show simple examples of use of definition.
- Try examples and exercises.

Topic Activities:

Type	Details	Duration	Progress
Watch video	9 : 52 minute video	3× video length	
Read Handout	Read page 70 to page 73	8 mins/page	
Try Examples	Work through Example 3.3	5 mins	
Practice Exercises	Exercise ??	15 mins	

http://media.ed.ac.uk/media/1_akng711x

Video Summary: This video builds on empirical probability and defines the classical definition by considering equally probable outcomes. The video discusses several examples using that can be easily studied with the classical definition.

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 determined *a priori* without actual experimentation. It is given by the ratio:

$$\Pr(A) = \frac{N_A}{N} \quad (3.2)$$

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*.

Examples include:

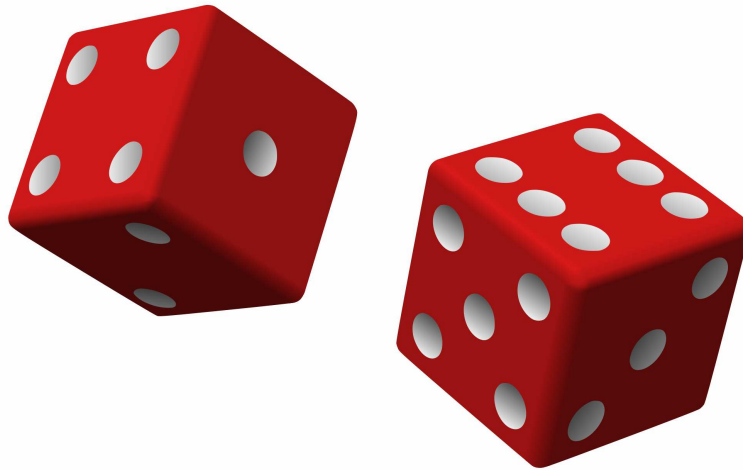


Figure 3.5: Two red dice: https://commons.wikimedia.org/wiki/File:Two_red_dice_01.svg

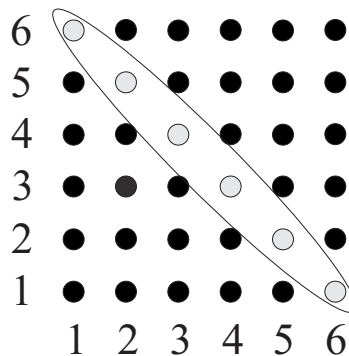


Figure 3.6: Two dice statespace, and (highlighted) the event of rolling a sum of 7.

1. Probability of a specific number being rolled on a six-sided die ($1/6$);
2. Probability of rolling an even number on a six-sided die ($3/6 = 1/2$).

This definition, however, has some difficulties when the number of possible outcomes is infinite, as illustrated in the detailed example in Section 3.3.3.

3.3.1 Using the Classical Definition

The classical definition is reasonably powerful, and is able to deal with many simple problems.

However, there are difficulties with the classical definition in Equation 3.2, as will be seen in Bertrand's Paradox in Section 3.3.3, is determining N and N_A .

It is important to ensure that the different possible outcomes are, in fact, equally probable. In this section, some examples are shown where the incorrect conclusion is obtained through the incorrect determination of an equally probable sample space. Other examples are provided in simple scenarios where the classical example does actually work.



Figure 3.7: Arranging cups and saucers randomly. See Example 3.2.

Example 3.1 (Rolling two dice). Two dice are rolled (see Figure 3.5); 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 $\{2, 3, \dots, 12\}$. 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, $(1, 1), (1, 2), \dots, (1, 6), (2, 1), \dots, (6, 6)$, of which there are three favourable pairs $(3, 4), (5, 2)$ and $(6, 1)$. However, again, 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 count all pairs of numbers distinguishing between the first and second die, as shown in the statespace in Figure 3.6. This will give the correct probability of $6/36 = 1/6$.

Note that many important problems involve counting the number of equally probable events.

Example 3.2 (Cups and Saucers). Six cups and saucers come in pairs: there are two cups and saucers which are red (R), two which are green (G), and two which are yellow (Y). If the cups are placed randomly onto the saucers (one each), find the probability that no cup is upon a saucer of the same colour.

This problem has parallels in **template matching** where, for example, the saucers represent a target sequence of symbols, and the cups represent an input symbol sequence. The problem is to calculate the probability that at random no input symbol is in the correct place compared with the target sequence.

SOLUTION. • Lay the saucers in order, say as $RRGGYY$. The ordering of the saucers is arbitrary in this instance.

- 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 is determined simply by counting, and perhaps by some insightful observation, and are:

$$\begin{aligned}
 &\underline{GGYYRR}, \quad \underline{GYRYGR}, \quad \underline{YGRYGR}, \quad \underline{YYRRGG} \\
 &\underline{GYRYRG}, \quad \underline{YGRYRG} \\
 &\underline{GYYRGR}, \quad \underline{YGYRGR} \\
 &\underline{GYYRRG}, \quad \underline{YGYRGR}
 \end{aligned} \tag{3.3}$$

□

Note that the underlining and bold fonts are to emphasize the ordering more clearly.

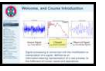
- Hence, the required probability is $10/90 = 1/9$.

Example 3.3 (Sampling). In sequences of k binary digits, 1's and 0's are equally likely. What is the probability of encountering a sequence with a single 1 in any position, and all other digits zero?

– End-of-Topic 10: Classical Definition of Probability and Examples of
How to Use It –



3.3.2 Difficulties with the Classical Definition



Topic Summary 11 Bertrand's Paradox

New slide

Topic Objectives:

- Discuss limitations of the classical definition of probability.
- Show limitations using the infamous Bertrand's Paradox.

Topic Activities:

Type	Details	Duration	Progress
Watch video	13 : 47 minute video	3× video length	
Read Handout	Read page 74 to page 77	8 mins/page	
Self-study	Read further on the paradox	20 mins	
Discussion Board	Share what you have discovered	10 mins	

Discussion: Bertrand's Paradox

Finally, in the random radius method, 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.

Random Midpoint
Random End point
Random Radius

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

Different selection methods.
There are three different reasonable solutions. Which is valid?

http://media.ed.ac.uk/media/0_3jxf1ljjc

Video Summary: This video highlights key difficulties with the classical definition of probability. It uses Bertrand's paradox as a problem in which to study the problems associated with classical probabilities.

The classical definition in Equation 3.2 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 3.2.

3. If the number of possible outcomes is infinite, then some other measure of infinity for determining the classical probability ratio in Equation 3.2 is needed, such as length, or area. This leads to difficulties, such as Bertrand's paradox.

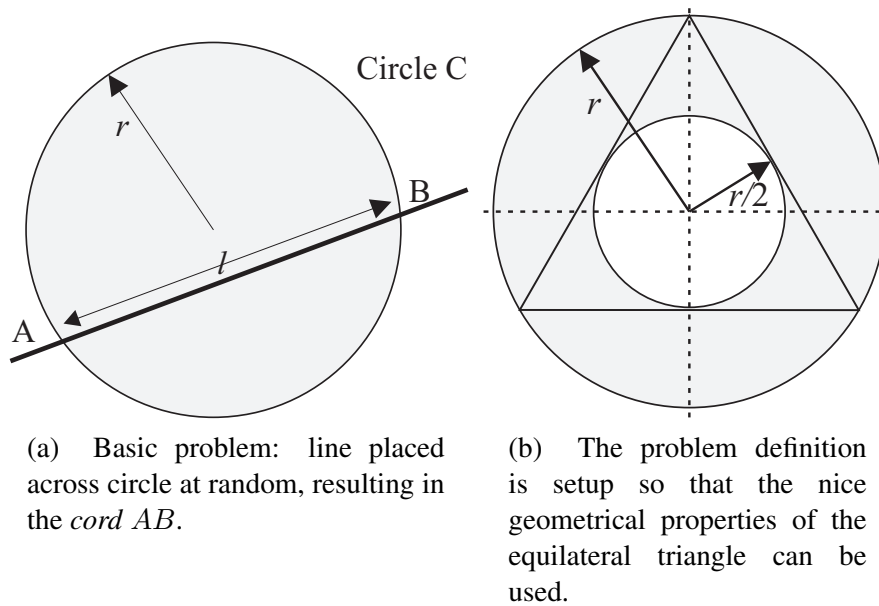


Figure 3.8: Bertrand's paradox, problem definition.

3.3.3 Discussion: Bertrand's Paradox

The Bertrand paradox is a problem within the classical interpretation of probability theory.

Consider a circle C of radius r ; what is the probability p that the length ℓ of a *randomly selected* cord AB ¹ is greater than the length, $r\sqrt{3}$, of the inscribed equilateral triangle? This problem is illustrated in Figure 3.8.

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 so that the calculations are very straightforward, 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 degs, then the resulting equilateral triangle fits inside a larger circle of radius r , as shown in Figure 3.8. The length of the sides of one of this equilateral triangle is $r\sqrt{3}$. The fact the sizes of the inscribed triangle are tangential to the circle of radius $r/2$ is also an important simplifying property that can be used.

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

- In the first method, the **random midpoints** method, a cord is selected by choosing a point M anywhere in the full circle, and two end-points A and B on the circumference of the circle, such that the resulting chord AB through these chosen points has M as its midpoint. There will only be a single cord which satisfies this constraint, and this is shown graphically in Figure 3.9a.

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 . This is because any point M in the inner-circle must have a cord that is at least of length $\sqrt{3}r$.

Therefore, using as a measure of these outcomes the corresponding areas, it follows that:

$$p = \frac{\pi \left(\frac{r}{2}\right)^2}{\pi r^2} = \frac{1}{4} \quad (3.4)$$

¹A cord is a line connecting two points on the circumference of the circle.

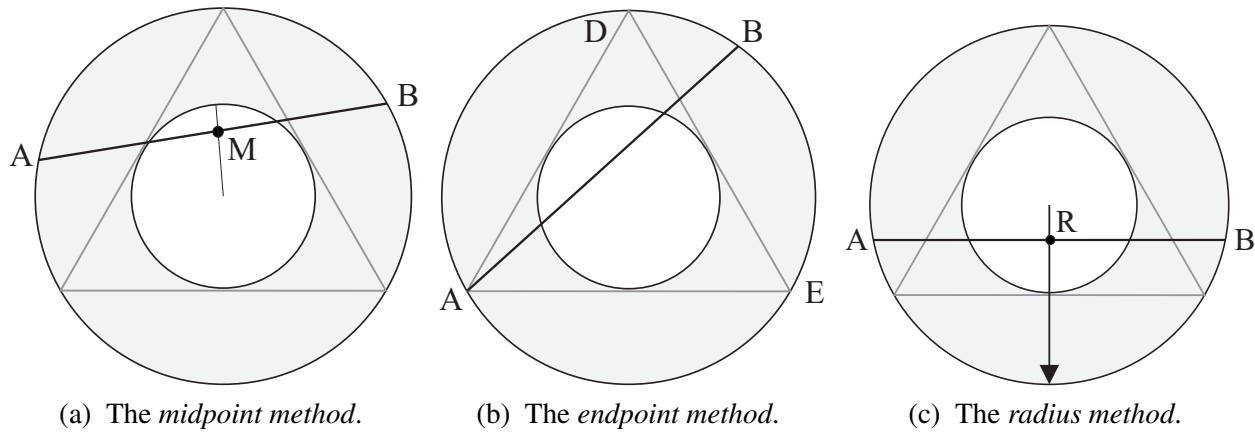
(a) The *midpoint method*.(b) The *endpoint method*.(c) The *radius method*.

Figure 3.9: Different selection methods.

- In the second method, 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 3.9b, 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:

$$p = \frac{\frac{2\pi r}{3}}{2\pi r} = \frac{1}{3} \quad (3.5)$$

- Finally, in the third method, 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 3.9c.

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

$$p = \frac{r}{2r} = \frac{1}{2} \quad (3.6)$$

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

Example 3.4 (Multi-choice: Bertrand's Paradox). Consider a circle of radius r . What is the probability that the length of a *randomly selected* cord is greater than the length, $r\sqrt{3}$, of the inscribed equilateral triangle?

- | | |
|------------------|---------------------------|
| 1. $\frac{1}{4}$ | 3. $\frac{1}{2}$ |
| 2. $\frac{1}{3}$ | 4. Need more information. |

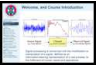
KEYPOINT! (Confused?). The solution to this paradox is indeed quite complicated, and has been discussed in a number of research papers! A discussion will take place in the hybrid classes, but if you are interested in finding out more, you are encouraged to look into this further.

One interesting solution by Jaynes exploits the fact that the position or size of the circle is not specified, and argues that any objective solution must be scale and translation invariant.

– End-of-Topic 11: Awareness of the difficulties with the Classical
Definition of Probability –



3.4 Axiomatic Definition



Topic Summary 12 Axiomatic Definition of Probability

New slide

Topic Objectives:

- Review Kolmogorov's Axioms.
- Derive results from these Axioms.
- Use addition law of probability.
- Examples of using these axioms.

Topic Activities:

Type	Details	Duration	Progress
Watch video	9 : 46 minute video	3 × video length	
Read Handout	Read page 78 to page 80	8 mins/page	
Try Example	Work through Example 3.5	10 minutes	

http://media.ed.ac.uk/media/1_5k714c8b

Video Summary: The Kolmogorov axioms are the foundations of probability theory introduced by Andrey Kolmogorov in 1933. Using these axioms, this video shows how many other familiar results can be derived from these axioms. These results are then applied to several problems which highlights the importance for introducing set theory, that is covered in Topic 13.

The Kolmogorov axioms are the foundations of probability theory introduced by Andrey Kolmogorov in 1933. These axioms remain central and have direct contributions to mathematics, the physical sciences, and real-world probability cases. An alternative approach to formalising probability, favoured by some Bayesians, is given by Cox's theorem.

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:

$$\Pr(A) \geq 0 \quad (3.7)$$

2. Defining the **certain event**, S , as the event that occurs in every trial, then:

$$\Pr(S) = 1 \quad (3.8)$$

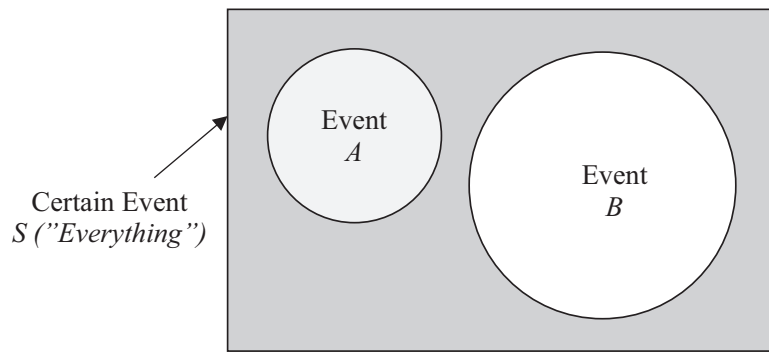


Figure 3.10: A Venn diagram for two mutually exclusive events.

3. If the events A and B are **mutually exclusive**, then:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) \quad (3.9)$$

This result is apparent from the Venn diagram shown in Figure 3.10. 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:

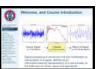
$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i) \quad (3.10)$$

Note that Equation 3.10 does not directly follow from Equation 3.9, even though it may appear to. Dealing with infinitely many sets requires further insight, and here the result of Equation 3.10 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.

3.4.1 Properties of Axiomatic Probability



Some simple consequences of the definition of probability defined in Section 3.4 follow immediately: *New slide*

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

$$\Pr(\emptyset) = 0 \quad (3.11)$$

Complements Since $A \cup \bar{A} = S$ and $A\bar{A} = \{\emptyset\}$, then :

$$\Pr(\bar{A}) = 1 - \Pr(A) \quad (3.12)$$

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

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) \quad (3.13)$$

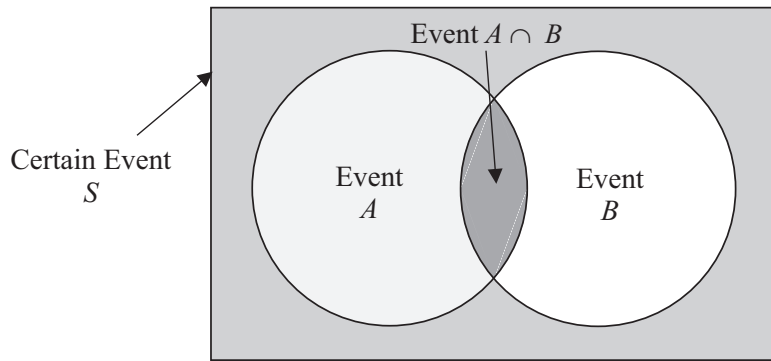


Figure 3.11: Venn diagram to prove the addition law of probability.

Example 3.5 (Sum Rule). Let A and B be events with probabilities $\Pr(A) = 3/4$ and $\Pr(B) = 1/3$. Show that $1/12 \leq \Pr(A \cap B) \leq 1/3$.

SOLUTION. Using the sum rule, that:

$$\Pr(A \cap B) = \Pr(A) + \Pr(B) - \Pr(A \cup B) \geq \Pr(A) + \Pr(B) - 1 = \frac{1}{12} \quad (3.14)$$

□

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

– End-of-Topic 12: **Properties of axiomatic probability theory, and an interesting example** –



3.4.2 Set Theory

Topic Summary 13 Set theory and its use in Probability Theory

Topic Objectives:

- Basic Definitions in Set Theory.
- Venn diagrams and set manipulations.
- Proof of the Sum Rule.

Topic Activities:

Type	Details	Duration	Progress
Watch video	16.28 min video	3× video length	
Study Handout	Read page 81 to page 84	8 mins/page	
Tutorial Exercise	Exercise ??	20 minutes	

The screenshot shows a video player interface for a lecture on Set Theory. The title is 'Set Theory' and the subtitle is 'De Morgan's Law Using Venn diagrams, it can be shown'. The main content displays the following mathematical expressions:

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

As an application of this, note that:

$$\begin{aligned} \overline{A \cup B \cap C} &= \overline{A} \overline{B \cap C} = \overline{A} (\overline{B} \cup \overline{C}) \\ &= (\overline{A} \overline{B}) \cup (\overline{A} \overline{C}) = \overline{A} \overline{B} \cup \overline{A} \overline{C} \\ &\Rightarrow \overline{A \cup B \cap C} = (\overline{A} \overline{B}) \cup (\overline{A} \overline{C}) \end{aligned}$$

A Venn diagram below shows three overlapping circles labeled A, B, and C. The region where A and B overlap but not C is shaded, and labeled 'Event A ∩ B \setminus C'. A red arrow points to this shaded region.

http://media.ed.ac.uk/media/1_v1wzihow

Video Summary: This video gives the background to set theory which is fundamental for dealing with probability more generally. The video discusses using Venn diagrams as a simple way of proving a number of results, such as De Morgan's law. However, we also discuss how to prove this formally using set theory results. An example is using various forms of De Morgan's law to derive the sum rule, or the addition law of probability. A tutorial exercise challenges you to derive the sum rule for three events.

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, pencil*” 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 are commutative, associative, and distributive, such that:

$$A \cup B = B \cup A, \quad (A \cup B) \cup C = A \cup (B \cup C) \quad (3.15)$$

$$AB = BA, \quad (AB)C = A(BC), \quad A(B \cup C) = AB \cup AC \quad (3.16)$$

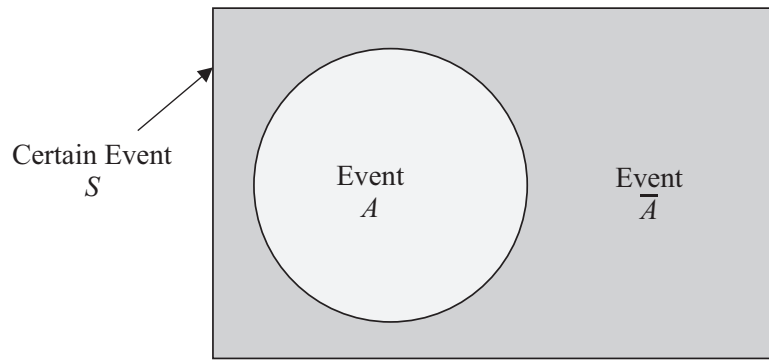


Figure 3.12: The complement \bar{A} of $A \subset S$ is the set of all elements of S not in A .

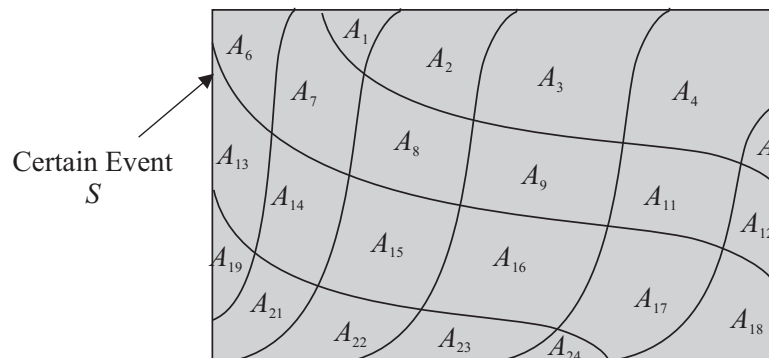


Figure 3.13: A partition of the certain event using mutually exclusive subsets A_i , whose union equates to S .

Complements The complement \bar{A} of a set $A \subset S$ is the set consisting of all elements of S not in A :

$$A \cup \bar{A} = S \quad \text{and} \quad A \cap \bar{A} \equiv A\bar{A} = \{\emptyset\} \quad (3.17)$$

This is shown graphically using a Venn diagram, as shown in Figure 3.12.

Partitions A partition U of a set S is a collection of mutually exclusive subsets A_i of S whose union equates to S , as shown in Figure 3.13, such that:

$$\bigcup_{i=1}^{\infty} A_i = S, \quad A_i \cap A_j = \{\emptyset\}, \quad i \neq j \quad \Rightarrow \quad U = [A_1, \dots, A_n] \quad (3.18)$$

De Morgan's Law Using Venn diagrams, it is relatively straightforward to show as in Figure 3.15 that:

$$\overline{A \cup B} = \bar{A} \cap \bar{B} \equiv \bar{A}\bar{B} \quad \text{and} \quad \overline{A \cap B} \equiv \overline{AB} = \bar{A} \cup \bar{B} \quad (3.19)$$

As an application of this, note that:

$$\overline{A \cup BC} = \bar{A}\bar{BC} = \bar{A}(\bar{B} \cup \bar{C}) \quad (3.20)$$

$$= (\bar{A}\bar{B}) \cup (\bar{A}\bar{C}) = \overline{A \cup B} \cup \overline{A \cup C} \quad (3.21)$$

$$\Rightarrow \quad A \cup BC = (A \cup B)(A \cup C) \quad (3.22)$$

This result can easily be derived by using Venn diagrams, as shown in Figure 3.15, and it is worth checking this result yourself. This latter identity will also be used later in Section 3.4.1.

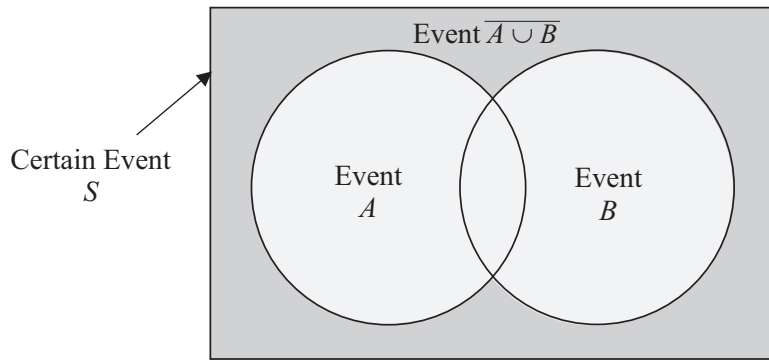


Figure 3.14: The event $\overline{A \cup B}$.

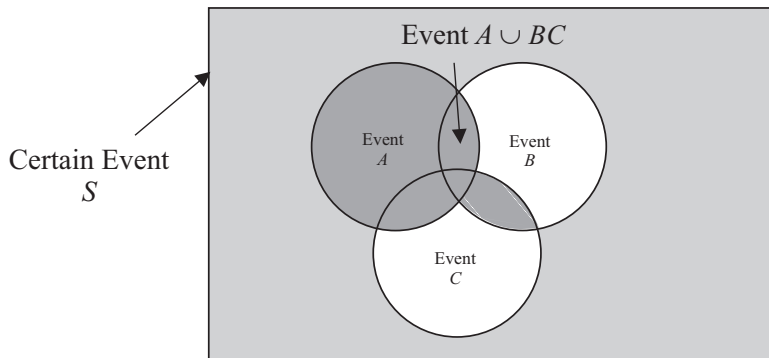


Figure 3.15: The event $\overline{A \cup BC}$.

Example 3.6 (Proof of the Sum Rule). Prove the result in Equation 3.13 regarding the addition law of probability (or sum rule), namely:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) \tag{3.23}$$

SOLUTION. To prove this, separately write *each of* $A \cup B$ and B as the union of two mutually exclusive events (using Equation 3.22 and the fact $A \cup \overline{A} = S$ and $S B = B$).

- First, to write $A \cup B$ in this way, use S :

$$A \cup B = S(A \cup B) = (A \cup \overline{A})(A \cup B) = A \cup (\overline{A}B) \tag{3.24}$$

Since the intersection $A \cap (\overline{A}B) = (A\overline{A})B = \{\emptyset\}B = \{\emptyset\}$, then A and $\overline{A}B$ are mutually exclusive events, as required.

- Second, and using a similar approach, note that:

$$B = SB = (A \cup \overline{A})B = (AB) \cup (\overline{A}B) \tag{3.25}$$

Since the intersection $(AB) \cap (\overline{A}B) = A\overline{A}B = \{\emptyset\}B = \{\emptyset\}$ and are therefore mutually exclusive events.

Using these two disjoint unions, then:

$$\Pr(A \cup B) = \Pr(A \cup (\overline{A}B)) = \Pr(A) + \Pr(\overline{A}B) \tag{3.26}$$

$$\Pr(B) = \Pr((AB) \cup (\overline{A}B)) = \Pr(AB) + \Pr(\overline{A}B) \tag{3.27}$$

Eliminating $\Pr(\bar{A}B)$ by subtracting these equations gives the desired result:

$$\Pr(A \cup B) - \Pr(B) = \Pr(A \cup (\bar{A}B)) = \Pr(A) - \Pr(\bar{A}B) \quad (3.28)$$

□

– End-of-Topic 13: **Set theory and its used in probability theory.** –



3.4.3 Countable Spaces and Principle of Total Probability

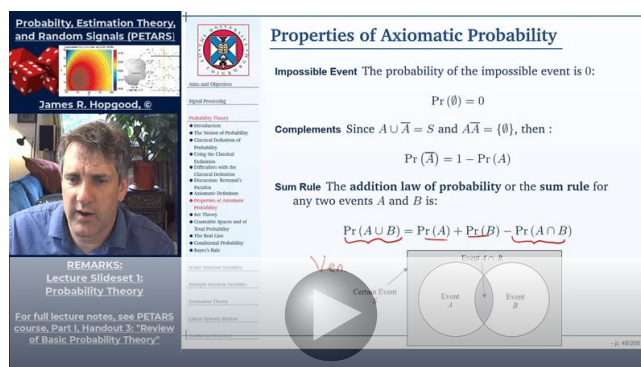
Topic Summary 14 Total Probability

Topic Objectives:

- Introduce uncertainty through a simple example.

Topic Activities:

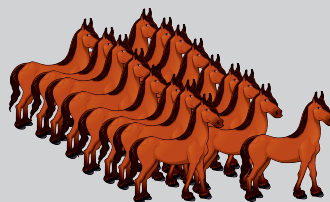
Type	Details	Duration	Progress
Watch video	9 : 46 min video	3 × video length	
Read Handout	Read page 85 to page 88	8 mins/page	
Try Example	Work through Examples 3.7 and 3.8	20 minutes	



http://media.ed.ac.uk/media/1_5k714c8b

Video Summary:

Example 3.7 (Farmer and his Will). A farmer leaves a will saying that they wish for their first child to get half of his property, the second child to get a third, and the third child to get a ninth. As seventeen horses have been left, the children are distressed because they don't want to cut any horses up.



However, a local statistician lends them a horse so that they have eighteen. The children then take nine, six, and two horses, respectively. This adds up to seventeen, so they give the statistician the horse back, and everyone is happy. What is wrong with this story?

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 $\{\zeta_i\}$.

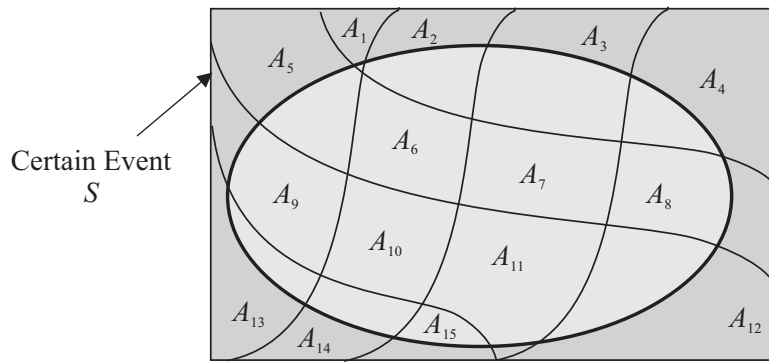


Figure 3.16: A Venn diagram clearly illustrates the principle of total probability.

From the basic axioms, it follows that $p_i \geq 0$ and that

$$\sum_{i=1}^N p_i = 1 \quad (3.29)$$

This can be used in obtaining the **principle of total probability**. Let A_1, A_2, A_3, \dots be a finite or countably infinite set of mutually exclusive and collectively exhaustive events, then from the Venn diagram in Figure 3.16,

$$\sum_i \Pr(A_i \cap B) = \Pr(B) \quad (3.30)$$

Example 3.8 (Detection and Classification). An acoustic scene analysis algorithm is monitoring an Edinburgh City park for animal sounds, and makes a large number of sound classifications on detected acoustic events, either being labelled as bird, fox, or pet sounds. Each labelled acoustic event is either a true detection of the corresponding animal sound, or is a false alarms. The false alarms can be considered as bad detections. Based on previous statistical analysis, it has been determined that in one (long) recording:

- 29% of the detected sounds are false alarms;
- 3% of labelled bird sounds are false alarm detections;
- 12% of detected bird sounds are correctly labelled;
- 5% of labelled fox sounds are false alarm detections;
- 32% are correct detections of domestic pet sounds.

The following events are defined: correctly classified – C ; mis-classified or false alarms – M ; bird sound – B ; fox sound – F ; domestic pet sound – D .

Draw a Venn diagram of the problem, and determine the following:

1. What is the probability that a detection is classified as a bird sound, either correctly or incorrectly?
2. What is the probability that a detection is a false alarm and/or a labelled bird sound?
3. What is the probability that a sound is correctly classified as a fox or domestic pet sound?

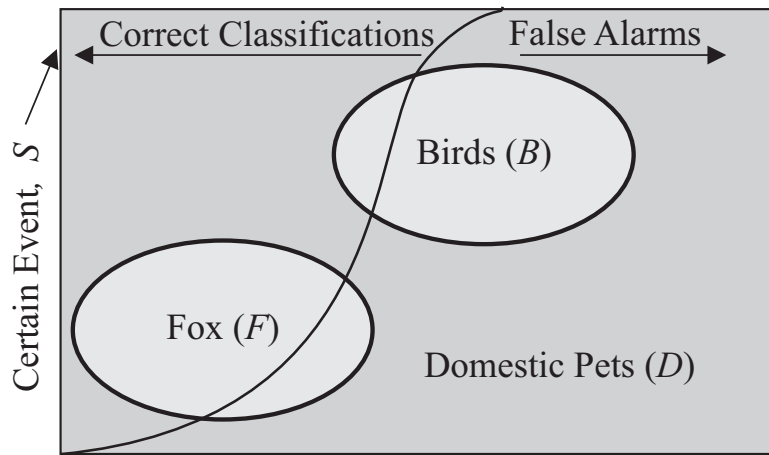


Figure 3.17: The Venn diagram for this problem, although the size of the events are not to scale.

4. What is the probability of a false alarm for a domestic pet sound?

The Venn diagram for this problem is sketched in Figure 3.17, where the three types of classification are shown for birds (B), foxes (F), and domestic pets (D). The cases where events are correctly classified (C) or mis-classified (M) are also indicated.

Writing out the known probabilities in terms of the events, we have:

Table 3.1: Known events and probabilities

Event	Notation	Probability
Detections are false alarms	M	0.29
Birds are mis-classifications	$B \cap M$	0.03
Birds are correctly classified	$B \cap C$	0.12
Foxes are mis-classifications	$F \cap M$	0.05
Pets are correctly classified	$D \cap C$	0.32

1. The probability that a detection is classified as a bird sound, either correctly or incorrectly, can be expressed by using total probability:

$$\Pr(B) = \Pr(B \cap C) + \Pr(B \cap M) = 0.12 + 0.03 = 0.15 \quad (3.31)$$

2. The probability that a detection is a false alarm and/or a labelled bird sound is obtained using the probability sum rule:

$$\Pr(B \cup M) = \Pr(B) + \Pr(M) - \Pr(B \cap M) = 0.15 + 0.29 - 0.03 = 0.41 \quad (3.32)$$

3. Considering the left hand side of the Venn diagram in Figure 3.17, the probability that a sound is correctly classified as a fox or domestic pet sound can be written as the complement of the event of being a false alarm or a bird. This is most easily seen from the Venn diagram in Figure 3.18.

Therefore:

$$\Pr((F \cap C) \cup (D \cup M)) = 1 - \Pr(F \cup B) = 1 - 0.41 = 0.59 \quad (3.33)$$

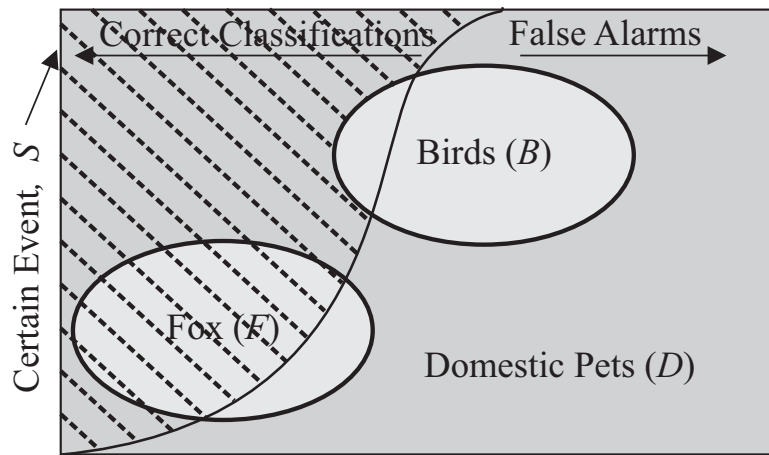


Figure 3.18: The Venn diagram with the event $1 - \Pr(B \cup M)$ highlighted.

4. Finally, the probability of a false alarm for a domestic pet, $\Pr(D \cap M)$, can be obtained from the Venn diagram and total probability:

$$\Pr(M) = \Pr(D \cap M) + \Pr(F \cap M) + \Pr(B \cap M) \quad (3.34)$$

$$0.29 = \Pr(D \cap M) + 0.05 + 0.03 \quad \Rightarrow \quad \Pr(D \cap M) = 0.21 \quad (3.35)$$

3.4.4 The Real Line

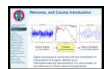
If the **certain event**, S , consists of a non-countable infinity of elements, then its probabilities cannot be determined in terms of the probabilities of elementary events. This is the case if S is the set of points in an n -dimensional space.

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, it suffices to assign probabilities to the events $\{x \leq x_i\}$.

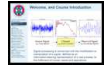
This notion leads to **cumulative distribution functions (cdfs)** and **probability density functions (pdfs)** in the next handout.

– End-of-Topic 14: **Countable Spaces, Total Probabilities, and Uncountable Spaces on the Real line** –



New slide

3.5 Conditional Probability



New slide

Topic Summary 15 Conditional Probability and Bayes Rule

Topic Objectives:

- Introduce conditional probability.
- Examples of applying conditional probability.
- Developing Bayes's Theorem.
- Bayes's Theorem and Inverse Problems.
- Prisoner's Problem and Monte Hall.
- Practical application of Bayes Theorem.

Topic Activities:

Type	Details	Duration	Progress
Watch video	21 : 59 minute video	3× video length	
Read Handout	Read page 89 to page 94	8 mins/page	
Try Example	Try Examples 3.9 and 3.10	20 minutes	
Practice Exercises	Exercises ?? and ??	30 mins	

Bayes's Rule

After this lecture, try the following example in the notes:

Example (Classification Accuracy). An algorithm using electrocardiogram (ECG) data is used to test for a certain irregular heartbeat and is 95% accurate. A person submits to the test and the results are positive. Suppose the person comes from a population of 10%, where 2000 people suffer the irregularity.

http://media.ed.ac.uk/media/1_7zsoflwm

Video Summary: This slightly longer than usual video covers conditional probability and gives some examples that are initially counter-intuitive. Bayes theorem is then developed from conditional probability, and the role of inverse problems in the context of Bayes theorem is discussed. Bayes theorem is then applied to a puzzle-type problem to demonstrate the counter-intuitive nature of probability. An example is then presented for you to consider, which will be answered in the handout.

To introduce conditional probability, consider the discussion about proportions in Section 3.1. If an experiment is repeated n times, and the occurrences or non-occurrences two events A and B are observed. Suppose that only those outcomes for which B occurs are considered.

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

$$\Pr(A|B) \approx \frac{n_{AB}}{n_B} = \frac{n_{AB}/n}{n_B/n} = \frac{\Pr(AB)}{\Pr(B)} \quad (3.36)$$

provided that n is sufficiently large.

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

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \quad (3.37)$$

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

Example 3.9 (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.

A simple method for solving this problem is to list all the possibilities:

C_1	C_2	Outcome	
Gender	Gender	Relevant?	Desired?
B	B	✓	✓
G	B	✓	
B	G	✓	
G	G		
Count		3	1

Therefore, using classical probability, since the events are all equally probable, the answer is $p = N_A/N = 1/3$.

A more formal solution is to consider the set of four possibilities for the gender of the children, namely:

$$S = \{GG, GB, BG, BB\} \quad (3.38)$$

where the four possibilities are equally probable:

$$\Pr(GG) = \Pr(GB) = \Pr(BG) = \Pr(BB) = \frac{1}{4} \quad (3.39)$$

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:

$$\Pr(BB|S_B) = \frac{\Pr(BB \cap (GB \cup BG \cup BB))}{\Pr(S_B)} \quad (3.40)$$

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

$$\Pr(BB|S_B) = \frac{\Pr(BB)}{1 - \Pr(GG)} = \frac{1/4}{3/4} = \frac{1}{3} \quad (3.41)$$

□

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.

Example 3.10 (Two Children (Variant)). A family has two children. One of the children is a boy born in an *even* month, where even months are defined as *February, April, June, August, October, and December*, while odd months are defined as *January, March, May, July, September, and November*. What is the probability that both are boys?

SOLUTION. The younger and older children may each be male or female, and it is assumed that each is equally likely. Moreover, the month in which each child is born is assumed to be equally likely. Denoting the first child as C_1 , and the second by C_2 , there are 16 different but equally likely possibilities, which are denoted given by:

C_1		C_2		Outcome	
Gender	Month	Gender	Month	Relevant?	Desired?
B	O	B	O		
B	O	B	E	✓	✓
B	E	B	O	✓	✓
B	E	B	E	✓	✓
G	O	B	O		
G	O	B	E	✓	
G	E	B	O		
G	E	B	E	✓	
B	O	G	O		
B	O	G	E		
B	E	G	O	✓	
B	E	G	E	✓	
G	O	G	O		
G	O	G	E		
G	E	G	O		
G	E	G	E		
Count				7	3

□

Therefore, the number of favourable outcomes to the question in hand is $3/7 = 0.428$, which is getting closer to one half than a third.

The example in Unknown `exmp.twoChildrew` 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 single package, you get two good CDs?

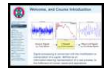
It should be apparent that this is the same problem as in Unknown `exmp.twoChildrew`. One further problem to consider is given below in Example 3.11.

A further example discussed in the lectures covers mobile phones; a company sells mobile phones in boxes, and are equally likely to be broken (B) or working (W). You are given two boxes and told that in one of the boxes there is a working phone. What is the probability that the other box also contains a working phone? Suppose now that all phones are manufactured by four companies: A , E , N , and

S . You are told that one of the boxes contains a working phone manufactured by company S . What is the probability that the other box contains a working phone?

Finally, to extend the discussion further, suppose all the phones are made between the years 1997 and 2016, and by the four companies above. One of the boxes contains a working phone made in 2007 by manufacturer A . What is the probability the other box contains a working phone? It should be apparent that by giving more information about one of the phones, the probability of the other box containing a working phone approaches a half.

3.6 Bayes's Rule



New slide

Conditional probability leads onto Bayes's theorem. Returning to Equation 3.37, then writing $\Pr(A \cap B) \equiv \Pr(AB)$ as follows:

$$\Pr(AB) = \Pr(A | B) \Pr(B) = \Pr(B | A) \Pr(A) \quad (3.42)$$

giving

$$\Pr(B | A) = \frac{\Pr(A | B) \Pr(B)}{\Pr(A)} \quad (3.43)$$

Bayes's rule will be used throughout this course, and commonly arises in the analysis of signal and communication systems, machine learning, and data science. Bayesian inference is typically a computationally expensive problem, but can be solved efficiently using graphical models, sparsity, and numerical Bayesian methods such as Monte Carlo and Message Passing techniques.

Example 3.11 (Prisoner's Problem). Three prisoners, A , B and C , are in separate cells and sentenced to remain there for a long time. The governor has selected one of them at random to be pardoned and therefore released. The warden knows which one is to be released, but is not allowed to say. Prisoner A begs the warden to be told the identity of one of the *others* who **will not** be released.

Prisoner A says: *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 will not be released.

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

SOLUTION. This problem is mathematically equivalent to the Monty Hall problem with the main prize and replaced with freedom. It can be solved using the principle of total probability and Bayes's theorem.

- Let A , B , and C be the events that the corresponding prisoner will be pardoned.
- Note that A , B , and C are independent events, *before* the warden has provided any information.
- Let b be the event that the warden tells A that prisoner B is **not** to be released.

Using Bayes's theorem, it follows that:

$$\Pr(A | b) = \frac{\Pr(b | A) \Pr(A)}{\Pr(b)} \quad (3.44)$$

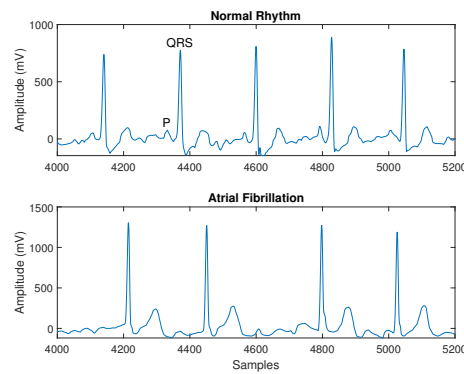


Figure 3.19: Regular (or normal) and irregular heartbeat rhythms.

Using the principal of total probability:

$$\Pr(b) = \sum_{i \in \{A, B, C\}} \Pr(b, i) \quad (3.45)$$

$$= \Pr(b, A) + \Pr(b, B) + \Pr(b, C) \quad (3.46)$$

$$= \Pr(b | A) \Pr(A) + \Pr(b | B) \Pr(B) + \Pr(b | C) \Pr(C) \quad (3.47)$$

$$= \frac{1}{2} \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = \frac{1}{2} \quad (3.48)$$

The crucial point here is that if A is actually to be released, the warden can tell A that either B or C will not be released through the toss of the coin, and therefore $\Pr(b | A) = \frac{1}{2}$. Whereas, if C is to be released, then the warden is now constrained to tell A that B will not be released, so $\Pr(b | C) = 1$.

Finally, returning to Bayes rule,

$$\Pr(A | b) = \frac{\Pr(b | A) \Pr(A)}{\Pr(b)} = \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3} \quad (3.49)$$

However, the same calculation for C is different in the numerator:

$$\Pr(C | b) = \frac{\Pr(b | C) \Pr(C)}{\Pr(b)} = \frac{1 \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3} \quad (3.50)$$

KEYPOINT! (Why the paradox). The tendency of people to provide the answer $\frac{1}{2}$ neglects to take into account that the warden may have tossed a coin before giving an answer. The warden may have answered B because either:

- A is to be released and the warden tossed a coin;
- or C is to be released.

The probabilities of these two events are not equal.

After this lecture, try the following example in the notes:

Example 3.12 (Classification Accuracy). A statistical signal processing and machine learning algorithm using electrocardiogram (ECG) data is used to test for a certain irregular heartbeat and is 95% accurate. A person submits to the test and the results are positive. Suppose that the person comes from a population of 10^5 , where 2000 people suffer from the irregularity.

What can we conclude about the probability that the person under test has that particular heartbeat irregularity?²

SOLUTION. The test is known to be 95% accurate, which means that 95% of all positive tests are correct, and 95% of all negative tests are correct. Let the events T_+ and T_- denote the test being positive and negative respectively. Let the events R and I denote a regular and irregular heartbeat in a patient. Hence, it is known:

$$\Pr(T_+ | I) = 0.95, \quad \Pr(T_+ | R) = 0.05 \quad (3.51)$$

$$\Pr(T_- | I) = 0.05, \quad \Pr(T_- | R) = 0.95 \quad (3.52)$$

The population space gives an empirical probability that a regular heartbeat occurs with probability $\Pr(R) = \frac{98,000}{100,000} = 0.98$ and $\Pr(I) = 0.02$. Hence, using total probability and Bayes's theorem, it follows that:

$$\Pr(I | T_+) = \frac{\Pr(T_+ | I) \Pr(I)}{\Pr(T_+)} \quad (3.53)$$

$$= \frac{\Pr(T_+ | I) \Pr(I)}{\Pr(T_+ | I) \Pr(I) + \Pr(T_+ | R) \Pr(R)} \quad (3.54)$$

$$= \frac{0.95 \times 0.02}{0.95 \times 0.02 + 0.05 \times 0.98} = 0.278 \quad (3.55)$$

The results states that if the test is taken by someone from this population *without knowing* whether that person has the irregular heartbeat or not, then even a positive test would only suggest there is a 27.8% chance of having an irregularity. However, if the person knows that they have the irregularity, then the test is 95% accurate.

KEYPOINT! (Influence of the prior). The resulting accuracy is due to a Bayesian update involving the prior on the population space, so $\Pr(R)$ and $\Pr(I)$. However, one key question is how are these probabilities known?

The question assumed that for a given population, the percentage of the population who suffer from this irregularity is known. But how is this known in practice if we don't have a reliable test? Can it be deduced in other ways? This is one of the key questions that influences the Bayesian posterior inference.

– End-of-Topic 15: **Conditional Probability, and a basic but important Introduction to Bayes Rule** –



²As an example of such an algorithm, see Figure 3.19, as described in: <https://uk.mathworks.com/help/signal/examples/classify-ecg-signals-using-long-short-term-memory-networks.html>

4

Scalar Random Variables

Every line is the perfect length if you don't measure it.

Marty Rubin

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.

4.1 Abstract

Topic Summary 16 Introduction to Random Variables and Cummulative Distribution Functions

Topic Objectives:

- Notion of a random variable.
- Formal definition involving experimental outcomes, sample space, probability of events, and assigned values.
- the concept of the cumulative distribution function (cdf).

Topic Activities:

Type	Details	Duration	Progress
Watch video	16 : 12 min video	3 × video length	
Read Handout	Read page 96 to page 100	8 mins/page	

Definition

Physical Experiment

Abstract sample space, S

real number line

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

Note that for continuous random variables, the outcomes are events, such as small intervals on the real axis as described in the previous lecture.

http://media.ed.ac.uk/media/1_6m2jkb8

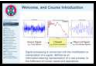
Video Summary: This video introduces and defines scalar real random variables, covering the sample/state space, probability of outcomes, and mapping to the real axis. Some simple examples are presented. The video then motivates the probability set function by considering the axiomatic interval of the random variable taking on a value less than or equal to a specific value. It also demonstrates using the Kolmogorov’s axioms and set theory, it is possible to determine the probability of being within an interval. In the limit, it is demonstrated that the gradient of the cumulative distribution function is important, which leads to the probability density function. This video sets the foundations for the rest of this Chapter and indeed course.

Ultimately, the purpose of this course is to move from probability theory through to random signals. Therefore, before introducing random variables, lets take a step back and consider the bigger picture.

- 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. 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.
- The aim of statistical signal processing is to develop the properties of stochastic processes, both in terms of an exact probabilistic description, but also characteristic features such as mean, variance, and other moments. This course begins by looking at the simplest description of random scalars, or random variables, on which the rest of statistical signal processing is developed.

4.2 Definition Random Variables



A **random variable (RV)** $X(\zeta)$ is a mapping that assigns a real number $X \in (-\infty, \infty)$ to every outcome, or elementary event, 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 $X(\zeta)$ 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, or a temperature measurement.

A more graphical representation of a discrete RV is shown in Figure 4.1. 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 , then has a probability $\Pr(\zeta_k)$ assigned to it. Additionally, each outcome ζ_k also has a real number assigned to that outcome, x_k . The RV is then defined as the collection of these three values; an outcome event, 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 4.2 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 4.1 (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:

$$X(\zeta_1) = X(\zeta_3) = X(\zeta_5) = 0 \quad \text{and} \quad X(\zeta_2) = X(\zeta_4) = X(\zeta_6) = 1 \quad (4.1)$$

✕

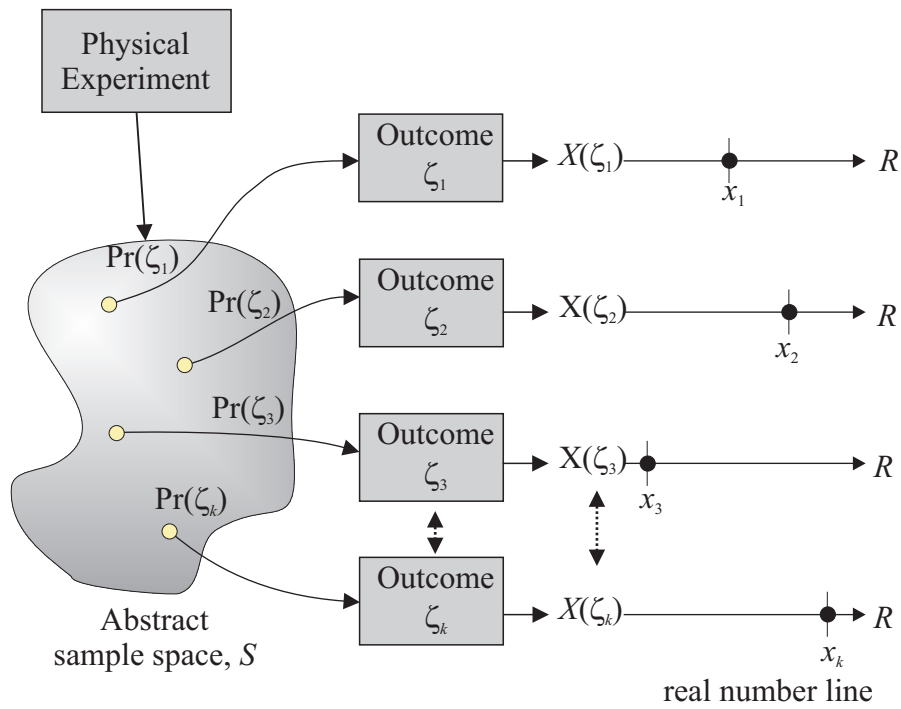


Figure 4.1: A graphical representation of a random variable.

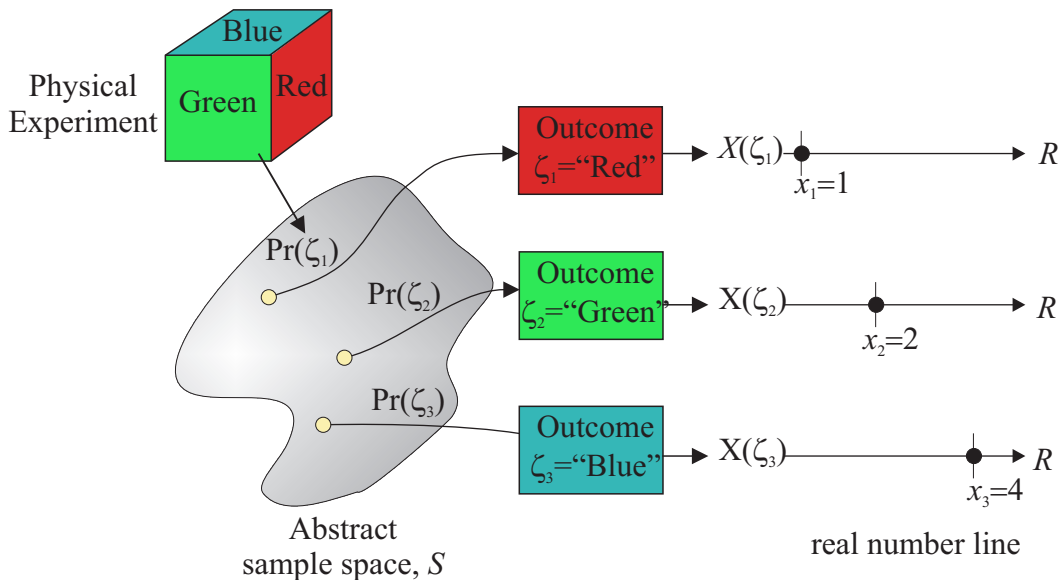


Figure 4.2: A graphical representation of a random variable for a more specific example. Note that for continuous random variables, the outcomes are **events**, such as small intervals on the real axis as described in the previous lecture handout.

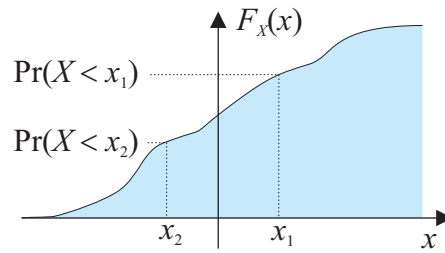


Figure 4.3: The cumulative distribution function.

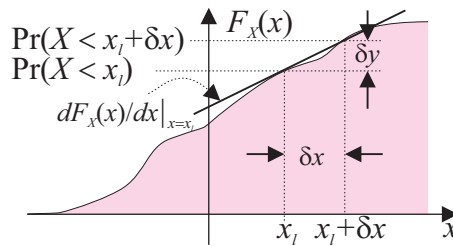
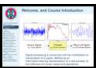


Figure 4.4: The gradient of the cdf is very important, and leads to the probability density function (pdf).

Example 4.2 (Letters of the alphabet). Suppose the outcome of an experiment is a letter A to Z, such that $X(A) = 1, X(B) = 2, \dots, X(Z) = 26$. Then the event $X(\zeta) \leq 5$ corresponds to the letters A, B, C, D, or E.

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

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

It is graphically shown in Figure 4.3.

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

$$\Pr(x_\ell < X(\zeta) \leq x_r) = \Pr(X(\zeta) \leq x_r) - \Pr(X(\zeta) \leq x_\ell) \tag{4.2}$$

$$= F_X(x_r) - F_X(x_\ell) \tag{4.3}$$

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

This can be seen by setting $x_r = x_l + \delta x$:

$$\Pr(x_\ell < X(\zeta) \leq x_\ell + \delta x) = \Pr(X(\zeta) \leq x_\ell + \delta x) - \Pr(X(\zeta) \leq x_\ell) \quad (4.4)$$

$$\approx \Pr(X(\zeta) \leq x_\ell) + \left. \frac{dF_X(x)}{dx} \right|_{x=x_\ell} \delta x - \Pr(X(\zeta) \leq x_\ell) \quad (4.5)$$

$$\approx \left. \frac{dF_X(x)}{dx} \right|_{x=x_\ell} \delta x \quad (4.6)$$

Shortly, it will be seen that $\frac{dF_X(x)}{dx}$ is indeed the pdf.

4.2.2 Kolmogorov's Axioms

The events $\{X(\zeta) \leq x_1\}$ and $\{x_1 < X(\zeta) \leq x_2\}$ are mutually exclusive events. Therefore, their union equals $\{X(\zeta) \leq x_2\}$, and thus:

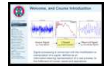
$$\Pr(X(\zeta) \leq x_1) + \Pr(x_1 < X(\zeta) \leq x_2) = \Pr(X(\zeta) \leq x_2) \quad (4.7)$$

$$\int_{-\infty}^{x_1} p(v) dv + \Pr(x_1 < X(\zeta) \leq x_2) = \int_{-\infty}^{x_2} p(v) dv \quad (4.8)$$

$$\Rightarrow \Pr(x_1 < X(\zeta) \leq x_2) = \int_{x_1}^{x_2} p(v) dv \quad (4.9)$$

where $p(v)$ is an probability density function (pdf) that will be described in more detail in the next section.

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



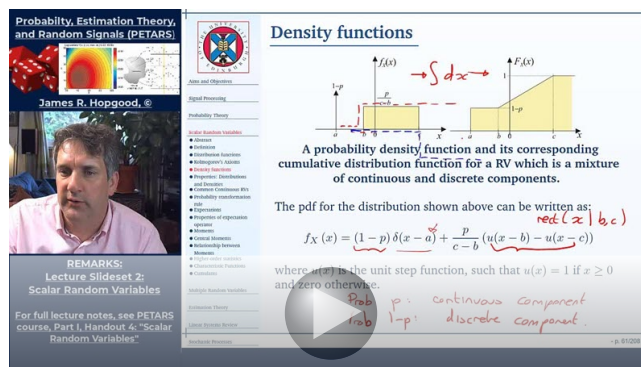
4.3 Density functions

Topic Summary 17 Introduction to probability density functions (pdfs) and their properties

- Topic Objectives:**
- The probability density function (pdf).
 - Formal properties of probability density functions (pdfs).
 - Discrete random variables (RVs), their probability mass function (pmf) the corresponding pdfs and cdfs, as well as mixtures of continuous and discrete random variables.
 - Examples of mixed density functions.

Topic Activities:

Type	Details	Duration	Progress
Watch video	14 : 19 minute video	3 × video length	
Read Handout	Read page 101 to page 104	8 mins/page	
Practice Exercises	Exercises ?? to ??	30 mins	



http://media.ed.ac.uk/media/1_1egxxc2x

Video Summary: This video discusses the probability density function (pdf) and how it is used, including how to deal with mixed discrete and continuous random variables. The key properties of the pdf are then defined, and the viewer should then undertake the exercises associated with this topic.

It was seen in the previous section that gradients of the cdf are important when determining the probability of being within small intervals.

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

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

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

$$f_X(x) \Delta x \approx \Delta F_X(x) \triangleq F_X(x + \Delta x) - F_X(x) \approx \Pr(x < X(\zeta) \leq x + \Delta x) \tag{4.10}$$

Sidebar 3 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:

$$\Pr(X(\zeta) = x_0) = \lim_{\Delta x_0 \rightarrow 0} \Pr(x_0 - \Delta x_0 \leq X(\zeta) \leq x_0 + \Delta x_0) \quad (4.13)$$

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

$$\Pr(X(\zeta) = x_0) = \lim_{\Delta x_0 \rightarrow 0} \int_{x_0 - \Delta x_0}^{x_0 + \Delta x_0} f_X(u) du \quad (4.14)$$

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:

$$f_X(x) = p_0 \delta(x - x_0) \quad (4.15)$$

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

$$\int_{\mathcal{R}} \phi(t) \delta(t - T) dt = \begin{cases} \phi(T) & \text{if } T \in \mathcal{R} \\ 0 & \text{otherwise} \end{cases}, \quad (4.16)$$

then it becomes clear that

$$\Pr(X(\zeta) = x_0) = \lim_{\Delta x_0 \rightarrow 0} \int_{x_0 - \Delta x_0}^{x_0 + \Delta x_0} p_0 \delta(x - x_0) du = p_0 \quad (4.17)$$

whereas for the continuous time case, the limit in Equation 4.14 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 continuous random variables is shown in Figure 4.6. Note the step function in the cumulative distribution function (cdf).

This can be written, more formally, as:

$$f_X(x) = \lim_{\Delta x \rightarrow 0} \frac{F_X(x + \Delta x) - F_X(x)}{\Delta x} \quad (4.11)$$

$$= \lim_{\Delta x \rightarrow 0} \frac{\Pr(x < X(\zeta) \leq x + \Delta x)}{\Delta x} \quad (4.12)$$

- It directly follows that:

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

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

The pmf for a discrete RVs can be written as a pdf through:

$$f_X(x) = \sum_k p_k \delta(x - x_k) \quad (4.18)$$

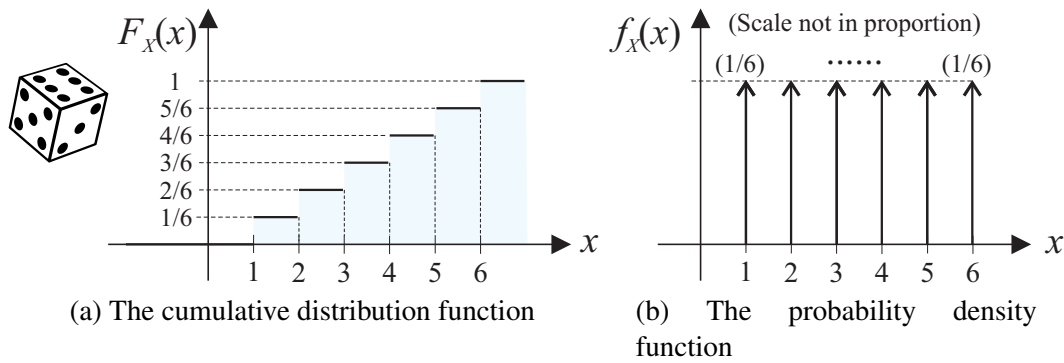


Figure 4.5: The cdf and pdf for a fair six-sided die.

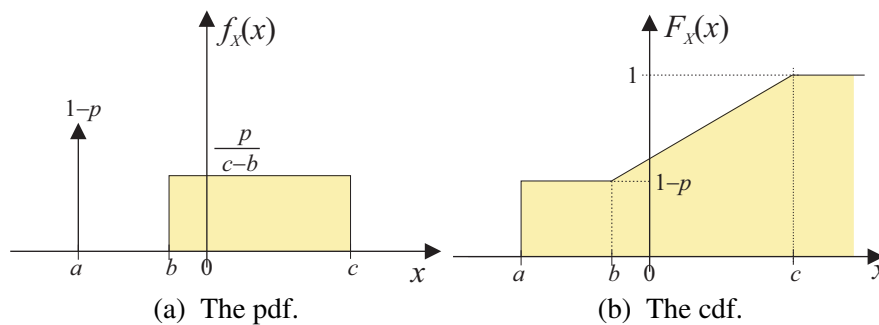


Figure 4.6: A probability density function and its corresponding cumulative distribution function for a RV which is a mixture of continuous and discrete components.

where $\delta(x)$ is the Dirac-delta function, and is given by:

$$\delta(x) = 0 \quad \text{if } x \neq 0 \tag{4.19a}$$

$$\int_{-\infty}^{\infty} \delta(x) dx = 1 \tag{4.19b}$$

Example 4.3 (6-sided die). Describe the cdf and pdf for a fair six-sided die.

SOLUTION. The probability mass function (pmf) is given by $p_i = \Pr(X(\zeta) = x_i) = \frac{1}{6}$, where $x_i = i, i \in \{1, \dots, 6\}$.

The cdf can be drawn by noting that $\Pr(X(\zeta) < x_1) = 0$ whereas $\Pr(X(\zeta) \leq x_1) = 1/6$. In other words, we need to carefully consider the probability of the events on an interval, not a discrete event, and hence when the cdf actually transitions values.

The pdf is obtained by differentiating the cdf:

$$f_X(x) = \sum_{i=1}^N p_i \delta(x - x_i) = \frac{1}{6} \sum_{i=1}^6 \delta(x - i) \tag{4.20} \quad \square$$

Moreover, a mixture of continuous and discrete components will have a pdf that is composed of delta functions as well as continuous functions:

$$f_{X,m}(x) = \sum_k p_k \delta(x - x_k) + f_{X,c}(x) \tag{4.21}$$

An example of a mixture is shown in Figure 4.6. The pdf for the distribution shown in Figure 4.6 can be written as:

$$f_X(x) = (1-p)\delta(x-a) + \frac{p}{c-b}(u(x-b) - u(x-c)) \quad (4.22)$$

where $u(x)$ is the unit step function, such that $u(x) = 1$ if $x \geq 0$ and zero otherwise.

Integrating, it can be shown that:

$$F_X(\infty) = \int_{-\infty}^{\infty} f_X(x) dx = (1-p) + \frac{p}{c-b} \times (c-b) = 1 \quad (4.23)$$

The result of a property of pdfs.

KEYPOINT! (Discussion Topic). Can you think of examples of a mixture of discrete and continuous random variables?

4.4 Properties of Distribution and Density Functions

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

- Properties of **cdf**:

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

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

$$F_X(a) \leq F_X(b) \quad \text{if } a \leq b \quad (4.24)$$

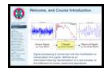
- Properties of **pdfs**:

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

- Probability of arbitrary events:

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

– End-of-Topic 17: Introduction to pdf and their properties –



New slide

4.5 Examples of Continuous random variables

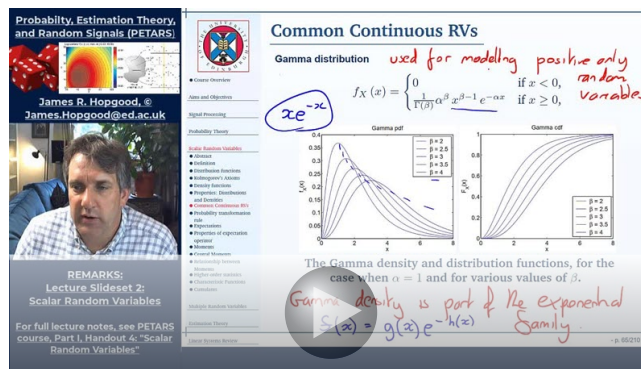
Topic Summary 18 Common density functions and their properties

Topic Objectives:

- Look at common pdfs used in signal processing algorithms.
- Consider pdfs across different intervals.
- Resources for finding out other density functions.

Topic Activities:

Type	Details	Duration	Progress
Watch video	12 : 54 minute video	3 × video length	
Read Handout	Read page 105 to page 109	8 mins/page	



http://media.ed.ac.uk/media/1_tfm5yn5

Video Summary: This video introduces a number of common probability density functions (pdfs) that are used in signal processing algorithms. Examples are given over finite-intervals, the entire real axis, and semi-infinite intervals. More significantly, this video shows how to use Wikipedia to discover other important densities as and when they arise in your work. Signal processing applications of the von-Mises and Voigt densities are mentioned.

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 4.7.

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 4.7. Roughly speaking, X takes on any value between a and b with equal probability.

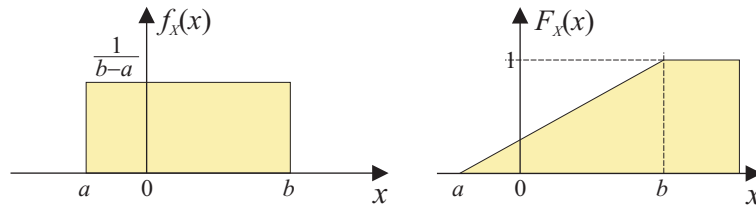


Figure 4.7: The uniform probability density function and cumulative distribution function.

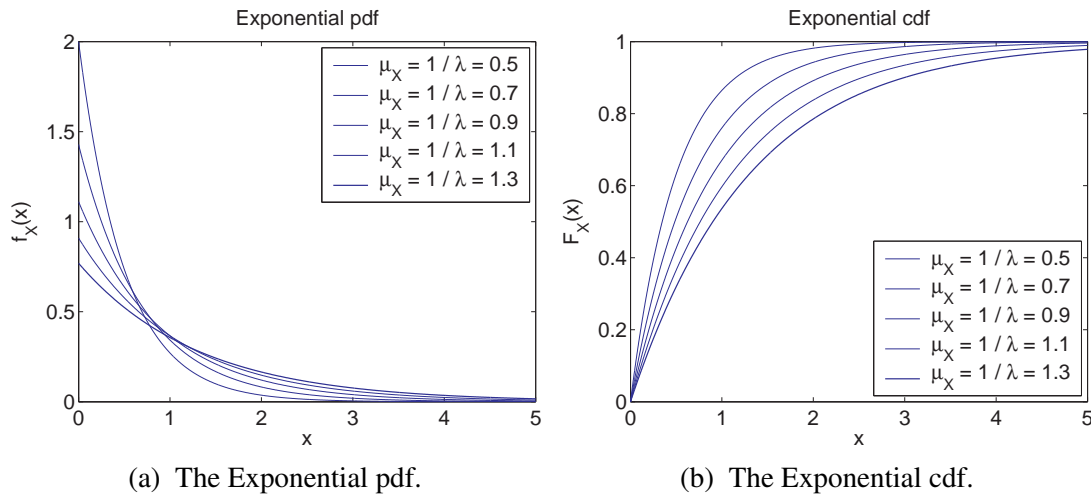


Figure 4.8: The exponential density and distribution functions, for various different values of the parameter λ .

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:

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \lambda e^{-\lambda x} & \text{if } x \geq 0, \end{cases} \tag{4.25}$$

Consequently, the cdf is given by:

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 - e^{-\lambda x} & \text{if } x \geq 0, \end{cases} \tag{4.26}$$

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 4.8, for various different values of the parameter λ .

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

$$\mu_X = \frac{1}{\lambda} \quad \text{and} \quad \sigma_X^2 = \mu_X^2 = \frac{1}{\lambda^2} \tag{4.27}$$

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

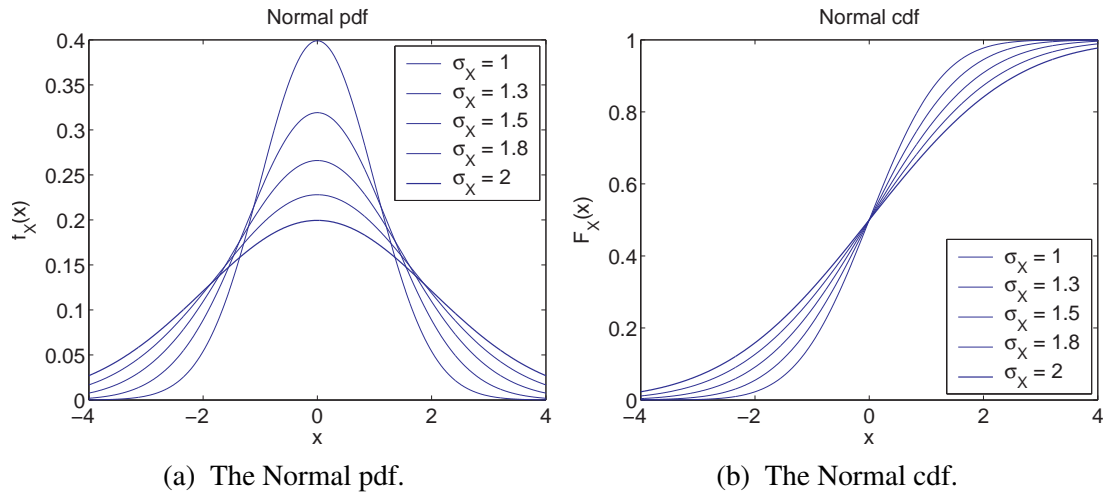


Figure 4.9: The Gaussian density and distribution functions; these plots are for a zero mean normal pdf, and are plotted for various different variances, σ_X^2 .

Normal distribution Arguably the most important continuous distribution is the *normal* or **Gaussian distribution**; these terms will be used interchangeably.

The pdf of a Gaussian distributed RV, $X(\zeta)$, with mean μ_X and standard deviation σ_X^2 , is given by:

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

It is common to denote this by:

$$f_X(x) = \mathcal{N}(x | \mu_X, \sigma_X^2) \quad (4.28)$$

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

$$\hat{x} \sim \mathcal{N}(\mu_X, \sigma_X^2) \quad (4.29)$$

The Gaussian pdf and cdf are shown in Figure 4.9 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:

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{1}{\Gamma(\beta)} \alpha^\beta x^{\beta-1} e^{-\alpha x} & \text{if } x \geq 0, \end{cases} \quad (4.30)$$

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

$$\Gamma(\beta) = \int_0^\infty x^{\beta-1} e^{-x} dx \quad (4.31)$$

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

The Gamma pdf and cdf are shown in Figure 4.10, for the case when $\alpha = 1$ and for various values of the parameter β .

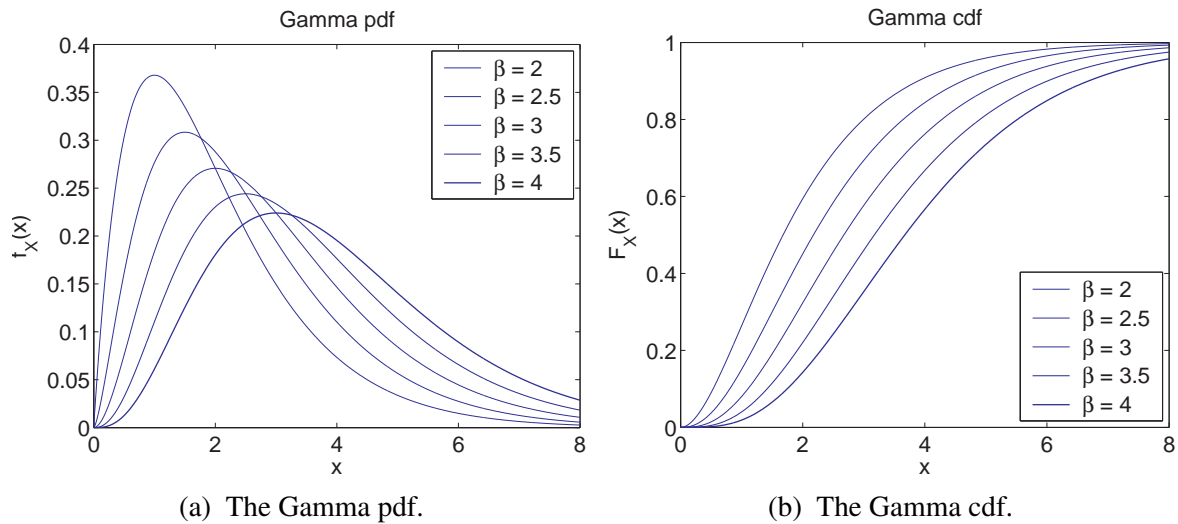


Figure 4.10: The Gamma density and distribution functions, for the case when $\alpha = 1$ and for various values of β .

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:

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{1}{\Gamma(\beta)} \alpha^\beta x^{-(\beta+1)} e^{-\frac{\alpha}{x}} & \text{if } x \geq 0, \end{cases} \quad (4.32)$$

It is common to denote this by:

$$f_X(x) = \mathcal{IG}(x | \alpha, \beta) \quad (4.33)$$

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

$$\hat{x} \sim \mathcal{IG}(\alpha, \beta) \quad (4.34)$$

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

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

The Cauchy random variable is symmetric around the value $x = \mu_X$, but its mean and variance (or other moments) do not exist. The corresponding cdf is given by:

$$F_X(x) = \frac{1}{2} + \frac{1}{\pi} \arctan \frac{x - \mu_X}{\beta} \quad (4.35)$$

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:

$$f_X(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1} & 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases} \quad (4.36)$$

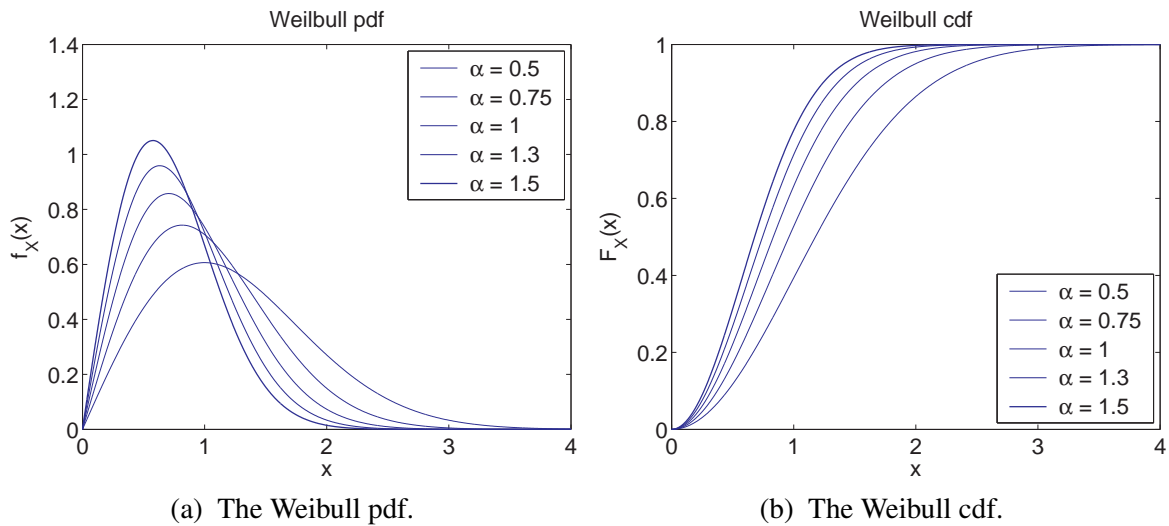


Figure 4.11: The Weibull density and distribution functions, for the case when $\alpha = 1$, and for various values of the parameter β .

where the **beta function** is given by

$$B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx \quad (4.37)$$

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:

$$f_X(x) = \begin{cases} \frac{\gamma^k (\gamma x)^{k-1} e^{-\gamma k x}}{(k-1)!} & x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \quad (4.38)$$

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

$$\mu_X = \frac{1}{\gamma} \quad \text{and} \quad \sigma_X^2 = \frac{1}{k\gamma^2} \quad (4.39)$$

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

$$f_X(x) = \begin{cases} 0 & x < 0 \\ \alpha\beta x^{\beta-1} e^{-\alpha x^\beta} & x \geq 0 \end{cases} \quad (4.40)$$

The corresponding the cdf is given by:

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\alpha x^\beta} & x \geq 0 \end{cases} \quad (4.41)$$

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

The Weibull pdf and cdf are shown in Figure 4.11, for the case when $\alpha = 1$, and for various values of the parameter β .



4.6 Probability transformation rule

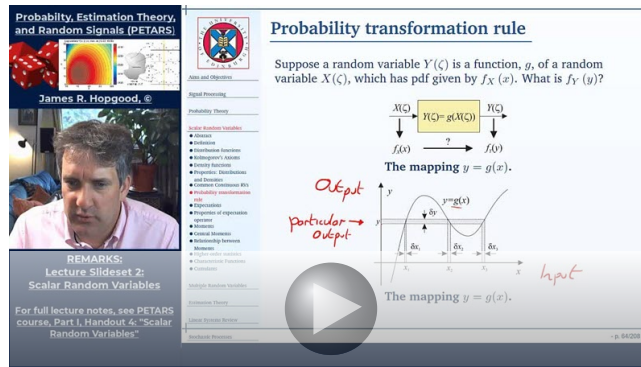
Topic Summary 19 Probability Transformation Rule and Its Applications

Topic Objectives:

- Need for the Probability Transformation Rule.
- Conceptual Proof.
- Examples and applications.

Topic Activities:

Type	Details	Duration	Progress
Watch video	12 : 25 min video	3 × length	
Read Handout	Read page 110 to page 113	8 mins/page	
Try Examples	Try Examples 4.4 and 4.5	15 minutes	
Practice Exercises	Exercise ?? to ?? (4 questions)	60 mins	



http://media.ed.ac.uk/media/1_asatl2ps

Video Summary: This video introduces the probability transformation rule, for finding the pdf of the mapping of another random variable. A derivation of the transformation rule is presented, by considering mutually exclusive small intervals, such that the rule is effectively an application of the axiomatic probability sum rule. An example with a single root is provided, leading to the log-normal distribution. The viewer is recommended to work through the example at the end of the inverse transformation of a random variable that is Cauchy distributed.

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

This functional relationship is shown diagrammatically in Figure 4.12, and an arbitrary function between $X(\zeta)$ and $Y(\zeta)$ is shown in Figure 4.13.

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

1. Its domain must include the range of the RV $X(\zeta)$.

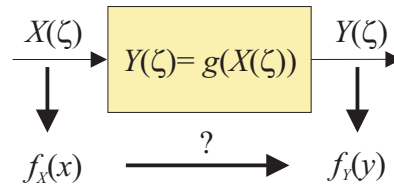


Figure 4.12: The mapping $y = g(x)$.

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

These properties are usually satisfied, but they are defined in order to avoid difficult cases, where the function $g(x)$ behaves in a way that mathematical technicalities arise.

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, $g : \mathbb{R} \rightarrow \mathbb{R}$. 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 4.1 (Probability Transformation Rule). Denote the real roots of $y = g(x)$ by $\{x_n, n \in \mathcal{N}\}$, such that:

$$y = g(x_1) = \dots = g(x_N) \tag{4.42}$$

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

$$f_Y(y) = \sum_{n=1}^N \frac{f_X(x_n)}{|g'(x_n)|} \tag{4.43}$$

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

PROOF. First consider the output **pdf** which, by definition, is given by:

$$f_Y(y) dy = \Pr(y < Y(\zeta) \leq y + dy) \tag{4.44}$$

KEYPOINT! (Informal proof). It would be more precise to use δx and δy instead of dx and dy , and then undertake a formal limiting operation as per the fundamental operations of calculus. However, this is a slightly more informal proof that is adequate for the scope of this course.

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

$$x_n < x \leq x_n + dx_n \tag{4.45}$$

It is easier to understand these proofs if you consider these intervals to be mutually exclusive (which is why the function $g(x)$ must satisfy the Baire property). This is shown in Figure 4.13 for the case when there are three mutually exclusive solutions to the equation $y = g(x)$.

The probability that x lies in this set is, of course:

$$f_X(x_n) dx_n = \Pr(x_n < X(\zeta) \leq x_n + dx_n) \tag{4.46}$$

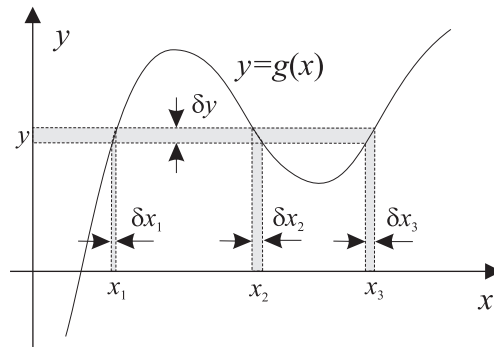


Figure 4.13: The mapping $y = g(x)$, and the effect of the mapping on intervals.

and, from the transformation from x to y , then

$$dx_n = \frac{dy}{|g'(x_n)|} \quad (4.47)$$

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

Finally, since these are N mutually exclusive sets corresponding to the N different roots to $y = g(x)$, then

$$\Pr(y < Y(\zeta) \leq y + dy) = \sum_{n=1}^N \Pr(x_n < X(\zeta) \leq x_n + dx_n) \quad (4.48)$$

$$\approx f_Y(y) dy \approx \sum_{n=1}^N f_X(x_n) dx_n \quad (4.49)$$

$$f_Y(y) dy = \sum_{n=1}^N f_X(x_n) \frac{dy}{|g'(x_n)|} \quad (4.50)$$

$$f_Y(y) = \sum_{n=1}^N \left. \frac{f_X(x_n)}{\left| \frac{dy}{dx} \right|_{x=x_n}} \right|_{x_n=g^{-1}(y)} \quad (4.51) \quad \square$$

where as a reminder $x_n = g^{-1}(y)$ is the roots of the equation $y = g[x]$, and thus the desired result is obtained after minor rearrangement.

Example 4.4 (Log-normal distribution). Let $Y = e^X$, where $X \sim \mathcal{N}(0, 1)$. Find the pdf for the RV Y .

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

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (4.52)$$

Considering the transformation $y = g(x) = e^x$, there is one root, given by $x = \ln y$. Therefore, the derivative of this expression is $g'(x) = \frac{d e^x}{dx} = e^x = y$. Hence, it follows:

$$f_Y(y) = \frac{f_X(x)}{g'(x)} = \frac{f_X(\ln y)}{y} = \frac{1}{y\sqrt{2\pi}} e^{-\frac{(\ln y)^2}{2}} \quad (4.53) \quad \square$$

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

$$f_X(x) = \frac{\alpha}{\pi} \frac{1}{x^2 + \alpha^2} \quad (4.54)$$

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:

$$f_Y(y) = \frac{1}{y^2} f_X\left(\frac{1}{y}\right) \quad (4.55)$$

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

$$f_X(x) = \frac{\alpha}{\pi} \frac{1}{x^2 + \alpha^2} \quad (4.56)$$

such that:

$$f_Y(y) = \frac{1}{y^2} f_X\left(\frac{1}{y}\right) = \frac{1}{y^2} \frac{\alpha}{\pi} \frac{1}{\frac{1}{y^2} + \alpha^2} \quad (4.57)$$

$$= \frac{1/\alpha}{\pi} \frac{1}{y^2 + \frac{1}{\alpha^2}} \quad (4.58)$$

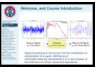
□

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

– End-of-Topic 19: Derivation of the Probability Transformation Rule,
and some examples –



4.7 Expectations



Topic Summary 20 Expectations and their Properties

New slide

Topic Objectives:

- Summary of key aspects of a pdf.
- Properties of the mean value of a random variable.
- Invariance of the Expectation Operator.
- Examples of finding expected value.

Topic Activities:

Type	Details	Duration	Progress
Watch video	18 : 08 min video	3× length	
Read Handout	Read page 114 to page 116	8 mins/page	
Try Example	Work through Example 4.6	10 minutes	

Expectations

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

Skewness
 - 2nd order statistic
 - Measure of asymmetry
 - Difference in tails

Kurtosis
 - 4th order statistic
 - Measure of size of tails

Mean
 - 1st order statistic
 - Centre of mass

Variance
 - 2nd order statistic
 - "spread of the pdf"

The four salient or key features or statistics of the pdf.

http://media.ed.ac.uk/media/0_j196xtbds

Video Summary: This video discusses why it is useful to characterise a pdf in terms of salient features which measure the location, spread, asymmetry, and the tails of the density. Other key statistics are also mentioned in relation to this characterisation. The expected value is then formally introduced both for continuous random variables, but also for discrete random variables. The properties of the mean value is then considered for even and symmetric densities. Next, the video looks at the invariance of the expectation operator for finding the expected value of a nonlinear function of another random variable, including a proof. The video finishes with an example showing the expected value of a trigonometric transformation of a uniform random variable.

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 four salient or key features are shown in Figure 4.14. These can be characterised by looking at the notion of expectation, which in turn defines moments.

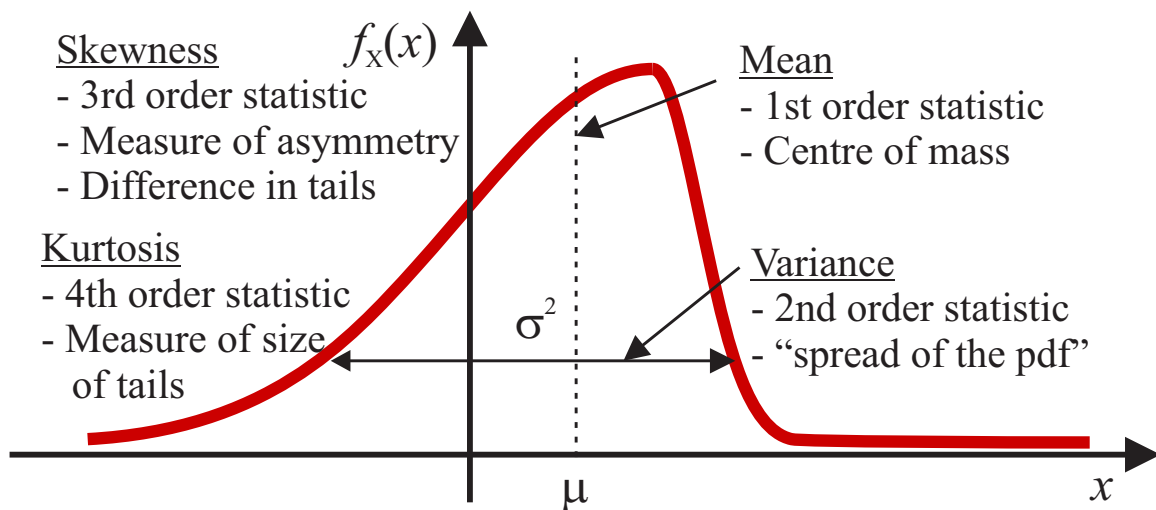


Figure 4.14: The four salient or key features or statistics of the pdf.

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

$$\mathbb{E}[X(\zeta)] = \int_{\mathbb{R}} x f_X(x) dx \quad (4.59)$$

- Recall: if $X(\zeta)$ is discrete then, as shown earlier in this handout, its corresponding **pdf** may be written in terms of its **pmf** as:

$$f_X(x) = \sum_k p_k \delta(x - x_k) \quad (4.60)$$

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

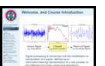
- Hence, for a discrete **RV**, the **expected** value is given by:

$$\mu_x = \int_{\mathbb{R}} x f_X(x) dx = \int_{\mathbb{R}} x \sum_k p_k \delta(x - x_k) dx = \sum_k x_k p_k \quad (4.61)$$

where the order of integration and summation have been interchanged because they do not depend on each other, and the sifting-property is applied such that:

$$\int_{\mathbb{R}} x \delta(x - x_k) dx = x_k \quad (4.62)$$

4.7.1 Properties of expectation operator



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

- 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 $\mu_X = a$ provided that the mean is finite (and therefore exists).

- The expectation operator is linear:

$$\mathbb{E} [\alpha X(\zeta) + \beta] = \alpha \mu_X + \beta \quad (\text{M:3.1.10})$$

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

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

This property is known as the **invariance of the expectation operator**.

KEYPOINT! (Invariance of the Expectation Operator). This property means that you don't need to keep track of which pdf the expectation is taken with respect to. Rather, you simply need to consider the RV inside the expectation, and the expectation is taken w. r. t. the pdf of that RV.

As an outline sketch, or simple proof, to prove this result, consider a monotonic one-to-one function $y = g(x)$, such that using the probability transformation rule $f_Y(y) = \frac{f_X(x)}{\frac{dy}{dx}}$. Then, it follows that:

$$\mathbb{E}_{f_Y} [Y(\zeta)] = \int y f_Y(y) dy = \int g(x) \frac{f_X(x)}{\frac{dy}{dx}} dy = \int g(x) f_X(x) dx \quad (4.63)$$

Note that *cancelling* the dy 's is not a formal mathematical process, but it gives an overview of the proposed approach. A more detailed proof for many-to-one functions with negative gradients is discussed in much more detail in Sidebar 4.

Example 4.6 (Trigonometric Transformation). The continuous random variable (RV), $\Theta(\zeta)$, is uniformly distributed between $-\pi$ and π .

1. Calculate the expected value of $\Theta(\zeta)$.
2. Now consider the RV, $Y(\zeta) = A \cos^2 \Theta(\zeta)$, where A is assumed to be a constant value. What is the expected value of $Y(\zeta)$?

SOLUTION. 1. The expected value of $\Theta(\zeta)$ is:

$$\mathbb{E} [\Theta(\zeta)] = \int_{-\infty}^{\infty} \theta f_{\Theta}(\theta) d\theta = \int_{-\pi}^{\pi} \theta \frac{1}{2\pi} d\theta \quad (4.68)$$

$$= \frac{\theta^2}{4\pi} \Big|_{-\pi}^{\pi} = 0 \quad (4.69)$$

2. Using the invariance of the expectation operator gives:

$$\mathbb{E} [Y(\zeta)] = \mathbb{E} [A \cos^2 \theta(\zeta)] = \int_{-\pi}^{\pi} [A \cos^2(\theta)] f_{\Theta}(\theta) d\theta \quad (4.70)$$

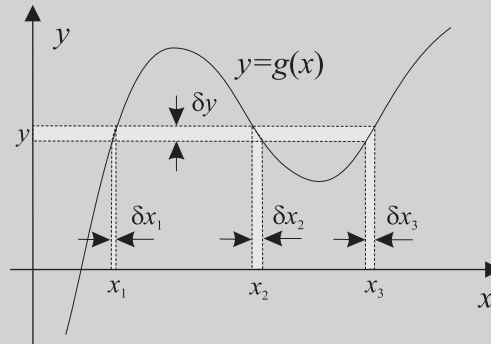
$$= \frac{A}{2\pi} \int_{-\pi}^{\pi} \cos^2(\theta) d\theta = \frac{A}{4\pi} \int_{-\pi}^{\pi} (1 + \cos 2\theta) d\theta = \frac{A}{2} \quad (4.71)$$

□



Sidebar 4 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 4.1.



Consider again Figure 4.13 on page 112, which is reproduced above. Let $Y(\zeta) = g(X(\zeta))$. Consider first the approximation for the expectation of $Y(\zeta)$:

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy \approx \sum_{\forall k} y_k f_Y(y_k) \delta y \quad (4.64)$$

where $f_Y(y_k) \delta y = \Pr(y_k < Y(\zeta) \leq y_k + \delta y)$ 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 4.1, can be written as the sum of the probabilities that $X(\zeta)$ is each of the corresponding small intervals shown in Figure 4.13 above, such that:

$$f_Y(y_k) \delta y = \sum_{n=1}^N \Pr(x_{k,n} < X(\zeta) \leq x_{k,n} + \delta x_{k,n}) = \sum_{n=1}^N f_X(x_{k,n}) \delta x_{k,n} \quad (4.65)$$

Substituting Equation 4.65 into Equation 4.67 gives:

$$\mathbb{E}[Y] \approx \sum_{\forall k} y_k \sum_{n=1}^N f_X(x_{k,n}) \delta x_{k,n} = \sum_{\forall k} \sum_{n=1}^N g(x_{k,n}) f_X(x_{k,n}) \delta x_{k,n} \quad (4.66)$$

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

$$\mathbb{E}[Y] \approx \sum_{\forall \ell} g(x_\ell) f_X(x_\ell) \delta x_\ell \quad (4.67)$$

which in the limit gives the integral Equation M:3.1.11, page 116. So, in summary, to compute the expectation of $Y(\zeta) = g(X(\zeta))$, it is not necessary to transform and find the pdf of $f_Y(y)$, but simply use this invariance of expectation property.

4.8 Moments

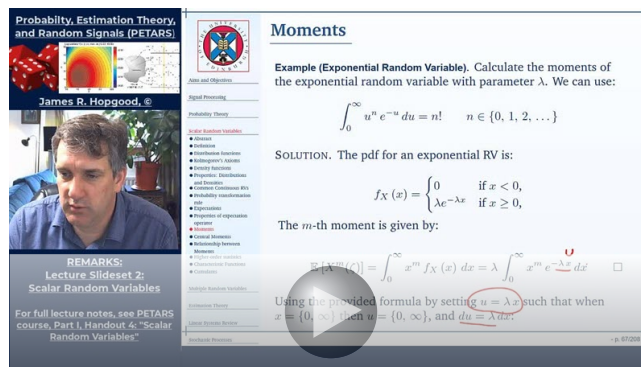
Topic Summary 21 Moments and Definitions

Topic Objectives:

- General definition of moments.
- Examples of calculating moments.
- Central moments and relationship with moments.

Topic Activities:

Type	Details	Duration	Progress
Watch video	17 : 52 min video	3 × length	
Read Handout	Read page 118 to page 122	8 mins/page	
Try Example	Work through Examples 4.7 and 4.8	20 minutes	



http://media.ed.ac.uk/media/1_8kwpp2.js

Video Summary: This video builds on Topic 20 by explicitly defining variance in terms of expectations, and the more general definition of moments. The video then considers calculating moments for a couple of simple examples, namely the exponential random variable, but also a property of moments for non-negative random variables. The second half of the video then considers central moments, and the relationship between moments and central moments (with an opportunity to mention Pascal’s triangle!).

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

$$\mathbb{E}[X(\zeta)] = \mu_X = \int_{\mathbb{R}} x f_X(x) dx \tag{4.72}$$

$$\text{var}[X(\zeta)] = \sigma_X^2 = \int_{\mathbb{R}} x^2 f_X(x) dx - \mu_X^2 = \mathbb{E}[X^2(\zeta)] - \mathbb{E}^2[X(\zeta)] \tag{4.73}$$

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

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

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

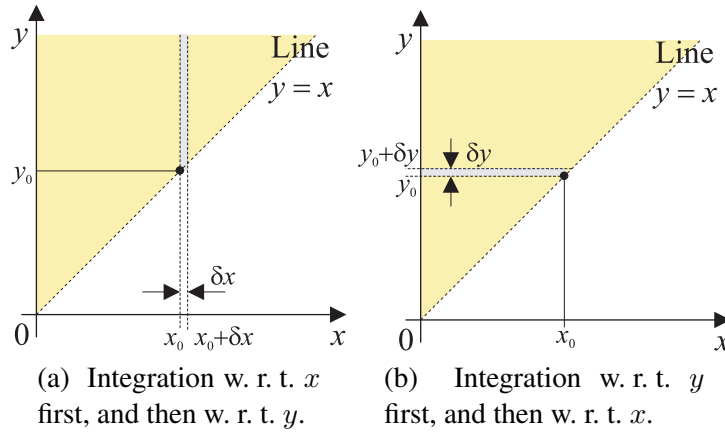


Figure 4.15: The region of integration for the integral in Equation 4.79.

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

Example 4.7 (Exponential Random Variable). Calculate the moments of the exponential random variable with parameter λ . We can make use of the formula (proof left as an exercise for the reader!):

$$\int_0^\infty u^n e^{-u} du = n! \quad n \in \{0, 1, 2, \dots\} \tag{4.74}$$

SOLUTION. The pdf for an exponential RV is (see Section 4.5 for full details):

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \lambda e^{-\lambda x} & \text{if } x \geq 0, \end{cases} \tag{4.75}$$

The m -th moment is given by:

$$\mathbb{E} [X^m(\zeta)] = \int_0^\infty x^m f_X(x) dx = \lambda \int_0^\infty x^m e^{-\lambda x} dx \tag{4.76}$$

Using the provided formula by setting $u = \lambda x$ such that when $x = \{0, \infty\}$ then $u = \{0, \infty\}$, and $du = \lambda dx$, it follows:

$$\mathbb{E} [X^m(\zeta)] = \frac{1}{\lambda^m} \int_0^\infty u^m e^{-u} du = \frac{m!}{\lambda^m} \quad \square \tag{4.77}$$

In particular, by setting $m = 1$, the mean is given by $\mu_X = \mathbb{E} [X(\zeta)] = 1/\lambda$.

Setting $m = 2$, the second-moment is $\mathbb{E} [X^2(\zeta)] = 2/\lambda^2$, which means the variance is given by $\sigma_X^2 = \text{var} [X(\zeta)] = 2/\lambda^2 - (1/\lambda)^2 = 1/\lambda^2 = \mu_X^2$.

Example 4.8 (Expectations of non-negative RVs). Let $X(\zeta)$ be a non-negative RV with pdf $f_X(x)$. Show that

$$\mathbb{E} [X^m(\zeta)] = \int_0^\infty m x^{m-1} \text{Pr} (X(\zeta) > x) dx \tag{4.78}$$

for any $m \geq 1$ 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,

$$\int_0^\infty m x^{m-1} \Pr(X(\zeta) > x) dx = \int_0^\infty m x^{m-1} \left\{ \int_{y=x}^\infty f_X(y) dy \right\} dx \quad (4.79)$$

and rearrange the order of integration, noting the region of integration as shown in Figure 4.15, and thus the change in the limits:

$$= \int_0^\infty f_X(y) \left\{ \int_{x=0}^y m x^{m-1} dx \right\} dy \quad (4.80)$$

$$= \int_0^\infty f_X(y) [x^m]_0^y dy = \int_0^\infty y^m f_X(y) dy = \mathbb{E}[X^m(\zeta)] \quad (4.81)$$

□

4.8.1 Central Moments

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

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

Some obvious properties that follow from these definitions are:

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

$$\text{var}[X(\zeta)] \triangleq \sigma_X^2 \triangleq \gamma_X^{(2)} = \mathbb{E}[(X(\zeta) - \mu_X)^2] \quad (4.82)$$

- **Standard deviation** is given by: $\sigma_X = \sqrt{\text{var}[X(\zeta)]}$.
- **Trivial moments:** $r_X^{(0)} = 1$ and $r_X^{(1)} = \mu_X$.
- **Trivial central moments:** $\gamma_X^{(0)} = 1$, $\gamma_X^{(1)} = 0$, and $\gamma_X^{(2)} = \sigma_X^2$.

The polynomial term in Equation M:3.1.14 can be expanded as

$$(x - \mu_X)^m = x^m - \mu_X^{m-1} x + \cdots - \mu_X x^{m-1} + \mu_X^m = \sum_{k=0}^m \binom{m}{k} (-1)^k \mu_X^k x^{m-k} \quad (4.83)$$

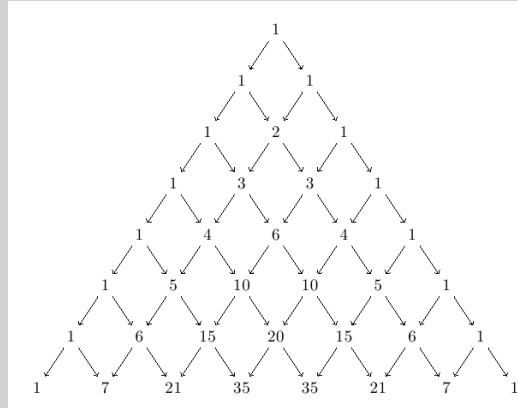
where the polynomial coefficients $\binom{m}{k}$ can be found using Pascal's Triangle, as shown in Sidebar 5. This leads onto the relationship between moments and central moments as discussed in the next section.

Sidebar 5 Combinatorial terms and Pascal's Triangle

A reminder of high-school maths that ${}^m C_k = \binom{m}{k}$ can be obtained via Pascal's triangle. These combinatorial terms are the coefficients of the polynomial expansion:

$$(a + b)^m = a^m + {}^m C_1 a^{m-1} b + {}^m C_2 a^{m-2} b^2 + \dots$$

which can be calculated from Pascal's triangle, as shown below.

**4.8.2 Relationship between Moments and Central Moments**

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

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

$$r_X^{(m)} = \sum_{k=0}^m \binom{m}{k} \mu_X^k \gamma_X^{(m-k)} \quad (4.84)$$

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

$${}^n C_r = \frac{n!}{r!(n-r)!} \quad (4.85)$$

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

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

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

$$\gamma_X^{(m)} = \mathbb{E}[(X(\zeta) - \mu_X)^m] \quad (4.86)$$

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

then using the binomial:

$$(x + a)^n = \sum_{k=0}^n \binom{n}{k} x^k a^{n-k} = \sum_{k=0}^n \binom{n}{k} a^k x^{n-k} \quad (4.87)$$

it follows:

$$\gamma_X^{(m)} = \int_{\mathbb{R}} \sum_{k=0}^m \binom{m}{k} x^{m-k} (-\mu_X)^k f_X(x) dx \quad (4.88)$$

$$= \sum_{k=0}^m \binom{m}{k} (-1)^k \mu_X^k \underbrace{\int_{\mathbb{R}} x^{m-k} f_X(x) dx}_{r_X^{(m-k)}} \quad (4.89)$$

as required. Similarly, note that

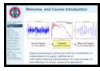
$$r_X^{(m)} = \int_{\mathbb{R}} [(x - \mu_X) + \mu_X]^m f_X(x) dx \quad (\text{M:3.1.12})$$

$$= \int_{\mathbb{R}} \sum_{k=0}^m \binom{m}{k} \mu_X^k (x - \mu_X)^{m-k} f_X(x) dx \quad (4.90)$$

$$= \sum_{k=0}^m \binom{m}{k} \mu_X^k \underbrace{\int_{\mathbb{R}} (x - \mu_X)^{m-k} f_X(x) dx}_{\gamma_X^{(m-k)}} \quad (4.91) \quad \square$$

giving the desired result. These expressions can also be obtained by using the linearity property of the expectation operator, rather than using the integral expressions above.





4.8.3 Higher-Order Statistics

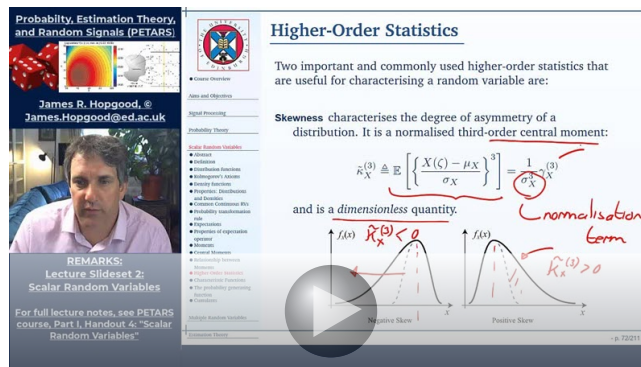
Topic Summary 22 Higher-Order Statistics

Topic Objectives:

- Skewness and its interpretation.
- Kurtosis and its interpretation.
- Examples of calculating skewness and kurtosis.

Topic Activities:

Type	Details	Duration	Progress
Watch video	11 : 44 min video	3× length	
Read Handout	Read page 123 to page 125	8 mins/page	
Try Examples	Try Examples 4.9 and 4.10	15 minutes	
Practice Exercise	Exercise ??	20 minutes	



http://media.ed.ac.uk/media/1_8kwpp2js

Video Summary: This video looks at two important and commonly used higher-order statistics that are useful for characterising a random variable, namely skewness and kurtosis. The video gives a physical meaning to each statistic and a mathematical definition. The video shows an example of calculating skewness for the exponential distribution, and kurtosis for the standard Laplacian distribution. The video then finishes with examples of using these higher-order statistics in signal processing applications.

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:

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

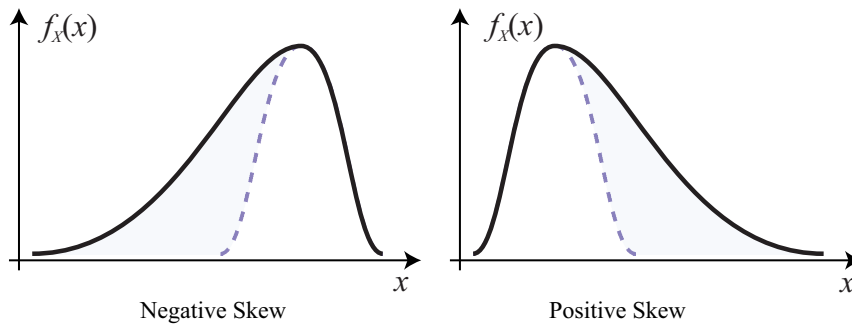


Figure 4.16: A graphical representation of the skewness of a pdf.

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

$$\tilde{\kappa}_X^{(3)} = \begin{cases} < 0 & \text{if the density leans or stretches out towards the left} \\ 0 & \text{if the density is symmetric about } \mu_X \\ > 0 & \text{if the density leans or stretches out towards the right} \end{cases} \quad (4.92)$$

In other words, if the left side or *left tail* of the distribution is more *stretched out* than the *right tail*, the function is said to have negative skewness (and is sometimes said to lean 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:

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

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 sometimes known as **kurtosis excess**, with **kurtosis proper** having the same definition but without the offset of 3.

Example 4.9 (Exponential distribution). Calculate the skewness of an exponential random variable with parameter λ .

SOLUTION. From earlier calculations in Example 4.7, it was shown that the m -th moment was given by $r_X^{(m)} = m!/\lambda^m$.

It can also be shown, by expanding the expression for skewness (see Unknown exer:skewness), that:

$$\tilde{\kappa}_X^{(3)} = \frac{r_X^{(3)} - 3r_X^{(1)}r_X^{(2)} + 2(r_X^{(1)})^3}{\sigma_X^3} \quad (4.93)$$

Hence, since it was also shown that $\sigma_X^2 = 1/\lambda^2$, then:

$$\tilde{\kappa}_X^{(3)} = \frac{\frac{3!}{\lambda^3} - 3\frac{1!}{\lambda}\frac{2!}{\lambda^2} + 2\frac{1}{\lambda^3}}{\frac{1}{\lambda^3}} = 2 \quad (4.94)$$

□

Positive skewness indicates leaning to the right, which it does!

Example 4.10 (Laplace distribution). Calculate the Kurtosis of the **standard Laplace distribution**, $f_X(x) = \frac{1}{2}e^{-|x|}$, $x \in \mathbb{R}$.

SOLUTION. Note that as the density is symmetric, the skewness is zero! Moreover, you can show that the odd moments are also equal to zero through symmetry (left as an exercise to the reader).

The even moments are given by:

$$r_X^{(m)} = \frac{1}{2} \int_{-\infty}^0 x^m e^x dx + \frac{1}{2} \int_0^{\infty} x^m e^{-x} dx = \int_0^{\infty} x^m e^{-x} dx = m! \quad (4.95)$$

Hence, using the formula for Kurtosis (noting $r_X^{(1)} = 0$):

$$\tilde{\kappa}_X^{(4)} = \mathbb{E} \left[\left\{ \frac{X(\zeta) - \mu_X}{\sigma_X} \right\}^4 \right] - 3 = \frac{r_X^{(4)}}{(r_X^{(2)})^2} - 3 = \frac{4!}{(2!)^2} - 3 = 3 \quad (4.96) \quad \square$$

Skewness and kurtosis are used in signal processing in the following applications:

Signal Separation is only possible if the signals are statistically distinctive and this requires non-Gaussianity; maximising kurtosis means that separated signals are ensured to be as non-Gaussian as possible.

Outlier detection As kurtosis is a measure of heaviness of the tails, it also provides a metric for the number of outliers. Outliers, for example positive values, can also lead to asymmetric densities, measured by skewness.

Features Skewness and kurtosis can be used in feature-based classification and machine learning algorithms.

– End-of-Topic 22: **Skewness, Kurtosis, and their Applications** –



4.9 Characteristic Functions

Topic Summary 23 Characteristic, Moment, Probability, and Cumulant Generating Functions

Topic Objectives:

- General definition of moments.
- Examples of calculating moments.
- Central moments and relationship with moments.

Topic Activities:

Type	Details	Duration	Progress
Watch video	23 : 20 min video	3× length	
Read Handout	Read page 126 to page 133	8 mins/page	
Try Examples	Try Examples 4.11, 4.11, and 4.13	30 minutes	
Practice Exercise	Exercises ?? to ??	100 minutes	

Characteristic Functions

The characteristic function of a rv $X(\xi)$ is defined by:

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

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

$$\Phi_X(s) \triangleq \mathbb{E} \left[e^{sX(\xi)} \right] = \int_{-\infty}^{\infty} f_X(x) e^{sx} dx$$

Using a series expansion for $e^{sX(\xi)}$ gives:

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

Given $f_X^{(n)}$ you can create $\Phi_X(s)$

Thus, if all moments of $X(\xi)$ are known upon inverse Laplace transformation, the pdf $f_X(x)$ can be determined.

http://media.ed.ac.uk/media/1_qo43cj0q

Video Summary: To readers familiar with Signal and System analysis in Engineering, it will be second nature to apply the Fourier and Laplace transforms as a powerful tool for mapping functions from one domain to another in order to simplify subsequent analysis. This video looks at using this trick for mapping the pdf into a characteristic or moment generating function (MGF), which can then easily be used for a number of probability analysis problems. The key application here is for calculating moments for continuous random variables. The probability generating function (PGF), which is the z -transform of the pdf, is used in the same way for dealing with discrete-random variables. Finally, cumulants are also mentioned. An example of calculating the PGF for a geometric distribution is presented.

The Fourier and Laplace transforms find many uses in probability theory through the concepts of **characteristic functions** and **MGFs**. They have similar useful applications in probabilistic analysis, where these transforms can be used to simplify manipulations of pdfs, and evaluating properties such as finding moments. Ultimately, as with all transform methods, the usefulness of these techniques depends very much on the availability of transform pairs, or whether numerical calculations of the transform is efficient.

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

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

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:

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

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

KEYPOINT! (Relationship to Moments). The MGF can be directly related to the moments by an expansion of the exponential term, and use of the *three R's*, namely: **replace**, **reorder**, and **recognise**. This will give us a relationship between the MGF and the moments, such that the moments can easily be obtained (or generated).

One of the most useful applications for the MGF is, as the name suggests, a technique for finding moments quickly and efficiently. To demonstrate this, consider the following analysis.

Using a series expansion for $e^{sX(\zeta)}$ gives an alternative expression for the moment generating function (MGF):¹

$$\bar{\Phi}_X(s) = \mathbb{E} [e^{sX(\zeta)}] = \mathbb{E} \left[\sum_{n=0}^{\infty} \frac{(sX(\zeta))^n}{n!} \right] \quad (4.98)$$

$$= \sum_{n=0}^{\infty} \frac{s^n}{n!} \mathbb{E} [X^n(\zeta)] \quad (4.99)$$

Noting that $\mathbb{E} [X^n(\zeta)] = r_X^{(n)}$, it follows that:

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

provided that every moment $r_X^{(n)}$ exists. A physical interpretation of this result is that the MGF is *the Laplace transform of the pdf is a weighted summation of all the moments of the RV*.

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. This is described in more detail in Sidebar 6.

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

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

¹It 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:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \quad (4.97)$$

Sidebar 6 Estimating pdfs from moments

The relationship between the MGF and the moments of a RV in Equation M:3.1.23 lead to a method for estimating probability density functions. Suppose, for example, that the first three moments of a RV have been estimated (using the techniques later in the estimation theory handout) as $\hat{r}_X^{(k)}$ for $k = \{1, 2, 3\}$.

For example, it will be seen that the first and second moments can be estimated from N data points, $\{x[n], n \in \{0, \dots, N-1\}\}$, as:

$$\hat{r}_X^{(1)} = \mu_X = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \quad \text{and} \quad \hat{r}_X^{(2)} = \frac{1}{N^2} \sum_{n=0}^{N-1} x^2[n] \quad (4.103)$$

The MGF can then be estimated by the approximation:

$$\hat{\Phi}_X(s) \approx 1 + s \hat{r}_X^{(1)} + \frac{s^2}{2} \hat{r}_X^{(2)} + \frac{s^3}{6} \hat{r}_X^{(3)} \quad (4.104)$$

The pdf can then be estimated by taking the inverse-Laplace transform to give:

$$\hat{f}_X(x) \approx \mathcal{L}^{-1} \left(\hat{\Phi}_X(s) \right) \quad (4.105)$$

This is shown by differentiating Equation M:3.1.23 term by term:

$$\bar{\Phi}_X(s) = 1 + s r_X^{(1)} + \frac{s^2}{2} r_X^{(2)} + \frac{s^3}{6} r_X^{(3)} + \dots \quad (4.100)$$

$$\frac{d\bar{\Phi}_X(s)}{ds} = r_X^{(1)} + s r_X^{(2)} + \frac{s^2}{2} r_X^{(3)} + \dots \Rightarrow \left. \frac{d\bar{\Phi}_X(s)}{ds} \right|_{s=0} = r_X^{(1)} \quad (4.101)$$

Similarly, differentiating again:

$$\frac{d^2\bar{\Phi}_X(s)}{ds^2} = r_X^{(2)} + s r_X^{(3)} + \dots \Rightarrow \left. \frac{d^2\bar{\Phi}_X(s)}{ds^2} \right|_{s=0} = r_X^{(2)} \quad (4.102)$$

and the proof continues for all moments.

Characteristic functions and MGFs have applications to:

Manipulations of distributions, and specifically linear functions of independent variables; the characteristic function helps obtain complex results in a simplified manner.

Used in proofs such as the central limit theorem (CLT) in Section 5.10.

Calculating moments in a much faster way than finding the expectations directly.

Theorem 4.2 (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:

$$\sum_j \sum_k \Phi_X(\xi_j - \xi_k) z_j z_k^* \geq 0 \quad (4.106)$$

for all real ξ_i and complex z_i .

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

$$|\Phi_X(\xi)| \leq \int f_X(x) |e^{j\xi x}| dx = \int f_X(x) dx = 1 \quad (4.107)$$

as required.

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

$$|\Phi_X(\xi + \delta\xi) - \Phi_X(\xi)| = |\mathbb{E}[e^{j(\xi+\delta\xi)X(\zeta)} - e^{j\xi X(\zeta)}]| \quad (4.108)$$

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:

$$|\Phi_X(\xi + \delta\xi) - \Phi_X(\xi)| \leq \mathbb{E}[|e^{j(\xi+\delta\xi)X(\zeta)} - e^{j\xi X(\zeta)}|] \quad (4.109)$$

$$\leq \mathbb{E}[|e^{j\xi X(\zeta)} (e^{j\delta\xi X(\zeta)} - 1)|] \quad (4.110)$$

$$\leq \mathbb{E}[|e^{j\delta\xi X(\zeta)} - 1|] \quad (4.111)$$

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

$$|\Phi_X(\xi + \delta\xi) - \Phi_X(\xi)| \rightarrow 0 \quad \text{as } \delta\xi \rightarrow 0 \quad (4.112)$$

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

3. Finally,

$$\sum_p \sum_q \Phi_X(\xi_p - \xi_q) z_p z_q^* = \sum_p \sum_q z_p z_q^* \int f_X(x) e^{j(\xi_p - \xi_q)x} dx \quad (4.113)$$

$$= \int f_X(x) \left\{ \sum_p \sum_q z_p e^{j\xi_p x} z_q^* e^{-j\xi_q x} \right\} dx \quad (4.114)$$

$$= \int f_X(x) \left| \sum_p z_p e^{j\xi_p x} \right|^2 dx = \mathbb{E} \left[\left| \sum_p z_p e^{j\xi_p X} \right|^2 \right] \geq 0 \quad (4.115) \quad \square$$

Example 4.11 ([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:

$$\bar{\Phi}_Y(s) \triangleq \mathbb{E}[e^{sY(\zeta)}] = \mathbb{E}[e^{s(aX(\zeta)+b)}] \equiv e^{sb} \mathbb{E}[e^{asX(\zeta)}] = e^{sb} \bar{\Phi}_X(sa) \quad (4.116)$$

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:

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

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

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

Substituting one into the other gives

$$\bar{\Phi}_X(s) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{x - \mu_X}{\sigma_X}\right)^2\right] e^{sx} dx \quad (4.117)$$

$$= \frac{1}{\sqrt{2\pi\sigma_X^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2 - 2(\mu_X + \sigma_X^2 s)x + \mu_X^2}{2\sigma_X^2}\right] dx \quad (4.118)$$

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

$$\bar{\Phi}_X(s) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{(x - \{\mu_X + \sigma_X^2 s\})^2 - (2\mu_X\sigma_X^2 s + \{\sigma_X^2 s\}^2)}{2\sigma_X^2}\right] dx \quad (4.119)$$

$$\bar{\Phi}_X(s) = \exp\left[\mu_X s + \frac{1}{2}\sigma_X^2 s^2\right] \underbrace{\frac{1}{\sqrt{2\pi\sigma_X^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{(x - \{\mu_X + \sigma_X^2 s\})^2}{2\sigma_X^2}\right] dx}_{=1} \quad (4.120)$$

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

$$\bar{\Phi}_X(s) = \exp\left[\mu_X s + \frac{1}{2}\sigma_X^2 s^2\right] \quad (4.121)$$

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

$$\bar{\Phi}_Y(s) = e^{sb} \bar{\Phi}_X(sa) = e^{sb} \exp\left[a\mu_X s + \frac{1}{2}\sigma_X^2 a^2 s^2\right] \quad (4.122)$$

$$= \exp\left[(a\mu_X + b)s + \frac{1}{2}(\sigma_X^2 a^2)s^2\right] = \exp\left[\mu_Y s + \frac{1}{2}\sigma_Y^2 s^2\right] \quad (4.123) \quad \square$$

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.

4.9.1 The probability generating function

- The **characteristic function** and MGF can be extended to deal with discrete random variables by replacing the Laplace and Fourier transforms with the z -transform and DTFT, respectively.
- It is, however, necessary to modify how moments are calculated from the PGF, as the following example shows.

Example 4.12 (PGF). Let $X(\zeta)$ be a discrete random variable taking non-negative integers, k , with pmf given by $p_k = \Pr(X(\zeta) = k)$ if $k \geq 0$, and zero otherwise. Its PGF is defined as

$$G_X(z) = \mathbb{E}[z^{X(\zeta)}] = \sum_{k=0}^{\infty} p_k z^k$$

1. Show that the expected value, μ_X , of $X(\zeta)$ can be written as:

$$\mu_X = \mathbb{E}[X(\zeta)] = \left. \frac{dG_X(z)}{dz} \right|_{z=z_0}$$

stating clearly the value of z_0 required for this to be true.

2. Find an expression for the variance σ_X^2 of $X(\zeta)$ in terms of $G_X(z)$.

SOLUTION. 1. Differentiating $G_X(z)$ w. r. t. z term by term gives:

$$\frac{dG_X(z)}{dz} = \sum_{k=0}^{\infty} p_k \frac{dz^k}{dz} = \sum_{k=0}^{\infty} p_k k z^{k-1} \quad (4.124)$$

and setting $z = z_0 = 1$ gives (by definition):

$$\left. \frac{dG_X(z)}{dz} \right|_{z=1} = \sum_{k=0}^{\infty} p_k k = \mu_X \quad (4.125)$$

2. To find an expression for the variance σ_X^2 of $X(\zeta)$ in terms of $G_X(z)$, then differentiating Equation 4.124 again gives:

$$\frac{d^2G_X(z)}{dz^2} = \sum_{k=0}^{\infty} p_k k \frac{dz^{k-1}}{dz} = \sum_{k=0}^{\infty} p_k k(k-1)z^{k-2} \quad (4.126)$$

Setting $z = 1$ gives

$$\left. \frac{d^2G_X(z)}{dz^2} \right|_{z=1} = \sum_{k=0}^{\infty} p_k k(k-1) = \mathbb{E}[k^2] - \mu_X \quad (4.127)$$

Since $\sigma_X^2 = \mathbb{E}[X(\zeta)^2] - \mu_X^2 = \mathbb{E}[k^2] - \mu_X^2$, it follows that

$$\sigma_X^2 = \left. \frac{d^2G_X(z)}{dz^2} \right|_{z=1} + \left. \frac{dG_X(z)}{dz} \right|_{z=1} - \left[\left. \frac{dG_X(z)}{dz} \right|_{z=1} \right]^2 \quad (4.128)$$

It is also acceptable to leave the first two terms as a combined derivative, so an equally valid answer would be:

$$\sigma_X^2 = \left[\frac{d}{dz} \left(z \frac{dG_X(z)}{dz} \right) \right]_{z=1} - \left[\left. \frac{dG_X(z)}{dz} \right|_{z=1} \right]^2 \quad (4.129) \quad \square$$

Example 4.13 (Applying PGF). The **geometric distribution** is used for modelling the number of consecutive independent successes before a failure, and its pmf is given by

$$p_k = \begin{cases} p(1-p)^{k-1} & k \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

where $0 \leq p \leq 1$ is an individual probability of failure.

1. Find the probability generating function (PGF) for this distribution, and write down conditions on z for when the PGF exists.
2. Using the probability generating function, or otherwise, find the mean of this distribution, and show that the variance is $\frac{1-p}{p^2}$.

SOLUTION. 1. To find the PGF of the Geometric density, then noting that $p_0 = 0$, and taking the “ z -transform”:

$$G_X(z) = \sum_{k=1}^{\infty} p(1-p)^{k-1} z^k \quad (4.130)$$

setting $n = k - 1$, so that when $k = 1$ then $n = 0$, so that:

$$G_X(z) = pz \sum_{n=0}^{\infty} [z(1-p)]^n \quad (4.131)$$

$$G_X(z) = \frac{pz}{1 - z(1-p)} \quad (4.132)$$

This series converges for $|z(1-p)| < 1$ or $|z| < \frac{1}{1-p}$.

2. The mean is given by differentiating the PGF which gives:

$$\frac{dG_X(z)}{dz} = \frac{p[1 - z(1-p)] - [-(1-p)]pz}{[1 - z(1-p)]^2} = \frac{p}{[1 - z(1-p)]^2} \quad (4.133)$$

and by setting $z = 1$, this gives the mean of $\mu_X = \frac{p}{p^2} = \frac{1}{p}$.

Differentiating for a second time, then

$$\frac{d^2G_X(z)}{dz^2} = p \frac{-2 \times -(1-p)}{[1 - z(1-p)]^3} \quad (4.134)$$

Setting $z = 1$ gives

$$\left. \frac{d^2G_X(z)}{dz^2} \right|_{z=1} = \frac{2(1-p)}{p^2} \quad (4.135)$$

Using the result:

$$\sigma_X^2 = \left. \frac{d^2G_X(z)}{dz^2} \right|_{z=1} + \left. \frac{dG_X(z)}{dz} \right|_{z=1} - \left[\left. \frac{dG_X(z)}{dz} \right|_{z=1} \right]^2 \quad (4.136)$$

and using Equation 4.128 gives the desired answer:

$$\sigma_X^2 = \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{2 - 2p + p - 1}{p^2} \quad (4.137)$$

□

4.9.2 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:

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

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:

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

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.

– End-of-Topic 23: **Characteristic, Moment, Probability, and Cumulant
Generating Functions** –



5

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.

5.1 Abstract

Topic Summary 24 Introduction to Random Vectors

Topic Objectives:

- Introduction to the concept of random vectors.
- Formal definition of random vectors.
- Definition of the joint-cumulative distribution function (cdf) and joint-probability density function (pdf).

Topic Activities:

Type	Details	Duration	Progress
Watch video	10 : 20 min video	3 × length	
Read Handout	Read page 135 to page 137	8 mins/page	

http://media.ed.ac.uk/media/1_asatl2ps

Video Summary: A short introduction to random vectors, why multiple random variables occur as a group, and some example applications of random vectors. A graphical representation of a random vector which builds on the same concept from scalar random variables is presented. A formal definition of the random vector is discussed, followed by the definition of the joint-cdf and joint-pdf.

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

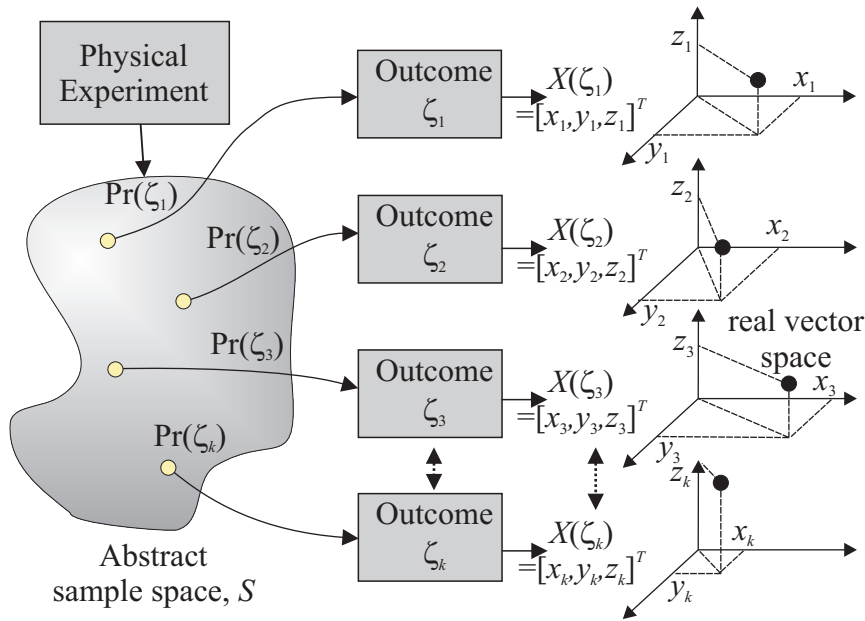


Figure 5.1: A graphical representation of a random vector.

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.

5.2 Definition of Random Vectors

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

$$\mathbf{X}(\zeta) = [X_1(\zeta) \quad X_2(\zeta) \quad \dots \quad X_N(\zeta)]^T \tag{M:3.2.1}$$

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:

$$\mathbf{X}(\zeta) = \begin{bmatrix} X_R(\zeta) \\ X_I(\zeta) \end{bmatrix} \tag{5.1}$$

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, as shown in the graphical representation in Figure 5.1.

Denote a specific value for a random vector as:

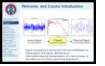
$$\mathbf{x} = [x_1 \quad x_2 \quad \dots \quad x_N]^T \tag{5.2}$$

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



New slide

5.2.1 Distribution and Density Functions



New slide

As with random variables, a random vector is completely characterised by its cdf and 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:

$$F_{\mathbf{X}}(\mathbf{x}) \triangleq \Pr(\{X_n(\zeta) \leq x_n, n \in \mathcal{N}\}) = \Pr(\mathbf{X}(\zeta) \leq \mathbf{x}) \quad (\text{M:3.2.2})$$

A random vector can also be characterised by its **joint pdf**:

$$f_{\mathbf{X}}(\mathbf{x}) = \lim_{\Delta \mathbf{x} \rightarrow \mathbf{0}} \frac{\Pr(\{x_n < X_n(\zeta) \leq x_n + \Delta x_n, n \in \mathcal{N}\})}{\Delta x_1 \cdots \Delta x_N} \quad (\text{M:3.2.4})$$

$$= \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} \cdots \frac{\partial}{\partial x_N} F_{\mathbf{X}}(\mathbf{x}) \quad (5.3)$$

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

Hence, it follows:

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



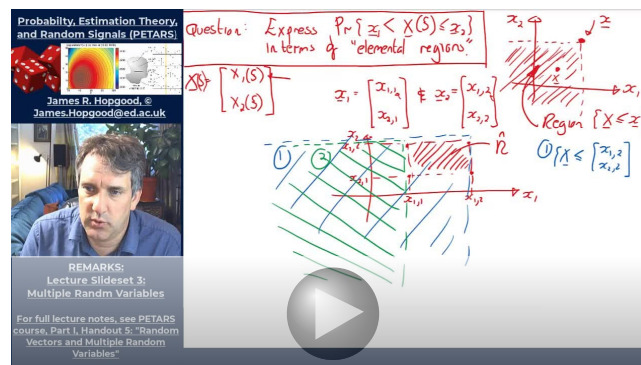
Topic Summary 25 Joint Distributions and Densities

Topic Objectives:

- Familiarise with properties of joint-cdf and joint-pdf.
- Example of finding joint-cdf from joint-pdf.
- Consider probability of arbitrary events.

Topic Activities:

Type	Details	Duration	Progress
Watch video	17 : 49 min video	3× length	
Read Handout	Read page 138 to page 140	8 mins/page	
Try Example	Try Example 5.1	15 minutes	
Practice Exercises	Exercises ?? and ??	40 mins	



http://media.ed.ac.uk/media/1_10k89edo

Video Summary: This video looks at the properties of joint-cdf and joint-pdf. It also looks at the probability of arbitrary events, and shows that the relationship to the axiomatic events that define the cdf is slightly more involved than the scalar case. The video then considers the example of finding the joint-cdf from a joint-pdf. Following the video, the viewer should consider the problems in the tutorial exercise sheet.

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

- Properties of **joint-cdf**:

$$0 \leq F_{\mathbf{X}}(\mathbf{x}) \leq 1, \quad \lim_{\mathbf{x} \rightarrow -\infty} F_{\mathbf{X}}(\mathbf{x}) = 0, \quad \lim_{\mathbf{x} \rightarrow \infty} F_{\mathbf{X}}(\mathbf{x}) = 1 \quad (5.4)$$

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

$$F_{\mathbf{X}}(\mathbf{a}) \leq F_{\mathbf{X}}(\mathbf{b}) \quad \text{if } \mathbf{a} \leq \mathbf{b} \quad (5.5)$$

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**:

$$f_{\mathbf{X}}(\mathbf{x}) \geq 0, \quad \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = 1 \quad (5.6)$$

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 5.6.

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

$$\Pr(\mathbf{x}_1 < \mathbf{X}(\zeta) \leq \mathbf{x}_2) = \int_{\mathbf{x}_1}^{\mathbf{x}_2} f_{\mathbf{X}}(\mathbf{v}) d\mathbf{v} \neq F_{\mathbf{X}}(\mathbf{x}_2) - F_{\mathbf{X}}(\mathbf{x}_1) \quad (5.7)$$

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

Example 5.1 ([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:

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x + 3y) & 0 \leq \{x, y\} \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (5.8)$$

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

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

$$\int_{-\infty}^{\infty} f_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z} = \int_0^1 \int_0^1 \frac{1}{2}(x + 3y) dx dy = \int_0^1 \frac{1}{2} \left[\frac{1}{2}x^2 + 3xy \right]_0^1 dy \quad (5.9)$$

$$= \int_0^1 \frac{1}{4} + \frac{3}{2}y dy = \left[\frac{y}{4} + \frac{3y^2}{4} \right]_0^1 = \frac{1}{4} + \frac{3}{4} = 1 \quad (5.10)$$

The pdf and the region over which it is non-zero is shown in Figure 5.2.

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:

$$F_{\mathbf{Z}}(\mathbf{z}) = \int_{-\infty}^{\mathbf{z}} f_{\mathbf{Z}}(\bar{\mathbf{z}}) d\bar{\mathbf{z}} = \int_0^y \int_0^x \frac{1}{2}(\bar{x} + 3\bar{y}) d\bar{x} d\bar{y} \quad (5.11)$$

$$= \int_0^y \frac{1}{2} \left(\frac{x^2}{2} + 3x\bar{y} \right) d\bar{y} = \frac{1}{2} \left(\frac{x^2}{2}y + \frac{3xy^2}{2} \right) = \frac{xy}{4}(x + 3y) \quad (5.12)$$

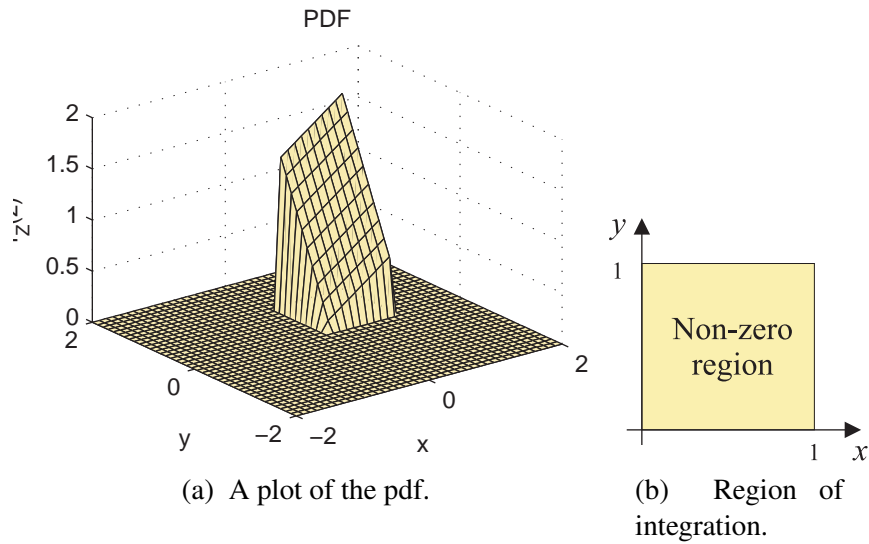


Figure 5.2: A plot of the probability density function, $f_{\mathbf{Z}}(\mathbf{z})$, for the problem in [Therrien:1992, Example 2.1, Page 20], and a figure showing the region over which the pdf is non-zero, which is the region of integration for calculating the cdf.

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:

$$F_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} 0 & x \leq 0 \text{ or } y \leq 0 \\ \frac{xy}{4}(x + 3y) & 0 < x, y \leq 1 \\ \frac{x}{4}(x + 3) & 0 < x \leq 1, 1 < y \\ \frac{y}{4}(1 + 3y) & 0 < y \leq 1, 1 < x \\ 1 & 1 < x, y < \infty \end{cases} \quad (5.13) \quad \square$$

The cdf is plotted in Figure 5.3.



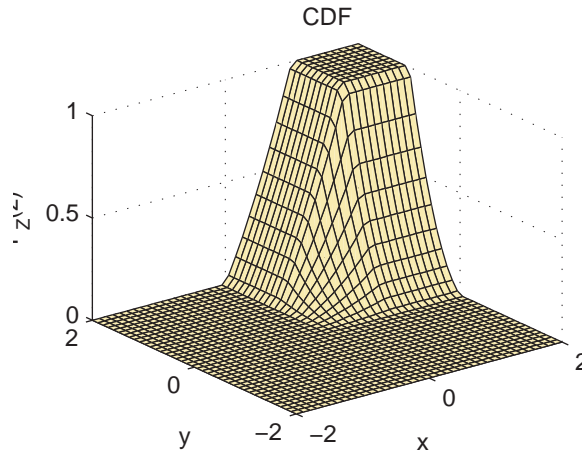


Figure 5.3: A plot of the cumulative distribution function, $F_{\mathbf{Z}}(\mathbf{z})$, for the problem in [Therrien:1992, Example 2.1, Page 20].

5.2.2 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:

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

and the variance is defined as:

$$\text{var}[X(\zeta)] = \sigma_X^2 = \mathbb{E}[|X(\zeta) - \mu_X|^2] \quad (\text{M:3.2.9})$$

which can be shown to equal

$$\text{var}[X(\zeta)] = \mathbb{E}[|X(\zeta)|^2] - |\mu_X|^2 \quad (\text{M:3.2.10})$$

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

$$\sigma_X^2 = \mathbb{E}[|X(\zeta) - \mu_X|^2] \quad (\text{M:3.2.9})$$

$$= \mathbb{E}[(X(\zeta) - \mu_X)^*(X(\zeta) - \mu_X)] \quad (5.14)$$

$$= \mathbb{E}[|X(\zeta)|^2 - \mu_X^* X(\zeta) - X^*(\zeta)\mu_X + |\mu_X|^2] \quad (5.15)$$

$$= \mathbb{E}[|X(\zeta)|^2] - \underbrace{\mu_X^* \mathbb{E}[X(\zeta)]}_{\mathbb{E}[|\mu_X|^2]} - \underbrace{\mathbb{E}[X^*(\zeta)] \mu_X}_{\mathbb{E}[|\mu_X|^2]} + |\mu_X|^2 \quad (5.16)$$

□

giving the desired result.

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

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

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 $\mathbf{X}(\zeta)$ is defined as:

$$F_{\mathbf{X}}(\mathbf{x}) \triangleq \Pr(\mathbf{X}(\zeta) \leq \mathbf{x}) \triangleq \Pr(\mathbf{X}_R(\zeta) \leq \mathbf{x}_r, \mathbf{X}_I(\zeta) \leq \mathbf{x}_i) \quad (\text{M:3.2.12})$$

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

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

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

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

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:

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

5.2.3 Marginal Density Function

Topic Summary 26 Marginal Distributions and Densities

Topic Objectives:

- Introduce notion of marginal density and distribution.
- Example of finding marginal-pdf and marginal-cdf from joint-pdf.
- Consider applications of marginals.

Topic Activities:

Type	Details	Duration	Progress
Watch video	11 : 38 min video	3× length	
Read Handout	Read page 143 to page 145	8 mins/page	
Try Example	Try Example 5.2	15 minutes	
Practice Exercises	Exercise ??	20 mins	

Marginal Density Function

Example (Marginalisation).

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

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

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

Note that the marginal-pdf is not a slice of the joint-pdf.
It is the integral of the joint-pdf over the other variable along a line whose position corresponds to the value of interest.

http://media.ed.ac.uk/media/1_abevu23q

Video Summary: This video discusses the marginal-pdf which describes the pdf of a subset of elements from the random vector. Examples of applications in which this is useful are described. A worked example of finding marginal-pdfs and marginal-cdfs from a joint-pdf is provided, along with plots of the functions. The viewer should verify the results presented in this video for themselves.

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 $M = 3$,

$$\mathbf{k} = [1 \quad 5 \quad 12]^T \quad (5.17)$$

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

$$\mathbf{k} = \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_M \end{bmatrix} \quad \text{then} \quad \mathbf{X}_{\mathbf{k}}(\zeta) = \begin{bmatrix} X_{k_1}(\zeta) \\ X_{k_2}(\zeta) \\ \vdots \\ X_{k_M}(\zeta) \end{bmatrix} \quad (5.18)$$

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

$$\mathbf{X}_{[1,5,12]}(\zeta) = \begin{bmatrix} X_1(\zeta) \\ X_5(\zeta) \\ X_{12}(\zeta) \end{bmatrix} \quad (5.19)$$

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

$$f_{\mathbf{X}_{\mathbf{k}}}(\mathbf{x}_{\mathbf{k}}) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{N-M \text{ integrals}} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}_{-\mathbf{k}} \quad (5.20)$$

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 :

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

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 5.2 (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:

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{2}(x + 3y) & 0 \leq \{x, y\} \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (5.21)$$

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:

$$f_X(x) = \int_{\mathbb{R}} f_{\mathbf{Z}}(\mathbf{z}) dy \quad (5.22)$$

$$f_Y(y) = \int_{\mathbb{R}} f_{\mathbf{Z}}(\mathbf{z}) dx \quad (5.23)$$

Taking $f_X(x)$, then:

$$f_X(x) = \begin{cases} \frac{1}{2} \int_0^1 (x + 3y) dy & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (5.24)$$

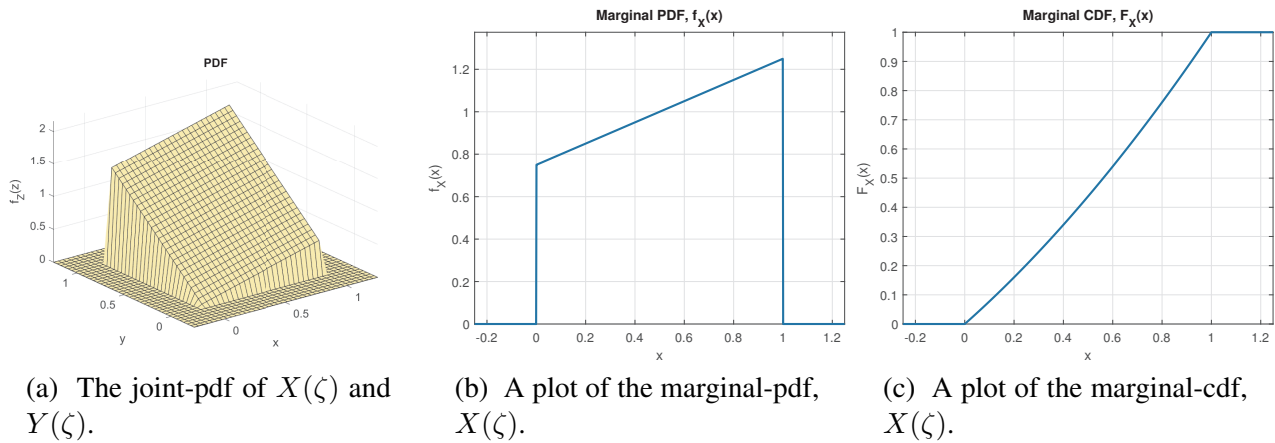


Figure 5.4: The marginal-pdf, $f_X(x)$, and cdf, $F_X(x)$, for the RV, $X(\zeta)$.

which after a simple integration gives:

$$f_X(x) = \begin{cases} \frac{1}{2} \left(x + \frac{3}{2}\right) & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (5.25)$$

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

$$F_X(x) = \int_{-\infty}^x f_X(u) du = \begin{cases} 0 & x \leq 0 \\ \frac{1}{2} \int_0^x \left(u + \frac{3}{2}\right) du & 0 \leq x \leq 1 \\ \frac{1}{2} \int_0^1 \left(u + \frac{3}{2}\right) du & x > 1 \end{cases} \quad (5.26)$$

Which after, again, a straightforward integration gives:

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ \frac{x}{4} (x + 3) & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases} \quad (5.27)$$

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

Similarly, it can be shown that:

$$f_Y(y) = \begin{cases} \frac{1}{2} \left(\frac{1}{2} + 3y\right) & 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (5.28)$$

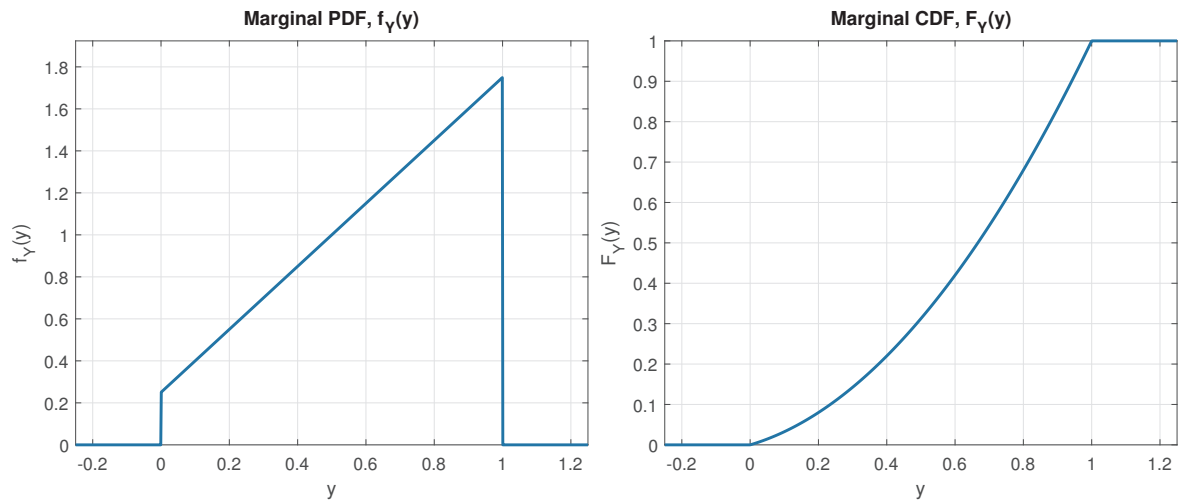
and

$$F_Y(y) = \begin{cases} 0 & y \leq 0 \\ \frac{y}{4} (1 + 3y) & 0 \leq y \leq 1 \\ 1 & y > 1 \end{cases} \quad (5.29)$$

The marginal-pdfs and cdfs are shown in Figure 5.4 and Figure 5.5 respectively.

KEYPOINT! (Interpretation). Note that the marginal-pdf is not a *slice* of the joint-pdf. Rather it is the integral of the joint-pdf over the other variable along a given line whose position corresponds to the value of the variable of interest.





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

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

Figure 5.5: The marginal-pdf, $f_Y(y)$, and cdf, $F_Y(y)$, for the RV, $Y(\zeta)$.

5.2.4 Independence

Topic Summary 27 Independence, Conditional, and Bayes's Theorem

Topic Objectives:

- The notion of independence and its applications.
- Conditional densities and Bayes Theorem.
- Examples of testing independence.
- Examples of using Bayes rule.

Topic Activities:

Type	Details	Duration	Progress
Watch video	18 : 59 min video	3× length	
Read Handout	Read page 147 to page 151	8 mins/page	
Try Example	Try Examples 5.4, 5.4, 5.5	30 minutes	
Practice Exercises	Exercises ?? and ??	40 mins	

Blind Signal Separation

Standard signal separation using the independent component assumption.

A powerful assumption is that the source signals are statistically independent of one another; independent component analysis (ICA).

REMARKS:
Lecture Slideset 3:
Multiple Random Variables
For full lecture notes, see PETARS course, Part I, Handout 5: "Random Vectors and Multiple Random Variables"

http://media.ed.ac.uk/media/1_gg8ef3ep

Video Summary: This video introduces the notions of independence, conditional densities, and Bayes's theorem. The use of independence in signal processing applications such as Blind Source Separation is introduced, although this will be expanded in future videos on Statistical Signal Processing. Analytical tests for independence given the pdf is considered for a couple of examples, including deriving the joint density for independent Gaussian random variables. Conditional densities are then introduced, and Bayes theorem for solving inverse problems is developed from this. The final section of the video then considers in detail the problem of estimating a parameter from a noisy observation.

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

$$\Pr(X_1(\zeta) \leq x_1, X_2(\zeta) \leq x_2) = \Pr(X_1(\zeta) \leq x_1) \Pr(X_2(\zeta) \leq x_2) \quad (5.30)$$

This then implies that

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

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

KEYPOINT! (Region of support). If the regions of support of the pdfs of $X(\zeta)$ and $Y(\zeta)$ are bounded, then $X(\zeta)$ and $Y(\zeta)$ cannot be independent if their ranges are dependent. Therefore, independence of $X(\zeta)$ and $Y(\zeta)$ requires the support of the joint-pdf, $f_{XY}(x, y)$ to be just the Cartesian product of the support of $f_X(x)$ and the support of $f_Y(y)$.

Example 5.3 (Testing independence). Suppose the joint-pdf of two RVs $X(\zeta)$ and $Y(\zeta)$ is given by $f_{XY}(x, y) = 1 + xy$ for $0 < x < 1$ and $0 < y < 1$. Are $X(\zeta)$ and $Y(\zeta)$ independent?

SOLUTION. The joint-pdf cannot be written in the form $g(x)h(y)$ for any functions g and h . Therefore, these RVs are not independent.

Example 5.4 (Testing independence). Let $f_{XY}(x, y) = 6x$ for $0 < x < y < 1$. Plot the region of support and determine if $X(\zeta)$ and $Y(\zeta)$ are independent.

As a side-up question, check that this is a valid pdf in the first place!

As an example that will be used many times in estimation theory, suppose that N RVs, $X_n(\zeta)$ for $n \in \{0, \dots, N-1\}$, are independent, and each have pdf given by $f_{X_n}(x_n)$. Then the joint-pdf of the random vector $\mathbf{X}(\zeta) = [X_0(\zeta), \dots, X_{N-1}(\zeta)]^T$ is given by:

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{n=0}^{N-1} f_{X_n}(x_n) \quad (5.31)$$

For example, suppose that $X_n(\zeta)$ is Gaussian distributed with zero-mean and unit variance, such that:

$$f_{X_n}(x_n) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_n^2}{2}} \quad (5.32)$$

then:

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_n^2}{2}} = \frac{1}{(2\pi)^{\frac{N}{2}}} e^{-\frac{1}{2} \sum_{n=0}^{N-1} x_n^2} \quad (5.33)$$

This form will be used extensively in developing **likelihood functions**.

5.2.5 Conditional Densities and Bayes's Theorem

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

$$\Pr(Y | X) = \frac{\Pr(X, Y)}{\Pr(X)} \quad (\text{T:2.35})$$



New slide

Defining the event X as:

$$X : \mathbf{x} \leq \mathbf{X}(\zeta) \leq \mathbf{x} + d\mathbf{x} \quad (\text{T:2.36})$$

and the event Y as:

$$Y : \mathbf{y} \leq \mathbf{Y}(\zeta) \leq \mathbf{y} + d\mathbf{y} \quad (\text{T:2.37})$$

then

$$\Pr(Y | X) = \frac{\Pr(\mathbf{x} \leq \mathbf{X}(\zeta) \leq \mathbf{x} + d\mathbf{x}, \mathbf{y} \leq \mathbf{Y}(\zeta) \leq \mathbf{y} + d\mathbf{y})}{\Pr(\mathbf{x} \leq \mathbf{X}(\zeta) \leq \mathbf{x} + d\mathbf{x})} \quad (5.34)$$

$$= \frac{f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) \prod d\mathbf{x} d\mathbf{y}}{f_{\mathbf{X}}(\mathbf{x}) \prod d\mathbf{x}} = \left\{ \frac{f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})} \right\} \prod d\mathbf{y} \quad (5.35)$$

$$\triangleq f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} | \mathbf{x}) \prod d\mathbf{y} \quad (5.36)$$

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

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

Note that

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

This emphasises that $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} | \mathbf{x})$ 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 $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} | \mathbf{x})$ 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: $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} | \mathbf{x}) = f_{\mathbf{Y}}(\mathbf{y})$. Hence, it follows that:

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

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:

$$\Pr(X, Y) = \Pr(X | Y) \Pr(Y) = \Pr(Y | X) \Pr(X) = \Pr(Y, X) \quad (5.37)$$

and therefore

$$\Pr(X | Y) = \frac{\Pr(Y | X) \Pr(X)}{\Pr(Y)} \quad (\text{T:2.42})$$

An analogous expression can be written for density functions. Since

$$f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} | \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} | \mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{YX}}(\mathbf{y}, \mathbf{x}) \quad (\text{T:2.43})$$

it follows

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} | \mathbf{y}) = \frac{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} | \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{f_{\mathbf{Y}}(\mathbf{y})} \quad (\text{T:2.44})$$

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

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

$$f_{\mathbf{Y}}(\mathbf{y}) = \int_{\mathbb{R}} f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) d\mathbf{x} = \int_{\mathbb{R}} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} | \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \quad (5.38)$$

then it follows

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \frac{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{\int_{\mathbb{R}} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}} \quad (\text{T:2.45})$$

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 $\mathbf{Y}(\zeta)$ is an observation of an experiment which depends on some unknown random vector $\mathbf{X}(\zeta)$; for example, $\mathbf{Y}(\zeta)$ is $\mathbf{X}(\zeta)$ observed in additive noise. Then given $\mathbf{X}(\zeta)$, it is easy to find the *likelihood* of $\mathbf{Y}(\zeta)$, which is represented by the density $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$; this is the **likelihood function**, and will again be introduced later in this course. The **prior density**, $f_{\mathbf{X}}(\mathbf{x})$, 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**, $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$, which is the density of the unseen random vector $\mathbf{X}(\zeta)$ given the observations $\mathbf{Y}(\zeta)$.

Example 5.5 (Bayes's Theorem (Papoulis, Example 6-42)). An unknown random phase $\Theta(\zeta)$ is *a priori* assumed to be uniformly distributed in the interval $[0, 2\pi)$. The phase is observed through a noisy sensor, such that $R(\zeta) = \Theta(\zeta) + N(\zeta)$, where $N(\zeta)$ is Gaussian distributed with zero mean and variance σ_N^2 .

What is the **posterior** pdf $f_{\Theta|R}(\theta|r)$, which gives the distribution of $\Theta(\zeta)$ given an observation?

SOLUTION. In practical situations, it is reasonable to assume that $\Theta(\zeta)$ and $N(\zeta)$ are independent. Using the probability transformation rule for scalar random variables, from $N(\zeta)$ to $R(\zeta) = \theta + N(\zeta)$ where $\Theta(\zeta) = \theta$ is considered fixed, it follows there is one inverse solution $n = r - \theta$, and the Jacobian of the transformation is unity. Therefore:

$$f_{R|\Theta}(r|\theta) = \frac{1}{1} f_N(r - \theta) = \frac{1}{\sqrt{2\pi\sigma_N^2}} e^{-\frac{(r-\theta)^2}{2\sigma_N^2}} \quad (5.39)$$

Using Bayes theorem, it directly follows that:

$$f_{\Theta|R}(\theta|r) = \frac{f_{R|\Theta}(r|\theta) f_{\Theta}(\theta)}{\int_0^{2\pi} f_{R|\Theta}(r|\hat{\theta}) f_{\Theta}(\hat{\theta}) d\hat{\theta}} \quad (5.40)$$

which, since $f_{\Theta}(\theta) = \frac{1}{2\pi}$ for $0 \leq \theta < 2\pi$, can be written as:

$$f_{\Theta|R}(\theta|r) = \frac{e^{-\frac{(r-\theta)^2}{2\sigma_N^2}}}{\int_0^{2\pi} e^{-\frac{(r-\theta)^2}{2\sigma_N^2}} d\theta} \quad 0 \leq \theta < 2\pi \quad (5.41) \quad \square$$

and zero otherwise, where it is noted that the factors $\frac{1}{2\pi}$ and $\frac{1}{\sqrt{2\pi\sigma_N^2}}$ have cancelled each other in the numerator and denominator.

Note the knowledge about the observation, r , is reflected in the posterior pdf of $\Theta(\zeta)$, as shown in Figure 5.6, and it shows higher probability density in the neighbourhood of $\Theta(\zeta) = r$.

Example 5.6 (Chapman-Kolmogorov Equation). Consider a state-space model with an unknown state \mathbf{x}_n and measurement vector \mathbf{y}_n .

Assume the Markov property that $p(\mathbf{x}_n | \mathbf{x}_{n-1}, \mathbf{y}_{1:n-1}) = p(\mathbf{x}_n | \mathbf{x}_{n-1})$ and $p(\mathbf{y}_n | \mathbf{x}_n, \mathbf{y}_{1:n-1}) = p(\mathbf{y}_n | \mathbf{x}_n)$.

Show that:

$$p(\mathbf{x}_n | \mathbf{y}_{1:n-1}) = \int p(\mathbf{x}_n | \mathbf{x}_{n-1}) p(\mathbf{x}_{n-1} | \mathbf{y}_{1:n-1}) d\mathbf{x}_{n-1} \quad (5.42)$$

$$p(\mathbf{x}_n | \mathbf{y}_{1:n}) = \frac{p(\mathbf{y}_n | \mathbf{x}_n) p(\mathbf{x}_n | \mathbf{y}_{1:n-1})}{p(\mathbf{y}_n | \mathbf{y}_{1:n-1})} \quad (5.43)$$

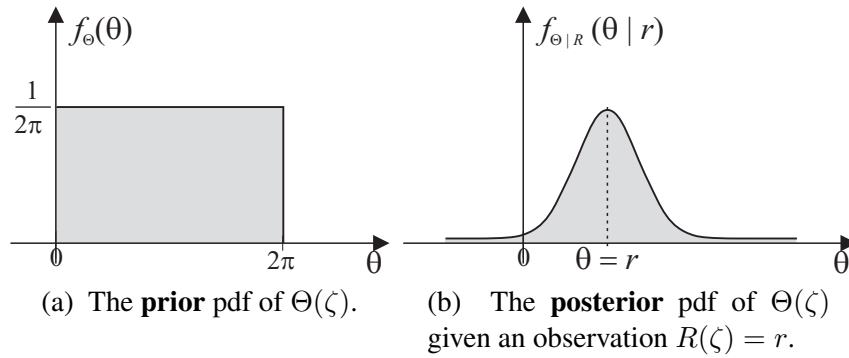


Figure 5.6: The knowledge about the observation r is reflected in the posterior pdf.

SOLUTION. The first equation is a direct application of marginalisation of a joint-pdf:

$$p(\mathbf{x}_n | \mathbf{y}_{1:n-1}) = \int p(\mathbf{x}_n, \mathbf{x}_{n-1} | \mathbf{y}_{1:n-1}) d\mathbf{x}_{n-1} \quad (5.44)$$

$$= \int p(\mathbf{x}_n | \mathbf{x}_{n-1}, \mathbf{y}_{1:n-1}) p(\mathbf{x}_{n-1} | \mathbf{y}_{1:n-1}) d\mathbf{x}_{n-1} \quad (5.45)$$

$$= \int p(\mathbf{x}_n | \mathbf{x}_{n-1}) p(\mathbf{x}_{n-1} | \mathbf{y}_{1:n-1}) d\mathbf{x}_{n-1} \quad (5.46)$$

using the Markov property.

The second equation is a direct application of Bayes's theorem keeping $\mathbf{y}_{1:n-1}$ a conditional in each term:

$$p(\mathbf{x}_n | \mathbf{y}_{1:n}) = p(\mathbf{x}_n | \mathbf{y}_n, \mathbf{y}_{1:n-1}) \quad (5.47)$$

$$= \frac{p(\mathbf{y}_n | \mathbf{x}_n, \mathbf{y}_{1:n-1}) p(\mathbf{x}_n | \mathbf{y}_{1:n-1})}{p(\mathbf{y}_n | \mathbf{y}_{1:n-1})} \quad (5.48)$$

□

and then using $p(\mathbf{y}_n | \mathbf{x}_n, \mathbf{y}_{1:n-1}) = p(\mathbf{y}_n | \mathbf{x}_n)$.

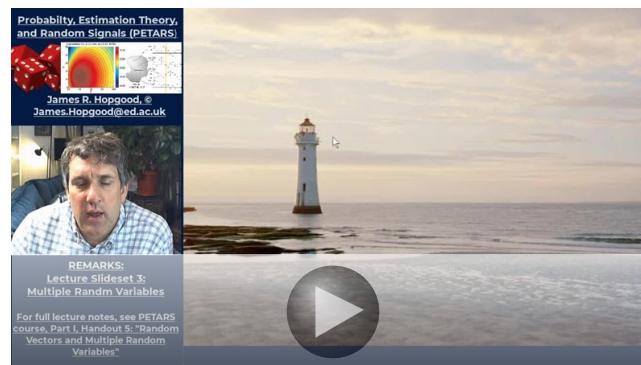


Topic Summary 28 Gull's Lighthouse Problem
Topic Objectives:

- Use all the techniques discussed in the course so far to address a simple inverse problem.
- Appreciate application of this techniques to localisation or tomography.
- Be aware of the importance of optimisation and integration in signal processing.

Topic Activities:

Type	Details	Duration	Progress
Watch video	22 : 29 min video	3× length	
Read Handout	Read page 152 to page 155	8 mins/page	
Try Example	Try Example 5.7	40 minutes	
Try Code	Use the MATLAB code	10 minutes	
Further Reading	Search for more on this problem	30 mins	



http://media.ed.ac.uk/media/1_yuj5go1d

Video Summary: Gull's lighthouse problem is a famous problem in tomography or localisation problem, that is an example of an inverse problem. This exercise uses all the knowledge gained so far in the course, including using probability transformations, conditional probabilities, independence, Bayes theorem, marginalisation and optimisation. This relatively simple problem is analysed systematically, with the various assumptions discussed. The video then finishes by discussing two key problems in signal processing: the problems of integration (for marginalisation of nuisance parameters), and optimisation (for finding estimators). Some example techniques for addressing these problems are then discussed.

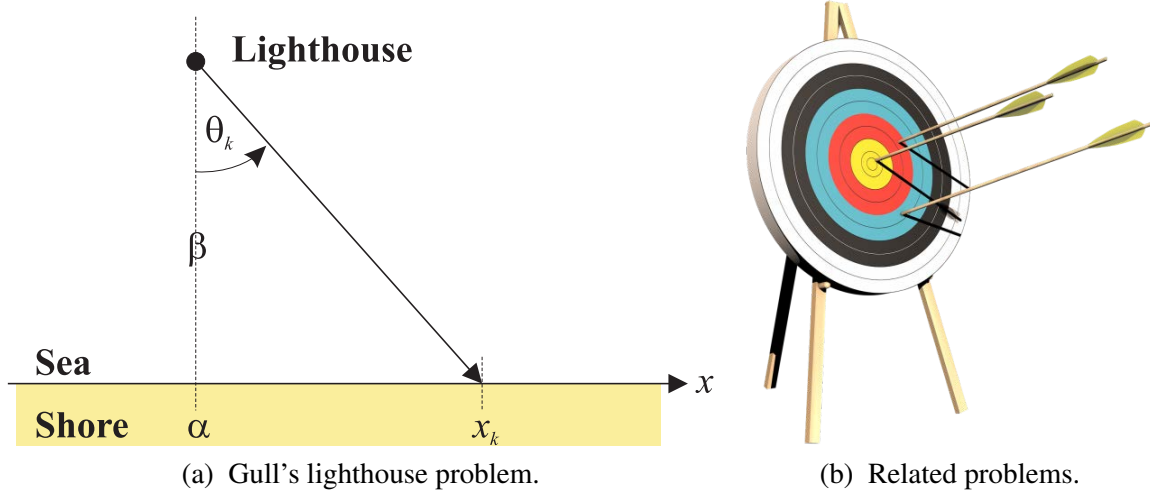


Figure 5.7: The geometry of the lighthouse problem, and related problems.

Example 5.7 (Gull's 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 of arrival from which it came. N flashes have so far been recorded at positions $\{x_k\}$. Where is the lighthouse?

KEYPOINT! (Other Forms). This problem can be phrased in a number of other ways, such as throwing darts randomly at a wall and so forth. It is essentially a tomography problem, and is a classic inverse problem.

It can also be phrased as a geolocation problem, and there are a number of articles on this topic if you search the web!

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 5.7.

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,

$$f_{\Theta}(\theta) = \begin{cases} \frac{1}{\pi} & -\frac{\pi}{2} < \theta < \frac{\pi}{2} \\ 0 & \text{otherwise} \end{cases} \quad (5.49)$$

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 5.7 shows that:

$$\beta \tan \theta = x - \alpha \quad (5.50)$$

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

$$f_X(x | \alpha) = \frac{\beta}{\pi [\beta^2 + (x - \alpha)^2]} \quad (5.51)$$

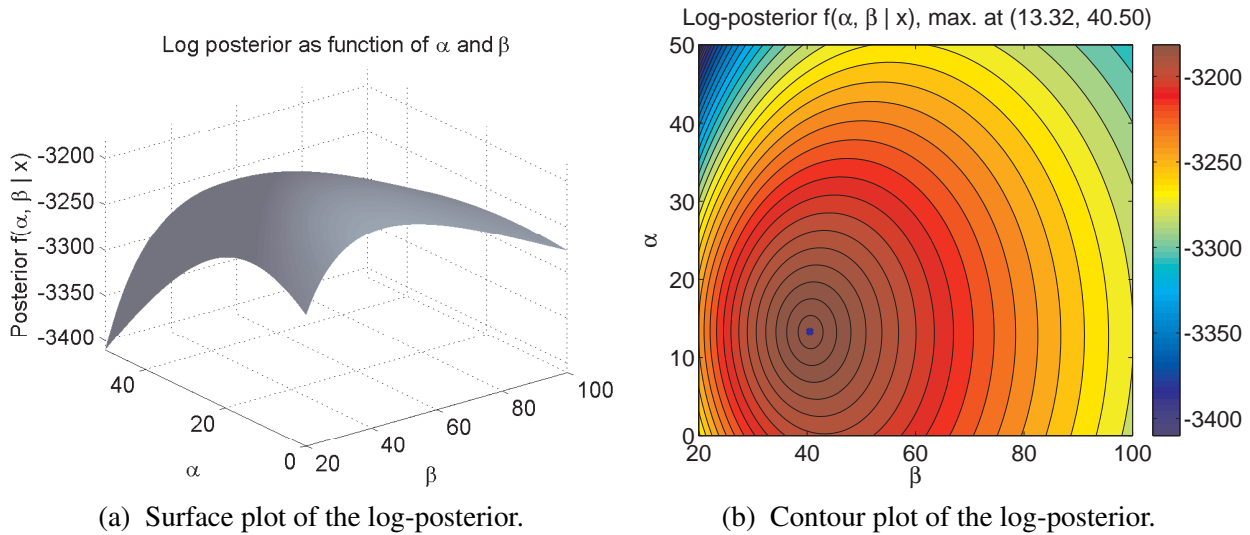


Figure 5.8: Visualising the log-posterior function described in Equation 5.56 when both α and β are unknown. In this case, the number of data-points used is $N = 500$. The actual lighthouse location is at $(\alpha, \beta) = (15, 45)$. Note the error in the estimate of the maximum value.

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:

$$\begin{aligned}
 f_{\mathbf{X}}(\mathbf{x} | \alpha) &= f_{\mathbf{X}}(x_1, \dots, x_N | \alpha) = \prod_{k=1}^N f_X(x_k | \alpha) \\
 &= \prod_{k=1}^N \frac{\beta}{\pi [\beta^2 + (x_k - \alpha)^2]}
 \end{aligned} \tag{5.52}$$

The position of the lighthouse is then expressed by:

$$f_A(\alpha | \mathbf{x}) = \frac{f_{\mathbf{X}}(\mathbf{x} | \alpha) f_A(\alpha)}{f_{\mathbf{X}}(\mathbf{x})} \tag{5.53}$$

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

$$f_A(\alpha) = \begin{cases} \frac{1}{\alpha_{\max} - \alpha_{\min}} & \alpha_{\min} \leq \alpha \leq \alpha_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{5.54}$$

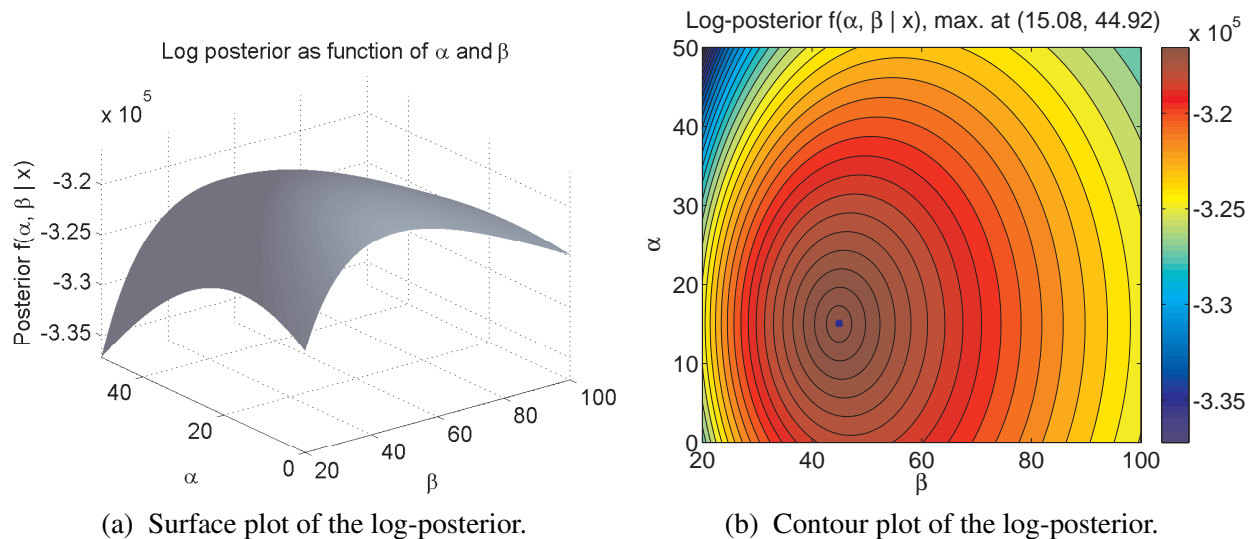
Hence, it follows that:

$$f_A(\alpha | \mathbf{x}) = \frac{f_{\mathbf{X}}(\mathbf{x} | \alpha) f_A(\alpha)}{f_{\mathbf{X}}(\mathbf{x})} \propto f_{\mathbf{X}}(\mathbf{x} | \alpha) f_A(\alpha) \tag{5.55}$$

$$\propto \frac{1}{\alpha_{\max} - \alpha_{\min}} \prod_{k=1}^N \frac{\beta}{\pi [\beta^2 + (x_k - \alpha)^2]}, \quad \text{for } \alpha_{\min} \leq \alpha \leq \alpha_{\max} \quad \square \tag{5.56}$$

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 5.56 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)



(a) Surface plot of the log-posterior.

(b) Contour plot of the log-posterior.

Figure 5.9: Visualising the log-posterior function described in Equation 5.56 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 $N = 500$.

function is shown in Figure 5.8 and Figure 5.9 for when the lighthouse 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.

KEYPOINT! (Key Problems). This example highlights two key problems in Signal Processing:

Integration Marginalising out nuisance parameters:

$$f_A(\alpha | \mathbf{x}) = \int f_A(\alpha, \beta | \mathbf{x}) d\beta \quad (5.57)$$

Optimisation Finding the maximum marginal *a posteriori* (MMAP) estimate:

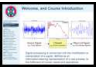
$$\hat{\alpha} = \arg_{\alpha} \max f_A(\alpha | \mathbf{x}) \quad (5.58)$$

□

– End-of-Topic 28: **Tomography: An Inverse Problem using Probability Transformations, Conditional Probability, Independence, Bayes Theorem, Marginalisation, and Optimisation.** –



5.3 Probability Transformation Rule



Topic Summary 29 Probability Transformation Rule for Random Vectors

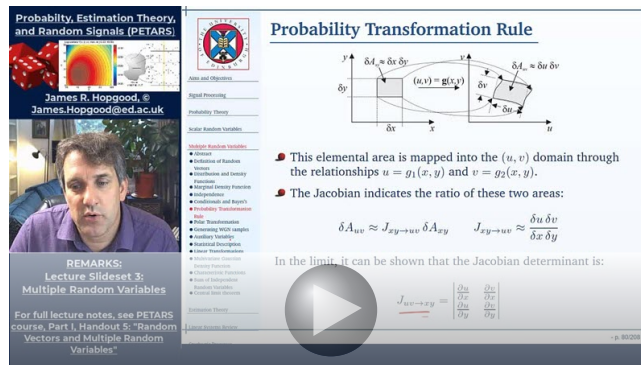
New slide

Topic Objectives:

- Extend probability transformation rule to random vectors.
- Understand what a Jacobian is and how to calculate it.
- Application to the Cartesian to Polar transformation.

Topic Activities:

Type	Details	Duration	Progress
Watch video	18 : 18 min video	3 × length	
Read Handout	Read page 156 to page 160	8 mins/page	
Practice Exercises	Exercises ?? and ??	20 mins	



http://media.ed.ac.uk/media/1_7rnjbf3t

Video Summary: This video extends the probability transformation rule from the scalar case to the vector case for vector functions of random vectors. The video discusses how the Jacobian determinant needs to be calculated instead of a simple gradient, and therefore this can influence whether the Jacobian or its inverse should be calculated depending on the ease of calculating the derivatives for the mapping or inverse mapping. The video provides a reminder of the physical interpretation of the Jacobian. Finally, the video considers the probability transformation for the Cartesian to Polar coordinate mappings.

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

Theorem 5.1 (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:

$$Y_n(\zeta) = g_n(\mathbf{X}(\zeta)), \quad n \in \mathcal{N} \tag{5.61}$$

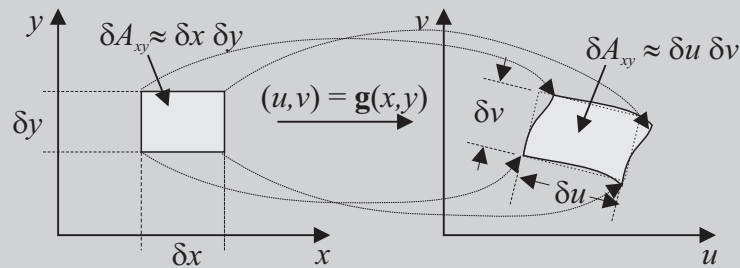
or, using an alternative notation,

$$\mathbf{Y}(\zeta) = \mathbf{g}(\mathbf{X}(\zeta)) \tag{5.62}$$

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

Sidebar 7 Jacobian

The Jacobian determinant is used to represent how an elemental region in one domain changes volume when it is mapped to another domain. Consider the elemental area $\delta A_{xy} = \delta x \delta y$ in the (x, y) domain:



This elemental area is mapped into the (u, v) domain through the relationships $u = g_1(x, y)$ and $v = g_2(x, y)$. The elemental area in the (u, v) domain is approximately given by $\delta A_{uv} = \delta u \delta v$. The Jacobian determinant indicates the ratio of these two elemental areas, namely:

$$\delta A_{uv} \approx J_{xy \rightarrow uv} \delta A_{xy} \quad J_{xy \rightarrow uv} \approx \frac{\delta u \delta v}{\delta x \delta y} \quad (5.59)$$

In the limit, it can be shown that the Jacobian determinant, or just the Jacobian, is given by :

$$J_{uv \rightarrow xy} = \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial y} & \frac{\partial v}{\partial y} \end{vmatrix} \quad (5.60)$$

Informally, the Jacobian can be considered as the multi-dimensional version of the scalar derivative $\left| \frac{dy}{dx} \right|$.

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

$$\mathbf{y} = \mathbf{g}(\mathbf{x}_1) = \cdots = \mathbf{g}(\mathbf{x}_M) \quad (5.63)$$

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

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{m=1}^M \frac{f_{\mathbf{X}}(\mathbf{x}_m)}{|J(\mathbf{x}_m)|} \quad (5.64)$$

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

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

It should also be noted, from vector calculus results, that the Jacobian can also be expressed as:

$$\frac{1}{J_{\mathbf{g}}(\mathbf{x})} \triangleq \frac{\partial(x_1, \dots, x_N)}{\partial(y_1, \dots, y_N)} = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} & \cdots & \frac{\partial x_N}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_N}{\partial y_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_1}{\partial y_N} & \frac{\partial x_2}{\partial y_N} & \cdots & \frac{\partial x_N}{\partial y_N} \end{vmatrix} \quad (\text{T:2.123})$$

For further background information on the Jacobian, see Sidebar 7.

PROOF. The proof follows a very similar line to that for the scalar RVs case. The definition of the joint-**pdf** is:

$$f_{\mathbf{Y}}(\mathbf{y}) \prod d\mathbf{y} = \Pr(\mathbf{y} < \mathbf{Y}(\zeta) \leq \mathbf{y} + d\mathbf{y}) \quad (5.65)$$

where $\prod d\mathbf{y} = dy_1 dy_2 \dots dy_N$. The set of values \mathbf{x} such that $\mathbf{y} < \mathbf{g}(\mathbf{x}) \leq \mathbf{y} + d\mathbf{y}$, consists of the intervals:

$$\mathbf{x}_m < \mathbf{x} \leq \mathbf{x}_m + d\mathbf{x}_m \quad (5.66)$$

The probability that \mathbf{x} lies in this set is, of course,

$$f_{\mathbf{X}}(\mathbf{x}_m) \prod d\mathbf{x}_m = \Pr(\mathbf{x}_m < \mathbf{X}(\zeta) \leq \mathbf{x}_m + d\mathbf{x}_m) \quad (5.67)$$

Moreover, the transformation from \mathbf{x} to \mathbf{y} is given by the Jacobian:

$$\prod d\mathbf{y} = J_{\mathbf{g}}(\mathbf{x}) \prod d\mathbf{x} \quad (5.68)$$

Since these are mutually exclusive sets, then

$$\Pr(\mathbf{y} < \mathbf{Y}(\zeta) \leq \mathbf{y} + d\mathbf{y}) = \sum_{m=1}^M \Pr(\mathbf{x}_m < \mathbf{X}(\zeta) \leq \mathbf{x}_m + d\mathbf{x}_m) \quad (5.69)$$

$$= \sum_{m=1}^M f_{\mathbf{X}}(\mathbf{x}_m) \frac{\prod d\mathbf{y}}{J_{\mathbf{g}}(\mathbf{x}_m)} \quad (5.70) \quad \square$$

and thus the desired result is obtained after minor rearrangement.

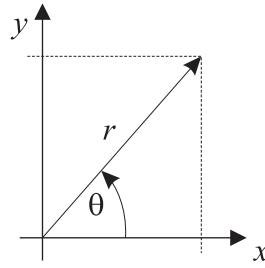


Figure 5.10: The Cartesian and polar coordinate systems.

5.3.1 Polar Transformation

An important transformation example is the mapping from Cartesian to polar coordinates. Each of these coordinates are shown in Figure 5.10.

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

$$\begin{aligned} r(\zeta) &= \sqrt{X^2(\zeta) + Y^2(\zeta)} \\ \theta(\zeta) &= \arctan \frac{Y(\zeta)}{X(\zeta)} \end{aligned} \quad (5.71)$$

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:

$$\left. \begin{aligned} x &= r \cos \theta \\ y &= r \sin \theta \end{aligned} \right\} \text{ for } r > 0 \quad (5.72)$$

The Jacobian is given by:

$$J_{\mathbf{g}}(\mathbf{c}) = \frac{\partial(r, \theta)}{\partial(x, y)} = \begin{vmatrix} \frac{\partial \theta}{\partial x} & \frac{\partial r}{\partial x} \\ \frac{\partial \theta}{\partial y} & \frac{\partial r}{\partial y} \end{vmatrix} = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix}^{-1} \quad (5.73)$$

In the case of polar transformations, $J_{\mathbf{g}}(\mathbf{c})$ simplifies to:

$$J_{\mathbf{g}}(\mathbf{c}) = \begin{vmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{vmatrix}^{-1} = \frac{1}{r} \quad (5.74)$$

Thus, it follows that:

$$f_{R, \Theta}(r, \theta) = r f_{XY}(r \cos \theta, r \sin \theta) \quad (5.75)$$

Example 5.8 (Cartesian to polar transformation of RVs). If $X(\zeta)$ and $Y(\zeta)$ are independent and identically distributed (i. i. d.) Gaussian distributed coordinates in Cartesian space, such that $X(\zeta), Y(\zeta) \sim \mathcal{N}(0, \sigma^2)$, find the distribution when these are transformed into polar coordinates.

SOLUTION. First, note:

$$f_{XY}(x, y) = f_X(x) f_Y(y) = \frac{1}{2\pi\sigma^2} \exp \left\{ -\frac{x^2 + y^2}{2\sigma^2} \right\} \quad (5.76)$$

Hence, applying the transformation $r = \sqrt{x^2 + y^2}$, $\theta = \arctan \frac{y}{x}$, it directly follows that

$$f_{R\Theta}(r, \theta) = \frac{r}{2\pi\sigma^2} \exp \left\{ -\frac{r^2}{2\sigma^2} \right\} \mathbb{I}_{[-\pi, \pi]}(\theta) \mathbb{I}_{\mathbb{R}^+}(r) \quad (5.77)$$

where, as a reminder, $\mathbb{I}_{\mathcal{A}}(a) = 1$ 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:

$$f_R(r) = \frac{r}{\sigma^2} \exp\left\{-\frac{r^2}{2\sigma^2}\right\} \mathbb{I}_{\mathbb{R}^+}(r) \quad \text{and} \quad f_{\Theta}(\theta) = \frac{1}{2\pi} \mathbb{I}_{[-\pi, \pi]}(\theta) \quad (5.78)$$

where the scaling factors have been apportioned such that these are proper densities, in the sense that $\int_{\mathbb{R}} f_R(r) dr = \int_{\mathbb{R}} f_{\Theta}(\theta) d\theta = 1$. Note that θ is uniformly distributed, while r has a **Rayleigh distribution**. □

– End-of-Topic 29: **Probability Transformation rule for Random Vectors** –



5.3.2 Generating Gaussian distributed samples

Topic Summary 30 Generating Gaussian Samples

Topic Objectives:

- Investigate Box-Muller transformations for generating Gaussians.
- Use probability transformation rule to prove this result.
- Understand importance of simulating random variables.

Topic Activities:

Type	Details	Duration	Progress
Watch video	14 : 04 min video	3× length	
Read Handout	Read page 161 to page 164	8 mins/page	
Try the code	Use MATLAB code on LEARN	20 mins	

http://media.ed.ac.uk/media/1_g0nvuf4r

Video Summary: In this video, the probability transformation rule is used to show that the Box-Muller transformation can convert two uniform random variables into two independent Gaussian random variables. Although one aim of this video is to provide as another example of how to use the probability transformation rule, it also motivates the discussion about tools that can be used for simulating random numbers from various distributions.

It is often important to generate samples from a Gaussian density, primarily for simulation studies. In practice, it is difficult for a computer to generate random numbers from an arbitrary density. However, it is possible to generate uniform random variates fairly easily. This will be seen in later handouts.

The **probability transformation rule** can be used to take random variables from one distribution as inputs, and outputs random variables in a new distribution function. 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,

$$f_{X_k}(x_k) = \mathbb{I}_{0,1}(x_k), \quad k = 1, 2 \quad (5.79)$$

where $\mathbb{I}_{\mathcal{A}}(x) = 1$ if $x \in \mathcal{A}$, and zero otherwise.

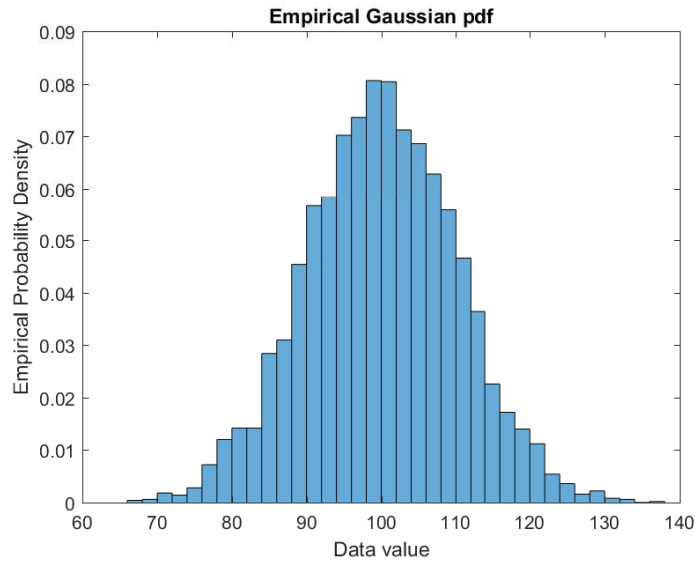


Figure 5.11: This histogram shows an empirical Gaussian probability density function, where the samples are drawn from a Gaussian density. But how are these samples drawn?

Now let two random variables y_1, y_2 be given by:

$$y_1 = \sqrt{-2 \ln x_1} \cos 2\pi x_2 \quad (5.80)$$

$$y_2 = \sqrt{-2 \ln x_1} \sin 2\pi x_2 \quad (5.81)$$

It follows, by rearranging these equations, that:

$$x_1 = \exp \left[-\frac{1}{2}(y_1^2 + y_2^2) \right] \quad (5.82)$$

$$x_2 = \frac{1}{2\pi} \arctan \frac{y_2}{y_1} \quad (5.83)$$

The Jacobian determinant can be calculated as:

$$J(x_1, x_2) = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{vmatrix} \quad (5.84)$$

$$= \begin{vmatrix} \frac{-1}{x_1 \sqrt{-2 \ln x_1}} \cos 2\pi x_2 & -2\pi \sqrt{-2 \ln x_1} \sin 2\pi x_2 \\ \frac{-1}{x_1 \sqrt{-2 \ln x_1}} \sin 2\pi x_2 & 2\pi \sqrt{-2 \ln x_1} \cos 2\pi x_2 \end{vmatrix} \quad (5.85)$$

$$= \frac{2\pi}{x_1} \quad (5.86)$$

Hence, it follows:

$$f_Y(y_1, y_2) = \frac{x_1}{2\pi} = \left[\frac{1}{\sqrt{2\pi}} e^{-y_1^2/2} \right] \left[\frac{1}{\sqrt{2\pi}} e^{-y_2^2/2} \right] \quad (5.87)$$

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 the pdfs of y_1 alone and y_2 alone, and therefore each y_k 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.

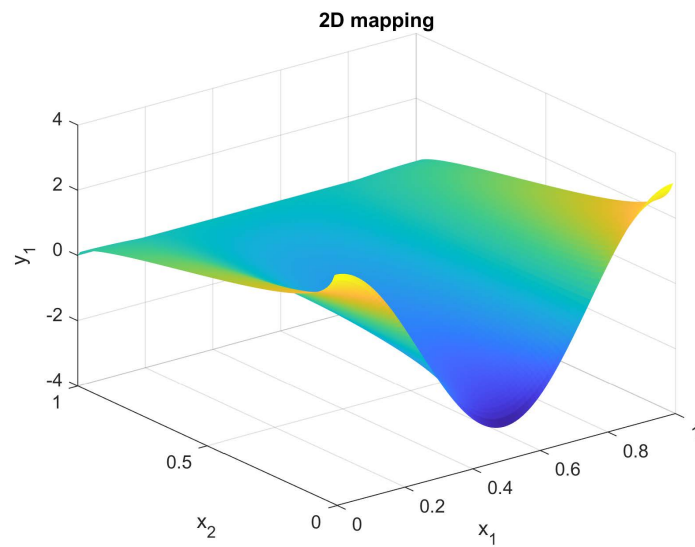


Figure 5.12: The first Box-Muller transformation.

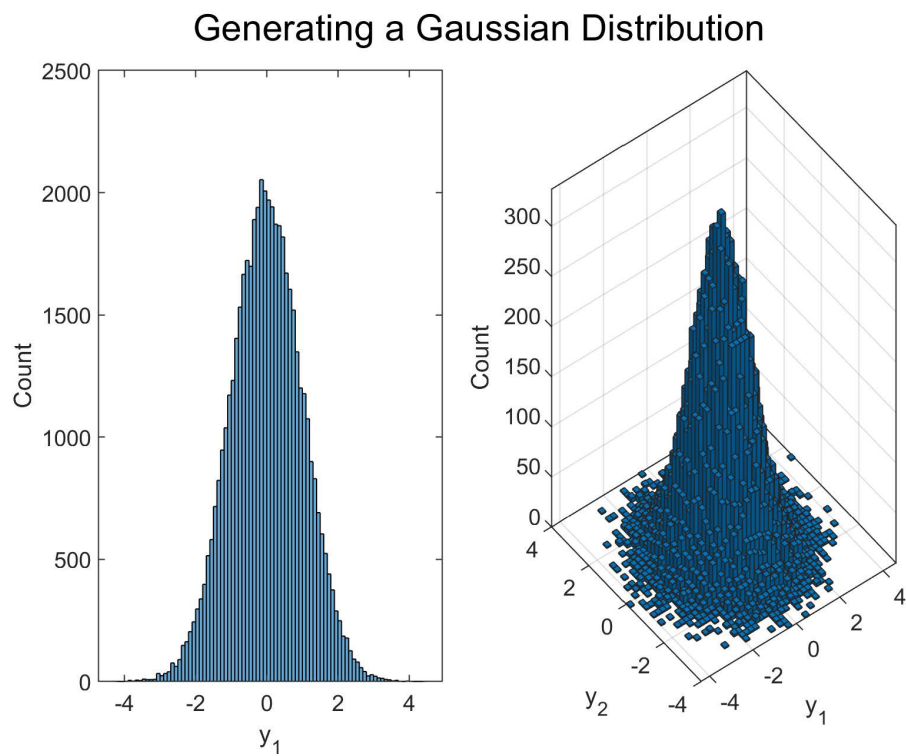


Figure 5.13: The resulting histogram from the generation of these Gaussian samples.

– End-of-Topic 30: **Generating Gaussian Samples** –



5.3.3 Auxiliary Variables

Topic Summary 31 Auxiliary Variables

Topic Objectives:

- Use auxiliary variables for functions of multiple random variables.
- Applications of one random variable as a function of two random variables.
- Choices of auxiliary variable.
- Detailed examples to demonstrate application of technique.

Topic Activities:

Type	Details	Duration	Progress
Watch video	21 : 25 min video	3× length	
Read Handout	Read page 165 to page 168	8 mins/page	
Try Example	Try Examples 5.9 and 5.10	25 minutes	

Auxiliary Variables $z = aX + bY$

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.

$$\begin{aligned} w &= X \\ x &= w \\ y &= \frac{z - aw}{b} \end{aligned}$$

$$J = \begin{vmatrix} a & b \\ 1 & 0 \end{vmatrix} = -b$$

$$f_{wz}(w, z) = \frac{1}{|b|} f_{xy}\left(w, \frac{z - aw}{b}\right)$$

$$f_z(z) = \frac{1}{|b|} \int f_{xy}\left(w, \frac{z - aw}{b}\right) dw$$

$$f_z(z) = \frac{1}{|b|} \int f_{xy}\left\{\frac{z - b\hat{w}}{a}, \hat{w}\right\} \frac{b}{a} d\hat{w}$$

http://media.ed.ac.uk/media/1_5n9ox5os

Video Summary: Auxiliary variables are introduced as a method for calculating a single function of multiple random variables, through a two-stage process of using the probability transformation rule followed by marginalisation. While there are alternative methods for calculating a single function of multiple random variables, the auxiliary variable method is very algorithmic. More generally, the auxiliary variable method is for transforming N random variables to M random variables, where $M < N$. The video presents several problems of varying complexities and choice of auxiliary variables.

So far, when considering functions of random variables, the problem of transforming from $NRVs$ to NRV s has been considered. However, what about the case of transforming from $NRVs$ to $MRVs$, where $M < N$; for example, $Z(\zeta) = g(X(\zeta), Y(\zeta))$?

Note that in the case of $M > N$ need not be considered, as in this case, it can be shown that multiple variables are deterministically related, or 100% correlated.

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**, $W(\zeta)$. 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 the **probability transformation rule**,

$$f_{WZ}(w, z) dw = \sum_{m=1}^M \frac{f_{\mathbf{XY}}(x_m, y_m)}{|J(x_m, y_m)|} \quad (5.88)$$

followed by **marginalisation**:

$$f_Z(z) = \int_{\mathbb{R}} f_{WZ}(w, z) dw = \sum_{m=1}^M \int_{\mathbb{R}} \frac{f_{\mathbf{XY}}(x_m, y_m)}{|J(x_m, y_m)|} dw \quad (5.89)$$

Example 5.9 (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:

$$J(x, y) = \begin{vmatrix} \frac{\partial w}{\partial x} & \frac{\partial z}{\partial x} \\ \frac{\partial w}{\partial y} & \frac{\partial z}{\partial y} \end{vmatrix} = \begin{vmatrix} 0 & a \\ 1 & b \end{vmatrix} = -a \quad (5.90)$$

Hence, it follows that:

$$f_{WZ}(w, z) = \frac{1}{|a|} f_{XY}\left(\frac{z - bw}{a}, w\right) \quad (5.91)$$

Thus, it follows that:

$$f_Z(z) = \frac{1}{|a|} \int_{\mathbb{R}} f_{XY}\left(\frac{z - bw}{a}, w\right) dw \quad (5.92) \quad \square$$

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 auxiliary 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 8 on page 167. Nevertheless, it usually pays to choose the auxiliary variable carefully to minimise any difficulties in evaluating the marginal-pdf.

As an example, consider using $W(\zeta) = X(\zeta) - Y(\zeta)$ in the previous example (Example 5.9).

Example 5.10 ([Papoulis:1991, Page 149, Problem 6-8]). The RVs $X(\zeta)$ and $Y(\zeta)$ are independent with Rayleigh densities:

$$f_X(x) = \frac{x}{\alpha^2} \exp\left\{-\frac{x^2}{2\alpha^2}\right\} \mathbb{I}_{\mathbb{R}^+}(x) \quad (5.102)$$

$$f_Y(y) = \frac{y}{\beta^2} \exp\left\{-\frac{y^2}{2\beta^2}\right\} \mathbb{I}_{\mathbb{R}^+}(y) \quad (5.103)$$

1. Show that if $Z(\zeta) = X(\zeta)/Y(\zeta)$, then:

$$f_Z(z) = \frac{2\alpha^2}{\beta^2} \frac{z}{\left(z^2 + \frac{\alpha^2}{\beta^2}\right)^2} \mathbb{I}_{\mathbb{R}^+}(z) \quad (5.104)$$

Sidebar 8 What if you chose a complicated auxiliary variable?

Consider Example 5.9 and suppose that rather than choosing $W(\zeta) = Y(\zeta)$, you accidentally chose something more complicated such as:

$$W(\zeta) = \frac{X(\zeta)}{Y(\zeta)} \quad (5.93)$$

Will the resulting expression for $f_Z(z)$ be the same as Equation 5.92? 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 5.91, it is the **marginalisation** stage that ensures the expressions for $f_Z(z)$ are the same. For the auxiliary variable shown in Equation 5.93, noting that $Z(\zeta) = aX(\zeta) + bY(\zeta)$, then

$$x = wy \Rightarrow z = awy + by = y(aw + b) \quad (5.94)$$

$$y = \frac{z}{aw + b}, \quad x = \frac{wz}{aw + b} \quad (5.95)$$

The Jacobian is given by:

$$J = \text{abs} \begin{bmatrix} \frac{\partial z}{\partial x} & \frac{\partial z}{\partial y} \\ \frac{\partial w}{\partial x} & \frac{\partial w}{\partial y} \end{bmatrix} = \text{abs} \begin{bmatrix} a & b \\ \frac{1}{y} & -\frac{x}{y^2} \end{bmatrix} \quad (5.96)$$

$$= \text{abs} \frac{ax + by}{y^2} = \text{abs} \frac{z}{y^2} = \text{abs} \frac{(aw + b)^2}{z} \quad (5.97)$$

For simplicity, assume that $(x, y) > 0$.^a Then, the joint-pdf is given by:

$$f_{WZ}(w, z) = \frac{z}{(b + aw)^2} f_{XY} \left(\frac{wz}{aw + b}, \frac{z}{aw + b} \right) \quad (5.98)$$

This is clearly different to that in Equation 5.91. However, the marginal for $Z(\zeta)$ is:

$$f_Z(z) = \int \frac{z}{(b + aw)^2} f_{XY} \left(\frac{wz}{aw + b}, \frac{z}{aw + b} \right) dw \quad (5.99)$$

Let $\theta = \frac{z}{aw + b}$, such that $d\theta = -\frac{az}{(aw + b)^2} dw$, and also note that

$$\frac{wz}{aw + b} = \theta w = \theta \left(\frac{z - b\theta}{\theta a} \right) = \frac{z - b\theta}{a} \quad (5.100)$$

Substituting into Equation 5.99, and noting that the minus sign in the differential term will get absorbed into the limits of the integral, then Equation 5.99 becomes:

$$f_Z(z) = \frac{1}{a} \int f_{XY} \left(\frac{z - b\theta}{a}, \theta \right) d\theta \quad (5.101)$$

which is indeed equivalent to Equation 5.92.

^aThis 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.

2. Using this result, show that for any $k > 0$,

$$\Pr(X(\zeta) \leq kY(\zeta)) = \frac{k^2}{k^2 + \frac{\alpha^2}{\beta^2}} \quad (5.105)$$

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:

$$J(x, y) = \text{abs} \begin{vmatrix} \frac{\partial w}{\partial x} & \frac{\partial z}{\partial x} \\ \frac{\partial w}{\partial y} & \frac{\partial z}{\partial y} \end{vmatrix} = \text{abs} \begin{vmatrix} 1 & \frac{1}{y} \\ 0 & -\frac{x}{y^2} \end{vmatrix} = \text{abs} \left| -\frac{x}{y^2} \right| = \left| \frac{z^2}{w} \right| \quad (5.106)$$

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

$$J(x, y) = \frac{z^2}{w} \quad (5.107)$$

Hence, since $X(\zeta)$ and $Y(\zeta)$ are independent,

$$f_{WZ}(w, z) = \frac{w}{z^2} f_X(w) f_Y\left(\frac{w}{z}\right) \quad (5.108)$$

$$= \frac{1}{\alpha^2 \beta^2} \frac{w^3}{z^3} \exp \left\{ -\frac{w^2}{2} \left(\frac{1}{\alpha^2} + \frac{1}{z^2 \beta^2} \right) \right\} \mathbb{I}_{\mathbb{R}^+ \times \mathbb{R}^+}(w, z) \quad (5.109)$$

$$= \frac{\hat{\alpha}^2}{z^3 \alpha^2 \beta^2} \left[w^2 \frac{w}{\hat{\alpha}^2} \exp \left\{ -\frac{w^2}{2\hat{\alpha}^2} \right\} \right] \mathbb{I}_{\mathbb{R}^+ \times \mathbb{R}^+}(w, z) \quad (5.110)$$

where $\hat{\alpha}^2 = \alpha^2 \frac{z^2}{z^2 + \frac{\alpha^2}{\beta^2}}$. Integrating over all values of w gives:

$$f_Z(z) = \int_{\mathbb{R}^+} f_{XZ}(w, z) dw = \frac{\hat{\alpha}^2}{z^3 \alpha^2 \beta^2} \int_0^\infty w^2 \frac{w}{\hat{\alpha}^2} \exp \left\{ -\frac{w^2}{2\hat{\alpha}^2} \right\} dw \quad (5.111)$$

The integral is the **second moment** of a Rayleigh distribution. It can be shown that

$$\int_0^\infty w^2 \frac{w}{\hat{\alpha}^2} \exp \left\{ -\frac{w^2}{2\hat{\alpha}^2} \right\} dw = 2\hat{\alpha}^2 \quad (5.112)$$

Finally, therefore,

$$f_Z(z) = \frac{2\hat{\alpha}^4}{z^3 \alpha^2 \beta^2} \mathbb{I}_{\mathbb{R}^+}(z) = \frac{2\alpha^2}{\beta^2} \frac{z}{\left(z^2 + \frac{\alpha^2}{\beta^2}\right)^2} \mathbb{I}_{\mathbb{R}^+}(z) \quad (5.113)$$

For the second part of the question, notice that:

$$\Pr(X(\zeta) \leq kY(\zeta)) = \Pr(Z(\zeta) \leq k) = \int_0^k f_Z(z) dz \quad (5.114)$$

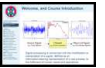
$$= \frac{\alpha^2}{\beta^2} \int_0^k \frac{2z}{\left(z^2 + \frac{\alpha^2}{\beta^2}\right)^2} dz = -\frac{\alpha^2}{\beta^2} \left[\frac{1}{z^2 + \frac{\alpha^2}{\beta^2}} \right]_0^k \quad (5.115)$$

$$= \frac{\alpha^2}{\beta^2} \left[\frac{1}{\frac{\alpha^2}{\beta^2}} - \frac{1}{k^2 + \frac{\alpha^2}{\beta^2}} \right] = 1 - \frac{\frac{\alpha^2}{\beta^2}}{k^2 + \frac{\alpha^2}{\beta^2}} \quad (5.116) \quad \square$$

which gives the desired result when these fractions are combined.



5.4 Statistical Description



Topic Summary 32 Statistical Description of Random Vectors

New slide

Topic Objectives:

- Appreciate the notion of correlation between random variables.
- Understand the details of the mean vector and correlation matrix.
- Calculate mean vector and correlation matrix from a joint-pdf.
- Awareness of statistical orthogonality.

Topic Activities:

Type	Details	Duration	Progress
Watch video	23 : 53 min video	3 × length	
Read Handout	Read page 169 to page 175	8 mins/page	
Try Example	Try Examples 5.11 and 5.12	20 minutes	
Practice Exercise	Exercise ??	25 mins	

http://media.ed.ac.uk/media/1_vxk6rqpd

Video Summary: This video extends the concept of statistical descriptors of pdfs to random vectors or multiple random variables. It introduces the concept of correlation, and how this relates to the dependency of the random variables. The mean vector and correlation matrix are introduced in detail with careful attention to the exact meaning of these expectations. An example of calculating these values for a given joint-pdf is covered carefully. Finally, the notion of statistical orthogonality is mentioned, although this will be covered in another video.

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.

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.

In particular, note that using second-order moments of individual RVs does not adequately capture the key characteristics of the joint-pdf. For example, as shown in Figure 5.14, two very different joint-pdfs

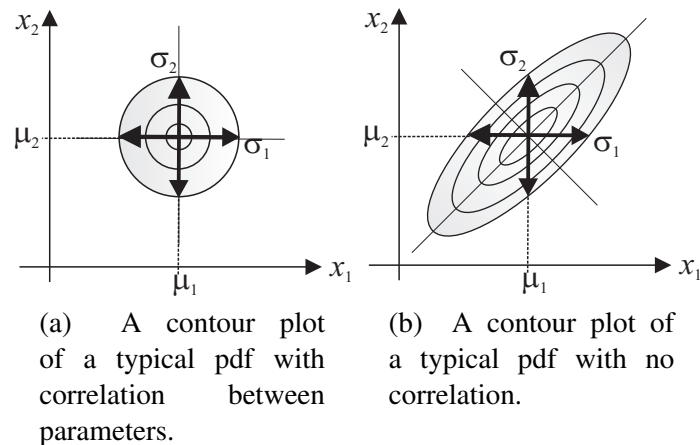
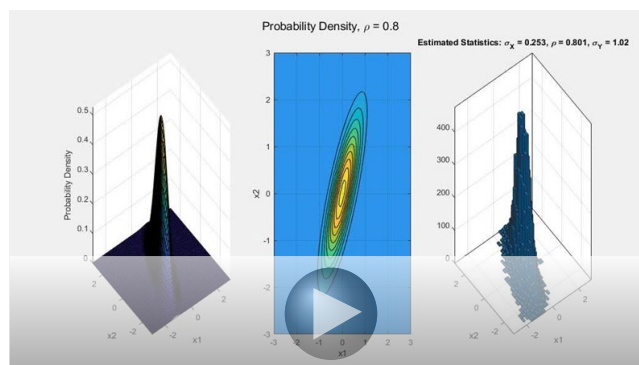


Figure 5.14: Mean and second-moments of individual RVs does not capture all of the information about the joint-pdf.

can have the same position and spread measurements, if only considered from the perspective of the cartesian axis representing the random variables. As will be see, other statistical descriptors are need to catch the richer information in a multi-dimensional pdf. Further examples of how the a joint-pdf relates to the key statistical feature of correlation is shown in Figure 5.15.

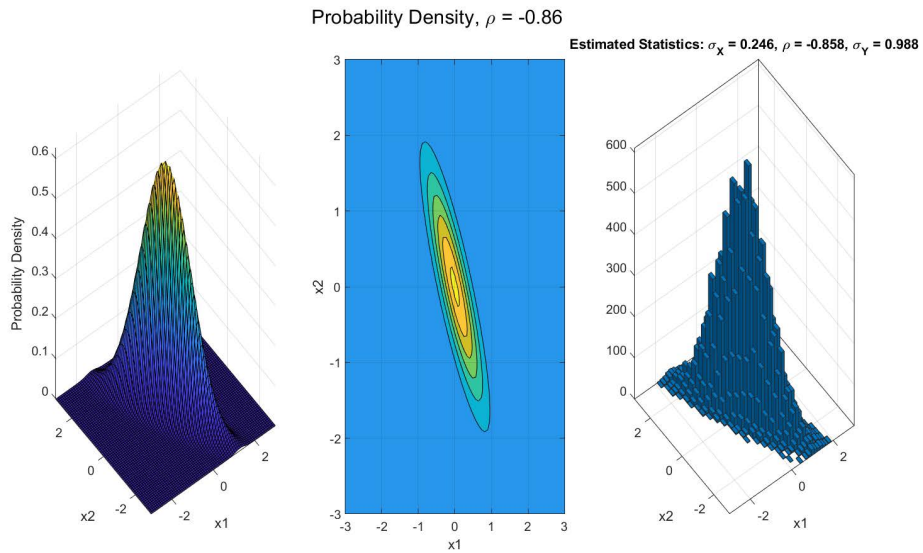


http://media.ed.ac.uk/media/1_gi3z1zp9

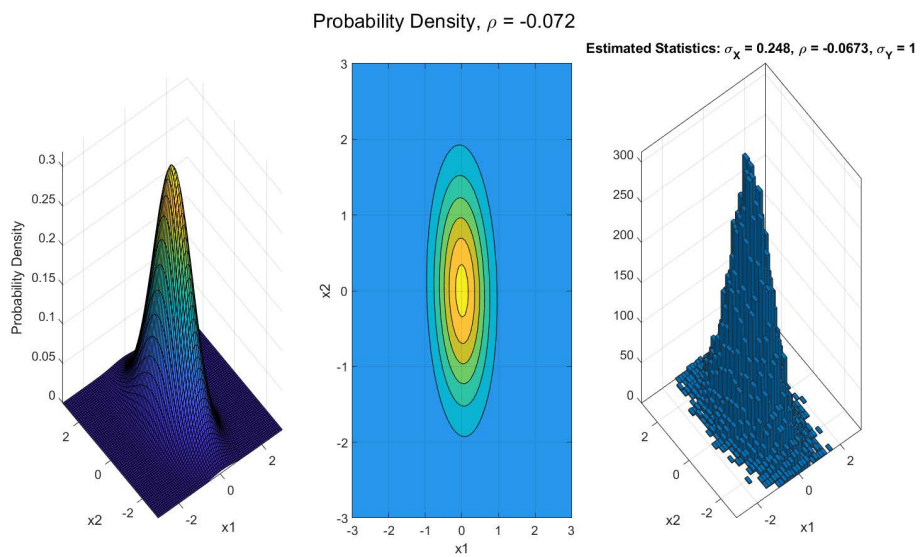
Video Summary: An animation showing how a bi-variate Gaussian changes with correlation term.

Consequently, 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**.



(a) A joint-pdf with strong correlation.



(b) A joint-pdf with very weak correlation, so almost independent.

Figure 5.15: Relating correlation to the description of the joint-pdf.

Sidebar 9 Elaborating on the Mean Vector for Real Random Vectors

The mean-vector, when written as the expectation $\boldsymbol{\mu}_{\mathbf{X}} = \mathbb{E}[\mathbf{X}(\zeta)]$, has a lot of hidden steps involved. This Sidebar deals with real random vectors; to deal with complex random vectors, it is necessary to extend this discussion by integrating over the real and imaginary elements separately. First, note that the mean vector can be written as:

$$\boldsymbol{\mu}_{\mathbf{X}} = \mathbb{E}_{f(\mathbf{x})}[\mathbf{X}(\zeta)] = \int \mathbf{x} f(\mathbf{x}) d\mathbf{x} \quad (5.117)$$

$$= \int \cdots \int \mathbf{x} f(x_1, x_2, \dots, x_N) dx_1 dx_2 \cdots dx_N \quad (5.118)$$

where $d\mathbf{x} \equiv \prod_{i=1}^N dx_i = dx_1 dx_2 \cdots dx_N$ and the multi-dimensional integral has been expanded. Therefore, note that this integral is a vector multiplied by a scalar function, and that $d\mathbf{x}$ isn't in this context considered as a vector. Thus, it follows that:

$$\boldsymbol{\mu}_{\mathbf{X}} = \int \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} f(\mathbf{x}) d\mathbf{x} \quad (5.119)$$

$$= \begin{bmatrix} \int x_1 f(\mathbf{x}) d\mathbf{x} \\ \vdots \\ \int x_N f(\mathbf{x}) d\mathbf{x} \end{bmatrix} \quad (5.120)$$

Note that the k -th row can be simplified as $\mu_k = [\boldsymbol{\mu}_{\mathbf{X}}]_k$:

$$\int x_k f(\mathbf{x}) d\mathbf{x} = \underbrace{\int \cdots \int}_{N \text{ integrals}} x_k f(\mathbf{x}) dx_1 \dots dx_{k-1} dx_{k+1} \dots dx_N \quad (5.121)$$

$$= \int x_k \left\{ \underbrace{\int \cdots \int}_{N-1 \text{ integrals}} f(\mathbf{x}) dx_1 \dots dx_{k-1} dx_{k+1} \dots dx_N \right\} dx_k$$

$$= \int x_k f(x_k) dx_k = \mathbb{E}[X_k(\zeta)] \quad (5.122)$$

which results from the marginalisation formula earlier. It therefore yields the results given in the definition of the mean vector, namely that the mean vector is the vector of the means of the individual elements.

Note that the element $d\mathbf{x}$ does depend on context, and in some cases, should in fact be interpreted as:

$$d\mathbf{x} = \begin{bmatrix} dx_1 \\ dx_2 \\ \vdots \\ dx_N \end{bmatrix} \quad (5.123)$$

This alternative definition will be introduced when appropriate.

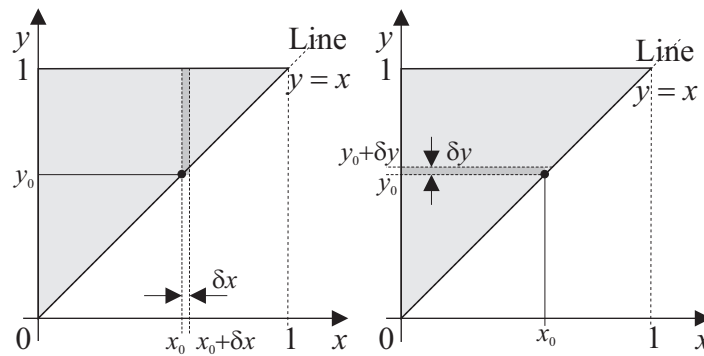


Figure 5.16: The region-of-support for the pdf in Example 5.11.

5.4.1 Mean Vectors and Correlation Matrices

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:

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

Further discussion on the mean-vector is given in Sidebar 9.

Example 5.11 (Mean Vector). This question follows up on Example 5.4 which introduced a simple pdf that clearly had dependency between the random variables. This example is similar, but different, so that a numerical example is easily generated in MATLAB.

Let $f_{XY}(x, y) = 2$ for $0 < x < y < 1$ and zero otherwise. Find the mean-vector.

SOLUTION. The calculation involves finding the marginals and then the expected value. Using the region-of-support for this problem as shown in Figure 5.16, then:

$$f_X(x) = \int_{y=x}^1 f_{XY}(x, y) dy = \int_x^1 2 dy = 2(1-x) \quad (5.124)$$

$$f_Y(y) = \int_{x=0}^y f_{XY}(x, y) dx = \int_0^y 2 dx = 2y \quad (5.125)$$

Taking expectations then gives:

$$\mu_X = \int_0^1 x f_X(x) dx = \int_0^1 2x(1-x) dx \quad (5.126)$$

$$\mu_X = 2 \left[\frac{x^2}{2} - \frac{x^3}{3} \right]_0^1 = \frac{1}{3} \quad (5.127)$$

$$\mu_Y = \int_0^1 y f_Y(y) dy = 2 \int_0^1 y^2 dy = 2 \left[\frac{y^3}{3} \right]_0^1 = \frac{2}{3} \quad (5.128)$$

□

This can be verified with the following very simple code:

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

$$\mathbf{R}_{\mathbf{X}} \triangleq \begin{bmatrix} \mathbb{E} [X_1(\zeta)X_1^*(\zeta)] & \cdots & \mathbb{E} [X_1(\zeta)X_N^*(\zeta)] \\ \vdots & \ddots & \cdots \\ \mathbb{E} [X_N(\zeta)X_1^*(\zeta)] & \cdots & \mathbb{E} [X_N(\zeta)X_N^*(\zeta)] \end{bmatrix} \quad (5.129)$$

or, more succinctly,

$$\mathbf{R}_{\mathbf{X}} \triangleq \mathbb{E} [\mathbf{X}(\zeta) \mathbf{X}^H(\zeta)] = \begin{bmatrix} r_{X_1X_1} & \cdots & r_{X_1X_N} \\ \vdots & \ddots & \vdots \\ r_{X_NX_1} & \cdots & r_{X_NX_N} \end{bmatrix} \quad (\text{M:3.2.17})$$

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

$$\mathbf{A}^H = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{2M} \\ \vdots & \vdots & \cdots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NM} \end{bmatrix}^H = \begin{bmatrix} a_{11}^* & a_{21}^* & \cdots & a_{N1}^* \\ a_{12}^* & a_{22}^* & \cdots & a_{N2}^* \\ \vdots & \vdots & \ddots & \vdots \\ a_{1M}^* & a_{2M}^* & \cdots & a_{NM}^* \end{bmatrix} \in \mathbb{C}^{M \times N} \quad (5.130)$$

The diagonal terms

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

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

The off-diagonal terms

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

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

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

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

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

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

Example 5.12 (Correlation Matrix). Following on from Example 5.11, find the correlation matrix for random variables with joint-pdf given by $f_{XY}(x, y) = 2$ for $0 < x < y < 1$ and zero otherwise.

SOLUTION. The second-moments can utilise the marginals calculated in Example 5.11, such that:

$$\mathbb{E} [X^2(\zeta)] = \int_0^1 x^2 f_X(x) dx = \int_0^1 2x^2(1-x) dx \quad (5.131)$$

$$= 2 \left[\frac{x^3}{3} - \frac{x^4}{4} \right]_0^1 = \frac{1}{6} \quad (5.132)$$

$$\mathbb{E} [Y^2(\zeta)] = \int_0^1 y^2 f_Y(y) dy = 2 \int_0^1 y^3 dy = 2 \left[\frac{y^4}{4} \right]_0^1 = \frac{1}{2} \quad (5.133)$$

The correlation terms are given by:

$$\mathbb{E} [X(\zeta) Y(\zeta)] = \int_0^1 \int_0^y xy f_{XY}(xy) dx dy \quad (5.134)$$

$$2 \int_0^1 y \int_0^y x dx dy = 2 \int_0^1 y \left[\frac{x^2}{2} \right]_0^y dy \quad (5.135)$$

$$= \int_0^1 y^3 dy = \left[\frac{y^4}{4} \right]_0^1 = \frac{1}{4} \quad (5.136)$$

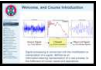
This correlation matrix can be evaluated by the MATLAB expression in addition to the code in Example 5.11,: Hence, putting all of these calculations together gives the correlation matrix:

$$\mathbf{R}_{XY} = \begin{bmatrix} r_{XX} & r_{XY} \\ r_{YX} & r_{YY} \end{bmatrix} = \begin{bmatrix} \frac{1}{6} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix} \quad (5.137)$$

□

– End-of-Topic 32: **Key Statistical definitions** –





5.4.2 Properties of Correlation Matrices

Topic Summary 33 Properties of Correlation Matrices

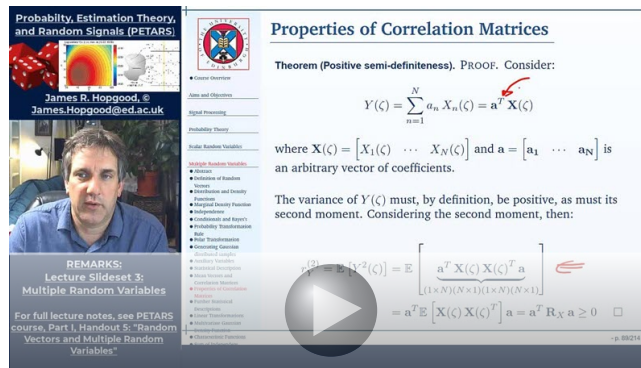
New slide

Topic Objectives:

- Understand properties that a valid correlation matrix must satisfy.
- Understand how to calculate positive semi-definiteness.
- Test several matrices to see if they are valid correlation matrices.

Topic Activities:

Type	Details	Duration	Progress
Watch video	15 : 50 min video	3× length	
Read Handout	Read page 176 to page 179	8 mins/page	
Try Example	Try Examples 5.13, 5.14, and 5.15	30 minutes	
Practice Exercise	Exercises ?? and ??	25 mins	



http://media.ed.ac.uk/media/1_xcdewkq9

Video Summary: This video considers the properties of valid correlation matrix, including the Hermitian property, positive semi-definiteness, and positive real-valued leading diagonal. This video proves these results, and shows how to test for positive semi-definiteness. The video then continues with a few examples, testing whether several matrices are valid correlation matrices or not.

It should be noticed that the **correlation** matrix is positive semidefinite; that is, the correlation matrices satisfies the relation:

$$\mathbf{a}^H \mathbf{R}_X \mathbf{a} \geq 0 \tag{T:2.65}$$

for any complex vector \mathbf{a} .

$$\mathbf{a}^H \mathbf{R}_X \mathbf{a} = \mathbf{a}^H \mathbb{E} [\mathbf{x}\mathbf{x}^H] \mathbf{a} = \mathbb{E} [|\mathbf{x}^H \mathbf{a}|^2] \tag{5.138}$$

A more detailed proof is given in Theorem 5.2. Note that a Hermitian matrix is semi-positive definite if all its eigenvalues are greater than or equal to zero. Moreover, note that if \mathbf{R}_X is real, then the expressions simplify somewhat to replacing \mathbf{a} with a real value, as shown in Sidebar 10. Hence, using the transpose rather than the Hermitian, such that $\mathbf{a}^T \mathbf{R}_X \mathbf{a} \geq 0$.

Sidebar 10 Positive semi-definiteness of Real Matrices

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

$$\mathbf{a} = \mathbf{a}_R + j\mathbf{a}_I \quad (5.142)$$

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

$$\mathcal{I} = \mathbf{a}^H \Gamma \mathbf{a} = (\mathbf{a}_R + j\mathbf{a}_I)^H (\Gamma \mathbf{a}_R + j\Gamma \mathbf{a}_I) \quad (5.143)$$

$$= \mathbf{a}_R^T (\Gamma \mathbf{a}_R + j\Gamma \mathbf{a}_I) - j\mathbf{a}_I^T (\Gamma \mathbf{a}_R + j\Gamma \mathbf{a}_I) \quad (5.144)$$

$$= \mathbf{a}_R^T \Gamma \mathbf{a}_R + j\mathbf{a}_R^T \Gamma \mathbf{a}_I - j\mathbf{a}_I^T \Gamma \mathbf{a}_R + \mathbf{a}_I^T \Gamma \mathbf{a}_I \quad (5.145)$$

$$= \mathbf{a}_R^T \Gamma \mathbf{a}_R + \mathbf{a}_I^T \Gamma \mathbf{a}_I + j(\mathbf{a}_R^T \Gamma \mathbf{a}_I - \mathbf{a}_I^T \Gamma \mathbf{a}_R) \quad (5.146)$$

Now, noting that \mathcal{I} is a scalar quantity, and with $\Gamma = \Gamma^T$, \mathcal{I} is also a real scalar quantity. Hence, as \mathcal{I} cannot have any imaginary terms, the last term above disappears and therefore $\mathbf{a}_R^H \Gamma \mathbf{a}_I = \mathbf{a}_I^H \Gamma \mathbf{a}_R$, such that:

$$\mathcal{I} = \mathbf{a}^T \Gamma \mathbf{a} = \mathbf{a}_R^T \Gamma \mathbf{a}_R + \mathbf{a}_I^T \Gamma \mathbf{a}_I \quad (5.147)$$

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 $\mathbf{a}_I = \mathbf{0}$.

Theorem 5.2 (Positive semi-definiteness of correlation matrix). Covariance and correlation matrices are positive semi-definite.

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

$$Y(\zeta) = \sum_{n=1}^N a_n X_n(\zeta) = \mathbf{a}^T \mathbf{X}(\zeta) \quad (5.139)$$

where $\mathbf{X}(\zeta) = [X_1(\zeta) \ \cdots \ X_N(\zeta)]^T \in \mathbb{R}^{N \times 1}$ and $\mathbf{a} = [a_1 \ \cdots \ a_N]^T \in \mathbb{R}^{N \times 1}$ is a non-random but arbitrary vector of coefficients.

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

$$r_Y^{(2)} = \mathbb{E} [Y^2(\zeta)] = \mathbb{E} [Y(\zeta) Y(\zeta)] = \mathbb{E} \left[\underbrace{\mathbf{a}^T \mathbf{X}(\zeta) \mathbf{X}(\zeta)^T \mathbf{a}}_{(1 \times N)(N \times 1)(1 \times N)(N \times 1)} \right] \quad (5.140)$$

$$= \mathbf{a}^T \mathbb{E} [\mathbf{X}(\zeta) \mathbf{X}(\zeta)^T] \mathbf{a} = \mathbf{a}^T \mathbf{R}_X \mathbf{a} \geq 0 \quad (5.141)$$

□

A similar expression can be obtained for the covariance matrix.

Example 5.13 (Valid correlation matrix). Determine whether the following is a valid correlation matrix:

$$\mathbf{R}_X = \begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix} \quad (5.148)$$

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

Example 5.14 (Valid correlation matrix). Determine whether the following is a valid correlation matrix:

$$\mathbf{R}_X = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \quad (5.149)$$

SOLUTION. Writing out the product $I = \mathbf{a}^T \mathbf{R}_X \mathbf{a}$ gives:

$$I = [\alpha \quad \beta] \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \quad (5.150)$$

$$= [\alpha \quad \beta] \begin{bmatrix} \alpha + 2\beta \\ 2\alpha + \beta \end{bmatrix} \quad (5.151)$$

$$= \alpha(\alpha + 2\beta) + \beta(2\alpha + \beta) \quad (5.152)$$

$$= \underbrace{\alpha^2 + 4\alpha\beta + \beta^2}_{\text{look to complete the square}} \quad (5.153)$$

$$I = \underbrace{\alpha^2 + 2\alpha\beta + \beta^2}_{\text{complete the square}} + 2\alpha\beta \quad (5.154)$$

$$= \underbrace{(\alpha + \beta)^2}_{\text{always positive}} + 2\alpha\beta \quad (5.155) \quad \square$$

Noting that the term $2\alpha\beta$ is not always positive, then selecting $\alpha = -\beta$, it follows that $I = -2\alpha^2 < 0$. Hence, \mathbf{R}_X is not a positive semi-definite matrix, and is therefore not a correlation matrix.

Example 5.15 ([Manolakis:2001, Exercise 3.14, Page 145]). Determine whether the following matrices are valid correlation matrices:

$$\mathbf{R}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \mathbf{R}_2 = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & 1 \end{bmatrix} \quad (5.156)$$

$$\mathbf{R}_3 = \begin{bmatrix} 1 & 1-j \\ 1+j & 1 \end{bmatrix} \quad \mathbf{R}_4 = \begin{bmatrix} 1 & \frac{1}{2} & 1 \\ \frac{1}{2} & 2 & \frac{1}{2} \\ 1 & 1 & 1 \end{bmatrix} \quad (5.157)$$

SOLUTION. Correlation 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

$$\mathbf{a}^T \mathbf{R}_1 \mathbf{a} = [a_1 \quad a_2] \begin{bmatrix} a_1 + a_2 \\ a_1 + a_2 \end{bmatrix} = a_1^2 + 2a_1a_2 + a_2^2 = (a_1 + a_2)^2 \geq 0 \quad (5.158)$$

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

$$\mathbf{a}^T \mathbf{R}_2 \mathbf{a} = [a_1 \quad a_2 \quad a_3] \begin{bmatrix} a_1 + \frac{a_2}{2} + \frac{a_3}{4} \\ \frac{a_1}{2} + a_2 + \frac{a_3}{2} \\ \frac{a_1}{4} + \frac{a_2}{2} + a_3 \end{bmatrix} \quad (5.159)$$

$$= a_1^2 + a_1 a_2 + \frac{1}{2} a_1 a_3 + a_2^2 + a_2 a_3 + a_3^2 \quad (5.160)$$

$$= \frac{1}{2} (a_1 + a_2 + a_3)^2 + \frac{1}{2} (a_1 - \frac{1}{2} a_3)^2 + \frac{1}{2} a_2^2 + \frac{3}{8} a_3^2 \geq 0 \quad (5.161)$$

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

$$\mathbf{a}^H \mathbf{R}_3 \mathbf{a} = [a_1^* \quad a_2^*] \begin{bmatrix} a_1 + (1-j)a_2 \\ (1+j)a_1 + a_2 \end{bmatrix} \quad (5.162)$$

$$= |a_1|^2 + (1-j)a_1^* a_2 + (1+j)a_2^* a_1 + |a_2|^2 \quad (5.163)$$

$$= |a_1 + (1-j)a_2|^2 - |a_2|^2 \quad (5.164)$$

□

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.

– End-of-Topic 33: **Positive Semi-Definiteness for Correlation Matrices**





5.4.3 Further Statistical Descriptions

Topic Summary 34 Further Statistical Descriptors

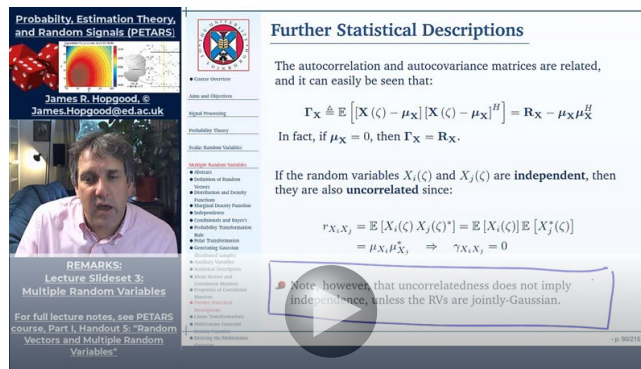
New slide

Topic Objectives:

- Define and understand the covariance matrix and its properties.
- Define and understand the correlation coefficient and its properties.
- Understand the notion of cross-correlation and cross-covariance matrices.
- Example of using cross-correlation for sum of random vectors.

Topic Activities:

Type	Details	Duration	Progress
Watch video	17 : 13 min video	3 × length	
Read Handout	Read page 180 to page 182	8 mins/page	
Try Example	Try Example 5.16	5 minutes	



http://media.ed.ac.uk/media/1_2vfmik1w

Video Summary: This video builds on the Statistical Descriptors introduced in Topic 32 by discussing the covariance matrix, the correlation coefficient, and then the cross-correlation and cross-covariance matrices for multiple random vectors. The properties of these matrices are discussed, including uncorrelatedness, followed by an example of calculating the cross-correlation of the sum of random vectors.

Building on the notes from the previous sections.

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

$$\Gamma_{\mathbf{X}} \triangleq \mathbb{E} \left[(\mathbf{X}(\zeta) - \boldsymbol{\mu}_{\mathbf{X}}) (\mathbf{X}(\zeta) - \boldsymbol{\mu}_{\mathbf{X}})^H \right] = \begin{bmatrix} \gamma_{X_1 X_1} & \cdots & \gamma_{X_1 X_N} \\ \vdots & \ddots & \vdots \\ \gamma_{X_N X_1} & \cdots & \gamma_{X_N X_N} \end{bmatrix} \quad (\text{M:3.2.20})$$

The diagonal terms

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

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

The off-diagonal terms

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

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

It can easily be shown that the **covariance** matrix, $\mathbf{\Gamma}_X$, must also be positive-semi definite, and is also a Hermitian matrix.

$$\mathbf{a}^H \mathbf{\Gamma}_X \mathbf{a} \geq 0\tag{T:2.65}$$

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

$$\rho_{X_i X_j} \triangleq \frac{\gamma_{X_i X_j}}{\sigma_{X_i} \sigma_{X_j}} = \rho_{X_j X_i}^*\tag{M:3.2.23}$$

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

Note that:

$$|\rho_{X_i X_j}| \leq 1, \quad i \neq j, \quad \text{and} \quad \rho_{X_i X_i} = 1\tag{M:3.2.24}$$

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

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

$$\mathbf{\Gamma}_X \triangleq \mathbb{E} [\mathbf{X}(\zeta) - \boldsymbol{\mu}_X [\mathbf{X}(\zeta) - \boldsymbol{\mu}_X]^H] = \mathbf{R}_X - \boldsymbol{\mu}_X \boldsymbol{\mu}_X^H\tag{M:3.3.25}$$

which shows that the two moments have essentially the same amount of information. In fact, if $\boldsymbol{\mu}_X = 0$, 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:

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

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

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

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

Cross-covariance is defined as

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

which too is a $N \times M$ matrix. The elements $\gamma_{X_i Y_j} = \mathbb{E} [(X_i(\zeta) - \mu_{X_i})(Y_j(\zeta) - \mu_{Y_j})^*]$ 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 $\Gamma_{\mathbf{XY}} = 0 \Rightarrow \mathbf{R}_{\mathbf{XY}} = \boldsymbol{\mu}_{\mathbf{X}} \boldsymbol{\mu}_{\mathbf{Y}}^H$.
- Orthogonal if $\mathbf{R}_{\mathbf{XY}} = 0$.

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

Example 5.16 (Sum of Random Vectors). Consider the sum of two zero-mean random vectors that are uncorrelated. What are the correlation and covariance matrices of the sum of random variables?

SOLUTION. Let $\mathbf{Z}(\zeta) = \mathbf{X}(\zeta) + \mathbf{Y}(\zeta)$. Then:

$$\mathbf{R}_{\mathbf{Z}} = \mathbb{E} [\mathbf{Z}(\zeta) \mathbf{Z}^H(\zeta)] = \mathbb{E} [(\mathbf{X}(\zeta) + \mathbf{Y}(\zeta)) (\mathbf{X}(\zeta) + \mathbf{Y}(\zeta))^H] \quad (5.165)$$

$$\begin{aligned}&= \mathbb{E} [\mathbf{X}(\zeta) \mathbf{X}^H(\zeta)] + \mathbb{E} [\mathbf{X}(\zeta) \mathbf{Y}^H(\zeta)] \\ &\quad + \mathbb{E} [\mathbf{Y}(\zeta) \mathbf{X}^H(\zeta)] + \mathbb{E} [\mathbf{Y}(\zeta) \mathbf{Y}^H(\zeta)]\end{aligned}\quad (5.166)$$

$$= \mathbf{R}_{\mathbf{X}} + \mathbf{R}_{\mathbf{XY}} + \mathbf{R}_{\mathbf{YX}} + \mathbf{R}_{\mathbf{YY}} \quad (5.167)$$

□

Since the random vectors are uncorrelated, then $\mathbf{R}_{\mathbf{XY}} = \mathbf{R}_{\mathbf{YX}} = \mathbf{0}$, and therefore $\mathbf{R}_{\mathbf{Z}} = \mathbf{R}_{\mathbf{X}} + \mathbf{R}_{\mathbf{Y}}$. Moreover, the covariance matrix is equal to the correlation matrix as the random vectors are zero-mean.



5.5 Linear Transformations

Topic Summary 35 Linear Transformations

Topic Objectives:

- Appreciate importance of linear transformations in probabilistic systems.
- Find transformed pdf in terms of the input pdf using the probability transformation rule.
- Calculate statistical descriptors for linearly transformed variables.
- Apply these results to a simple example.

Topic Activities:

Type	Details	Duration	Progress
Watch video	15 : 36 min video	3× length	
Read Handout	Read page 183 to page 186	8 mins/page	
Try Example	Try Example 5.17	10 minutes	
Practice Exercise	Exercises ??, ??, and ??	90 mins	

http://media.ed.ac.uk/media/1_k6muyz6h

Video Summary: Since linear transformations is such an important class of signal processing systems, this video looks at considering linear transformations of random vectors. After discussing various types of linear transformations, the video considers the relationships from the approach of using the probability transformation rule, but notes this is a rather tedious process in most cases. A more practical approach is to manipulate the statistical descriptors, which leads to a set of elegant results. An example of a 3-to-2 linear transformation is presented, and the viewer is encouraged to try the corresponding self-study exercises.

Since linear systems represent such an important class of 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} :

$$\mathbf{Y}(\zeta) = \mathbf{A} \mathbf{X}(\zeta) \quad (\text{M:3.2.32})$$

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).

Sidebar 11 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:

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{\mathbf{Y}(\zeta)} = \underbrace{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}}_{\mathbf{X}(\zeta)} \quad (5.168)$$

or in the scalar form by the linear equation:

$$y_i = \sum_{k=1}^N a_{ik} x_k \quad (5.169)$$

where a_{ij} is the i th row and j th column of the matrix \mathbf{A} . The Jacobian is obtained by calculating:

$$\frac{\partial y_i}{\partial x_j} = \sum_{k=1}^N a_{ik} \frac{\partial x_k}{\partial x_j} = a_{ij} \quad (5.170)$$

using the fact that

$$\frac{\partial x_k}{\partial x_j} = \begin{cases} 1 & \text{if } k = j \\ 0 & \text{if } k \neq j \end{cases} \quad (5.171)$$

Hence, constructing the Jacobian matrix using Equation 5.170 gives the matrix \mathbf{A} .

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,

$$Y_n(\zeta) = X_n(\zeta), \quad \text{for } n > m \quad (\text{M:3.2.33})$$

These additional auxiliary variables must then be marginalised out to obtain the joint-pdf for the original N -vector, $\mathbf{Y}(\zeta)$.

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 11. Thus, assuming $\mathbf{X}(\zeta)$, $\mathbf{Y}(\zeta)$, and \mathbf{A} are all real, then:

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{f_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y})}{|\det \mathbf{A}|} \quad (\text{M:3.2.34})$$

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 $\boldsymbol{\mu}_{\mathbf{X}}$, $\boldsymbol{\mu}_{\mathbf{Y}}$, $\boldsymbol{\Gamma}_{\mathbf{Y}}$, $\boldsymbol{\Gamma}_{\mathbf{X}\mathbf{Y}}$ or $\boldsymbol{\Gamma}_{\mathbf{Y}\mathbf{X}}$ is sufficient information for many algorithms.

Taking expectations of both sides of Equation M:3.2.32, $\mathbf{Y}(\zeta) = \mathbf{A} \mathbf{X}(\zeta)$, the following relations are found:

Mean vector:

$$\boldsymbol{\mu}_Y = \mathbb{E} [\mathbf{A} \mathbf{X}(\zeta)] = \mathbf{A} \boldsymbol{\mu}_X \quad (\text{M:3.2.38})$$

Autocorrelation matrix:

$$\begin{aligned} \mathbf{R}_Y &= \mathbb{E} [\mathbf{Y}(\zeta) \mathbf{Y}^H(\zeta)] = \mathbb{E} [\mathbf{A} \mathbf{X}(\zeta) \mathbf{X}^H(\zeta) \mathbf{A}^H] \\ &= \mathbf{A} \mathbb{E} [\mathbf{X}(\zeta) \mathbf{X}^H(\zeta)] \mathbf{A}^H = \mathbf{A} \mathbf{R}_X \mathbf{A}^H \end{aligned} \quad (\text{M:3.2.39})$$

Autocovariance matrix:

$$\boldsymbol{\Gamma}_Y = \mathbf{A} \boldsymbol{\Gamma}_X \mathbf{A}^H \quad (\text{M:3.2.40})$$

Cross-correlation matrix:

$$\begin{aligned} \mathbf{R}_{XY} &= \mathbb{E} [\mathbf{X}(\zeta) \mathbf{Y}^H(\zeta)] = \mathbb{E} [\mathbf{X}(\zeta) \mathbf{X}^H(\zeta) \mathbf{A}^H] \\ &= \mathbb{E} [\mathbf{X}(\zeta) \mathbf{X}^H(\zeta)] \mathbf{A}^H = \mathbf{R}_X \mathbf{A}^H \end{aligned} \quad (\text{M:3.2.42})$$

and hence $\mathbf{R}_{YX} = \mathbf{A} \mathbf{R}_X$.

Cross-covariance matrices:

$$\boldsymbol{\Gamma}_{XY} = \boldsymbol{\Gamma}_X \mathbf{A}^H \quad \text{and} \quad \boldsymbol{\Gamma}_{YX} = \mathbf{A} \boldsymbol{\Gamma}_X \quad (\text{M:3.2.43})$$

These results will be used to show what happens to a Gaussian random vector under a linear transformation in Section 5.6.

Example 5.17 (Linear Transformation). A random vector $\mathbf{X}(\zeta) = [X_1(\zeta) \ X_2(\zeta) \ X_3(\zeta)]^T$ has correlation matrix

$$\mathbf{R}_X = \begin{bmatrix} 9 & 3 & 1 \\ 3 & 9 & 3 \\ 1 & 3 & 9 \end{bmatrix} \quad (5.172)$$

This vector is transformed to another random vector $\mathbf{Y}(\zeta)$ by the following linear transformation:

$$\begin{bmatrix} Y_1(\zeta) \\ Y_2(\zeta) \end{bmatrix} = \begin{bmatrix} 3 & 2 & 1 \\ 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} X_1(\zeta) \\ X_2(\zeta) \\ X_3(\zeta) \end{bmatrix} \quad (5.173)$$

1. Find the correlation matrix \mathbf{R}_Y for $\mathbf{Y}(\zeta)$
2. Find the cross-correlation matrix \mathbf{R}_{XY} .

SOLUTION. The linear transformation can be written in the matrix-vector form as:

$$\mathbf{Y} = \mathbf{A} \mathbf{X} \quad (5.174)$$

1. Using Equation M:3.2.39, the autocorrelation matrix is given by:

$$\mathbf{R}_Y = \mathbf{A} \mathbf{R}_X \mathbf{A}^H = \begin{bmatrix} 3 & 2 & 1 \\ 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} 9 & 3 & 1 \\ 3 & 9 & 3 \\ 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} 3 & 1 \\ 2 & -2 \\ 1 & 1 \end{bmatrix} \quad (5.175)$$

Feel free to test your maths, or just use MATLAB:

```
A = [3 2 1; 1 -2 1]
RX = [9 3 1; 3 9 3; 1 3 9]
RY = A*RX*A.'
```

giving

$$\mathbf{R}_Y = \begin{bmatrix} 180 & -8 \\ -8 & 32 \end{bmatrix} \quad (5.176)$$

2. Finally, Equation M:3.2.42, then the cross-correlation is given by:

$$\mathbf{R}_{\mathbf{XY}} = \mathbf{R}_{\mathbf{X}} \mathbf{A}^H = \begin{bmatrix} 9 & 3 & 1 \\ 3 & 9 & 3 \\ 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} 3 & 1 \\ 2 & -2 \\ 1 & 1 \end{bmatrix} \quad (5.177)$$

$$= \begin{bmatrix} 34 & 4 \\ 30 & -12 \\ 18 & 4 \end{bmatrix} \quad (5.178) \quad \square$$

– End-of-Topic 35: **Linear Transformations and the Resulting Statistics**



5.6 Multivariate Gaussian Density Function

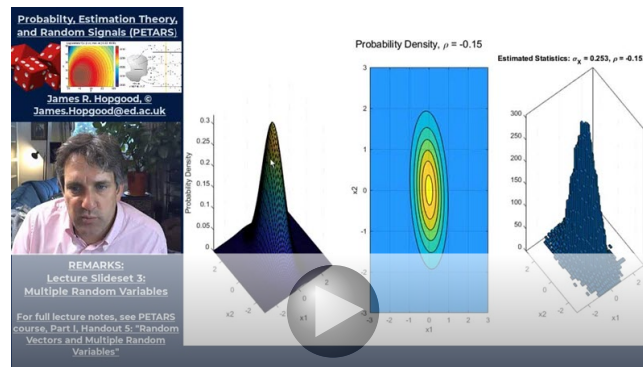
Topic Summary 36 The Multivariate Gaussian Distribution

Topic Objectives:

- Derive an expression for the pdf of a multivariate Gaussian.
- Understand how the Gaussian depends on the correlation coefficient.
- Key properties of the multivariate Gaussian.

Topic Activities:

Type	Details	Duration	Progress
Watch video	20 : 20 min video	3× length	
Read Handout	Read page 187 to page 191	8 mins/page	
Try Code	Use the MATLAB code	10 minutes	
Practice Exercise	Revisit Exercise ??	30 minutes	



http://media.ed.ac.uk/media/1_0pfz5b11

Video Summary: This video reviews the pdf for the multivariate Gaussian random variable. This pdf is then derived by developing the isotropic multivariate Gaussian, and then transforming through a linear transformation. The effect of the linear transformation on both the pdf and second-order statistical descriptors are considered. The video considers how the bivariate Gaussian density depends on the correlation coefficient, and how its orientation changes with this coefficient. Finally, the video considers key properties of the multivariate Gaussian, such as the fact that the linear transformation of a Gaussian is a Gaussian; that the marginal of a Gaussian is a Gaussian; and that the conditional distribution of a Gaussian is a Gaussian. The role of the multivariate Gaussian distribution within statistical signal processing is also discussed.

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:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{\Gamma}_{\mathbf{X}}|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})^T \mathbf{\Gamma}_{\mathbf{X}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}) \right] \quad (\text{M:3.2.44})$$

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

$$f_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Gamma}_{\mathbf{X}}) \quad (5.179)$$

Note the difference between the notation used here, and the notation used to indicate when a random vector is distributed, or sampled, from a normal distribution:

$$\mathbf{X}(\zeta) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Gamma}_{\mathbf{X}}) \quad (5.180)$$

The complex-valued normal random vector has pdf:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\pi^N |\boldsymbol{\Gamma}_{\mathbf{X}}|} \exp \left[-(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})^H \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}) \right] \quad (\text{M:3.2.47})$$

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].

5.6.1 Deriving the Multivariate Gaussian

The pdf for the multivariate Gaussian is often quoted, but where does it come from? It is most easily obtained by reusing some results from Section 5.2.4 and more specifically Topic 27.

Suppose that NRVs, $X_n(\zeta)$ for $n \in \{0, \dots, N-1\}$, are independent zero-mean unit variance Gaussian densities, and each have pdf given by $f_{X_n}(x_n)$. Then the joint-pdf of the multivariate random vector $\mathbf{X}(\zeta) = [X_0(\zeta), \dots, X_{N-1}(\zeta)]^T$ is given by:

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{n=0}^{N-1} f_{X_n}(x_n) \quad (5.181)$$

Since $X_n(\zeta)$ is Gaussian distributed with zero-mean and unit variance, such that $x_k \sim \mathcal{N}(0, 1)$ or:

$$f_{X_n}(x_n) = \mathcal{N}(x_n | 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_n^2}{2}} \quad (5.182)$$

and hence, as previously developed, it follows that:

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_n^2}{2}} = \frac{1}{(2\pi)^{\frac{N}{2}}} e^{-\frac{1}{2} \sum_{n=0}^{N-1} x_n^2} \quad (5.183)$$

Defining the vector $\mathbf{x} = [x_0, \dots, x_{N-1}]^T$, then it follows that

$$\mathbf{x}^T \mathbf{x} = [x_1 \quad \dots \quad x_N] \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \sum_{n=0}^{N-1} x_n^2 \quad (5.184)$$

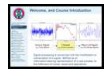
Using this relationship, it is possible to write Equation 5.183 as:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{N}{2}}} e^{-\frac{1}{2} \mathbf{x}^T \mathbf{x}} \quad (5.185)$$

This is an **isotropic Gaussian**, which is circularly symmetric.

A non-isotropic Gaussian can be obtained by a linear shift, scale, and rotation using the linear transformations from Topic 35. Hence, set:

$$\mathbf{y} = \mathbf{A} \mathbf{x} + \boldsymbol{\mu} \quad (5.186)$$



New slide

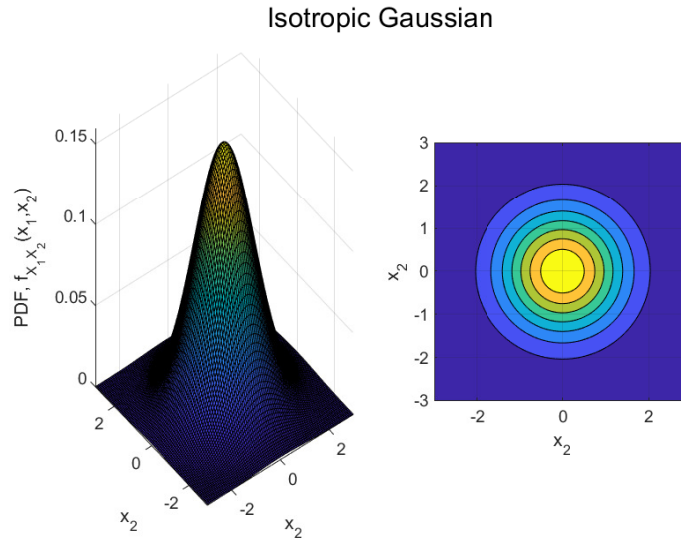


Figure 5.17: A graphical representation of an isotropic Gaussian random vector.

Similar to Topic 35, apply the probability transformation rule, noting there is one solution $\mathbf{x} = \mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ and the Jacobian $J_{\mathbf{x} \rightarrow \mathbf{y}} = \det \mathbf{A}$. Hence:

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{f_{\mathbf{X}}(\mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{\mu}))}{|\det \mathbf{A}|} \quad (5.187)$$

and

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|\det \mathbf{A}|} \frac{1}{(2\pi)^{\frac{N}{2}}} \exp \left[-\frac{1}{2} (\mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{\mu}))^T (\mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{\mu})) \right] \quad (5.188)$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{A}^T \mathbf{A}|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^T \mathbf{A}^{-T} \mathbf{A}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right] \quad (5.189)$$

where it has been noted that $|\mathbf{A} \mathbf{A}^T|^{\frac{1}{2}} = \det \mathbf{A}$.

Finally, writing $\boldsymbol{\Gamma}_{\mathbf{Y}} = \mathbf{A} \mathbf{A}^T$ and $\boldsymbol{\mu}_{\mathbf{Y}} = \boldsymbol{\mu}$, then:

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\boldsymbol{\Gamma}_{\mathbf{Y}}|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}})^T \boldsymbol{\Gamma}_{\mathbf{Y}}^{-1} (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}) \right] \quad (5.190)$$

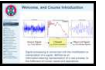
$$= \mathcal{N}(\mathbf{y} | \boldsymbol{\mu}_{\mathbf{Y}}, \boldsymbol{\Gamma}_{\mathbf{Y}}) \quad (5.191)$$

This is the expression for the multivariate Gaussian, which can be seen as a linear transformation of an isotropic Gaussian, which is derived from first principles. By calculating the second central moment for this density, it can be shown that $\boldsymbol{\Gamma}_{\mathbf{Y}}$ is indeed the covariance matrix, and it is also evident that $\boldsymbol{\mu}_{\mathbf{Y}}$ is the mean vector.

Using the definition of the correlation coefficient in Topic 34, for a bivariate Gaussian, the covariance matrix can be written as:

$$\boldsymbol{\Gamma}_{\mathbf{Y}} = \begin{bmatrix} \sigma_{Y_1}^2 & \rho_{Y_1 Y_2} \sigma_{Y_1} \sigma_{Y_2} \\ \rho_{Y_1 Y_2} \sigma_{Y_1} \sigma_{Y_2} & \sigma_{Y_2}^2 \end{bmatrix} \quad (5.192)$$

The pdf can then be plotted as $\rho_{Y_1 Y_2}$ changes. This can be seen in the animation video shown in Topic 32



5.6.2 Properties of Multivariate Gaussians

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: New slide

$$(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})^T \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}) = \sum_{i=1}^N \sum_{j=1}^N \langle \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} \rangle_{ij} (x_i - \mu_i)(x_j - \mu_j) \quad (\text{M:3.2.45})$$

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

1. $f_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Gamma}_{\mathbf{X}})$ 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 5.3 (Moments of a Gaussian RV). The moments of a Gaussian RV $X(\zeta) \sim \mathcal{N}(0, \sigma_x^2)$, are given by:

$$\mathbb{E}[X^k(\zeta)] = \begin{cases} 1 \cdot 3 \cdots (k-1) \sigma_x^k & k \text{ even} \\ 0 & k \text{ odd} \end{cases} \quad (5.193)$$

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 $\mathbb{E}[X^k(\zeta)]$ and $\mathbb{E}[X^{k+2}(\zeta)]$. 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 result builds on the results derived earlier for obtaining the standard expression of a multivariate Gaussian. It can be readily extended 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,

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{f_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y})}{|\det \mathbf{A}|} \quad (\text{M:3.2.34})$$

then if $f_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Gamma}_{\mathbf{X}})$, it follows:

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|\det \mathbf{A}|} \frac{1}{(2\pi)^{\frac{N}{2}} |\boldsymbol{\Gamma}_{\mathbf{X}}|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (\mathbf{A}^{-1}\mathbf{y} - \boldsymbol{\mu}_{\mathbf{X}})^T \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} (\mathbf{A}^{-1}\mathbf{y} - \boldsymbol{\mu}_{\mathbf{X}}) \right] \quad (5.194)$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{A}\boldsymbol{\Gamma}_{\mathbf{X}}\mathbf{A}^T|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu}_{\mathbf{X}})^T \mathbf{A}^{-T} \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} \mathbf{A}^{-1} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu}_{\mathbf{X}}) \right] \quad (5.195)$$

where it has been noted that $|\mathbf{A}\Gamma_{\mathbf{X}}\mathbf{A}^T|^{\frac{1}{2}} = |\mathbf{A}||\Gamma_{\mathbf{X}}|^{\frac{1}{2}}$. Thus, using the expressions for $\boldsymbol{\mu}_{\mathbf{Y}}$ and $\Gamma_{\mathbf{Y}}$ above, it directly follows that

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Gamma_{\mathbf{Y}}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}})^T \Gamma_{\mathbf{Y}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}})\right] \quad (5.196)$$

$$= \mathcal{N}(\mathbf{y} | \boldsymbol{\mu}_{\mathbf{Y}}, \Gamma_{\mathbf{Y}}) \quad (5.197)$$

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:

$$\mathbf{Z}(\zeta) = \begin{bmatrix} \mathbf{X}(\zeta) \\ \mathbf{Y}(\zeta) \end{bmatrix} \quad (\text{T:2.101})$$

then the corresponding mean vector and covariance matrix is given by:

$$\boldsymbol{\mu}_{\mathbf{Z}} = \mathbb{E} \left[\begin{bmatrix} \mathbf{X}(\zeta) \\ \mathbf{Y}(\zeta) \end{bmatrix} \right] = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{X}} \\ \boldsymbol{\mu}_{\mathbf{Y}} \end{bmatrix} \quad (\text{T:2.102})$$

$$\Gamma_{\mathbf{Z}} = \mathbb{E} \left[\begin{bmatrix} \mathbf{X}(\zeta) - \boldsymbol{\mu}_{\mathbf{X}} \\ \mathbf{Y}(\zeta) - \boldsymbol{\mu}_{\mathbf{Y}} \end{bmatrix} \begin{bmatrix} \mathbf{X}(\zeta) - \boldsymbol{\mu}_{\mathbf{X}} & \mathbf{Y}(\zeta) - \boldsymbol{\mu}_{\mathbf{Y}} \end{bmatrix}^H \right] = \begin{bmatrix} \Gamma_{\mathbf{X}} & \Gamma_{\mathbf{XY}} \\ \Gamma_{\mathbf{XY}}^H & \Gamma_{\mathbf{Y}} \end{bmatrix} \quad (\text{T:2.103})$$

Hence, the **joint-pdf** is given by:

$$f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{Z}}(\mathbf{z}) = \mathcal{N}(\mathbf{z} | \boldsymbol{\mu}_{\mathbf{Z}}, \Gamma_{\mathbf{Z}}) \quad (5.198)$$

$$= \frac{1}{(2\pi)^{\frac{N+M}{2}} |\Gamma_{\mathbf{Z}}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{Z}})^T \Gamma_{\mathbf{Z}}^{-1}(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{Z}})\right] \quad (5.199)$$

But by substituting for \mathbf{z} , $\boldsymbol{\mu}_{\mathbf{z}}$ and $\Gamma_{\mathbf{z}}$ 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:

$$f_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{\mathbf{X}}, \Gamma_{\mathbf{X}}) \quad (5.200)$$

$$f_{\mathbf{Y}}(\mathbf{y}) = \mathcal{N}(\mathbf{y} | \boldsymbol{\mu}_{\mathbf{Y}}, \Gamma_{\mathbf{Y}}) \quad (5.201)$$

Moreover, since

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

then the conditional density is also Gaussian, given by:

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} | \mathbf{x}) = \frac{1}{(2\pi)^{\frac{M}{2}} |\Gamma_{\mathbf{Y}|\mathbf{X}}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}})^T \Gamma_{\mathbf{Y}|\mathbf{X}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}})\right] \quad (\text{T:2.106})$$

where

$$\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}} = \boldsymbol{\mu}_{\mathbf{Y}} + \Gamma_{\mathbf{XY}}^H \Gamma_{\mathbf{X}}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}) \quad (\text{T:2.108})$$

$$\Gamma_{\mathbf{Y}|\mathbf{X}} = \Gamma_{\mathbf{Y}} - \Gamma_{\mathbf{XY}}^H \Gamma_{\mathbf{X}}^{-1} \Gamma_{\mathbf{XY}} \quad (\text{T:2.109})$$



5.7 Characteristic Functions

Topic Summary 37 The Multivariate Characteristic Function

Topic Objectives:

- Concept of extending the characteristic and moment generating function (MGF) to random vectors.
- Example of calculating the characteristic function of a multivariate Gaussian.
- Conceptual use of these transform domain operators.

Topic Activities:

Type	Details	Duration	Progress
Watch video	12 : 45 min video	3 × length	
Read Handout	Read page 192 to page 195	8 mins/page	
Try Example	Try Example 5.18	20 minutes	

The screenshot shows a video player interface. On the left is a video thumbnail of James R. Hoggood. The main content area is titled 'Characteristic Functions' and contains the following text:

Example (Multivariate Gaussian). Calculate the characteristic function for a multivariate Gaussian.

SOLUTION. Using the integral identity:

$$\int_{\mathbb{R}^P} \exp\left\{-\frac{1}{2}[\alpha + 2\mathbf{y}^T\boldsymbol{\beta} + \mathbf{y}^T\boldsymbol{\Gamma}\mathbf{y}]\right\} d\mathbf{y}$$

$$= \frac{(2\pi)^{\frac{P}{2}}}{|\boldsymbol{\Gamma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}[\alpha - \boldsymbol{\beta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\beta}]\right\}$$

where $\mathbf{y} \in \mathbb{R}^P$. Then it follows, by setting $\alpha = \boldsymbol{\mu}_X^T\boldsymbol{\Gamma}_X^{-1}\boldsymbol{\mu}_X$, $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_X^{-1}$, $\mathbf{y} = \mathbf{x}$ and $P = N$, that:

$$\Phi_X(\boldsymbol{\xi}) = \exp\left[j\boldsymbol{\xi}^T\boldsymbol{\mu}_X - \frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\Gamma}_X\boldsymbol{\xi}\right]$$

http://media.ed.ac.uk/media/1_kv2kkbar

Video Summary: The concept of the characteristic function for scalar random variables is extended to multivariate densities of random vectors. This is defined as the multi-dimensional Fourier transform, or the multi-dimensional Laplace transform for MGFs. There is a discussion that the multi-dimensional transforms are perhaps more useful as a conceptual rather than practical tool. The video then considers an example of finding the characteristic function of the multivariate Gaussian. As an Appendix to this Topic, a derivation of a key integral identify used in the example is provided.

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 and Laplace transforms of the joint-pdf. Hence, the characteristic function is:

$$\Phi_{\mathbf{X}}(\boldsymbol{\xi}) \triangleq \mathbb{E}\left[e^{j\boldsymbol{\xi}^T\mathbf{X}(\zeta)}\right] = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) e^{j\boldsymbol{\xi}^T\mathbf{x}} d\mathbf{x} \quad (5.202)$$

Here, as \mathbf{x} is a vector, so is the variable $\boldsymbol{\xi}$ which is defined as:

$$\boldsymbol{\xi} = [\xi_1 \quad \xi_2 \quad \cdots \quad \xi_N]^T \quad (5.203)$$

such that $\boldsymbol{\xi}^T\mathbf{x}$ is a scalar, $\boldsymbol{\xi}^T\mathbf{x} = \sum_{n=1}^N \xi_n x_n$, and the differential $d\mathbf{x} = \prod_{n=1}^N dx_n$.

Similarly, the moment generating function is given by:

$$\bar{\Phi}_{\mathbf{X}}(\mathbf{s}) \triangleq \mathbb{E} \left[e^{\mathbf{s}^T \mathbf{X}(\zeta)} \right] = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) e^{\mathbf{s}^T \mathbf{x}} d\mathbf{x} \quad (5.204)$$

Example 5.18 (Multivariate Gaussian). Calculate the characteristic function for a multivariate Gaussian.

SOLUTION. This problem is an interesting exercise in multi-dimensional integration, and some of the identities used will be used again in later Topics.

The characteristic function for a real-valued Gaussian random vector is given by:

$$\Phi_{\mathbf{X}}(\boldsymbol{\xi}) = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) e^{j\boldsymbol{\xi}^T \mathbf{x}} d\mathbf{x} \quad (5.205)$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}} |\boldsymbol{\Gamma}_{\mathbf{X}}|^{\frac{1}{2}}} \int_{-\infty}^{\infty} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})^T \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}) \right] e^{j\boldsymbol{\xi}^T \mathbf{x}} d\mathbf{x} \quad (5.206)$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}} |\boldsymbol{\Gamma}_{\mathbf{X}}|^{\frac{1}{2}}} \int_{-\infty}^{\infty} \exp \left[-\frac{\mathbf{x}^T \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} \mathbf{x} + 2\mathbf{x}^T \boldsymbol{\beta} + \boldsymbol{\mu}_{\mathbf{X}}^T \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} \boldsymbol{\mu}_{\mathbf{X}}}{2} \right] d\mathbf{x} \quad (5.207)$$

where $\boldsymbol{\beta} = -(\boldsymbol{\Gamma}_{\mathbf{X}}^{-1} \boldsymbol{\mu}_{\mathbf{X}} + j\boldsymbol{\xi})^T$, and the relationship that $\boldsymbol{\beta}^T \mathbf{x} = (\mathbf{x}^T \boldsymbol{\beta})^T$ both equals scalar values have been used.

Using the integral identity:

$$\begin{aligned} \int_{\mathbb{R}^P} \exp \left\{ -\frac{1}{2} [\alpha + 2\mathbf{y}^T \boldsymbol{\beta} + \mathbf{y}^T \boldsymbol{\Gamma} \mathbf{y}] \right\} d\mathbf{y} \\ = \frac{(2\pi)^{\frac{P}{2}}}{|\boldsymbol{\Gamma}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} [\alpha - \boldsymbol{\beta}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{\beta}] \right\} \end{aligned} \quad (5.208)$$

where $\mathbf{y} \in \mathbb{R}^P$ is a P -dimensional column vector. This result is proved in Sidebar 12. Then it follows, by setting $\alpha = \boldsymbol{\mu}_{\mathbf{X}}^T \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} \boldsymbol{\mu}_{\mathbf{X}}$, $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_{\mathbf{X}}^{-1}$, $\mathbf{y} = \mathbf{x}$ and $P = N$, that:

$$\Phi_{\mathbf{X}}(\boldsymbol{\xi}) = \exp \left[-\frac{1}{2} \left\{ \boldsymbol{\mu}_{\mathbf{X}}^T \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} \boldsymbol{\mu}_{\mathbf{X}} - (\boldsymbol{\mu}_{\mathbf{X}}^T \boldsymbol{\Gamma}_{\mathbf{X}}^{-1} + j\boldsymbol{\xi}^T) \boldsymbol{\Gamma}_{\mathbf{X}} (\boldsymbol{\Gamma}_{\mathbf{X}}^{-1} \boldsymbol{\mu}_{\mathbf{X}} + j\boldsymbol{\xi}) \right\} \right] \quad (5.209)$$

which, after multiplying out, gives:

$$\Phi_{\mathbf{X}}(\boldsymbol{\xi}) = \exp \left[j\boldsymbol{\xi}^T \boldsymbol{\mu}_{\mathbf{X}} - \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\Gamma}_{\mathbf{X}} \boldsymbol{\xi} \right] \quad (\text{M:3.2.46})$$

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

$$\Phi_{\mathbf{X}}(\boldsymbol{\xi}) = \exp \left[j\Re\{\boldsymbol{\xi}^H \boldsymbol{\mu}_{\mathbf{X}}\} - \frac{1}{4} \boldsymbol{\xi}^H \boldsymbol{\Gamma}_{\mathbf{X}} \boldsymbol{\xi} \right] \quad (\text{M:3.2.50})$$

□

The multivariate characteristic function is perhaps more useful as a powerful conceptual tool than a practical method for manipulation, as there are only a few cases where analytical results exist. The concept can also be extended to a multi-dimensional version of the probability generating function (PGF) for discrete random variables.

The result for the characteristic function of a multivariate Gaussian yields some interesting consequences:

Sidebar 12 Proof of the Multivariate Gaussian Identity

The identity in Equation 5.208 can be derived by using another axiomatic identity, which is the fact that a multivariate Gaussian pdf integrates to one, such that:

$$\int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{P}{2}} |\Sigma_{\mathbf{Y}}|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}})^T \Sigma_{\mathbf{Y}}^{-1} (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}) \right] d\mathbf{y} = 1 \quad (5.210)$$

where $\mathbf{y} \in \mathbb{R}^{P \times 1}$. Set $\boldsymbol{\Gamma} = \Sigma_{\mathbf{Y}}^{-1}$; this substitution arguably looks as though it is confusing things further, but it does keep the manipulations simpler when working towards the required identity. This means that $\det \boldsymbol{\Gamma} = |\boldsymbol{\Gamma}| = 1/|\Sigma_{\mathbf{Y}}|$. Hence, substituting and rearranging by bringing the constant term outside the exponent and to the other side of the equation gives:

$$\int_{-\infty}^{\infty} \exp \left[-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}})^T \boldsymbol{\Gamma} (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}) \right] d\mathbf{y} = \frac{(2\pi)^{\frac{P}{2}}}{|\boldsymbol{\Gamma}|^{\frac{1}{2}}} \quad (5.211)$$

Expanding the exponent, noting that $\boldsymbol{\mu}_{\mathbf{Y}}^T \boldsymbol{\Gamma} \mathbf{y} = \mathbf{y}^T \boldsymbol{\Gamma} \boldsymbol{\mu}_{\mathbf{Y}}$, as $\boldsymbol{\Gamma}^T = \boldsymbol{\Gamma}$, then:

$$\int_{-\infty}^{\infty} \exp \left[-\frac{1}{2} (\mathbf{y}^T \boldsymbol{\Gamma} \mathbf{y} - 2\mathbf{y}^T \boldsymbol{\Gamma} \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\mu}_{\mathbf{Y}}^T \boldsymbol{\Gamma} \boldsymbol{\mu}_{\mathbf{Y}}) \right] d\mathbf{y} = \frac{(2\pi)^{\frac{P}{2}}}{|\boldsymbol{\Gamma}|^{\frac{1}{2}}} \quad (5.212)$$

Setting $\boldsymbol{\beta} = -\boldsymbol{\Gamma} \boldsymbol{\mu}_{\mathbf{Y}}$, such that $\boldsymbol{\mu}_{\mathbf{Y}} = -\boldsymbol{\Gamma}^{-1} \boldsymbol{\beta}$ gives:

$$\int_{-\infty}^{\infty} \exp \left[-\frac{1}{2} (\mathbf{y}^T \boldsymbol{\Gamma} \mathbf{y} + 2\mathbf{y}^T \boldsymbol{\beta} + (-\boldsymbol{\beta}^T \boldsymbol{\Gamma}^{-1}) \boldsymbol{\Gamma} (-\boldsymbol{\Gamma}^{-1} \boldsymbol{\beta})) \right] d\mathbf{y} = \frac{(2\pi)^{\frac{P}{2}}}{|\boldsymbol{\Gamma}|^{\frac{1}{2}}} \quad (5.213)$$

Simplifying $(-\boldsymbol{\beta}^T \boldsymbol{\Gamma}^{-1}) \boldsymbol{\Gamma} (-\boldsymbol{\Gamma}^{-1} \boldsymbol{\beta}) = \boldsymbol{\beta}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{\beta}$, splitting up the exponent, and taking the term in $\boldsymbol{\beta}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{\beta}$ to the other side gives:

$$\int_{-\infty}^{\infty} \exp \left[-\frac{1}{2} (\mathbf{y}^T \boldsymbol{\Gamma} \mathbf{y} + 2\mathbf{y}^T \boldsymbol{\beta}) \right] d\mathbf{y} = \frac{(2\pi)^{\frac{P}{2}}}{|\boldsymbol{\Gamma}|^{\frac{1}{2}}} \exp \left[\frac{1}{2} \boldsymbol{\beta}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{\beta} \right] \quad (5.214)$$

Finally, multiplying both sides by $\exp(-\frac{\alpha}{2})$ gives the desired identity:

$$\int_{-\infty}^{\infty} \exp \left[-\frac{1}{2} (\mathbf{y}^T \boldsymbol{\Gamma} \mathbf{y} + 2\mathbf{y}^T \boldsymbol{\beta} + \alpha) \right] d\mathbf{y} = \frac{(2\pi)^{\frac{P}{2}}}{|\boldsymbol{\Gamma}|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (\alpha - \boldsymbol{\beta}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{\beta}) \right] \quad (5.215)$$

There are many variants of this proof, and indeed of this identity, but they all broadly follow the same starting point.

1. The Fourier transform of a Gaussian function is still a Gaussian function.
2. The Fourier transform is a linear transformation, and therefore this result is a natural consequence.

– End-of-Topic 37: **Characteristic Functions** –



5.8 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

$$\mathbf{X}(\zeta) = [X_1(\zeta) \ X_2(\zeta) \ X_3(\zeta) \ X_4(\zeta)]^T \quad (5.216)$$

can be expressed in terms of its second order moments. For the real case when $\mathbf{X}(\zeta) \sim \mathcal{N}(\mathbf{0}, \Gamma_{\mathbf{X}})$, then:

$$\begin{aligned} \mathbb{E} [X_1(\zeta)X_2(\zeta)X_3(\zeta)X_4(\zeta)] &= \mathbb{E} [X_1(\zeta)X_2(\zeta)] \mathbb{E} [X_3(\zeta)X_4(\zeta)] \\ &\quad + \mathbb{E} [X_1(\zeta)X_3(\zeta)] \mathbb{E} [X_2(\zeta)X_4(\zeta)] \\ &\quad + \mathbb{E} [X_1(\zeta)X_4(\zeta)] \mathbb{E} [X_2(\zeta)X_3(\zeta)] \end{aligned} \quad (\text{M:3.2.53})$$

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.

5.9 Sum of Independent Random Variables

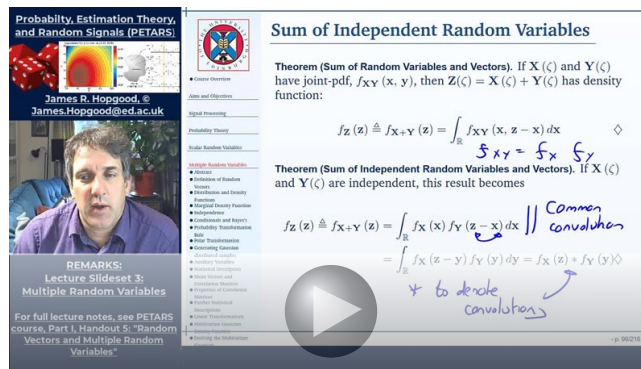
Topic Summary 38 Sum of Independent Random Variables

Topic Objectives:

- Investigate pdfs of sum of random variables.
- Apply to simple examples.
- Understand the role of characteristic functions in this case.

Topic Activities:

Type	Details	Duration	Progress
Watch video	18 : 23 min video	3 × length	
Read Handout	Read page 196 to page 198	8 mins/page	
Try Example	See video	10 mins	
Practice Exercise	Try Exercises ?? to ??	80 minutes	



http://media.ed.ac.uk/media/1_kxi2oy5p

Video Summary: This Topic considers further the sum of random variables that was introduced in Topic 31 on auxiliary variables. The case of independent random variables and vectors is considered specifically, where it is seen that the pdf of the sum is the convolution of the individual pdfs. An example shown in the video, but not in the notes, is the probability mass function (pmf) of the sum of two fair dice. The video then shows how the sum of independent random variables can be elegantly dealt with using characteristic functions, because convolution in the pdf space becomes multiplication in the characteristic function space. This becomes useful in proofs such as the central limit theorem (CLT) in Topic 39.

Theorem 5.4 (Sum of Random Variables and Vectors). If $X(\zeta)$ and $Y(\zeta)$ have joint-pdf, $f_{XY}(x, y)$, then $Z(\zeta) = X(\zeta) + Y(\zeta)$ has density function:

$$f_Z(z) \triangleq f_{X+Y}(z) = \int_{\mathbb{R}} f_{XY}(x, z - x) dx \tag{5.217}$$

PROOF. This can easily be obtained using the probability rule and an appropriate auxiliary variable, as in Section 5.3.3, and indeed is a simplification of the result already proved there. However, an alternative proof which avoids the use of auxiliary variables is given here for completeness.

Define the event $Z = \{(x, y) : x + y \leq z\}$. Then:

$$\Pr(\mathbf{X} + \mathbf{Y} \leq \mathbf{x}) = \iint_Z f_{\mathbf{XY}}(\mathbf{u}, \mathbf{v}) d\mathbf{u} d\mathbf{v} = \int_{\mathbf{v} \in \mathbb{R}} \int_{\mathbf{u} = -\infty}^{z-\mathbf{v}} f_{\mathbf{XY}}(\mathbf{u}, \mathbf{v}) d\mathbf{u} d\mathbf{v} \quad (5.218)$$

and by making the substitution $\mathbf{w} = \mathbf{u} + \mathbf{v}$.

$$= \int_{\mathbf{v} \in \mathbb{R}} \int_{\mathbf{w} = -\infty}^z f_{\mathbf{XY}}(\mathbf{w} - \mathbf{u}, \mathbf{v}) d\mathbf{w} d\mathbf{v} \quad (5.219)$$

$$= \int_{\mathbf{w} = -\infty}^z \int_{\mathbf{v} \in \mathbb{R}} f_{\mathbf{XY}}(\mathbf{u}, \mathbf{w} - \mathbf{v}) d\mathbf{u} d\mathbf{v} \triangleq \int_{\mathbf{w} = -\infty}^z f_{\mathbf{X}}(\mathbf{v}) d\mathbf{v} \quad (5.220)$$

□

giving the result as required.

Theorem 5.5 (Sum of Independent Random Variables and Vectors). If $\mathbf{X}(\zeta)$ and $\mathbf{Y}(\zeta)$ are independent, this result becomes

$$f_Z(\mathbf{z}) \triangleq f_{\mathbf{X}+\mathbf{Y}}(\mathbf{z}) = \int_{\mathbb{R}} f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}}(\mathbf{z} - \mathbf{x}) d\mathbf{x} \quad (5.221)$$

$$= \int_{\mathbb{R}} f_{\mathbf{X}}(\mathbf{z} - \mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} = f_{\mathbf{X}}(\mathbf{z}) * f_{\mathbf{Y}}(\mathbf{y}) \quad (5.222)$$

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** or moment generating functions (MGFs) as introduced in the lecture on scalar random variables.

If $Z(\zeta) = X(\zeta) + Y(\zeta)$, then its characteristic function is:

$$\Phi_Z(\xi) \triangleq \mathbb{E}[e^{j\xi Z(\zeta)}] = \mathbb{E}[e^{j\xi[X(\zeta)+Y(\zeta)}]] = \mathbb{E}[e^{j\xi X(\zeta)}] \mathbb{E}[e^{j\xi Y(\zeta)}] \quad (\text{M:3.2.59})$$

where the last inequality follows from independence. More explicitly, observe that:

$$\Phi_Z(\xi) = \mathbb{E}[e^{j\xi[X(\zeta)+Y(\zeta)}]] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) e^{j\xi[x+y]} dx dy \quad (5.223)$$

and noting that due to independence $f_{XY}(x, y) = f_X(x) f_Y(y)$, then

$$\Phi_Z(\xi) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_X(x) f_Y(y) e^{j\xi x} e^{j\xi y} dx dy \quad (5.224)$$

$$= \left\{ \int_{-\infty}^{\infty} f_X(x) e^{j\xi x} dx \right\} \left\{ \int_{-\infty}^{\infty} f_Y(y) e^{j\xi y} dy \right\} \quad (5.225)$$

giving the desired result.

Hence, from the convolution property of the Fourier transform, it follows directly from this result that

$$f_Z(z) = f_X(x) * f_Y(y) \quad (\text{M:3.2.61})$$

This result can be generalised to the summation of M independent RVs:

$$Y(\zeta) = \sum_{k=1}^M c_k X_k(\zeta) \quad (\text{M:3.2.55})$$

where $\{c_k\}_1^M$ is a set of fixed (deterministic) coefficients.

It follows straightforwardly that:

$$\Phi_Y(\xi) \triangleq \mathbb{E} [e^{j\xi Y(\zeta)}] = \prod_{k=1}^M \mathbb{E} [e^{j\xi c_k X_k(\zeta)}] = \prod_{k=1}^M \Phi_{X_k}(c_k \xi) \quad (\text{M:3.2.72})$$

Hence, the pdf of $Y(\zeta)$ is given by:

$$f_Y(y) = \frac{1}{|c_1|} f_{X_1}\left(\frac{y}{c_1}\right) * \frac{1}{|c_2|} f_{X_2}\left(\frac{y}{c_2}\right) * \cdots * \frac{1}{|c_M|} f_{X_M}\left(\frac{y}{c_M}\right) \quad (\text{M:3.2.73})$$

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 5.6 (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:

$$\mu_Y = \sum_{k=1}^M c_k \mu_{X_k} \quad (\text{M:3.2.56})$$

Moreover, assuming independence, then the variance of $Y(\zeta)$ is given by:

$$\sigma_Y^2 = \mathbb{E} \left[\left| \sum_{k=1}^M c_k \mu_{X_k} - \mu_{X_k} \right|^2 \right] = \sum_{k=1}^M |c_k|^2 \sigma_{X_k}^2 \quad (\text{M:3.2.57})$$

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 n th-order cumulants for $Y(\zeta)$.

Recall that

$$\Psi_X(\xi) \triangleq \ln \Phi_X(\xi) = \ln \mathbb{E} [e^{j\xi X(\zeta)}] \quad (5.226)$$

Then, from Equation M:3.2.72,

$$\Psi_Y(\xi) \triangleq \ln \mathbb{E} [e^{j\xi Y(\zeta)}] = \sum_{k=1}^M \ln \mathbb{E} [e^{j\xi c_k X_k(\zeta)}] = \sum_{k=1}^M \Psi_{X_k}(c_k \xi) \quad (\text{M:3.2.74})$$

Therefore, it can readily be shown that the cumulants of $Y(\zeta)$ are given by:

$$\kappa_Y^{(n)} = \sum_{k=1}^M c_k^n \kappa_{X_k}^{(n)} \quad (\text{M:3.2.75})$$

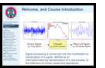
It is left as an exercise for the reader to demonstrate this.

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.



5.10 Central limit theorem



Topic Summary 39 The Central Limit Theorem

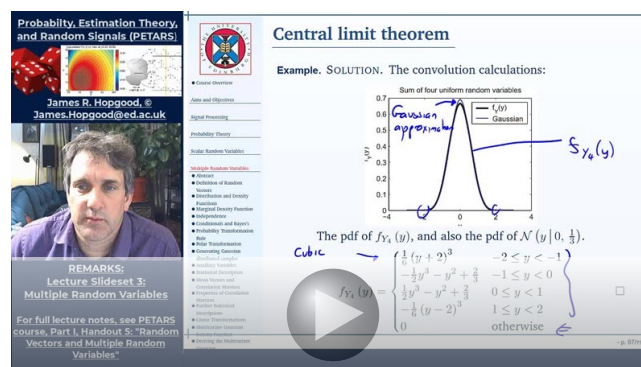
New slide

Topic Objectives:

- Motivate the central limit theorem (CLT) using an example.
- Demonstrate the CLT using a simulated numerical example.
- Formally define the CLT.
- Give an outline proof of the CLT.

Topic Activities:

Type	Details	Duration	Progress
Watch video	23 : 02 min video	3× length	
Read Handout	Read page 199 to page 202	8 mins/page	
Try Example	Try Example 5.19	15 minutes	
Try Code	Use the MATLAB code	10 minutes	
Practice Exercise	Try Exercise ??	10 minutes	



http://media.ed.ac.uk/media/1_795s4i9h

Video Summary: This Topic motivates the CLT by considering the pdf of the sum of uniform random variables, from two through to just four variables. It is shown that this pdf approaches a Gaussian very rapidly. In addition to a mathematical development, the video also shows simulated numerical results (in this case using MATLAB, but easily done in any language). The video considers the formal definition of the CLT in terms of normalised random variables. Finally, for completeness, the video gives an outline sketch of the CLT using characteristic functions.

To motivate the central limit theorem, consider the following example.

Example 5.19. In Exercise ??, the problem considers the sum of four independent random variables. Suppose $\{X_k(\zeta)\}_{k=1}^4$ are four i. i. d. random variables uniformly distributed over $[-0.5, 0, 5]$. Compute and plot the pdfs of $Y_M(\zeta) \triangleq \sum_{k=1}^M X_k(\zeta)$ for $M = \{2, 3, 4\}$.

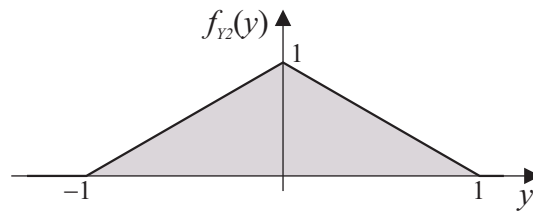


Figure 5.18: The pdf $f_{Y_2}(u)$, the sum of two uniform random variables.

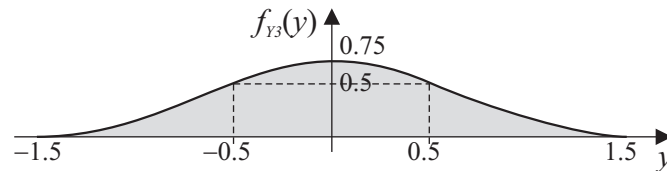
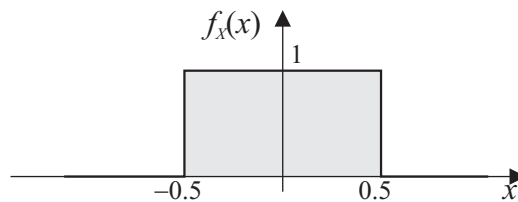


Figure 5.19: The pdf $f_{Y_3}(u)$, the sum of three uniform random variables.



SOLUTION. The zero-mean uniform pdf for $f_{X_k}(x_k)$ as specified is shown in Figure ??, where the subscripts have been dropped for clarity. Using the convolution result for the sum of independent random variables from Section 5.9, it follows:

$$f_{Y_2}(y) = f_{X_1}(y) * f_{X_2}(y) = f_X(y) * f_X(y) \quad (5.227)$$

$$f_{Y_3}(y) = f_{Y_2}(y) * f_{X_3}(y) = f_{Y_2}(y) * f_X(y) \quad (5.228)$$

$$f_{Y_4}(y) = f_{Y_3}(y) * f_{X_4}(y) = f_{Y_3}(y) * f_X(y) \quad (5.229)$$

The convolution calculations to this problem Example ?? should yield the following pdfs:

$$f_{Y_2}(y) = \begin{cases} 1 + y & -1 \leq y < 0 \\ 1 - y & 0 \leq y < 1 \\ 0 & \text{otherwise} \end{cases} \quad (5.230)$$

$$f_{Y_3}(y) = \begin{cases} \frac{1}{2} \left(y + \frac{3}{2}\right)^2 & -\frac{3}{2} \leq y < -\frac{1}{2} \\ \frac{3}{4} - y^2 & -\frac{1}{2} \leq y < \frac{1}{2} \\ \frac{1}{2} \left(y - \frac{3}{2}\right)^2 & \frac{1}{2} \leq y < \frac{3}{2} \\ 0 & \text{otherwise} \end{cases} \quad (5.231)$$

$$f_{Y_4}(y) = \begin{cases} \frac{1}{6} (y + 2)^3 & -2 \leq y < -1 \\ -\frac{1}{2} y^3 - y^2 + \frac{2}{3} & -1 \leq y < 0 \\ \frac{1}{2} y^3 - y^2 + \frac{2}{3} & 0 \leq y < 1 \\ -\frac{1}{6} (y - 2)^3 & 1 \leq y < 2 \\ 0 & \text{otherwise} \end{cases} \quad (5.232) \quad \square$$

These pdfs are plotted in Figure 5.18, Figure 5.19, and Figure 5.20, respectively.

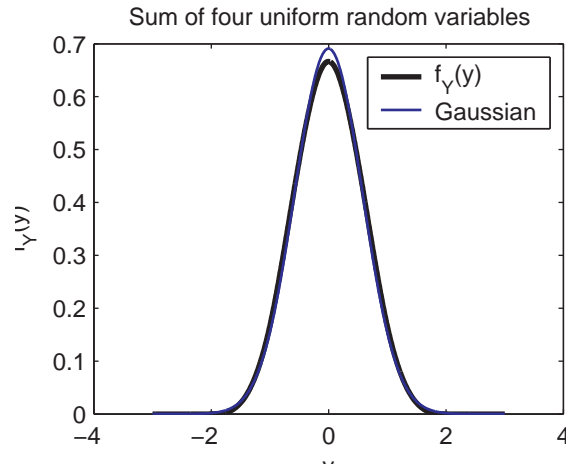


Figure 5.20: The pdf of $f_{Y_4}(y)$, and also the pdf of $\mathcal{N}(y | 0, \frac{1}{3})$.

Consider the random variable $Y(\zeta)$ given by:

$$Y_M(\zeta) = \sum_{k=1}^M X_k(\zeta) \quad (\text{M:3.2.55})$$

What is the distribution of $Y_M(\zeta)$ as $M \rightarrow \infty$?

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 $\lim_{M \rightarrow \infty} Y_M(\zeta)$.

Informally, the CLT is well known, and the answer is a Gaussian. However, more care is needed. Assume that the $X_M(\zeta)$'s are i. i. d., and the mean and variance of $X_m(\zeta)$ are finite and given by μ_X and σ_X^2 . Then:

- the mean of $Y_M(\zeta)$ is

$$\mathbb{E}[Y_M] = \mathbb{E}\left[\sum_{m=1}^M X_m(\zeta)\right] = \sum_{m=1}^M \mathbb{E}[X_m(\zeta)] \quad (5.233)$$

$$\mu_Y = M\mu_X \quad \text{What is } \mu_Y \text{ as } M \rightarrow \infty? \quad (5.234)$$

- the variance of $Y_M(\zeta)$ is

$$\text{var}[Y_M] = \text{var}\left[\sum_{m=1}^M X_m(\zeta)\right] = \sum_{m=1}^M \text{var}[X_m(\zeta)] \quad (5.235)$$

$$\sigma_Y^2 = M\sigma_X^2 \quad \text{Similarly, what is } \sigma_Y^2 \text{ as } M \rightarrow \infty? \quad (5.236)$$

Theorem 5.7 (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 $k = \{1, \dots, M\}$. Define the normalised random variable:

$$\hat{Y}_M(\zeta) = \frac{Y_M(\zeta) - \mu_{Y_M}}{\sigma_{Y_M}} \quad \text{where} \quad Y_M(\zeta) = \sum_{k=1}^M X_k(\zeta) \quad (\text{M:3.2.55})$$

Then the distribution of $\hat{Y}_M(\zeta)$ approaches that of a normal random variable with zero mean and unit standard deviation as $M \rightarrow \infty$; in other words,

$$\lim_{M \rightarrow \infty} f_{\hat{Y}_M}(y) = \mathcal{N}(y | 0, 1) \quad (5.237)$$

PROOF. Since the $X_k(\zeta)$'s are i. i. d., then $\mu_{Y_M} = M\mu_X$ and $\sigma_{Y_M}^2 = M\sigma_X^2$. Let

$$Z_k(\zeta) = \frac{X_k(\zeta) - \mu_X}{\sigma_X} \quad (5.238)$$

such that $\mu_{Z_k} = \mu_Z = 0$, $\sigma_{Z_k}^2 = \sigma_Z^2 = 1$ and the normalised random variable can be written as:

$$\hat{Y}_M(\zeta) = \frac{1}{\sqrt{M}} \sum_{k=1}^M Z_k(\zeta) \quad (5.239)$$

Noting that if $V(\zeta) = aU(\zeta)$ for some real-scalar a then

$$\Phi_V(\xi) = \mathbb{E} [e^{j\xi aU(\zeta)}] = \Phi_U(a\xi) \quad (5.240)$$

Hence, from Equation M:3.2.72, the characteristic function for $\hat{Y}_M(\zeta)$ is given by:

$$\Phi_{\hat{Y}_M}(\xi) = \prod_{k=1}^M \Phi_{Z_k} \left(\frac{\xi}{\sqrt{M}} \right) \quad (5.241)$$

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:

$$\Phi_{\hat{Y}_M}(\xi) = \Phi_Z^M \left(\frac{\xi}{\sqrt{M}} \right) \quad (5.242)$$

From the previous chapter on scalar random variables,

$$\Phi_Z(\xi) = \mathbb{E} [e^{j\xi Z(\zeta)}] = \sum_{n=0}^{\infty} \frac{(j\xi)^n}{n!} \mathbb{E} [Z^n(\zeta)] \quad (5.243)$$

and therefore, the characteristic function for $\hat{Y}_M(\zeta)$ becomes:

$$\Phi_{\hat{Y}_M}(\xi) = \left\{ \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{j\xi}{\sqrt{M}} \right)^n \mathbb{E} [Z^n(\zeta)] \right\}^M \quad (5.244)$$

$$= \left\{ 1 + \frac{j\xi\mu_Z}{\sqrt{M}} - \frac{\xi^2\sigma_Z^2}{2M} + \mathcal{O} \left(\left\{ \frac{\xi}{\sqrt{M}} \right\}^3 \right) \right\}^M \quad (5.245)$$

Using the moments $\mu_Z = 0$ and $\sigma_Z^2 = 1$,

$$\Phi_{\hat{Y}_M}(\xi) = \left\{ 1 - \frac{\xi^2}{2M} + \mathcal{O} \left(\left\{ \frac{\xi}{\sqrt{M}} \right\}^3 \right) \right\}^M \rightarrow e^{-\frac{1}{2}\xi^2} \quad \text{as } M \rightarrow \infty \quad (5.246)$$

where the following limit is used:

$$\lim_{n \rightarrow \infty} \left(1 + \frac{x}{n} \right)^n = e^x \quad (5.247)$$

□

This last term is the characteristic function of the $\mathcal{N}(y | 0, 1)$ distribution.



6

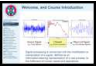
Principles of Estimation Theory

An approximate answer to the right
problem is worth a good deal more than an
exact answer to an approximate problem.

John Tukey

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.

6.1 Introduction



Topic Summary 40 Introduction to Estimation Theory

New slide

Topic Objectives:

- Motivation for Estimating Parameters from Data.
- Examples of Parameter Estimation.
- Properties of Statistical Estimations.
- Conceptual difference between a point-estimate and the estimator as a random variable or vector.
- Numerical example showing the sampling distribution.

Topic Activities:

Type	Details	Duration	Progress
Watch video	18 : 31 min video	3 × length	
Read Handout	Read page 204 to page 209	8 mins/page	
Try Code	Use the MATLAB code	10 minutes	
Try Example	Try Example 6.1	15 mins	

Properties of Estimators

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

As an example of an estimator, consider estimating the mean μ_X of a random Variate, $X(c)$, from N observations $\mathcal{X} = \{x[n]\}_0^{N-1}$. The most natural estimator is a simple arithmetic average of these observations, the **sample mean**:

$$\hat{\mu}_X = \hat{\theta}[\mathcal{X}] = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$$

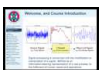
http://media.ed.ac.uk/media/1_1_tynr4nm

Video Summary: In this video, Estimation Theory is introduced in which unknown parameters are estimated from data, rather than assuming that problems can be described by fully known distributions or statistics. The taxi-cab problem from Topic 9 is highlighted as an example. Examples of parameter estimation problems are discussed, followed by the concept of an estimator. The concept of distinguishing point-estimators from an estimator as a random variable before a set of observations is introduced, and a numerical example of the sampling distribution of the sample mean is presented in depth.

- 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 know 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.

6.1.1 A (Confusing) Note on Notation



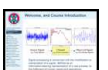
Note that, unfortunately, from this point onwards, a slightly different (and abusive use of) notation for random quantities is used than what was presented in the first set of handouts. *New slide*

So far, as in the literature, the n th-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 also used in some literature to refer to the **random variable** itself with the consequence that, in different contexts, $x[n]$ can refer both to 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** later in the 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.

6.1.2 Examples of parameter estimation



To motivate this handout, this section lists a number of potential problems in which parameters might wish to be estimated. *New slide*

Frequency Estimation Consider estimating the spectral content of a harmonic process, $x[n]$, consisting of a single-tone, given by

$$x[n] = A_0 \cos(\omega_0 n + \phi_0) + w[n] \quad (6.1)$$

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, giving rise to observations $x[n]$.

Sidebar 13 The taxi-cab problem (Repeated)

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 was initially discussed as the Venice Water-Taxi problem in Chapter 2.

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.

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 θ , such that:

$$x[n] \sim f_X(x | \theta) \quad (6.2)$$

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:

$$x[n] = A + w[n], \quad n \in \{0, \dots, N-1\} \quad (6.3)$$

This list isn't exhaustive, but gives an example of the type of **parameter estimation** problems that need to be addressed.

6.2 Properties of Estimators

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

$$\hat{\theta} = \hat{\theta}[\mathcal{X}] = \hat{\theta}[\{x[n]\}_0^{N-1}] \quad (6.4)$$



Sidebar 14 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 enemy 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 $\hat{\theta}[\mathcal{X}]$ 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**:

$$\hat{\mu}_X = \hat{\theta}[\mathcal{X}] = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \quad (\text{M:3.6.1})$$

Similarly, a natural estimator of the variance, σ_X^2 , of the random variable $X(\zeta)$, $x[n]$, would be:

$$\hat{\sigma}_X^2 = \hat{\theta}'[\mathcal{X}] = \frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \hat{\mu}_X)^2 \quad (\text{M:3.6.2})$$

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.

Example 6.1 (Numerical Example). Suppose that $N = 1000$ observations are generated from a Gaussian density with mean $\mu = 5$ and variance $\sigma^2 = 1$. Use MATLAB and a Monte Carlo experiment to find the distribution of the sample mean.

SOLUTION. One realisation of the experiment would generate $N = 1000$ data points generated from $x[n] \sim \mathcal{N}(\mu = 5, \sigma^2 = 1)$ using the code:

```
mu = 5; sigma = 1; N = 1000;
x = mu + sigma * randn(N, 1);
muEst = sum(x)/N
```

The second line of the code utilises a probability transformation rule from a Gaussian density of unit variance and zero mean. This experiment can be repeated $K = 100000$ times to produce a Monte Carlo estimate. This can be achieved with the following code:

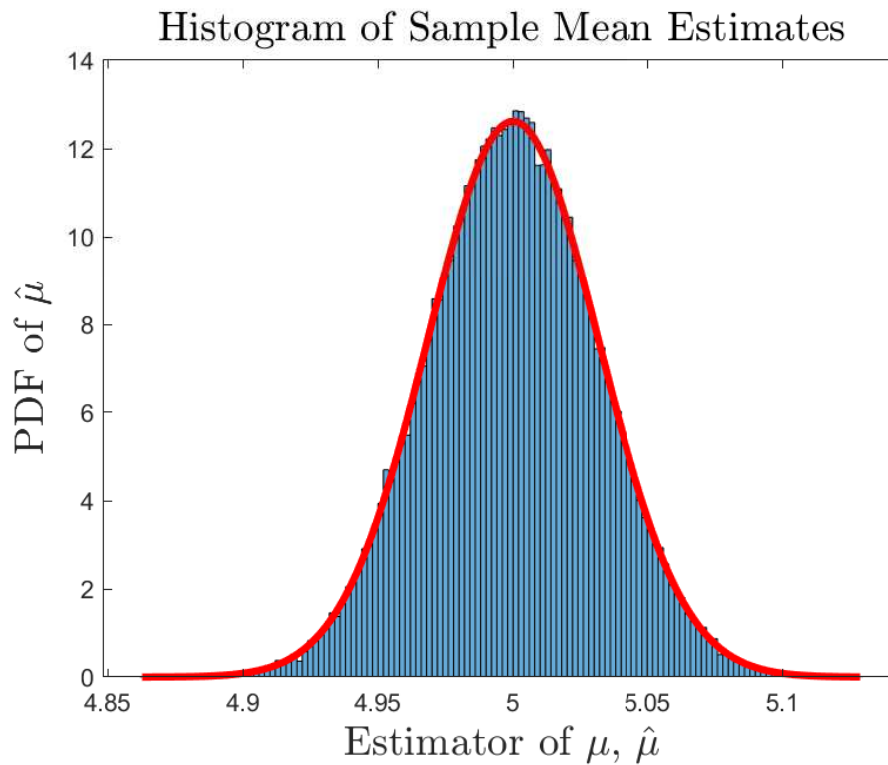


Figure 6.1: A Monte Carlo experiment showing the distribution of the Sample Mean estimator.

```

clear var; close all
N = 1000; K = 100000;
mu = 5; sigma = 1;

muEst = zeros(1, K);
for k = 1 : K
    x = mu + sigma * randn(N, 1);
    muEst(k) = sum(x) / N;
end
mean(muEst)

figure; histogram(muEst, 'Normalization', 'pdf');

L = 1000; % Number of points to plot
muPlot = linspace(min(muEst), max(muEst), L);
muPDF = normpdf(muPlot, mu, sigma/sqrt(N));

hold on; plot(muPlot, muPDF, 'r-', 'linewidth', 3);

```

The results of this Monte Carlo experiment are hence shown in Figure 6.12.

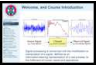
The set of N observations, $\{x[n]\}_{n=0}^{N-1}$ can be regarded as one realisation of the random process $\{x[n, \zeta]\}_{n=0}^{N-1}$ which, technically, is defined on an N -dimensional sample space. Hence, the estimator $\theta[\{x[n, \zeta]\}_{n=0}^{N-1}]$ becomes a RV whose probability density function can be obtained from the joint-pdf of the random variables $\{x[n, \zeta]\}_{n=0}^{N-1}$ 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, 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.

– End-of-Topic 40: **Introduction to Estimation Theory and the Definition of an Estimator** –



6.2.1 What makes a good estimator?



Topic Summary 41 Measuring Performance of an Estimator

New slide

Topic Objectives:

- Understanding how good an estimator is.
- Concepts and definitions of bias and variance.
- Calculating bias and variance.
- Understanding the bias-variance tradeoff for an estimator.

Topic Activities:

Type	Details	Duration	Progress
Watch video	20 : 20 min video	3 × length	
Read Handout	Read page 210 to page 213	8 mins/page	
Try Example	Try Example 6.2 and Example 6.2	15 mins	
Practice Exercises	Exercises ?? to ??	40 mins	

What makes a good estimator?

$Bias: E\{\hat{p}\} - \mu = 0$

$var(\hat{p}) = E\{(\hat{p} - \mu)^2\}$

True value μ

Here, the pdf of the estimated value, $\hat{\mu}$, is centered on the true value, μ . However, the spread of the estimated value around the true value is very large.

With a high probability the estimator could be a long way from the true value

Bias-variance - Trade-off

http://media.ed.ac.uk/media/1_7isroiw3

Video Summary: In this video, the question of measuring and quantifying the performance of an estimator is discussed. The video focusses on the concepts and definitions of bias and variance of the pdf of the estimator, and highlights the bias-variance trade-off; namely, that by introducing a small amount of bias in an estimator, the variance can be reduced. The normalised bias and normalised variance are also defined. Assuming the observations are independent, then the bias of the sample mean is calculated and shown to be unbiased. Similarly, the variance of the sample mean is calculated, using two similar but different calculations.

Figure 6.2 and Figure 6.3 illustrate properties of the sampling distribution, and how they might inform how to choose a *good estimator*.

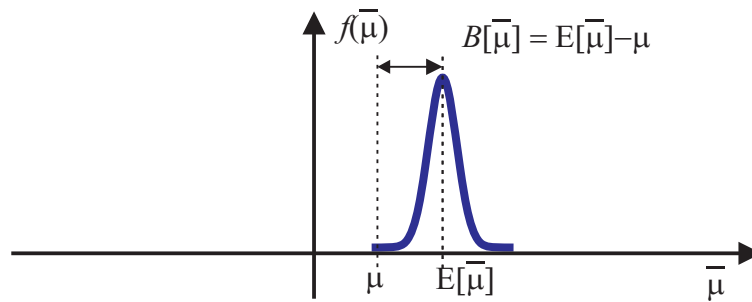


Figure 6.2: Here, the pdf of the estimated value, $\bar{\mu}$, is biased away from the true value, μ . However, the spread of the estimated value around the true value is small.

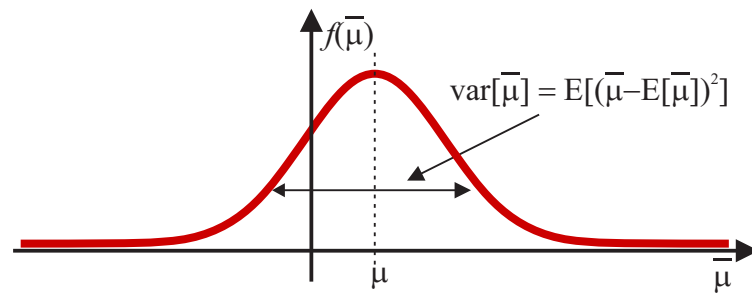


Figure 6.3: Here, the pdf of the estimated value, $\bar{\mu}$, is centered on the true value, μ . However, the spread of the estimated value around the true value is very large.

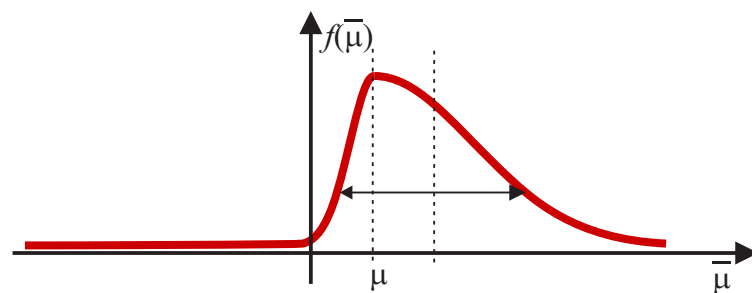


Figure 6.4: It is important to note that higher-order statistics can also play a part in quantifying the performance of an estimator, although that won't be considered further here.

Sidebar 15 Expectation w. r. t. what?

Note that the expectation is taken with respect to the pdf of the data \mathcal{X} , denoted by $p(\mathcal{X} | \theta)$. Thus, more precisely one would write:

$$B(\hat{\theta}) \triangleq \mathbb{E}_{p(\mathcal{X} | \theta)} [\hat{\theta}] - \theta \quad (6.5)$$

where

$$\mathbb{E}_{p(\mathcal{X} | \theta)} [\hat{\theta}] \triangleq \int_{\Theta} \hat{\theta}(\mathcal{X}) p(\mathcal{X} | \theta) d\mathcal{X} \quad (6.6)$$

However, often in textbooks and the literature, the pdf with which the expectation is taken against is omitted.

6.2.2 Bias of estimator

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

$$B(\hat{\theta}) \triangleq \mathbb{E} [\hat{\theta}] - \theta \quad (\text{M:3.6.3})$$

It is important to appreciate that the expectation is taken with respect to (w. r. t.) the observed data *given* the true parameter θ .

Therefore, the **normalised bias** is often used:

$$\epsilon_b(\hat{\theta}) \triangleq \frac{B(\hat{\theta})}{\theta} = \frac{\mathbb{E} [\hat{\theta}]}{\theta} - 1, \quad \theta \neq 0 \quad (\text{M:3.6.4})$$

Example 6.2 (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 $\mathbb{E} [\hat{\mu}_x] = \mathbb{E} \left[\frac{1}{N} \sum_{n=0}^{N-1} x[n] \right] = \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E} [x[n]] = \frac{N\mu_X}{N} = \mu_X$.

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:

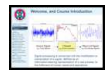
Definition 6.1 (Bias of an estimator). An estimate $\hat{\theta}_N$, based on N data observations, of a parameter θ is **unbiased** if

$$\mathbb{E} [\hat{\theta}_N] = \theta \quad (6.7)$$

Otherwise, the estimate is **biased** with bias $B(\hat{\theta}_N) = \mathbb{E} [\hat{\theta}_N] - \theta$. An estimate is **asymptotically unbiased** if

$$\lim_{N \rightarrow \infty} \mathbb{E} [\hat{\theta}_N] = \theta \quad (6.8)$$

◇



New slide

6.2.3 Variance of estimator

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

$$\text{var} [\hat{\theta}] = \sigma_{\hat{\theta}}^2 \triangleq \mathbb{E} \left[\left| \hat{\theta} - \mathbb{E} [\hat{\theta}] \right|^2 \right] \quad (\text{M:3.6.5})$$

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 in the next topic.

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

$$\epsilon_r \triangleq \frac{\sigma_{\hat{\theta}}}{\theta}, \quad \theta \neq 0 \quad (\text{M:3.6.6})$$

Example 6.3 (Variance of Sample Mean). Calculate the variance of the sample mean, assuming the observations are independent.

SOLUTION. Noting that the samples $\{x[n]\}_{n=0}^{N-1}$ are independent and identically distributed (i. i. d.) with variance σ_x^2 , then there are two approaches to calculating the variance.

The first is to use the result that the variance of a sum of independent random variables, is equal to the sum of the variances, or generalised to:

$$\text{var} \left[\sum_{n=0}^{N-1} c_n X_n(\zeta) \right] = \sum_{n=0}^{N-1} c_n^2 \text{var} [X_n(\zeta)] \quad (6.9)$$

Therefore,

$$\text{var} [\hat{\mu}_x] = \text{var} \left[\frac{1}{N} \sum_{n=0}^{N-1} x[n] \right] = \frac{1}{N^2} \sum_{n=0}^{N-1} \text{var} [x[n]] = \frac{\sigma_x^2}{N} \quad (6.10)$$

The second approach uses the result that $\mathbb{E} [x[n] x[m]] = \sigma_x^2 \delta(n - m) + \mu_x^2$. The sample mean estimator is unbiased, and therefore writing $\theta = \mu_x$, then $\mathbb{E} [\hat{\mu}_x] = \mu_x$. Therefore:

$$\text{var} [\hat{\mu}_x] = \mathbb{E} \left[\left| \left\{ \frac{1}{N} \sum_{n=0}^{N-1} x[n] \right\} - \mu_x \right|^2 \right] \quad (6.11)$$

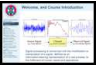
$$= \mathbb{E} \left[\frac{1}{N^2} \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} x[n] x[m] - 2 \frac{\mu_x}{N} \sum_{n=0}^{N-1} x[n] + \mu_x^2 \right] \quad (6.12)$$

$$= \frac{1}{N^2} \{ N [\sigma_x^2 + N \mu_x^2] - 2 N^2 \mu_x^2 + N^2 \mu_x^2 \} = \frac{\sigma_x^2}{N} \quad (6.13)$$

□



6.2.4 Mean square error



Topic Summary 42 Minimum Mean Square Error Estimators

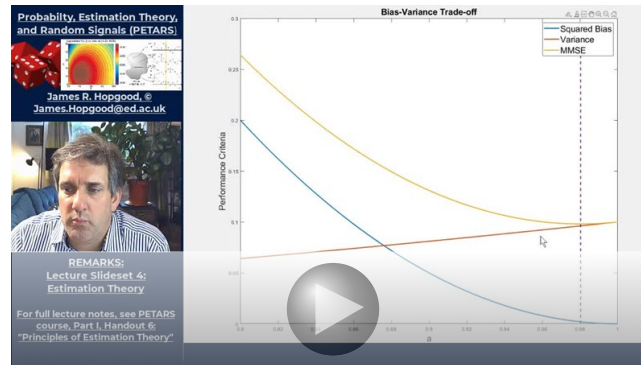
New slide

Topic Objectives:

- Definition of MSE and the MSE estimator.
- Relationship to bias and variance.
- Example of calculating MSE estimator.

Topic Activities:

Type	Details	Duration	Progress
Watch video	14 : 50 min video	3 × length	
Read Handout	Read page 214 to page 217	8 mins/page	
Try Code	Use the MATLAB code	10 minutes	
Try Example	Try Example 6.4	15 mins	



http://media.ed.ac.uk/media/1_4h9u0wfx

Video Summary: This video introduces the simple MSE as a criterion which trades-off bias and variance for an estimator. The relationship between the MSE and bias and variance is defined. The minimum MSE is introduced as an estimator which would appear to produce an improved design. However, through an example, it is shown that such estimators are sometimes unrealisable if there is bias. Nevertheless, there are applications where the MSE can produce results, or indeed inspire other estimators, such as estimators for variance (examples will be given later in the course).

Minimising variance can increase bias. A compromise criterion, and a natural one at that, is the MSE:

$$\text{MSE}(\hat{\theta}) = \mathbb{E} \left[\left| \hat{\theta} - \theta \right|^2 \right] = \sigma_{\hat{\theta}}^2 + |B(\hat{\theta})|^2 \quad (\text{M:3.6.7})$$

Again, it is important to remember that the expectation in the MSE term is w. r. t. the data, \mathbf{x} , as discussed in Sidebar 15 page 212.

PROOF (RELATIONSHIP BETWEEN MSE, VARIANCE AND BIAS OF AN ESTIMATOR.). Rewriting

Equation M:3.6.7 by subtracting and adding the mean of the estimator gives:

$$\text{MSE}(\hat{\theta}) = \mathbb{E} \left[|\hat{\theta} - \mathbb{E} [\hat{\theta}] - (\theta - \mathbb{E} [\hat{\theta}])|^2 \right] \quad (6.14)$$

$$= \mathbb{E} \left[|\hat{\theta} - \mathbb{E} [\hat{\theta}]|^2 \right] - \mathbb{E} \left[(\hat{\theta} - \mathbb{E} [\hat{\theta}])^* (\theta - \mathbb{E} [\hat{\theta}]) \right] \quad (6.15)$$

$$- \mathbb{E} \left[(\theta - \mathbb{E} [\hat{\theta}]) (\hat{\theta} - \mathbb{E} [\hat{\theta}])^* \right] + \mathbb{E} \left[|\theta - \mathbb{E} [\hat{\theta}]|^2 \right] \quad (6.16)$$

Now, note that $\mathbb{E} \left[|\theta - \mathbb{E} [\hat{\theta}]|^2 \right] = |\theta - \mathbb{E} [\hat{\theta}]|^2$, since both θ and $\mathbb{E} [\hat{\theta}]$ are deterministic values. Moreover,

$$\mathbb{E} \left[(\theta - \mathbb{E} [\hat{\theta}])^* (\hat{\theta} - \mathbb{E} [\hat{\theta}]) \right] = (\theta - \mathbb{E} [\hat{\theta}])^* \mathbb{E} \left[\hat{\theta} - \mathbb{E} [\hat{\theta}] \right] \quad (6.17)$$

$$= (\theta - \mathbb{E} [\hat{\theta}])^* \left\{ \mathbb{E} [\hat{\theta}] - \mathbb{E} [\hat{\theta}] \right\} = 0 \quad (6.18)$$

giving:

$$\text{MSE}(\hat{\theta}) = \underbrace{\mathbb{E} \left[|\hat{\theta} - \mathbb{E} [\hat{\theta}]|^2 \right]}_{\sigma_{\hat{\theta}}^2} + \underbrace{|\theta - \mathbb{E} [\hat{\theta}]|^2}_{B(\hat{\theta})} \quad (\text{M:3.6.9}) \quad \square$$

as required.

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

$$\hat{\theta}_{\text{MSE}} = \arg_{\hat{\theta}} \min \text{MSE}(\hat{\theta}) \quad (6.19)$$

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 6.19, 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 examples demonstrates these issues.

Example 6.4 ([Kay:1993, Example 2.1, Pages 16 and 19]). Consider the observations

$$x[n] = A + w[n], \quad n \in \{0, \dots, N-1\} \quad (\text{K:2.2})$$

where A is the parameter to be estimated, and $w[n]$ is white Gaussian noise (WGN) with variance σ^2 . The parameter A can take on any value in the interval $-\infty < A < \infty$. A reasonable estimator for the average value of $x[n]$, A , is:

$$\hat{A}_a = a \frac{1}{N} \sum_{n=0}^{N-1} x[n] \quad (6.20)$$

If $a = 1$, then this is just the sample mean. Find the optimal (modified) estimator \hat{A}_a by finding the value of a that minimises the MSE.

SOLUTION. Due to the linearity properties of the expectation operator, then it can be seen, as in the previous example, that:

$$\mathbb{E} [\hat{A}_a] = \mathbb{E} \left[a \frac{1}{N} \sum_{n=0}^{N-1} x[n] \right] = aA \quad (6.21)$$

for all A . Therefore, this is a **biased estimate** with bias $B(\hat{A}_a) = A(a - 1)$. As in the previous example, then:

$$\text{var} [\hat{A}_a] = \text{var} \left[a \frac{1}{N} \sum_{n=0}^{N-1} x[n] \right] \quad (6.22)$$

$$= \frac{a^2}{N^2} \sum_{n=0}^{N-1} \text{var} [x[n]] = \frac{a^2 \sigma^2}{N} \quad (6.23)$$

Hence, the MSE is given by:

$$\text{MSE}(\hat{A}_a) = \text{var} [\hat{A}_a] + |B(\hat{A}_a)|^2 = \frac{a^2 \sigma^2}{N} + (a - 1)^2 A^2 \quad (6.24)$$

In order to find the minimum mean-square error (MMSE), then differentiate this and set to zero:

$$\frac{d\text{MSE}(\hat{A}_a)}{da} = \frac{2a\sigma^2}{N} + 2(a - 1)A^2 \quad (6.25)$$

which is equal to zero when

$$a_{\text{opt}} = \frac{A^2}{A^2 + \frac{\sigma^2}{N}} \quad (6.26)$$

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, generally, lead to an unrealisable estimator. Although this is generally true, on occasion realisable MMSE estimators can be found.

Despite the unrealisable estimator, the result in Equation 6.27 can still be informative. First, note that Equation 6.27 can be written in the form:

$$a_{\text{opt}} = \frac{1}{1 + \frac{1}{N} \left(\frac{\sigma^2}{A^2} \right)} = \frac{1}{1 + \frac{1}{N \text{SNR}}} \quad (6.27)$$

where the signal-to-noise ratio (SNR) is the signal power, which in this case is the mean value squared, divided by the noise power, which in this case is the variance: $\text{SNR} = \frac{A^2}{\sigma^2}$. It is apparent that when N and the SNR are low, some value less than $a = 1$ may be appropriate.

Substituting Equation 6.27 into Equation 6.24, the minimum MSE can be calculated as:

$$\text{MSE} (a_{\text{opt}}) = \frac{a_{\text{opt}}^2 \sigma^2}{N} + (a_{\text{opt}} - 1)^2 A^2 \quad (6.28)$$

$$= \frac{\sigma^2}{N} [a_{\text{opt}}^2 + (a_{\text{opt}} - 1)^2 (N \text{SNR})] \quad (6.29)$$

$$= \frac{\sigma^2}{N} \left[\left(\frac{1}{1 + \frac{1}{N \text{SNR}}} \right)^2 + \left(\frac{1}{1 + \frac{1}{N \text{SNR}}} - 1 \right)^2 (N \text{SNR}) \right] \quad (6.30)$$

$$= \frac{\sigma^2}{N} \left[\left(\frac{1}{1 + \frac{1}{N \text{SNR}}} \right)^2 + \left(\frac{\frac{1}{N \text{SNR}}}{1 + \frac{1}{N \text{SNR}}} \right)^2 (N \text{SNR}) \right] \quad (6.31)$$

$$\text{MSE} (a_{\text{opt}}) = \frac{\sigma^2}{N} \left(\frac{1}{1 + \frac{1}{N \text{SNR}}} \right) \quad (6.32)$$

□

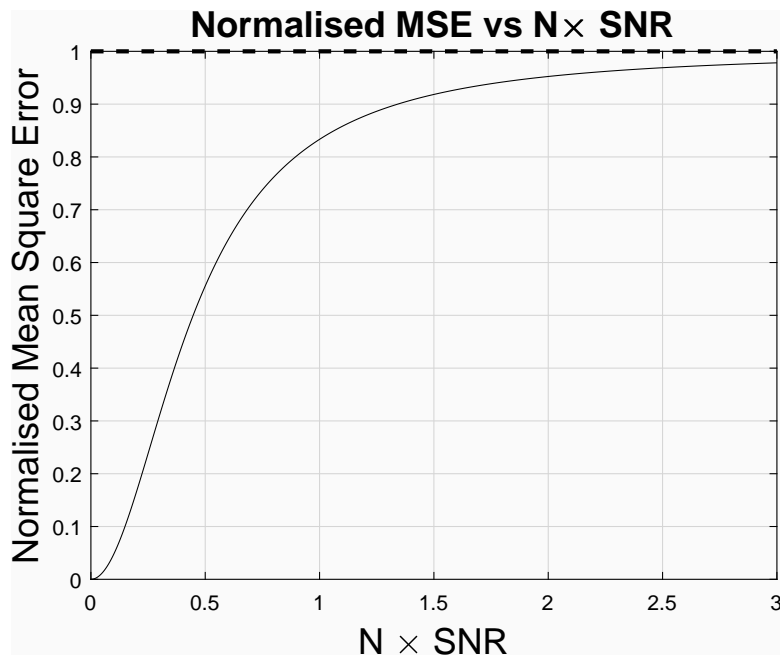


Figure 6.5: MSE vs $N \times \text{SNR}$ for the sample mean estimation problem.

This MSE is therefore the MSE of the sample mean, multiplied by a factor dependent on the SNR. This can therefore be plotted against this value, as shown in Figure 6.5, and indicates that for low SNR or a low number of samples, the estimator can do better than just the sample mean. Moreover, by plotting the bias, variance, and MSE separately, Figure 6.6 ultimately shows the bias-variance trade-off. Here, as the parameter a approaches 1, the bias reduces but the variance increases. Figure 6.6 also shows that a slightly lower value of a than unity gives a lower MSE.

Moreover, by plotting the bias, variance, and MSE as shown in Figure 6.6, we can see how the bias-variance trade-off occurs.

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.

– End-of-Topic 42: **Mean Square Error and MSE Estimators** –



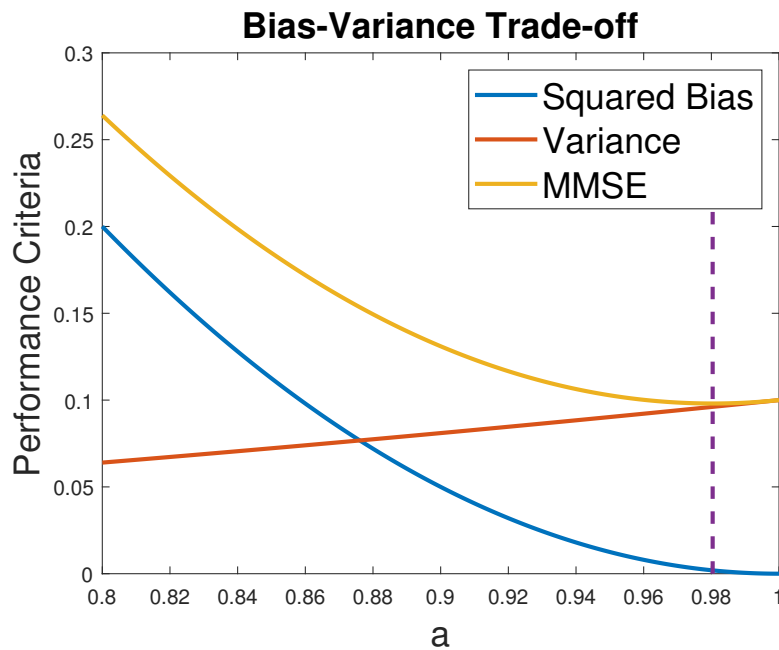
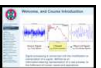


Figure 6.6: Plotting the bias, variance, and MSE.

6.2.5 Consistency of an Estimator



Topic Summary 43 Cramer Rao Lower Bound

New slide

Topic Objectives:

- Understanding the concept of a lower bound as a performance benchmark.
- Introduce the concept of the MVUE.
- Define and use the Cramér-Rao lower-bound (CRLB).
- Apply to the example of the sample mean.

Topic Activities:

Type	Details	Duration	Progress
Watch video	20 : 20 min video	3 × length	
Read Handout	Read page 221 to page 227	8 mins/page	
Try Example	Try Example 6.6	15 mins	

If the MSE of the estimator,

$$\text{MSE}(\hat{\theta}) = \mathbb{E} \left[|\hat{\theta} - \theta|^2 \right] = \sigma_{\hat{\theta}}^2 + |B(\hat{\theta})|^2 \quad (\text{M:3.6.7})$$

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 $N \rightarrow \infty$, 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 6.2 (Consistent Estimator). An estimate $\hat{\theta}_N$, based on N data observations, is **consistent** if

$$\lim_{N \rightarrow \infty} \Pr \left(\left| \hat{\theta}_N - \theta \right| < \epsilon \right) = 1 \quad (6.33)$$

◇

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 6.5 ([Manolakis:2001, Exercise 3.32, Page 147]). The Cauchy distribution with mean μ is given by:

$$f_X(x) = \frac{1}{\pi} \frac{1}{1 + (x - \mu)^2}, \quad x \in \mathbb{R} \quad (6.34)$$

Let $\{x_k\}_{k=0}^{N-1}$ be N i. i. d. RVs with this distribution. Consider the mean estimator based on these samples:

$$\hat{\mu} = \frac{1}{N} \sum_{k=0}^{N-1} x_k \quad (6.35)$$

Determine whether $\hat{\mu}$ is a consistent estimator of μ .

SOLUTION. It is simplest to use the definition that an estimator is consistent if $\lim_{N \rightarrow \infty} \text{MSE}(\theta) = 0$, where

$$\text{MSE}(\theta) = \mathbb{E} \left[|\hat{\theta} - \theta|^2 \right] = \sigma_{\hat{\theta}}^2 + |B(\hat{\theta})|^2 \quad (\text{M:3.6.7})$$

and

$$\sigma_{\hat{\theta}}^2 \triangleq \mathbb{E} \left[\left| \hat{\theta} - \mathbb{E} [\hat{\theta}] \right|^2 \right] \equiv \mathbb{E} \left[|\hat{\theta}|^2 \right] - \mathbb{E}^2 [\hat{\theta}] \quad (\text{M:3.6.5})$$

Hence, by noting that $\mathbb{E} [\hat{\mu}] = \mu$, such that $|B(\hat{\theta})|^2 = 0$, then the MSE is given by:

$$\text{MSE}(\theta) = \sigma_{\hat{\theta}}^2 = \mathbb{E} [|\hat{\mu}|^2] - \mathbb{E}^2 [\hat{\mu}] \quad (6.36)$$

$$\equiv \mathbb{E} [|\hat{\mu} - \mathbb{E} [\hat{\mu}]|^2] = \mathbb{E} \left[\left| \frac{1}{N} \sum_{k=0}^{N-1} x_k - \mu \right|^2 \right] \quad (6.37)$$

$$\equiv \frac{1}{N^2} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} \mathbb{E} [x_k x_l] - \mu^2 \quad (6.38)$$

Since the samples are independent and identically distributed (i. i. d.), then the autocorrelation function is given by:

$$\mathbb{E} [x_k x_l] = \begin{cases} \mathbb{E} [x_k] \mathbb{E} [x_l] & k \neq l \\ \mathbb{E} [x_k^2] & k = l \end{cases} \quad (6.39)$$

$$= \begin{cases} \mu^2 & k \neq l \\ \mu^2 + \sigma^2 & k = l \end{cases} \quad (6.40)$$

$$= \sigma^2 \delta(k - l) + \mu^2 \quad (6.41)$$

Hence,

$$\text{MSE}(\theta) = \frac{1}{N^2} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} (\sigma^2 \delta(k-l) + \mu^2) - \mu^2 \quad (6.42)$$

$$= \frac{1}{N^2} \sum_{k=0}^{N-1} (\sigma^2 + N\mu^2) - \mu^2 \quad (6.43)$$

$$= \frac{1}{N} (\sigma^2 + N\mu^2) - \mu^2 = \frac{\sigma^2}{N} \quad (6.44)$$

□

Since the variance for a Cauchy distribution is unbounded, such that $\sigma^2 \rightarrow \infty$, then $\lim_{N \rightarrow \infty} \text{MSE}(\theta)$ does not converge to zero, and is therefore **not consistent**.

Definition 6.3 (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.

– End-of-Topic 43: Consistency of Estimator –



6.2.6 Cramer-Rao Lower Bound

Topic Summary 44 Cramer Rao Lower Bound

Topic Objectives:

- Understanding the concept of a lower bound as a performance benchmark.
- Introduce the concept of the MVUE.
- Define and use the CRLB.
- Apply to the example of the sample mean.

Topic Activities:

Type	Details	Duration	Progress
Watch video	20 : 20 min video	3 × length	
Read Handout	Read page 221 to page 227	8 mins/page	
Try Example	Try Example 6.6	15 mins	
Practice Exercises	Exercises ?? and ??	50 mins	

Cramer-Rao Lower Bound

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

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

Alternatively, it may also be expressed as:

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

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

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

http://media.ed.ac.uk/media/1_r6cqib2g

Video Summary: In this video, the question of finding the lower bound on the performance of all estimators for a particular probabilistic problem, as a benchmark with which to compare the performance of a given estimator. The CRLB is introduced in this video for this benchmark, for the class of unbiased estimators. The Fisher Information is discussed, and it is shown how to test for the existence of a MVUE which attains the CRLB. An example is shown for deriving the sample mean, which is a MVUE (and as a result also the MSE estimator). In the example, the minimum variance is found through the two alternate but equivalent expressions for the CRLB.

In the previous sections, the performance of a given estimator has been considered; what is the bias, and what is the variance? The MSE criterion gives a possible design method for finding the structural form of an optimal estimator, but isn't always realisable. This leads to the general question of whether there is a particular methodology for designing an estimator for a given probabilistic problem.

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

$$\text{var} [\hat{\theta}] \geq \frac{1}{\mathbb{E} \left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} \right)^2 \right]} \quad (\text{M:3.6.17})$$

Alternatively, it may also be expressed as:

$$\text{var} [\hat{\theta}] \geq - \frac{1}{\mathbb{E} \left[\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta^2} \right]} \quad (\text{M:3.6.18})$$

The function $\ln f_{\mathbf{X}}(\mathbf{x} | \theta)$ is called the **log-likelihood** function of θ . A discussion about the likelihood-function is given in Sidebar 16.

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

$$\hat{\theta} - \theta = K(\theta) \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} \quad (\text{K:3.7})$$

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:

$$\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} = I(\theta) (\hat{\theta} - \theta), \quad \text{where } I(\theta) = K^{-1}(\theta). \quad (6.49)$$

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 $\mathbb{E} [\hat{\theta} - \theta] = 0$, which may be expressed as:

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} (\hat{\theta} - \theta) f_{\mathbf{X}}(\mathbf{x} | \theta) d\mathbf{x} = 0 \quad (\text{M:3.6.11})$$

Differentiating w. r. t. the true parameter θ , and assuming a real-value $\hat{\theta}$, then:

$$0 = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} [(\hat{\theta} - \theta) f_{\mathbf{X}}(\mathbf{x} | \theta)] d\mathbf{x} \quad (6.50)$$

$$= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} (\hat{\theta} - \theta) \frac{\partial f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} d\mathbf{x} - \underbrace{\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x} | \theta) d\mathbf{x}}_{=1} \quad (\text{M:3.6.12})$$

Sidebar 16 The likelihood function

The likelihood function is discussed in detail in Section 6.3. 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 $f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})$, 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:

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) \equiv f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta}) \Big|_{\text{fixed } \mathbf{x}, \text{ variable } \boldsymbol{\theta}} \quad (6.45)$$

Thus, the likelihood function $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$ should be interpreted as a function of $\boldsymbol{\theta}$ given \mathbf{x} . However, it is important to note that $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) \equiv f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ is not necessarily a pdf since, in general, it does not integrate to one over $\boldsymbol{\theta}$:

$$\int \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) d\boldsymbol{\theta} = \int f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta}) d\boldsymbol{\theta} \neq 1 \quad (6.46)$$

Note, however, that according to Bayes's theorem:

$$\int f_{\Theta}(\boldsymbol{\theta} | \mathbf{x}) d\boldsymbol{\theta} = \int \frac{f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta}) f_{\Theta}(\boldsymbol{\theta})}{f_{\mathbf{X}}(\mathbf{x})} d\boldsymbol{\theta} = 1 \quad (6.47)$$

or alternatively, a weighted version of the likelihood gives rise to the probability of the observations:

$$\int \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) f_{\Theta}(\boldsymbol{\theta}) d\boldsymbol{\theta} = f_{\mathbf{X}}(\mathbf{x}) \quad (6.48)$$

In other words, it is simply important to not interpret the likelihood function as a pdf, and simply to be careful with the manipulations.

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

$$\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} = \frac{1}{f_{\mathbf{X}}(\mathbf{x} | \theta)} \frac{\partial f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} \quad (6.51)$$

or,

$$\frac{\partial f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} = \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} f_{\mathbf{X}}(\mathbf{x} | \theta) \quad (\text{M:3.6.13})$$

then substituting into Equation M:3.6.12 gives:

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left\{ (\hat{\theta} - \theta) \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} \right\} f_{\mathbf{X}}(\mathbf{x} | \theta) d\mathbf{x} = 1 \quad (\text{M:3.6.14})$$

which can be written using the expectation operator as:

$$\mathbb{E} \left[(\hat{\theta} - \theta) \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} \right] = 1 \quad (\text{M:3.6.15})$$

Now, using the **Cauchy-Schwartz inequality** (see [Papoulis:1991]), which states that:

$$|\mathbb{E} [\mathbf{X}(\zeta)\mathbf{Y}(\zeta)]|^2 \leq \mathbb{E} [|\mathbf{X}(\zeta)|^2] \mathbb{E} [|\mathbf{Y}(\zeta)|^2] \quad (6.52)$$

then squaring both sides of Equation M:3.6.15 gives

$$1 = \mathbb{E}^2 \left[(\hat{\theta} - \theta) \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} \right] \leq \mathbb{E} [(\hat{\theta} - \theta)^2] \mathbb{E} \left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} \right)^2 \right] \quad (\text{M:3.6.16})$$

Note that the Cauchy-Schwartz inequality becomes an 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:

$$(\hat{\theta} - \theta)^2 f_{\mathbf{X}}(\mathbf{x} | \theta) = K(\theta) \left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} \right)^2 f_{\mathbf{X}}(\mathbf{x} | \theta) \quad (6.53)$$

or, alternatively,

$$\hat{\theta} - \theta = K(\theta) \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} \quad (\text{K:3.7})$$

This is the minimum variance unbiased estimator. Since the estimator is unbiased, then $\text{var} [\hat{\theta}] = \mathbb{E} [(\hat{\theta} - \theta)^2]$, and therefore:

$$\text{var} [\hat{\theta}] \geq \frac{1}{\mathbb{E} \left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} \right)^2 \right]} \quad (\text{M:3.6.17})$$

To derive the second form by starting with the simple condition that:

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x} | \theta) d\mathbf{x} = 1 \quad (6.54)$$

Differentiating once w. r. t. to θ and using Equation M:3.6.13 gives

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} d\mathbf{x} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} f_{\mathbf{X}}(\mathbf{x} | \theta) d\mathbf{x} = 0 \quad (6.55)$$

and differentiating again gives:

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta^2} f_{\mathbf{X}}(\mathbf{x} | \theta) + \left\{ \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} \right\}^2 f_{\mathbf{X}}(\mathbf{x} | \theta) \right) d\mathbf{x} = 0 \quad (6.56)$$

which gives the desired result

$$\mathbb{E} \left[\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta^2} \right] = -\mathbb{E} \left[\left\{ \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} \right\}^2 \right] \quad (6.57)$$

This can then be substituted into Equation M:3.6.17.

Note that a generalisation of the CRLB for biased estimates is given by:

$$\text{var} [\hat{\theta}] \geq \frac{\left(1 + \frac{\partial B(\hat{\theta})}{\partial \theta}\right)^2}{\mathbb{E} \left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} \right)^2 \right]} \quad (6.58)$$

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 6.6 ([Kay:1993, Example 3.3, Page 31]). Consider again the observations:

$$x[n] = A + w[n], \quad n \in \{0, \dots, N-1\} \quad (\text{K:2.2})$$

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 $w[n]$ and $x[n]$ is linear, with a multiplication factor of 1, the *likelihood function* can be written down as:

$$f_{\mathbf{X}}(\mathbf{x} | A) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2\sigma^2} (x[n] - A)^2 \right] \quad (6.66)$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \right] \quad (6.67)$$

Note, a more detailed derivation of this likelihood is given in Sidebar 17 on page 226. Taking the first derivative of the **log-likelihood** gives:

$$\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | A)}{\partial A} = \frac{\partial}{\partial A} \left[-\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \right] \quad (6.68)$$

$$= \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A) = \frac{N}{\sigma^2} \left(\left\{ \frac{1}{N} \sum_{n=0}^{N-1} x[n] \right\} - A \right) \quad (6.69)$$

$$= \frac{N}{\sigma^2} (\hat{\mu}_X - A) \quad (\text{K:3.8})$$

where $\hat{\mu}_X$ is the sample mean.

Sidebar 17 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:

$$x[n] = s[n; \boldsymbol{\theta}] + w[n], \quad w[n] \sim \mathcal{N}(0, \sigma_w^2) \quad (6.59)$$

where $s[n; \boldsymbol{\theta}]$ denotes a parametric model for the underlying signal, and is dependent on a parameter (vector) $\boldsymbol{\theta}$. 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 $\boldsymbol{\theta}$ 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:

$$J_{w[n] \rightarrow x[n]} = \frac{\partial x[n]}{\partial w[n]} = 1 \quad (6.60)$$

it follows that

$$f_X(x[n] | \boldsymbol{\theta}) = \frac{f_W(x[n] - s[n; \boldsymbol{\theta}])}{J_{w[n] \rightarrow x[n]}} = \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left\{-\frac{(x[n] - s[n; \boldsymbol{\theta}])^2}{2\sigma_w^2}\right\} \quad (6.61)$$

where it is implicitly understood that $f_X(x[n] | \boldsymbol{\theta}) = f_X(x[n] | \boldsymbol{\theta}, \sigma_w^2)$ 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:

$$f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta}) = f_{\mathbf{X}}(x[0], \dots, x[N-1] | \boldsymbol{\theta}) \quad (6.62)$$

$$= \prod_{n=0}^{N-1} f_X(x[n] | \boldsymbol{\theta}) \quad (6.63)$$

$$= \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left\{-\frac{(x[n] - s[n; \boldsymbol{\theta}])^2}{2\sigma_w^2}\right\} \quad (6.64)$$

$$= \frac{1}{(2\pi\sigma_w^2)^{\frac{N}{2}}} \exp\left\{-\frac{\sum_{n=0}^{N-1} (x[n] - s[n; \boldsymbol{\theta}])^2}{2\sigma_w^2}\right\} \quad (6.65)$$

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.

Differentiating again, then:

$$\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} | A)}{\partial A^2} = -\frac{N}{\sigma^2} \quad (6.70)$$

and noting that this second derivative is constant, then the CRLB is given by:

$$\text{var} [\hat{A}] \geq \frac{\sigma^2}{N} \quad (\text{K:3.9})$$

Comparing Equation K:3.7 and Equation K:3.8, where it is noted the first derivative of the log-likelihood is in the form:

$$\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} = I(\theta) (\hat{\theta} - \theta) = \frac{N}{\sigma^2} \left(\left\{ \frac{1}{N} \sum_{n=0}^{N-1} x[n] \right\} - A \right) \quad (6.71)$$

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 $\text{var} [\hat{A}] = \frac{\sigma^2}{N}$. □

– End-of-Topic 44: **Introduction to the CRLB and how to identify MVUE that satisfy the bound** –



Sidebar 18 Alternative Solution to Example 6.6

The solution to Example 6.6 used the second derivative form of the CRLB. But what if, in fact, the first version of the CRLB had been used, which calculates the square of the first derivative? What would the calculation look like?

Returning to Equation 6.69 and using the first form of the CRLB:

$$\text{var} [\hat{\theta}] \geq \frac{1}{\mathbb{E} \left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} \right)^2 \right]} \quad (\text{M:3.6.17})$$

Then note that

$$\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} \right)^2 = \left[\frac{1}{\sigma^2} \left(\sum_{n=0}^{N-1} x[n] - A \right) \right]^2 \quad (6.72)$$

$$= \frac{1}{\sigma^4} \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} (x[n] - A) (x[m] - A) \quad (6.73)$$

Taking expectations, then note that

$$\mathbb{E} [(x[n] - A) (x[m] - A)] = \begin{cases} \mathbb{E} [(x[n] - A)^2] = \sigma^2 & n = m \\ \mathbb{E} [(x[n] - A)] \mathbb{E} [(x[m] - A)] = 0 & m \neq n \end{cases} \quad (6.74)$$

where the independence of $x[n]$ and $x[m]$ have been used for $n \neq m$, and the fact that the first and second central moments are zero and the variance, respectively. Hence, in the double summation, $n = m$ occurs N times (giving rise to $N \sigma^2$ terms), and $n \neq m$ occurs $N^2 - N$ times (giving rise to $N^2 - N$ zero terms). Therefore:

$$\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \theta)}{\partial \theta} \right)^2 = \frac{1}{\sigma^2} \times N \times \sigma^2 = \frac{N}{\sigma^2} \quad (6.75)$$

which gives the same answer as determined in Unknown Exercise:Taxi2.

6.2.7 Estimating Multiple Parameters

Topic Summary 45 Cramer-Rao Lower Bound for Parameter Vectors

Topic Objectives:

- Extending the properties of scalar estimators to parameter vectors.
- Defining the Fisher information matrix (FIM) and the multi-parameter CRLB.
- Apply to the example of fitting a straight line.

Topic Activities:

Type	Details	Duration	Progress
Watch video	24 : 39 min video	3 × length	
Read Handout	Read page 229 to page 234	8 mins/page	
Try Example	Try Example 6.7	25 mins	
Practice Exercises	Exercises ??	30 mins	

http://media.ed.ac.uk/media/1_sxx68ats

Video Summary: In this video, the concepts in estimation theory introduced so far for scalar random variables are extended to deal with estimating multiple parameters, for example the mean and variance of a distribution simultaneously. The definition of a vector parameter estimator is introduced, and the example of extending the definition of bias. The principal focus of the video is on extending the CRLB to real parameter vectors, by placing a bound on the covariance matrix of the estimator. Parallels with the scalar CRLB are made throughout, but the emphasis is on the key calculation of the Fisher information matrix (FIM). This is the expectation of functions of the derivatives of the log-likelihood function, but considering the derivatives with respect to all the elements of the parameter vector. Finally, the line-fitting example of estimating the parameters of a straight line to fit a set of data that is assumed to follow a linear model. The FIM and CRLB are calculated, and it is shown that in this case the MVUE can be found as before. Numerical simulations are also provided to demonstrate the correctness of the calculations.

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 or pdf 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, θ , of a random event $X(\zeta)$ can be estimated from a set of observations, $\mathcal{X} = \{x[n]\}_0^{N-1}$, using some function:

$$\hat{\theta} = \hat{\theta}[\mathcal{X}] = \hat{\theta}[\{x[n]\}_0^{N-1}] \quad (6.76)$$

The definitions of **unbiasedness**, **consistency**, **efficiency**, and the **CRLB** are all straightforward extensions of the definitions and results for scalar parameter estimates.

Assuming θ is a $P \times 1$ parameter vector, these properties are:

Unbiased Estimator An estimate $\hat{\theta}_N$ is **unbiased** if

$$\mathbb{E}[\hat{\theta}_N] = \theta \quad (6.77)$$

Otherwise, the estimate is **biased** with bias $\mathbf{b}(\hat{\theta}_N) = \mathbb{E}[\hat{\theta}_N] - \theta$. An estimate is **asymptotically unbiased** if:

$$\lim_{N \rightarrow \infty} \mathbb{E}[\hat{\theta}_N] = \theta \quad (6.78)$$

Consistent Estimator An estimate $\hat{\theta}_N$, based on N data observations, is **consistent** if

$$\lim_{N \rightarrow \infty} \Pr\left(\left|\hat{\theta}_N - \theta\right| < \epsilon\right) = 1 \quad (6.79)$$

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 .

Efficient Estimator An estimate $\hat{\theta}$ is said to be **efficient** w. r. t. another estimate $\hat{\theta}'$ if the difference of their covariance matrices $\Gamma_{\hat{\theta}'} - \Gamma_{\hat{\theta}}$ is positive definite. This implies that the variance of every component of $\hat{\theta}$ must be smaller than the variance of the corresponding component of $\hat{\theta}'$. If $\hat{\theta}_N$ is unbiased and efficient with respect to $\hat{\theta}_{N-1}$ for all N , then $\hat{\theta}_N$ is a **consistent estimate**.

Theorem 6.2 (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 $\hat{\theta}$ is unbiased, then the vector parameter CRLB will place a bound on the variance of each element, as well as all the elements of the covariance matrix. 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:

$$\mathbf{s} \equiv \mathbf{s}(\mathbf{x}; \theta) \triangleq \nabla_{\theta} \ln f_{\mathbf{X}}(\mathbf{x} | \theta) \quad (\text{T:6.43})$$

The vector \mathbf{s} is called the **score** for θ based on \mathbf{x} . If $\hat{\theta}$ is substituted for θ , 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). The covariance of the score vector is known as the **FIM**, and is assumed to be nonsingular:

$$\mathbf{J}(\theta) = \mathbb{E}[\mathbf{s}(\mathbf{x}; \theta) \mathbf{s}^T(\mathbf{x}; \theta)] \quad (\text{T:6.42})$$

This form is equivalent to the first form of the scalar CRLB shown in Equation M:3.6.17 on page 228.

The Fisher information matrix can also be written in the following equivalent form:

$$[\mathbf{J}(\boldsymbol{\theta})]_{ij} = -\mathbb{E} \left[\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right] \quad (\text{K:3.21})$$

If $\hat{\boldsymbol{\theta}}$ is any unbiased estimate, and $\boldsymbol{\Gamma}_{\hat{\boldsymbol{\theta}}}$ is the covariance matrix of $\hat{\boldsymbol{\theta}}$, then the CRLB can be stated as:

$$\boldsymbol{\Gamma}_{\hat{\boldsymbol{\theta}}} \geq \mathbf{J}^{-1}(\boldsymbol{\theta}) \quad (6.80)$$

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:

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \mathbf{J}^{-1}(\boldsymbol{\theta}) \mathbf{s}(\mathbf{x}; \boldsymbol{\theta}) \quad (\text{T:6.47})$$

where $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{x})$ 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.

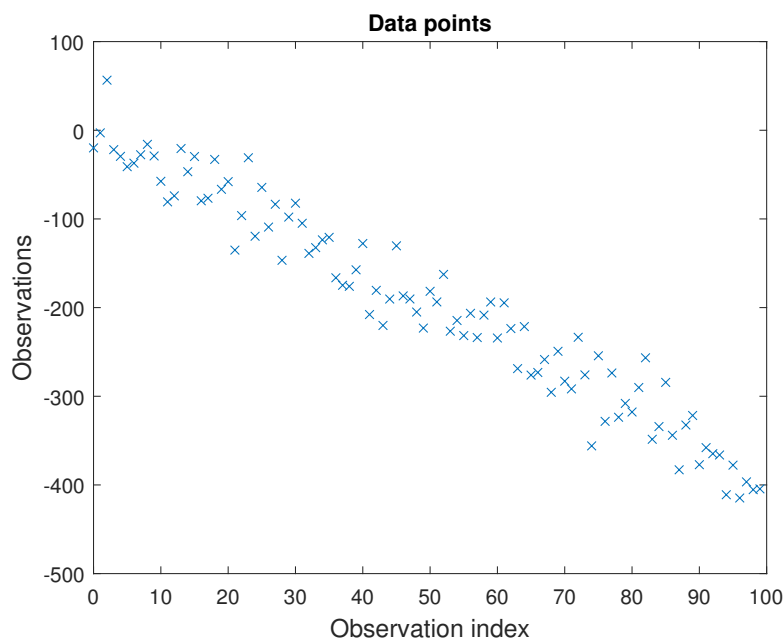
PROOF. For a full proof, see [Therrien:1992, Page 298], or [Kay:1993]. However, the proof is relatively straightforward and is analogous to the proof for the case of the scalar real parameter. It currently omitted from this document.

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 6.7 ([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:

$$x[n] = A + Bn + w[n], \quad n \in \{0, 1, \dots, N-1\} \quad (6.81)$$

where $w[n]$ is WGN with variance σ^2 . Determine the CRLB for the slope B and the intercept A , assuming σ^2 is known.



SOLUTION. The 2×2 Fisher information matrix (FIM) is given by:

$$\mathbf{J}(\boldsymbol{\theta}) = \mathbb{E} [\mathbf{s}(\mathbf{x}; \boldsymbol{\theta}) \mathbf{s}^T(\mathbf{x}; \boldsymbol{\theta})] \quad (\text{T:6.42})$$

$$= \mathbb{E} [\nabla_{\boldsymbol{\theta}} \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})] \quad (6.82)$$

$$= \begin{bmatrix} \mathbb{E} \left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial A} \right)^2 \right] & \mathbb{E} \left[\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial A} \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial B} \right] \\ \mathbb{E} \left[\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial B} \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial A} \right] & \mathbb{E} \left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial B} \right)^2 \right] \end{bmatrix} \quad (6.83)$$

where the notation $\boldsymbol{\theta} = [A, B]^T$ is used as a shorthand.

Alternatively, the elements of the Fisher information matrix be found using:

$$[\mathbf{J}(\boldsymbol{\theta})]_{ij} = -\mathbb{E} \left[\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right] \quad (\text{K:3.21})$$

or as a matrix:

$$\mathbf{J}(\boldsymbol{\theta}) = - \begin{bmatrix} \frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial A^2} & \frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial A \partial B} \\ \frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial A \partial B} & \frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial B^2} \end{bmatrix} \quad (6.84)$$

This alternative expression is often a more straightforward method for evaluating the Fisher information matrix, and will be used here. Similar to the derivation in the case of a DC signal in WGN, the likelihood function can be written as

$$f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A - Bn)^2 \right] \quad (6.85)$$

from which the following derivatives follow:

$$\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial A} = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A - Bn) \quad (6.86)$$

$$\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial B} = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A - Bn) n \quad (6.87)$$

and

$$\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial A^2} = -\frac{N}{\sigma^2} \quad (6.88)$$

$$\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial A \partial B} = -\frac{1}{\sigma^2} \sum_{n=0}^{N-1} n \quad (6.89)$$

$$\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial B^2} = -\frac{1}{\sigma^2} \sum_{n=0}^{N-1} n^2 \quad (6.90)$$

where it is noted that

$$\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial A \partial B} = \frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial B \partial A} \quad (6.91)$$

Using the identities that

$$\sum_{n=1}^N n = \frac{1}{2} N(N+1) \quad \text{and} \quad \sum_{n=1}^N n^2 = \frac{1}{6} N(N+1)(2N+1) \quad (6.92)$$

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:

$$\mathbf{J}(\boldsymbol{\theta}) = \frac{1}{\sigma^2} \begin{bmatrix} N & \frac{N(N-1)}{2} \\ \frac{N(N-1)}{2} & \frac{N(N-1)^2(2N-1)}{6} \end{bmatrix} \quad (6.93)$$

Inverting this FIM yields:

$$\mathbf{J}^{-1}(\boldsymbol{\theta}) = \sigma^2 \begin{bmatrix} \frac{2(2N-1)}{N(N+1)} & -\frac{6}{N(N+1)} \\ -\frac{6}{N(N+1)} & \frac{12}{N(N^2-1)} \end{bmatrix} \quad (6.94)$$

or, equivalently, the covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by:

$$\boldsymbol{\Gamma}_{\hat{\boldsymbol{\theta}}} \geq \frac{2\sigma^2}{N(N+1)} \begin{bmatrix} (2N-1) & -3 \\ -3 & \frac{6}{N-1} \end{bmatrix} \quad (6.95)$$

Hence, it can be deduced that the variances for the individual parameters are given by the CRLB or:

$$\text{var} [\hat{A}] \geq \frac{2(2N-1)\sigma^2}{N(N+1)} \quad (6.96)$$

$$\text{var} [\hat{B}] \geq \frac{12\sigma^2}{N(N^2-1)} \quad (6.97)$$

Finally, note that a MVUE, if it exists, satisfies the relationship:

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \mathbf{J}(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta}) \quad (\text{T:6.47})$$

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, it follows that using the expressions for the terms in the RHS

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \mathbf{J}(\boldsymbol{\theta})^{-1} \begin{bmatrix} \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial A} \\ \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial B} \end{bmatrix} \quad (6.98)$$

$$\begin{bmatrix} \hat{A} - A \\ \hat{B} - B \end{bmatrix} = \sigma^2 \begin{bmatrix} \frac{2(2N-1)}{N(N+1)} & -\frac{6}{N(N+1)} \\ -\frac{6}{N(N+1)} & \frac{12}{N(N^2-1)} \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A - Bn) \\ \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A - Bn) n \end{bmatrix} \quad (6.99)$$

$$= \begin{bmatrix} \frac{2(2N-1)}{N(N+1)} \sum_{n=0}^{N-1} (x[n] - A - Bn) - \frac{6}{N(N+1)} \sum_{n=0}^{N-1} (x[n] - A - Bn) n \\ -\frac{6}{N(N+1)} \sum_{n=0}^{N-1} (x[n] - A - Bn) + \frac{12}{N(N^2-1)} \sum_{n=0}^{N-1} (x[n] - A - Bn) n \end{bmatrix} \quad (6.100)$$

$$= \frac{2}{N(N+1)} \begin{bmatrix} (2N-1) \sum_{n=0}^{N-1} x[n] - 3 \sum_{n=0}^{N-1} n x[n] \\ -3 \sum_{n=0}^{N-1} x[n] + \frac{6}{(N-1)} \sum_{n=0}^{N-1} n x[n] \end{bmatrix} - \begin{bmatrix} A \\ B \end{bmatrix} \quad (6.101)$$

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.

This gives the final result that:

$$\begin{bmatrix} \hat{A} \\ \hat{B} \end{bmatrix} = \frac{2}{N(N+1)} \begin{bmatrix} (2N-1) \sum_{n=0}^{N-1} x[n] - 3 \sum_{n=0}^{N-1} n x[n] \\ -3 \sum_{n=0}^{N-1} x[n] + \frac{6}{(N-1)} \sum_{n=0}^{N-1} n x[n] \end{bmatrix} \quad (6.102)$$

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.

A numerical result is show to finish off this example.

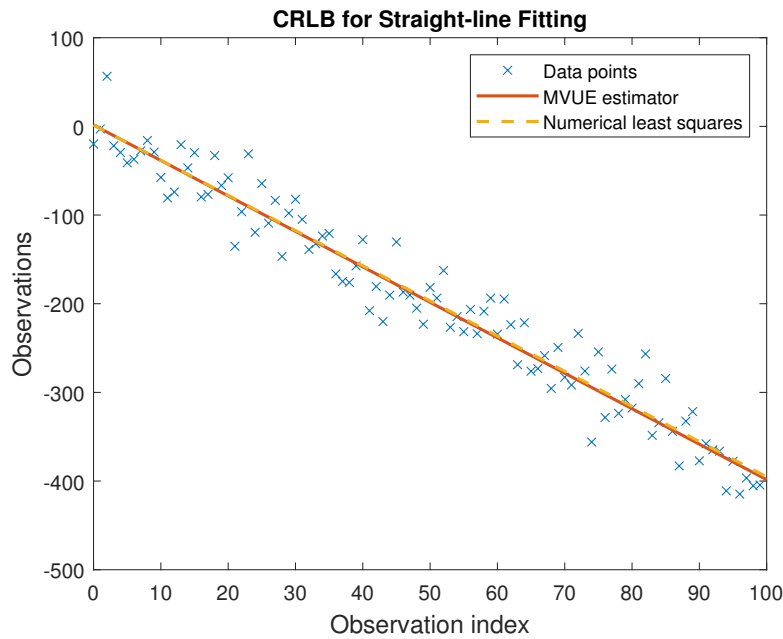


Figure 6.7: Data drawn from the model in Equation 6.81.

- Let $\sigma^2 = 1000$, $A = 4$, $B = -4$, and generate 100 data points.
- This gives the covariance matrix of $\hat{\theta}$ as:

$$\mathbf{\Gamma}_{\hat{\theta}} = \begin{bmatrix} 39.2079 & -0.5941 \\ -0.5941 & 0.0120 \end{bmatrix} \quad (6.103)$$

□

- Remember $\hat{\theta}$ only depends on N and σ^2 , and is not actually related to the value of A and B . So for a given N and σ^2 , the uncertainty is always the same.
- The estimates of the intercept, \hat{A} , will have a lot higher variance than the estimates for the gradient, \hat{B} .

A given realisation is shown in Figure 6.7, and the validation of the results is shown in Figure 6.8 which is a histogram of parameter estimates drawn from a Monte Carlo estimate of 1000 different noise realisations. The sample variances are also shown, and sample variance will be discussed elsewhere.

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 $\text{var}[\hat{A}] \geq \frac{\sigma^2}{N}$, 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*.



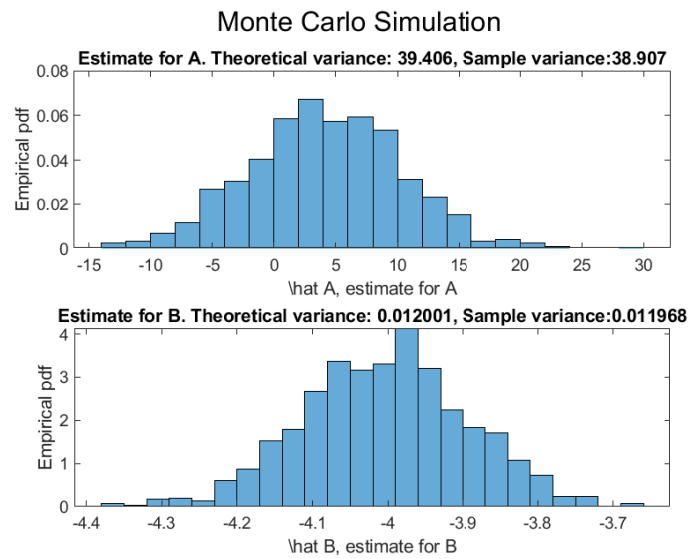


Figure 6.8: Data drawn from the model in Equation 6.81.

6.3 Maximum Likelihood Estimation

Topic Summary 46 Maximum Likelihood Estimation

Topic Objectives:

- Introduction to the notion of maximising the likelihood function.
- The maximum-likelihood estimate (MLE) techniques and the properties of the MLE.
- Example of applying the MLE technique.
- The invariance property of the MLE.

Topic Activities:

Type	Details	Duration	Progress
Watch video	15 : 26 min video	3 × length	
Read Handout	Read page 236 to page 241	8 mins/page	
Try Example	Try Example 6.8	15 mins	
Practice Exercises	Exercises ?? to ??	60 mins	

DC Level in white Gaussian noise

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

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

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

SOLUTION. Since this is a memoryless system, and $w[n]$ are i. i. d., then so is $x[n]$, and therefore:

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

http://media.ed.ac.uk/media/1_t7jwroia

Video Summary: This video introduces the MLE technique as a way of determining a good estimator for a given probabilistic problem. This method is very straightforward and intuitive, and the video motivates the approach by considering again how the likelihood function is formed. The properties of the MLE is discussed, and it is noted that many of the caveats and tricks used in optimisation theory simply apply to maximising the likelihood function. An example is shown for finding the MLE for estimating the mean of a Gaussian distributed set of data. This, of course, equals the MVUE since, as we know from a previous video, the MVUE exists. Finally, the video considers the MLE for a transformed parameter, and its application to, for example, calculating the SNR (although a detailed solution is saved for other exercises for the viewers).

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 $f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})$. 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 $f_{\mathbf{X}}(\hat{\mathbf{x}} | \boldsymbol{\theta})$. 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 $f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})$ as a function of \mathbf{x} ; now it is necessary to turn this thinking around, and view $f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$, for a given \mathbf{x} .

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

$$\hat{\boldsymbol{\theta}}_{ml}(\mathbf{x}) = \arg_{\boldsymbol{\theta}} \max f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta}) \quad (\text{T:6.40})$$

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

$$\begin{bmatrix} \frac{\partial f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial \theta_P} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \quad (6.104)$$

or, more simply,

$$\nabla_{\boldsymbol{\theta}} f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta}) \triangleq \frac{\partial f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}_{P \times 1} \quad (\text{K:7.35})$$

where $\mathbf{0}_{P \times 1}$ denotes the $P \times 1$ vector of zero elements. 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 $f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})$, 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.

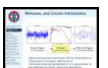
6.3.1 Properties of the MLE

1. The MLE satisfies

$$\nabla_{\boldsymbol{\theta}} f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ml}} = \mathbf{0}_{P \times 1} \quad (\text{T:6.41a})$$

$$\nabla_{\boldsymbol{\theta}} \ln f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ml}} = \mathbf{0}_{P \times 1} \quad (\text{T:6.41b})$$

where $\boldsymbol{\theta} \in \mathbb{R}^{P \times 1}$.



New slide

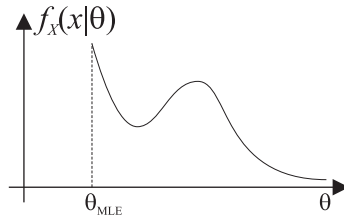


Figure 6.9: A single parameter MLE that occurs at a boundary, and therefore for which $\left. \frac{\partial f_{\mathbf{X}}(\mathbf{x}|\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}_{ml}} \neq 0$. Hence, in this case, a MLE and the MVUE are not necessarily equal.

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

Note that in the case of a scalar parameter, θ , then these expressions reduce:

$$\left. \frac{\partial f_{\mathbf{X}}(\mathbf{x}|\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}_{ml}} = 0 \quad (\text{T:6.10a})$$

$$\left. \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x}|\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}_{ml}} = 0 \quad (\text{T:6.10b})$$

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.

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:

$$\hat{\theta} - \theta = K(\theta) \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x}|\theta)}{\partial \theta} \quad (6.105)$$

The MLE satisfies

$$\left. \frac{\partial f_{\mathbf{X}}(\mathbf{x}|\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}_{ml}} = 0 \quad (6.106)$$

$$\left. \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x}|\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}_{ml}} = 0 \quad (6.107)$$

Hence, setting $\theta = \hat{\theta}_{ml}$ and substituting these into one another, gives:

$$\hat{\theta} - \hat{\theta}_{ml} = K(\hat{\theta}_{ml}) \left. \frac{\partial \ln f_{\mathbf{X}}(\mathbf{x}|\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}_{ml}} = 0 \quad (6.108)$$

Hence,

$$\hat{\theta} = \hat{\theta}_{ml} \quad (6.109)$$

□

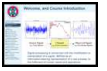
3. If the pdf, $f_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\theta})$, of the data \mathbf{x} satisfies certain *regularity* conditions, then the MLE of the unknown parameter $\boldsymbol{\theta}$ is asymptotically distributed (for large data records) according to a Gaussian distribution:

$$\hat{\boldsymbol{\theta}}_{ml} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{J}^{-1}(\boldsymbol{\theta})) \quad (6.110)$$

where $\mathbf{J}(\boldsymbol{\theta})$ is 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*.

6.3.2 DC Level in white Gaussian noise



An example of the maximum likelihood principle begins with the scalar case, and again deals with a DC level in WGN. New slide

Example 6.8 ([Therrien:1991, Example 6.1, Page 282]). A constant but unknown signal is observed in additive WGN. That is,

$$x[n] = A + w[n] \quad \text{where} \quad w[n] \sim \mathcal{N}(0, \sigma_w^2) \quad (6.111)$$

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

$$f_{\mathbf{X}}(x[n] | A) = f_W(w[n] | A) = f_W(x[n] - A) \quad (6.112)$$

Moreover, since this is a memoryless system, and $w[n]$ are i. i. d., then so is $x[n]$, and therefore:

$$f_{\mathbf{X}}(\mathbf{x} | A) = \prod_{n \in \mathcal{N}} f_W(x[n] - A) = \frac{1}{(2\pi\sigma_w^2)^{\frac{N}{2}}} \exp \left\{ -\frac{\sum_{n \in \mathcal{N}} (x[n] - A)^2}{2\sigma_w^2} \right\} \quad (6.113)$$

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:

$$\ln f_{\mathbf{X}}(\mathbf{x} | A) = -\frac{N}{2} \ln(2\pi\sigma_w^2) - \frac{\sum_{n \in \mathcal{N}} (x[n] - A)^2}{2\sigma_w^2} \quad (6.114)$$

Differentiating this expression w. r. t. A gives

$$\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x} | A)}{\partial A} = \frac{\sum_{n \in \mathcal{N}} (x[n] - A)}{\sigma_w^2} \quad (6.115)$$

and setting this to zero yields the MLE:

$$\hat{A}_{ml} = \frac{1}{N} \sum_{n \in \mathcal{N}} x[n] \quad (6.116)$$

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.

To complete the solution, note that it is worth checking that Equation 6.116 does, in fact, correspond to a maximum rather than a minimum or other stationary point. This can be verified by differentiating Equation 6.115 for a second time:

$$\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x} | A)}{\partial A^2} = \frac{\sum_{n \in \mathcal{N}} (-1)}{\sigma_w^2} = \frac{-N}{\sigma_w^2} < 0 \quad (6.117)$$

□

which is always negative and therefore corresponds to a minimum.

Example 6.9 ([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:

$$x[n] = A + w[n] \quad \text{where} \quad w[n] \sim \mathcal{N}(0, A\sigma_w^2) \quad (6.118)$$

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:

$$f_{\mathbf{x}}(\mathbf{x} | A) = \frac{1}{(2\pi A\sigma_w^2)^{\frac{N}{2}}} \exp \left\{ -\frac{\sum_{n \in \mathcal{N}} (x[n] - A)^2}{2A\sigma_w^2} \right\} \quad (6.119)$$

and thus the log-likelihood function is given by:

$$\ln f_{\mathbf{x}}(\mathbf{x} | A) = -\frac{N}{2} \ln(2\pi A\sigma_w^2) - \frac{\sum_{n \in \mathcal{N}} (x[n] - A)^2}{2A\sigma_w^2} \quad (6.120)$$

Differentiating the log-likelihood function w. r. t. A gives:

$$\frac{\partial \ln f_{\mathbf{x}}(\mathbf{x} | A)}{\partial A} = -\frac{N}{2A} + \frac{4A\sigma_w^2 \sum_{n \in \mathcal{N}} (x[n] - A) + 2\sigma_w^2 \sum_{n \in \mathcal{N}} (x[n] - A)^2}{4A^2\sigma_w^4} \quad (6.121)$$

$$= -\frac{N}{2A} + \frac{\sum_{n \in \mathcal{N}} (x[n] - A)}{A\sigma_w^2} + \frac{\sum_{n \in \mathcal{N}} (x[n] - A)^2}{2A^2\sigma_w^2} \quad (6.122)$$

and setting this equal to zero produces:

$$AN\sigma_w^2 = \sum_{n \in \mathcal{N}} \{(x[n] - A)^2 + 2A(x[n] - A)\} \quad (6.123)$$

$$A^2 + A\sigma_w^2 = \frac{1}{N} \sum_{n \in \mathcal{N}} x^2[n] \quad (6.124)$$

Solving for $\hat{A} > 0$ gives:

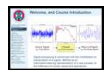
$$\hat{A} = -\frac{\sigma_w^2}{2} + \sqrt{\frac{\sigma_w^4}{4} + \frac{1}{N} \sum_{n \in \mathcal{N}} x^2[n]} \quad (6.125) \quad \square$$

Finally, that \hat{A} indeed maximises the log-likelihood function can be verified by examining the second derivative.

6.3.3 MLE for Transformed Parameter

Theorem 6.3 (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 $\alpha = \mathbf{g}(\theta)$, where \mathbf{g} is an r -dimensional function of the $P \times 1$ parameter θ , and the pdf, $f_{\mathbf{x}}(\mathbf{x} | \theta)$ is parameterised by θ , is given by

$$\hat{\alpha}_{ml} = \mathbf{g}(\hat{\theta}_{ml}) \quad (6.126)$$



New slide

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

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

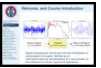
$$\bar{p}_T(\mathbf{x} | \alpha) = \max_{\theta: \alpha = g(\theta)} f_{\mathbf{X}}(\mathbf{x} | \theta) \quad (6.127)$$

◇

– End-of-Topic 46: **Introduction to MLE** –



6.4 Least Squares



Topic Summary 47 Least Squares Estimation

New slide

Topic Objectives:

- Understanding the principle of least squares estimation.
- Comparing least-squares principle with probabilistic approaches.
- Example of calculating the least-squares estimator.
- Understanding nonlinear least squares.

Topic Activities:

Type	Details	Duration	Progress
Watch video	18 : 38 min video	3 × length	
Read Handout	Read page 242 to page 245	8 mins/page	
Try Example	Try Example 6.10 and Example 6.11	10 mins	
Practice Exercises	Exercises ?? and ??	30 mins	

The Least Squares Approach

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

$$x[n] = \sum s[n] \theta + e_n[n]$$

$e[n]$ = modelling error

- In contrast to the MLE method, the least squares method considers $x[n]$ to be the sum of a known signal model, $s[n; \theta]$, plus an error term $e[n]$.
- This error term really consists of two components: the modelling error and an observation error.

http://media.ed.ac.uk/media/1_cza5m1gf

Video Summary: The least squares approach is presented as a non-probabilistic method for designing an estimator of a set of parameters, assuming a model is provided for describing the data. This is presented as an approach which makes *good sense* as opposed to being optimal. The least squares approach seeks to minimise the squared difference between the observed data and an assumed signal model. This is in contrast to the MLE which also assumes a statistical model on the excitation variable. Other norms, such as the L_1 norm is also mentioned as a comparison. The video considers a simple example to complement the techniques discussed in previous topics. Nonlinear least squares is also presented as a general approach, although needing more sophisticated optimisation techniques.

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. For some techniques, this means that the pdf of the data must be known somehow. 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**

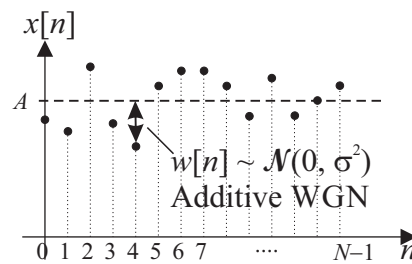


Figure 6.10: In the MLE method, the observed data $x[n, \zeta]$ is considered to be a random variable consisting of a known signal model, denoted by $s[n; \theta]$, where θ is a set of unknown model parameters, plus a noise term $w[n, \zeta]$ which has a given pdf.

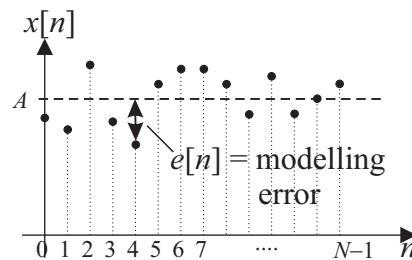


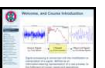
Figure 6.11: In contrast to Figure 6.10, the least squares method considers the observed data $x[n]$ to be the sum of the known signal model, $s[n; \theta]$, plus an error term, $e[n]$. The least squares method aims to minimise the total error term.

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.

6.4.1 The Least Squares Approach



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. New slide

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.

To clarify this further, consider the following difference between the MLE considered in Section 6.3, and the proposed approach. In the MLE method, the observed data $x[n] \equiv x[n, \zeta]$ is considered to be

a random variable consisting of a known signal model, denoted by $s[n; \boldsymbol{\theta}]$, where $\boldsymbol{\theta}$ is a set of unknown model parameters which define the functional form of the model, plus a noise term, $w[n, \zeta]$, which has a given pdf. In contrast to the MLE method, the least squares method considers $x[n]$ to be the sum of a known signal model, $s[n; \boldsymbol{\theta}]$, plus an error term $e[n]$. This error term really consists of two components: the modelling error, and an observation error.

The modelling error accounts for the fact that the proposed signal model may indeed just be wrong; for example, fitting a straight line to a set of data that is better described by a higher-order polynomial. The observation error or sensor error models the fact that any sensor will add noise to the measurement, and that the measurement therefore is itself not a true representation of the underlying signal model even if the signal model were perfectly accurate. In this chapter, these two errors are lumped together, but it should be noted that in general they should be considered as different concepts.

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 $s[n]$.

Now, one approach to finding the estimator is to minimise the sum of the absolute errors:

$$\hat{\boldsymbol{\theta}}_{L_1} = \arg_{\boldsymbol{\theta}} \min J_1(\boldsymbol{\theta}) \quad \text{where} \quad J_1(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} |x[n] - s[n, \boldsymbol{\theta}]| \quad (6.128)$$

However, in practice, while this is a good optimisation problem to try and solve, this is a difficult calculation to do in many cases.

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:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} (x[n] - s[n])^2 \quad (\text{K:8.1})$$

where $s[n] = s[n; \boldsymbol{\theta}]$ is a function of $\boldsymbol{\theta}$. The LSE is given by:

$$\hat{\boldsymbol{\theta}}_{LSE} = \arg_{\boldsymbol{\theta}} \min J(\boldsymbol{\theta}) \quad (6.129)$$

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.

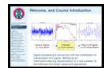
6.4.2 DC Level

Again, start by considering an example with a scalar parameter. The case with vector parameters follows a similar line.

Example 6.10 (Sample mean revisited: [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:

$$\hat{A}_{LSE} = \arg_A \min J(A) \quad \text{where} \quad J(A) = \sum_{n=0}^{N-1} (x[n] - A)^2 \quad (6.130)$$



New slide

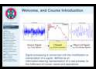
Differentiating w. r. t. A and setting the result to zero produces

$$\hat{A}_{LSE} = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \quad (6.131) \quad \square$$

which is the sample mean estimator. Differentiating for a second time shows this indeed minimises the squared error.

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.

6.4.3 Nonlinear Least Squares



Example 6.11 (Sinusoidal Frequency Estimation). Again, it is assumed that an observed signal, *New slide* $x[n]$, is a perturbed version of an unknown signal, $s[n]$, which is modelled as

$$s[n] = \cos 2\pi f_0 n \quad (6.132)$$

in which the frequency f_0 is to be estimated. The LSE can be found by minimising:

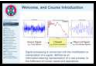
$$J(f_0) = \sum_{n=0}^{N-1} (x[n] - \cos 2\pi f_0 n)^2 \quad (6.133) \quad \times$$

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.

– End-of-Topic 47: **Introduction to Least Squares Estimation** –



6.4.4 Linear Least Squares



Topic Summary 48 Linear Least Squares Estimation

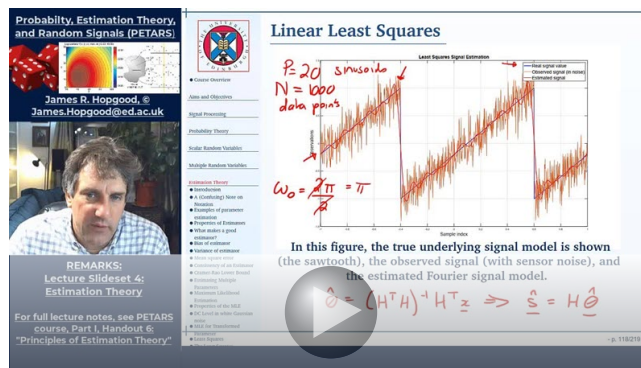
New slide

Topic Objectives:

- Awareness of linear in the parameters (LITP) signal model.
- Linear least square theory.
- Example of applying the linear least-squares estimator.

Topic Activities:

Type	Details	Duration	Progress
Watch video	20 : 15 min video	3× length	
Read Handout	Read page 246 to page 248	8 mins/page	
Try Code	Use the MATLAB code	20 minutes	
Try Example	Try Example 6.12	10 mins	
Practice Exercises	Exercise ??	30 mins	



http://media.ed.ac.uk/media/1_1wtlkjn8

Video Summary: The special case of linear least squares is presented as an extremely useful estimation approach, in cases when the signal model can be written as a linear combination of known basis functions, with unknown weighting parameters. The linear least squares problem can be written as a matrix vector formulation and solved to yield the so-called normal equations. The video considers an example of estimating the Fourier coefficients of a signal modelled as a linear combinations of trigonometric functions. Finally, a numerical example is shown. The linear algebra manipulations are shown throughout in order to help the viewer manipulate similar types of equations, although a full geometric interpretation is not considered here.

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:

$$\mathbf{s} = [s[0] \quad s[1] \quad \dots \quad s[N-1]]^T \quad \text{and} \quad \mathbf{x} = [x[0] \quad x[1] \quad \dots \quad x[N-1]]^T \quad (6.134)$$

In a linear signal model, it is assumed 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:

$$s[n] = \sum_{k=1}^P \theta_k h_k[n] \quad (6.135)$$

Writing this in matrix-vector notation, it follows that:

$$\underbrace{\begin{bmatrix} s[0] \\ s[1] \\ \vdots \\ s[N-1] \end{bmatrix}}_{\mathbf{s}} = \underbrace{\begin{bmatrix} h_1[0] & h_2[0] & \cdots & h_P[0] \\ h_1[1] & h_2[1] & \cdots & h_P[1] \\ \vdots & \vdots & \ddots & \vdots \\ h_1[N-1] & h_2[N-1] & \cdots & h_P[N-1] \end{bmatrix}}_{\mathbf{H}} \underbrace{\begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_P \end{bmatrix}}_{\boldsymbol{\theta}} \quad (6.136)$$

Thus, the unknown random-vector \mathbf{s} is linear in the unknown parameter vector $\boldsymbol{\theta} = [\theta_1, \dots, \theta_P]$, and can be written as:

$$\mathbf{s} = \mathbf{H}\boldsymbol{\theta} \quad (\text{K:8.8})$$

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:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} |x[n] - s[n]|^2 = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \quad (\text{K:8.9})$$

This can be written as:

$$J(\boldsymbol{\theta}) = \mathbf{x}^T \mathbf{x} - 2\mathbf{x}^T \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\theta}^T \mathbf{H}^T \mathbf{H}\boldsymbol{\theta} \quad (6.137)$$

and using the two identities that:

$$\frac{\partial \mathbf{b}^T \mathbf{a}}{\partial \mathbf{a}} = \mathbf{b} \quad \text{and} \quad \frac{\partial \mathbf{a}^T \mathbf{B} \mathbf{a}}{\partial \mathbf{a}} = (\mathbf{B} + \mathbf{B}^T) \mathbf{a} \quad (6.138)$$

then observing in this case $\mathbf{B} = \mathbf{H}^T \mathbf{H} = \mathbf{B}^T$ it follows that

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2\mathbf{H}^T \mathbf{x} + 2\mathbf{H}^T \mathbf{H}\boldsymbol{\theta} \quad (6.139)$$

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

$$\hat{\boldsymbol{\theta}}_{LSE} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} \quad (\text{K:8.10})$$

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**.

The minimum LS error is found from Equation K:8.9 and Equation K:8.10:

$$J_{\min} = J(\hat{\boldsymbol{\theta}}) = (\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}})^T (\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}) \quad (6.140)$$

$$= \left(\mathbf{x} - \mathbf{H} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} \right)^T \left(\mathbf{x} - \mathbf{H} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} \right) \quad (6.141)$$

or alternatively

$$J_{\min} = \mathbf{x}^T \left(\mathbf{I}_N - \mathbf{H} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \right) \left(\mathbf{I}_N - \mathbf{H} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \right) \mathbf{x} \quad (6.142)$$

Now, the matrix $\mathbf{A} = \mathbf{I}_N - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ is an **idempotent matrix** in that it has the property $\mathbf{A}^2 = \mathbf{A}$. This follows from noting that:

$$\mathbf{A}^2 = \mathbf{I}_N - 2\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T + \underbrace{\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T}_{=\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T} = \mathbf{A} \quad (6.143)$$

Hence,

$$J_{\min} = \mathbf{x}^T \left(\mathbf{I}_N - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T \right) \mathbf{x} \quad (\text{K:8.11})$$

Other forms for J_{\min} are:

$$J_{\min} = \mathbf{x}^T \mathbf{x} - \mathbf{x}^T \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T \mathbf{x} \quad (\text{K:8.12})$$

$$= \mathbf{x}^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \quad (\text{K:8.13})$$

Example 6.12 (Fourier Series Estimation). An application of the general linear model is in spectral estimation. Suppose that a signal, $s[n]$, is modelled as the sum of sinusoids:

$$s[n] = \sum_{p=1}^P a_p \sin(p\omega_0 n) + b_p \cos(p\omega_0 n) \quad (6.144)$$

where the coefficients $\{a_p, b_p\}_{p=1}^P$ are the unknown amplitudes to be estimated, and the fundamental frequency, ω_0 , and model order P , are assumed to be known. It is implicitly assumed that the sampling period $T = 1$ and that the fundamental ω_0 is normalised to between 0 and π .

The signal, $s[n]$, is observed in noise. Write down the least squares solution.

SOLUTION. Writing the relationship between the observation, signal model, and modelling error:

$$x[n] = s[n] + e[n] = \sum_{p=1}^P (a_p \sin \omega_p n + b_p \cos \omega_p n) + e[n] \quad (6.145)$$

This model can be written in a so-called LITP form by defining the matrix, where $\ell \triangleq N - 1$:

$$\mathbf{H} = \begin{bmatrix} 0 & 1 & 0 & 1 & \cdots & 0 & 1 \\ \sin \omega_0 & \cos \omega_0 & \sin 2\omega_0 & \cos 2\omega_0 & \cdots & \sin P\omega_0 & \cos P\omega_0 \\ \sin 2\omega_0 & \cos 2\omega_0 & \sin 4\omega_0 & \cos 4\omega_0 & \cdots & \sin 2P\omega_0 & \cos 2P\omega_0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \sin \ell\omega_0 & \cos \ell\omega_0 & \sin 2\ell\omega_0 & \cos 2\ell\omega_0 & \cdots & \sin P\ell\omega_0 & \cos P\ell\omega_0 \end{bmatrix} \quad (6.146)$$

Hence, with the parameter vector defined as:

$$\boldsymbol{\theta} = [a_1 \ b_1 \ a_2 \ b_2 \ \cdots \ a_P \ b_P]^T \quad (6.147)$$

the signal model is $\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$, and the linear LSE estimator is then given by:

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} \quad (6.148)$$

□

where the parameter vector, $\hat{\boldsymbol{\theta}}$, is of dimension $2P$, and therefore the size of \mathbf{H} is $N \times 2P$.

Using the orthogonality of the Fourier basis, it is possible to show that this relationship can simplify further, and this is left as an exercise.



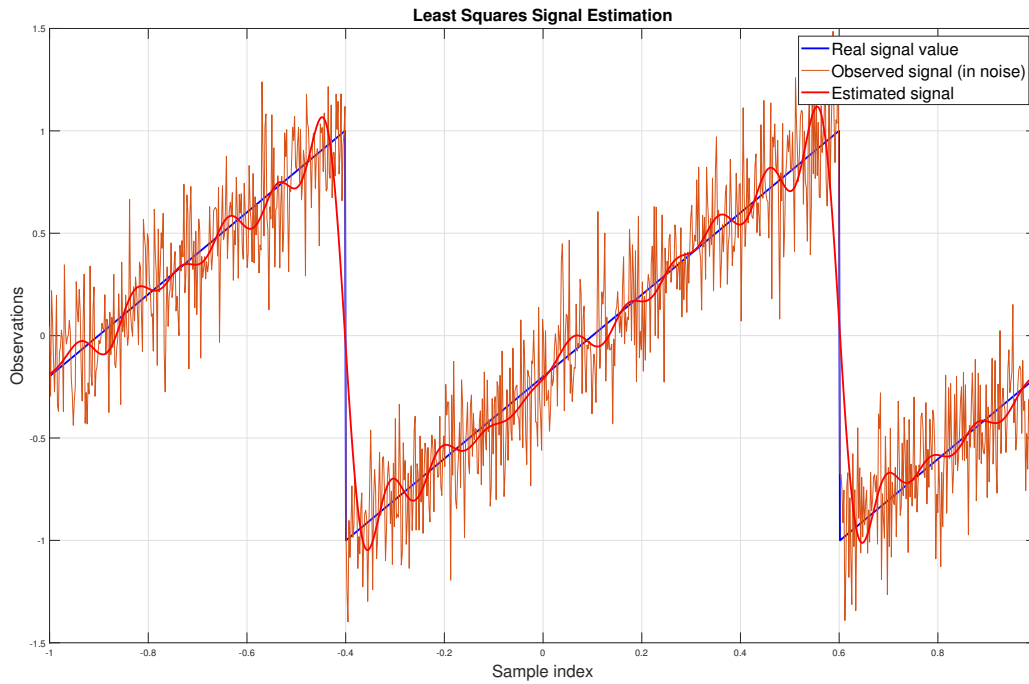


Figure 6.12: In this figure, the true underlying signal model is shown (the sawtooth), the observed signal (with sensor noise), and the estimated Fourier signal model.

6.4.5 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

$$J(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{W} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \quad (\text{K:8.14})$$

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:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} w_n (x[n] - s[n])^2 \quad (6.149)$$

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 6.10 on page 244, 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 6.13 ([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:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} \frac{1}{\sigma_n^2} (x[n] - A)^2 \quad (6.150)$$

Differentiating w. r. t. A , and setting to zero gives:

$$0 = \sum_{n=0}^{N-1} \frac{1}{\sigma_n^2} (x[n] - A) \quad (6.151)$$

Rearranging gives straightforwardly:

$$\hat{A}_{LSE} = \frac{\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2} x[n]}{\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}} \quad (6.152) \quad \square$$

The general form of the weighted LSE is readily shown to be:

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{W} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{W} \mathbf{x} \quad (\text{K:8.16})$$

and its minimum LS error is

$$J_{\min} = \mathbf{x}^T \left(\mathbf{W} - \mathbf{W} \mathbf{H} (\mathbf{H}^T \mathbf{W} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{W} \right) \mathbf{x} \quad (\text{K:8.17})$$

6.5 Bayesian Parameter Estimation

Topic Summary 49 Introduction to Advanced Bayesian Estimation Theory

Topic Objectives:

- Introduction to Bayesian Parameter Estimation.
- The Removal of Nuisance Parameters and Prior Probabilities.
- The General Linear Model.

Topic Activities:

Type	Details	Duration	Progress
Read Handout	Read page 250 to page 257	8 mins/page	
Try Example	Try Example 6.14	15 mins	

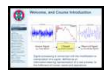
Using the method of maximum likelihood (or least squares) to infer the values of a parameter has several significant limitations:

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.

6.5.1 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**:

$$\mathcal{L}_k(\boldsymbol{\theta}_k; \mathbf{x}) = p_{\mathbf{X}|\boldsymbol{\theta}_k}(\mathbf{x} | \boldsymbol{\theta}_k, \mathcal{I}_k) \quad (6.153)$$

where $\boldsymbol{\theta}_k$ is an unknown parameter vector which characterises the k -th signal model, \mathcal{I}_k . Suppose knowledge *prior* to observing the data regarding the probability of the values of the parameters of \mathcal{I}_k is summarised by the probability density function, $p_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k | \mathcal{I}_k)$. Then **Bayes's theorem** gives:

$$p_{\boldsymbol{\theta}_k|\mathbf{x}}(\boldsymbol{\theta}_k | \mathbf{x}, \mathcal{I}_k) = \frac{p_{\mathbf{X}|\boldsymbol{\theta}_k}(\mathbf{x} | \boldsymbol{\theta}_k, \mathcal{I}_k) p_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k | \mathcal{I}_k)}{p_{\mathbf{X}}(\mathbf{x} | \mathcal{I}_k)} \quad (6.154)$$

Equation 6.154 is composed of the following terms:

- Prior:** $p_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k | \mathcal{I}_k)$ summarises all the knowledge of the values of the parameters $\boldsymbol{\theta}_k$ *prior* to observing the data;
- Likelihood:** $p_{\mathbf{X}|\boldsymbol{\theta}_k}(\mathbf{x} | \boldsymbol{\theta}_k, \mathcal{I}_k)$, is determined by the signal model \mathcal{I}_k ;
- Evidence:** $p_{\mathbf{X}}(\mathbf{x} | \mathcal{I}_k)$, which is the normalising expression in Equation 6.154, is known as the **Bayesian evidence**. Since the left hand side (LHS) must integrate to unity to be a valid pdf, then it follows:

$$p_{\mathbf{X}}(\mathbf{x} | \mathcal{I}_k) = \int_{\boldsymbol{\theta}_k} p_{\mathbf{X}|\boldsymbol{\theta}_k}(\mathbf{x} | \boldsymbol{\theta}_k, \mathcal{I}_k) p_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k | \mathcal{I}_k) d\boldsymbol{\theta}_k \quad (6.155)$$

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:** $p_{\boldsymbol{\theta}_k|\mathbf{x}}(\boldsymbol{\theta}_k | \mathbf{x}, \mathcal{I}_k)$ is the joint **posterior pdf** for the unknown parameters $\boldsymbol{\theta}_k$ given the observations \mathbf{x} .

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:

$$\hat{\boldsymbol{\theta}}_k = \arg_{\boldsymbol{\theta}_k} \max p_{\boldsymbol{\theta}_k | \mathbf{x}}(\boldsymbol{\theta}_k | \mathbf{x}, \mathcal{I}_k) \quad (6.156)$$

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:

$$p(\boldsymbol{\theta}_k | \mathbf{x}, \mathcal{I}_k) = \frac{p(\mathbf{x} | \boldsymbol{\theta}_k, \mathcal{I}_k) p(\boldsymbol{\theta}_k | \mathcal{I}_k)}{p(\mathbf{x} | \mathcal{I}_k)} \quad (6.157)$$

It is understood in Equation 6.157 that the probability density functions, $p(\cdot | \cdot)$, are identified based on its context. In other-words, it is important to realise that each term in Equation 6.157 represents a different functional form for the pdfs.

In cases where there is only one model in consideration, Equation 6.157 simplifies further to:

$$p(\boldsymbol{\theta} | \mathbf{x}, \mathcal{I}) = \frac{p(\mathbf{x} | \boldsymbol{\theta}, \mathcal{I}) p(\boldsymbol{\theta} | \mathcal{I})}{p(\mathbf{x} | \mathcal{I})} \quad (6.158)$$

6.5.2 The Removal of Nuisance Parameters

Consider a signal model, \mathcal{I} , that involves two parameters, α and β :

$$p(\alpha, \beta | \mathbf{x}, \mathcal{I}) = \frac{p(\mathbf{x} | \alpha, \beta, \mathcal{I}) p(\alpha, \beta | \mathcal{I})}{p(\mathbf{x} | \mathcal{I})} \quad (6.159)$$

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 :

$$\begin{aligned} p(\alpha | \mathbf{x}, \mathcal{I}) &= \int p(\alpha, \beta | \mathbf{x}, \mathcal{I}) d\beta \\ &= \frac{1}{p(\mathbf{x} | \mathcal{I})} \int p(\mathbf{x} | \alpha, \beta, \mathcal{I}) p(\alpha, \beta | \mathcal{I}) d\beta \end{aligned} \quad (6.160)$$

Marginalisation, also known as **marginal inference**, is an appealing procedure when the integral in Equation 6.160 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 $p(\alpha | \mathbf{x}, \mathcal{I})$. Note that marginalisation necessitates a loss of information; the integration in Equation 6.160 means that all the information about the value of β is lost.

If the marginal is used for parameter estimation, then the value of α that maximises the marginal:

$$\hat{\alpha} = \arg_{\alpha} \max p(\alpha | \mathbf{x}, \mathcal{I}) = \arg_{\alpha} \max \int p(\alpha, \beta | \mathbf{x}, \mathcal{I}) d\beta \quad (6.161)$$

is known as the maximum marginal *a posteriori* (MMAP) estimate.

6.5.3 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:

$$p(\boldsymbol{\theta}_k | \mathcal{I}_k) = k \quad (6.162)$$

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 6.162 is that it 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:

$$p(\boldsymbol{\theta}_k | \mathcal{I}_k) = \frac{1}{(2\pi\delta^2)^{\frac{P}{2}}} \exp\left[-\frac{\boldsymbol{\theta}_k^T \boldsymbol{\theta}_k}{2\delta^2}\right] \quad (6.163)$$

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.

6.5.4 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:

$$x[n] = \sum_{p=1}^P a_p g_p[n] + e[n], \quad \text{where} \quad 0 \leq n \leq N-1 \quad (6.164)$$

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 6.164 for all values of n :

$$\underbrace{\begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} g_1[0] & g_2[0] & \cdots & g_P[0] \\ g_1[1] & g_2[1] & \cdots & g_P[1] \\ \vdots & \vdots & \ddots & \vdots \\ g_1[N-1] & g_2[N-1] & \cdots & g_P[N-1] \end{bmatrix}}_{\mathbf{G}} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_P \end{bmatrix}}_{\mathbf{a}} + \underbrace{\begin{bmatrix} e[0] \\ e[1] \\ \vdots \\ e[N-1] \end{bmatrix}}_{\mathbf{e}} \quad (6.165)$$

In other-words, Equation 6.164 may be written as:

$$\mathbf{x} = \mathbf{G} \mathbf{a} + \mathbf{e} \quad (6.166)$$

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 6.166 is linear in the parameter vector \mathbf{a} ; hence, the model in Equation 6.166 is often called the **LITP** model. Now, consider finding the likelihood function $p(\mathbf{x} | \boldsymbol{\theta}, \mathbf{a}, \sigma_e^2, \mathcal{I})$, 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:

$$p(\mathbf{e} | \sigma_e^2) = \frac{1}{(2\pi\sigma_e^2)^{\frac{N}{2}}} \exp\left[-\frac{\mathbf{e}^T \mathbf{e}}{2\sigma_e^2}\right] \quad (6.167)$$

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:

$$p(\mathbf{x} | \boldsymbol{\theta}, \mathbf{a}, \sigma_e^2, \mathcal{I}) = \frac{1}{(2\pi\sigma_e^2)^{\frac{N}{2}}} \exp\left[-\frac{(\mathbf{x} - \mathbf{G}\mathbf{a})^T (\mathbf{x} - \mathbf{G}\mathbf{a})}{2\sigma_e^2}\right] \quad (6.168)$$

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:

$$p(\sigma_e^2 | \alpha_e, \beta_e) = \mathcal{IG}(\sigma_e^2 | \alpha_e, \beta_e) = \begin{cases} 0 & \text{if } \sigma_e^2 < 0, \\ \frac{\alpha_e^{\beta_e}}{\Gamma(\beta_e)} (\sigma_e^2)^{-(\beta_e+1)} e^{-\frac{\alpha_e}{\sigma_e^2}} & \text{if } \sigma_e^2 \geq 0, \end{cases} \quad (6.169)$$

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 6.163:

$$p(\mathbf{a} | \sigma_e^2, \mathcal{I}) = \mathcal{N}(\mathbf{a} | 0, \delta^2 \sigma_e^2 \mathbf{I}_P) = \frac{1}{(2\pi\delta^2 \sigma_e^2)^{\frac{P}{2}}} \exp\left[-\frac{\mathbf{a}^T \mathbf{a}}{2\delta^2 \sigma_e^2}\right] \quad (6.170)$$

where \mathbf{I}_P is the $P \times P$ identity matrix. Note that the prior $p(\mathbf{a} | \sigma_e^2, \delta, \mathcal{I})$ 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:

$$p(\boldsymbol{\theta}, \mathbf{a}, \sigma_e^2 | \mathbf{x}, \mathcal{I}) \propto p(\mathbf{x} | \boldsymbol{\theta}, \mathbf{a}, \sigma_e^2, \mathcal{I}) p(\boldsymbol{\theta}, \mathbf{a}, \sigma_e^2 | \mathcal{I}) \quad (6.171)$$

where the evidence term is considered as a constant and therefore omitted, and \propto indicates proportionality. The prior term factorises as:

$$p(\boldsymbol{\theta}, \mathbf{a}, \sigma_e^2) = p(\boldsymbol{\theta}) p(\mathbf{a} | \sigma_e^2) p(\sigma_e^2) \quad (6.172)$$

where the dependence on the model \mathcal{I} has been dropped for convenience. Thus, the joint posterior density is given by:

$$p(\boldsymbol{\theta}, \mathbf{a}, \sigma_e^2 | \mathbf{x}, \mathcal{I}) \propto p(\boldsymbol{\theta}) \frac{1}{(2\pi\sigma_e^2)^{\frac{N}{2}}} \exp\left[-\frac{(\mathbf{x} - \mathbf{G}\mathbf{a})^T (\mathbf{x} - \mathbf{G}\mathbf{a})}{2\sigma_e^2}\right] \times \frac{1}{(2\pi\delta^2\sigma_e^2)^{\frac{P}{2}}} \exp\left[-\frac{\mathbf{a}^T \mathbf{a}}{2\delta^2\sigma_e^2}\right] \frac{\alpha_e^{\beta_e}}{\Gamma(\beta_e)} (\sigma_e^2)^{-(\beta_e+1)} e^{-\frac{\alpha_e}{\sigma_e^2}} \quad (6.173)$$

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

$$p(\boldsymbol{\theta}, \mathbf{a}, \sigma_e^2 | \mathbf{x}, \mathcal{I}) \propto \frac{p(\boldsymbol{\theta})}{(\sigma_e^2)^{\frac{N+P}{2} + \beta_e + 1}} \exp\left[-\frac{\mathbf{a}^T (\mathbf{G}^T \mathbf{G} + \delta^{-2} \mathbf{I}_P) \mathbf{a} - 2\mathbf{x}^T \mathbf{G} \mathbf{a} + \mathbf{x}^T \mathbf{x} + 2\alpha_e}{2\sigma_e^2}\right] \quad (6.174)$$

The linear parameters \mathbf{a} can be marginalised out using the identity:

$$\int_{\mathbb{R}^P} \exp\left\{-\frac{1}{2} [\alpha + 2\mathbf{y}^T \boldsymbol{\beta} + \mathbf{y}^T \boldsymbol{\Gamma} \mathbf{y}]\right\} d\mathbf{y} = \frac{(2\pi)^{\frac{P}{2}}}{|\boldsymbol{\Gamma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2} [\alpha - \boldsymbol{\beta}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{\beta}]\right\} \quad (6.175)$$

To perform this, set $\mathbf{y} = \mathbf{a}$, $\boldsymbol{\Gamma} = \frac{1}{\sigma_e^2} (\mathbf{G}^T \mathbf{G} + \delta^{-2} \mathbf{I}_P)$, $\alpha = \frac{\mathbf{x}^T \mathbf{x} + 2\alpha_e}{\sigma_e^2}$, and $\boldsymbol{\beta} = -\frac{1}{\sigma_e^2} \mathbf{G}^T \mathbf{x}$, so that

$$p(\boldsymbol{\theta}, \sigma_e^2 | \mathbf{x}, \mathcal{I}) = \int_{-\infty}^{\infty} p(\boldsymbol{\theta}, \mathbf{a}, \sigma_e^2 | \mathbf{x}, \mathcal{I}) d\mathbf{a} \quad (6.176)$$

$$\propto \frac{p(\boldsymbol{\theta})}{\sqrt{\det |\mathbf{G}^T \mathbf{G} + \delta^{-2} \mathbf{I}_P|} (\sigma_e^2)^{R+1}} \exp\left[-\frac{\mathbf{x}^T \mathbf{x} + 2\alpha_e - \mathbf{x}^T \mathbf{G} (\mathbf{G}^T \mathbf{G} + \delta^{-2} \mathbf{I}_P)^{-1} \mathbf{G}^T \mathbf{x}}{2\sigma_e^2}\right] \quad (6.177)$$

where $R = \frac{N+2\beta_e}{2}$. Finally, the variance can be marginalised using the fact that the inverse-Gamma pdf implies:

$$1 = \int_0^{\infty} \mathcal{IG}(\sigma^2 | \alpha, \beta,) d\sigma^2 = \int_0^{\infty} \frac{\alpha^\beta}{\Gamma(\beta)} (\sigma^2)^{-(\beta+1)} e^{-\frac{\alpha}{\sigma^2}} d\sigma^2 \quad (6.178)$$

and therefore:

$$\int_0^{\infty} (\sigma^2)^{-(\beta+1)} e^{-\frac{\alpha}{\sigma^2}} d\sigma^2 = \frac{\Gamma(\beta)}{\alpha^\beta} \quad (6.179)$$

Hence, this gives the **marginal a posterior pdf** for the parameters $\boldsymbol{\theta}$ as

$$p(\boldsymbol{\theta} | \mathbf{x}, \mathcal{I}) = \int_0^{\infty} p(\boldsymbol{\theta}, \sigma_e^2 | \mathbf{x}, \mathcal{I}) d\sigma_e^2 \quad (6.180)$$

$$\propto p(\boldsymbol{\theta}) \frac{\left[\mathbf{x}^T \mathbf{x} + 2\alpha_e - \mathbf{x}^T \mathbf{G} (\mathbf{G}^T \mathbf{G} + \delta^{-2} \mathbf{I}_P)^{-1} \mathbf{G}^T \mathbf{x}\right]^{-\left(\frac{N}{2} + \beta_e\right)}}{\sqrt{\det |\mathbf{G}^T \mathbf{G} + \delta^{-2} \mathbf{I}_P|}}$$

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 6.180 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 6.14 (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:

$$s[n] = \sum_{p=1}^P (a_p \sin \omega_p n + b_p \cos \omega_p n) \quad (6.181)$$

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 :

$$x[n] = s[n] + e[n] = \sum_{p=1}^P (a_p \sin \omega_p n + b_p \cos \omega_p n) + e[n] \quad (6.182)$$

This model can be written in the LITP form by defining the matrix:

$$\mathbf{G} = \begin{bmatrix} 0 & 1 & 0 & 1 & \cdots & 0 & 1 \\ \sin \omega_1 & \cos \omega_1 & \sin \omega_2 & \cos \omega_2 & \cdots & \sin \omega_P & \cos \omega_P \\ \sin 2\omega_1 & \cos 2\omega_1 & \sin 2\omega_2 & \cos 2\omega_2 & \cdots & \sin 2\omega_P & \cos 2\omega_P \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \sin \ell\omega_1 & \cos \ell\omega_1 & \sin \ell\omega_2 & \cos \ell\omega_2 & \cdots & \sin \ell\omega_P & \cos \ell\omega_P \end{bmatrix} \quad (6.183)$$

where $\ell = N - 1$. Hence, with the parameter vector defined as:

$$\mathbf{a} = [a_1 \ b_1 \ a_2 \ b_2 \ \cdots \ a_P \ b_P]^T \quad (6.184)$$

the **marginal a posterior pdf** for the unknown frequencies $\{\omega_p\}_1^P$ is given by:

$$p(\{\omega_p\}_1^P | \mathbf{x}) \propto p(\{\omega_p\}_1^P) \frac{[\mathbf{x}^T \mathbf{x} + 2\alpha_e - \mathbf{x}^T \mathbf{G} (\mathbf{G}^T \mathbf{G} + \delta^{-2} \mathbf{I}_{2P})^{-1} \mathbf{G}^T \mathbf{x}]^{-\left(\frac{N}{2} + \beta_e\right)}}{\sqrt{\det |\mathbf{G}^T \mathbf{G} + \delta^{-2} \mathbf{I}_{2P}|}} \quad (6.185)$$

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.

6.5.4.1 Model Selection using Bayesian Evidence

Next, the Bayesian evidence term is considered:

$$p_{\mathbf{x}}(\mathbf{x} | \mathcal{I}_k) = \int_{\Theta_k} p_{\mathbf{x}|\Theta_k}(\mathbf{x} | \boldsymbol{\theta}_k, \mathcal{I}_k) p_{\Theta_k}(\boldsymbol{\theta}_k | \mathcal{I}_k) d\boldsymbol{\theta}_k \quad (6.186)$$

This term can be used to select signal models and noise statistics appropriate to the observed data. To clarify, in this equation, Θ_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 6.186 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:

$$p_{\mathcal{I}|\mathbf{x}}(\mathcal{I}_k | \mathbf{x}) = \frac{p_{\mathbf{x}|\mathcal{I}}(\mathbf{x} | \mathcal{I}_k) p_{\mathcal{I}}(\mathcal{I}_k)}{p_{\mathbf{x}}(\mathbf{x})} \quad (6.187)$$

where the probability of the observations is given by:

$$p_{\mathbf{x}}(\mathbf{x}) = \sum_{k=1}^M p_{\mathbf{x}|\mathcal{I}}(\mathbf{x} | \mathcal{I}_k) p_{\mathcal{I}}(\mathcal{I}_k) \quad (6.188)$$

If all the models are equally likely *a priori*, then

$$p_{\mathcal{I}}(\mathcal{I}_k) = \frac{1}{M} \quad (6.189)$$

Therefore, the posterior probability of a model is given by the **relative evidence**:

$$p_{\mathcal{I}|\mathbf{x}}(\mathcal{I}_k | \mathbf{x}) = \frac{p_{\mathbf{x}|\mathcal{I}}(\mathbf{x} | \mathcal{I}_k)}{\sum_{k=1}^M p_{\mathbf{x}|\mathcal{I}}(\mathbf{x} | \mathcal{I}_k)} \quad (6.190)$$

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.

– End-of-Topic 49: **Introduction to Advanced Bayesian Parameter Estimation** –



7

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.

7.1 Introduction

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

Optimisation: involves finding the solution to

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \Theta} h(\boldsymbol{\theta}) \quad (7.1)$$

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:

$$h(x) = (\cos 50x + \sin 20x)^2, \quad 0 \leq x \leq 1 \quad (7.2)$$

This function is plotted in Figure 7.1.

Integration: involves evaluating an integral,

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (7.3)$$

that cannot explicitly be calculated in *closed form*. For example, the Gaussian-error function:

$$\Phi(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}} d\theta \quad (7.4)$$

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

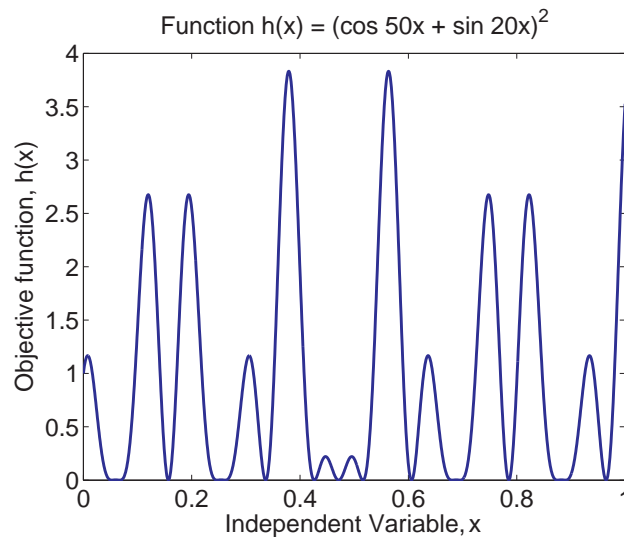


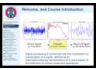
Figure 7.1: Plot of the function in Equation 7.2.

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

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \int_{\Omega} f(\theta, \omega) d\omega \quad (7.5)$$

The reader is encouraged to honestly consider how many problems they solve reduce to either an integration or an optimisation problem.

7.1.1 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: *New slide*

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-uniformly 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 7.1.1.1 and Section 7.1.1.2.

7.1.1.1 Deterministic Optimisation

The **Nelder-Mead Downhill Simplex method** simply crawls downhill in a straightforward fashion that makes almost no special assumptions about your function. This can be extremely slow, but 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:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - (\nabla \nabla^T h(\boldsymbol{\theta}_n))^{-1} \nabla h(\boldsymbol{\theta}_n) \quad (7.6)$$

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 7.6 is the computation of the Hessian term $\nabla \nabla^T h(\boldsymbol{\theta}_n)$. 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.

7.1.1.2 Deterministic Integration

Numerical computation of the scalar case of the integral in Equation 7.7 can be done using simple **Riemann integration**, or by improved methods such as the **trapezoidal rule**. For example, the

$$\mathcal{I} = \int_a^b f(\theta) d\theta, \quad (7.7)$$

can be solved with the trapezoidal rule using:

$$\hat{I} = \frac{1}{2} \sum_{k=0}^{N-1} (\theta_{k+1} - \theta_k) (f(\theta_k) + f(\theta_{k+1})) \quad (7.8)$$

where the θ_k 's constitute an ordered partition of $[a, b]$. Another formula is **Simpson's rule**:

$$\hat{I} = \frac{\delta}{3} \left\{ f(a) + 4 \sum_{k=1}^N f(\theta_{2k-1}) + 2 \sum_{k=1}^N f(\theta_{2k}) + f(b) \right\} \quad (7.9)$$

in the case of equally spaced samples with $\delta = \theta_{k+1} - \theta_k$.

7.1.2 Monte Carlo Numerical Methods

Monte Carlo methods are stochastic techniques, in which random numbers are generated and use to examine some problem.

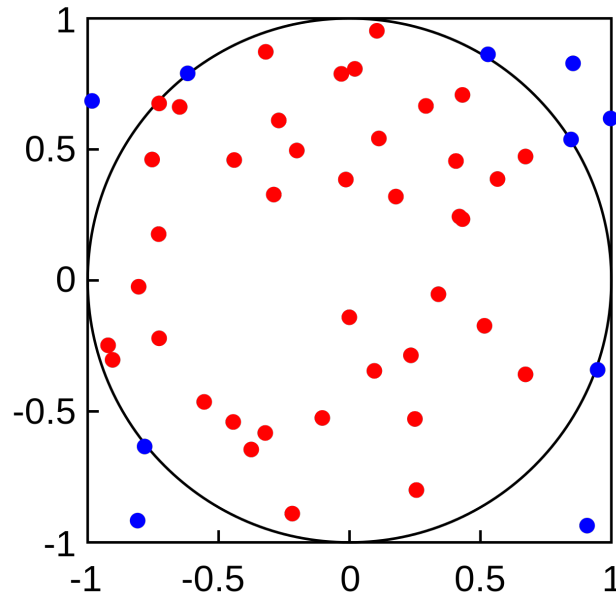


Figure 7.2: Estimating the value of π through Monte Carlo integration.

7.1.2.1 Monte Carlo Integration

Consider the integral,

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (7.10)$$

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:

$$\mathcal{I} = \int_{\Theta} \frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (7.11)$$

where the function $\pi(\boldsymbol{\theta}) > 0$, $\boldsymbol{\theta} \in \Theta$ is a probability density function (pdf) which satisfies the normalised expression:

$$\int_{\Theta} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1 \quad (7.12)$$

It can now be seen that Equation 7.57 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 7.57 becomes

This may be written as an expectation:

$$\mathcal{I} = \mathbb{E}_{\pi} \left[\frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \right] \quad (7.13)$$

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 $\pi(\boldsymbol{\theta})$,

$$\boldsymbol{\theta}^{(k)} \sim \pi(\boldsymbol{\theta}), \quad k \in \mathcal{N} = \{0, \dots, N-1\} \quad (7.14)$$

2. Calculate the sample average of the expectation in Equation 7.13 using

$$\hat{\mathcal{I}} = \frac{1}{N} \sum_{k=0}^{N-1} \frac{f(\boldsymbol{\theta}^{(k)})}{\pi(\boldsymbol{\theta}^{(k)})} \approx \mathbb{E}_{\pi} \left[\frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \right] \quad (7.15)$$

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.

7.1.2.2 Stochastic Optimisation

There are two distinct approaches to the Monte Carlo optimisation (here, maximisation) of the objective function $h(\boldsymbol{\theta})$:

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \Theta} h(\boldsymbol{\theta}) \quad (7.16)$$

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 7.16 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:

$$\hat{\boldsymbol{\theta}} \approx \arg \max_{\{\boldsymbol{\theta}^{(k)}\}} h(\boldsymbol{\theta}^{(k)}) \quad (7.17)$$

Typically, when no specific features regarding the function $h(\boldsymbol{\theta})$, are taken into account, $\pi(\boldsymbol{\theta})$ will take on a uniform distribution over Θ . Although this method converges as $N \rightarrow \infty$, the method is very slow: one can usually do better by finding a density $\pi(\boldsymbol{\theta})$ that is related to $h(\boldsymbol{\theta})$, but this requires some additional insight into the function $h(\boldsymbol{\theta})$.

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 7.6 is:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + \mathbf{G}_n \nabla h(\boldsymbol{\theta}_n) \quad (7.18)$$

where \mathbf{G}_n is a sequence which may approximate the Hessian of $h(\boldsymbol{\theta}_n)$ in order to ensure the algorithm converges.

7.1.2.3 Implementation issues

Monte Carlo methods rely on the assumption that it is possible to simulate **samples** or **variates** $\{\boldsymbol{\theta}^{(k)}\}$ from the density $\pi(\boldsymbol{\theta})$.

The next sections address how such samples can be obtained.

7.2 Generating Random Variables

This section discusses a variety of techniques for generating random variables from a different distributions.

7.2.1 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:

$$x_{n+1} = (a x_n + b) \pmod{m} \quad (7.19)$$

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 7.19 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.

7.2.2 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 7.1 (Probability transformation rule). Denote the real roots of $y = g(x)$ by $\{x_n, n \in \mathcal{N}\}$, such that:

$$y = g(x_1) = \dots = g(x_N) \quad (7.20)$$

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

$$f_Y(y) = \sum_{n=1}^N \frac{f_X(x_n)}{|g'(x_n)|} \quad (7.21)$$

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.

7.2.3 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 7.2 (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:

$$f_Y(y_1, \dots, y_n) = \frac{1}{|J(x_1, \dots, x_n)|} f_X(x_1, \dots, x_n) \quad (7.22)$$

where $J(x_1, \dots, x_n)$ is the **Jacobian** of the transformation given by:

$$J(x_1, \dots, x_n) = \frac{\partial(y_1, \dots, y_n)}{\partial(x_1, \dots, x_n)} \quad (7.23)$$

◇

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,

$$f_{X_k}(x_k) = \mathbb{I}_{0,1}(x_k), \quad k = 1, 2 \quad (7.24)$$

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:

$$y_1 = \sqrt{-2 \ln x_1} \cos 2\pi x_2 \quad (7.25)$$

$$y_2 = \sqrt{-2 \ln x_1} \sin 2\pi x_2 \quad (7.26)$$

It follows, by rearranging these equations, that:

$$x_1 = \exp \left[-\frac{1}{2}(y_1^2 + y_2^2) \right] \quad (7.27)$$

$$x_2 = \frac{1}{2\pi} \arctan \frac{y_2}{y_1} \quad (7.28)$$

The Jacobian determinant can be calculated as:

$$J(x_1, x_2) = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{vmatrix} = \begin{vmatrix} \frac{-1}{x_1 \sqrt{-2 \ln x_1}} \cos 2\pi x_2 & -2\pi \sqrt{-2 \ln x_1} \sin 2\pi x_2 \\ \frac{-1}{x_1 \sqrt{-2 \ln x_1}} \sin 2\pi x_2 & 2\pi \sqrt{-2 \ln x_1} \cos 2\pi x_2 \end{vmatrix} = \frac{2\pi}{x_1} \quad (7.29)$$

Hence, it follows:

$$f_Y(y_1, y_2) = \frac{x_1}{2\pi} = \left[\frac{1}{\sqrt{2\pi}} e^{-y_1^2/2} \right] \left[\frac{1}{\sqrt{2\pi}} e^{-y_2^2/2} \right] \quad (7.30)$$

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 7.1 (MSc. Exam Question, 2005). 1. Let U be a random variable generated from a uniform pdf on the interval $[0, 1]$, such that

$$f_U(u) = \begin{cases} 1, & \text{if } 0 \leq u \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

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:

$$f_Y(y) = \begin{cases} \frac{1}{B(\alpha, 1-\alpha)} y^{\alpha-1} (1-y)^{-\alpha}, & 0 \leq y \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

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:

$$f_{WZ}(w, z) = \begin{cases} \frac{\lambda}{B(\alpha, 1-\alpha)} e^{-\lambda w} z^{\alpha-1} (w-z)^{-\alpha}, & \text{if } (w, z) \in \mathcal{R} \\ 0, & \text{otherwise} \end{cases}$$

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:

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad \text{where} \quad \Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx \quad \times$$

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 $0 \leq u \leq 1$ has a single root:

$$u = \begin{cases} e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (7.31)$$

The derivative of the function $X = g(U)$ for $0 \leq u \leq 1$ is given by:

$$g'(u) = \frac{dg(u)}{du} = -\frac{1}{\lambda u} \quad (7.32)$$

Hence, noting that the pdf for the RV U is uniform, then the pdf for X is:

$$f_X(x) = \sum_{n=1}^N \frac{f_U(u_n)}{|g'(u_n)|} = \begin{cases} \frac{1}{\lambda u} = \lambda u & \text{if } 0 \leq u \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (7.33)$$

which gives the desired exponential distribution with pdf:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (7.34)$$

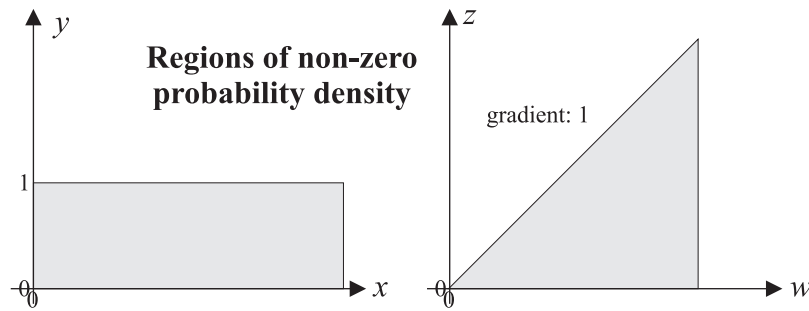


Figure 7.3: Region of non-zero probability density

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:

$$J = \frac{\partial(w, z)}{\partial(x, y)} = \begin{vmatrix} \frac{\partial w}{\partial x} & \frac{\partial w}{\partial y} \\ \frac{\partial z}{\partial x} & \frac{\partial z}{\partial y} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ y & x \end{vmatrix} = x = w \quad (7.35)$$

Moreover, note that there is one root of the transformation, and this is given by:

$$x = w \quad \text{and} \quad y = \frac{z}{w} \quad (7.36)$$

Since X and Y are independent RVs, the joint-pdf of W and Z is therefore:

$$f_{WZ}(w, z) = \frac{1}{J} f_{XY}(x, y) = \frac{1}{w} f_X(w) f_Y\left(\frac{z}{w}\right) \quad (7.37)$$

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 7.3

Substituting for $f_X(x)$ and $f_Y(y)$ in the non-zero region gives:

$$f_{WZ}(w, z) = \frac{1}{w} \lambda e^{-\lambda w} \frac{1}{B(\alpha, 1-\alpha)} \left(\frac{z}{w}\right)^{\alpha-1} \left(1 - \frac{z}{w}\right)^{-\alpha} \quad (7.38)$$

$$= \frac{\lambda}{B(\alpha, 1-\alpha)} e^{-\lambda w} z^{\alpha-1} w^{-\alpha} \left(\frac{w-z}{w}\right)^{-\alpha} \quad (7.39)$$

which gives the desired result:

$$f_{WZ}(w, z) = \begin{cases} \frac{\lambda}{B(\alpha, 1-\alpha)} e^{-\lambda w} z^{\alpha-1} (w-z)^{-\alpha} & w \geq 0 \quad \text{and} \quad 0 \leq z \leq w \\ 0 & \text{otherwise} \end{cases} \quad (7.40)$$

3. The marginal-pdf of Z is given by integrating over w :

$$f_Z(z) = \int_z^\infty f_{WZ}(w, z) dw \quad (7.41)$$

The limits of this integration are obtained by looking back at Figure 7.3, and considering the values of w for a fixed value of z . Hence, for $z > 0$,

$$f_Z(z) = \int_z^\infty \frac{\lambda}{B(\alpha, 1-\alpha)} e^{-\lambda w} z^{\alpha-1} (w-z)^{-\alpha} dw \quad (7.42)$$

$$= \frac{\lambda}{B(\alpha, 1-\alpha)} z^{\alpha-1} \int_z^\infty e^{-\lambda w} (w-z)^{-\alpha} dw \quad (7.43)$$

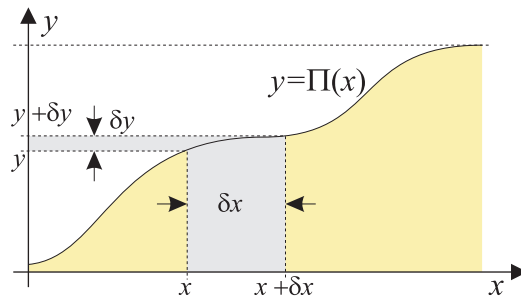


Figure 7.4: A simple derivation of the inverse transform method

Making the substitution $g = \lambda(w - z)$, such that when $w = z$, $g = 0$, and when $w \rightarrow \infty$, $g \rightarrow \infty$. Further, $dg = \lambda dw$. Therefore,

$$f_Z(z) = \frac{\lambda}{B(\alpha, 1 - \alpha)} z^{\alpha-1} \int_0^\infty e^{-(g+\lambda z)} \left(\frac{g}{\lambda}\right)^{-\alpha} \frac{dg}{\lambda} \quad (7.44)$$

$$= \frac{\lambda^\alpha}{B(\alpha, 1 - \alpha)} z^{\alpha-1} e^{-\lambda z} \int_0^\infty e^{-g} g^{-\alpha} dg \quad (7.45)$$

Finally, using the identities given in the question:

$$B(\alpha, 1 - \alpha) = \frac{\Gamma(\alpha)\Gamma(1 - \alpha)}{\Gamma(1)} \quad \text{where} \quad \Gamma(1 - \alpha) = \int_0^\infty x^{1-\alpha-1} e^{-x} dx \quad (7.46)$$

where $\Gamma(1) = 1$, then it follows that:

$$f_Z(z) = \frac{\lambda^\alpha}{\Gamma(\alpha)\Gamma(1 - \alpha)} z^{\alpha-1} e^{-\lambda z} \Gamma(1 - \alpha) = \frac{\lambda^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\lambda z}, \quad z \geq 0 \quad (7.47)$$

and zero otherwise, which, using the definition given in the notes, is a Gamma distribution with parameters λ and α : $f_Z(z) = \Gamma(z | \lambda, \alpha)$.

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 α , $1 - \alpha$.
 - (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*.

7.2.4 Inverse Transform Method

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. New slide

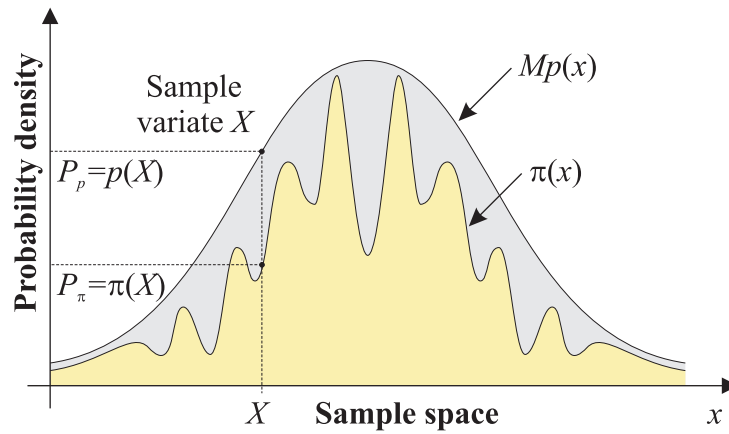


Figure 7.5: Rejection sampling

Referring to Figure 7.4, 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:

$$f_X(x) = \frac{d\Pi(x)}{dx} f_Y(y) \quad (7.48)$$

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

$$\pi(x) = \frac{d\Pi(x)}{dx} \quad (7.49)$$

it follows that

$$f_X(x) = \pi(x), \quad \text{where } x = \Pi^{-1}(y) \quad (7.50)$$

In other words, if

$$U(\zeta) \sim \mathcal{U}_{[0,1]}, \quad X(\zeta) = \Pi^{-1}U(\zeta) \sim \pi(x) \quad (7.51)$$

Example 7.2 (Exponential variable generation). If $X(\zeta) \sim \text{Exp}(1)$, 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).

7.2.5 Acceptance-Rejection Sampling

For most distributions, it is often difficult or even impossible to directly simulate using either the inverse transform or probability transformations. 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 $p(x)$ from which it is *easy* to sample from, where $Mp(x) > \pi(x)$.



New slide

The basic idea of acceptance-rejection sampling is shown in Figure 7.5. It is desired to sample from the distribution $\pi(x)$ which cannot be sampled from using the transform methods above. However, assume it has been possible to find a proper density $p(x)$ and a constant M such that $Mp(x) > \pi(x)$. This is shown in Figure 7.5 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 $p(x)$. This sample has been drawn with probability $P_g \delta x$ where $P_g = p(X)$. However, if the sample were really to have been drawn from the desired distribution, it should have probability $P_\pi \delta x$ where $P_\pi = \pi(X)$. Hence, on average, you would expect to have too many variates that take on the value X by a factor of

$$u(X) = \frac{P_p}{P_\pi} = \frac{p(X)}{\pi(X)} \quad (7.52)$$

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:

$$P_a = \Pr(\text{accept variate } X) = \frac{\pi(X)}{Mp(x)} \quad (7.53)$$

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.

7.2.5.1 Envelope and Squeeze Methods

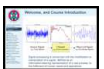
A problem with many sampling methods, which can make the density $\pi(x)$ difficult to simulate, is down to the complexity of the function $\pi(x)$ itself; the function may require substantial computing time at each evaluation.

It is possible to reduce the algorithmic complexity of the accept-reject algorithm by looking for another computationally simple function, $q(x)$ which *bounds* $\pi(x)$ from below.

In the case that the proposed variate X satisfies $q(X) \leq \pi(X)$, 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 7.7.

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.



New slide

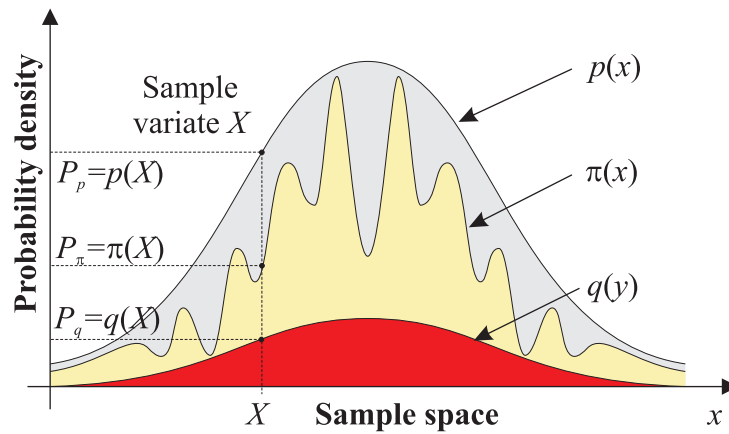


Figure 7.6: Envelope Rejection sampling

By construction of a lower envelope on $\pi(x)$, the number of function evaluations is potentially decreased by a factor of

$$P_{\bar{\pi}} = \frac{1}{M} \int q(x) dx \quad (7.54)$$

which is the probability that $\pi(x)$ is not evaluated.

7.2.6 Importance Sampling

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:

$$\mathcal{I} = \int_{\Theta} f(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (7.55)$$

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:

$$\hat{\mathcal{I}} = \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{I}_{\Theta}(\boldsymbol{\theta}^{(k)}), \quad \text{where } \boldsymbol{\theta}^{(k)} \sim f(\boldsymbol{\theta}) \quad (7.56)$$

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 7.56. A best estimate is as follows:

Defining an *easy-to-sample-from* density $\pi(\boldsymbol{\theta}) > 0, \forall \boldsymbol{\theta} \in \Theta$:

$$\mathcal{I} = \int_{\Theta} \frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathbb{E}_{\pi} \left[\frac{f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \right], \quad (7.57)$$

leads to an estimator based on the **sample expectation**:

$$\hat{\mathcal{I}} = \frac{1}{N} \sum_{k=0}^{N-1} \frac{f(\boldsymbol{\theta}^{(k)})}{\pi(\boldsymbol{\theta}^{(k)})} \quad (7.58)$$

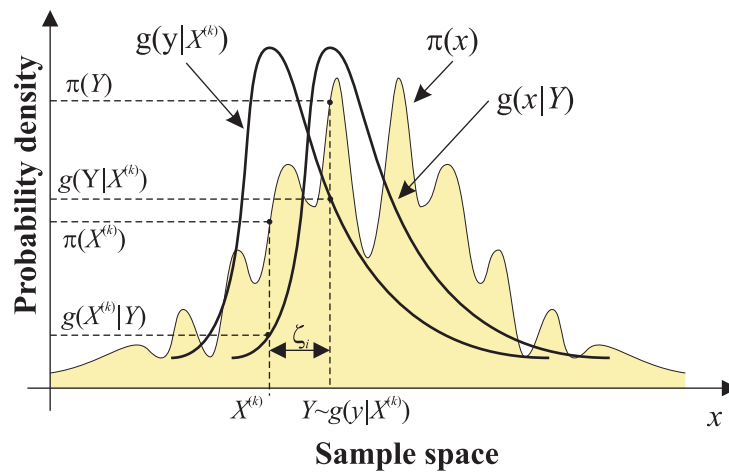


Figure 7.7: Graphical representation of the Metropolis-Hastings algorithm.

7.2.7 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).

7.3 Markov chain Monte Carlo Methods

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.

7.3.1 The Metropolis-Hastings algorithm

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 $\pi(x)$ 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**: $Y \sim g(y | X^{(k)})$.

2. Set the new random variate to be:

$$X^{(k+1)} = \begin{cases} Y & \text{with probability } \rho(X^{(k)}, Y) \\ X^{(k)} & \text{with probability } 1 - \rho(X^{(k)}, Y) \end{cases} \quad (7.59)$$

where the acceptance ratio function $\rho(x, y)$ is given by:

$$\rho(x, y) = \min \left\{ \frac{\pi(y)}{g(y|x)} \left(\frac{\pi(x)}{g(x|y)} \right)^{-1}, 1 \right\} \equiv \min \left\{ \frac{\pi(y)}{\pi(x)} \frac{g(x|y)}{g(y|x)}, 1 \right\} \quad (7.60)$$

This calculation is represented graphically in Figure 7.7.

7.3.1.1 Gibbs Sampling

Gibbs sampling is a Monte Carlo method that facilitates sampling from a multivariate density function, $\pi(\theta_0, \theta_1, \dots, \theta_M)$ by drawing successive samples from marginal densities of smaller dimensions.

Using the probability chain rule,

$$\pi(\{\theta_m\}_{m=1}^M) = \pi(\theta_\ell | \{\theta_m\}_{m=1, m \neq \ell}^M) \pi(\{\theta_m\}_{m=1, m \neq \ell}^M) \quad (7.61)$$

The Gibbs sampler works by drawing random variates from the marginal densities $\pi(\theta_\ell | \{\theta_m\}_{m=1, m \neq \ell}^M)$ in a cyclic iterative pattern.

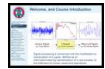
This proceeds as follows assuming the components are initialised with values $\theta_0^{(0)}, \theta_1^{(0)}, \dots, \theta_M^{(0)}$

First iteration:

$$\begin{aligned} \theta_1^{(1)} &\sim \pi(\theta_1 | \theta_2^{(0)}, \theta_3^{(0)}, \theta_4^{(0)}, \dots, \theta_M^{(0)}) \\ \theta_2^{(1)} &\sim \pi(\theta_2 | \theta_1^{(1)}, \theta_3^{(0)}, \theta_4^{(0)}, \dots, \theta_M^{(0)}) \\ \theta_3^{(1)} &\sim \pi(\theta_3 | \theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(0)}, \dots, \theta_M^{(0)}) \\ &\vdots \\ \theta_M^{(1)} &\sim \pi(\theta_M | \theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(1)}, \dots, \theta_{M-1}^{(1)}) \end{aligned} \quad (7.62)$$

Second iteration:

$$\begin{aligned} \theta_1^{(2)} &\sim \pi(\theta_1 | \theta_2^{(1)}, \theta_3^{(1)}, \theta_4^{(1)}, \dots, \theta_M^{(1)}) \\ \theta_2^{(2)} &\sim \pi(\theta_2 | \theta_1^{(2)}, \theta_3^{(1)}, \theta_4^{(1)}, \dots, \theta_M^{(1)}) \\ \theta_3^{(2)} &\sim \pi(\theta_3 | \theta_1^{(2)}, \theta_2^{(2)}, \theta_4^{(1)}, \dots, \theta_M^{(1)}) \\ &\vdots \\ \theta_M^{(2)} &\sim \pi(\theta_M | \theta_1^{(2)}, \theta_2^{(2)}, \theta_4^{(2)}, \dots, \theta_{M-1}^{(2)}) \end{aligned} \quad (7.63)$$



New slide

$k + 1$ -th iteration:

$$\begin{aligned}
 \theta_1^{(k+1)} &\sim \pi \left(\theta_1 \mid \theta_2^{(k)}, \theta_3^{(k)}, \theta_4^{(k)}, \dots, \theta_M^{(k)} \right) \\
 \theta_2^{(k+1)} &\sim \pi \left(\theta_2 \mid \theta_1^{(k+1)}, \theta_3^{(k)}, \theta_4^{(k)}, \dots, \theta_M^{(k)} \right) \\
 \theta_3^{(k+1)} &\sim \pi \left(\theta_3 \mid \theta_1^{(k+1)}, \theta_2^{(k+1)}, \theta_4^{(k)}, \dots, \theta_M^{(k)} \right) \\
 &\vdots \\
 \theta_M^{(k+1)} &\sim \pi \left(\theta_M \mid \theta_1^{(k)}, \theta_2^{(k)}, \theta_4^{(k)}, \dots, \theta_{M-1}^{(k)} \right)
 \end{aligned} \tag{7.64}$$

At the end of the j -th iteration, the samples $\theta_0^{(j)}, \theta_1^{(j)}, \dots, \theta_M^{(j)}$ are considered to be drawn from the joint-density $\pi(\theta_0, \theta_1, \dots, \theta_M)$.

Part III
Stochastic Processes

8

Review of Fourier Transforms and Discrete-Time Systems



This handout will review complex Fourier series and Fourier transforms, followed by a review of discrete-time systems. It covers complex Fourier series, Fourier transforms, Discrete-time Fourier transforms, Discrete Fourier Transforms, Parseval's Theorem, the bilateral Z-transform, frequency response, and rational transfer functions.

8.1 Introduction

This handout will review complex **Fourier series** and **Fourier transforms**, followed by a review of **discrete-time systems**. The reader is expected to have previously covered most of the concepts in this handout, although it is likely that the reader might need to revise the material if it's been a while since it's been studied. Nevertheless, this revision material is included in the module as review material purely for completeness and reference. It is not intended as a full introduction, although some parts of the review cover the subject in detail.

As discussed in the first handout, if the reader wishes to revise these topics in more detail, the following book comes *highly* recommended:

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

For undergraduate level text books covering signals and systems theory, which it is assumed you have covered, the following book is recommended:

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.homepages.ed.ac.uk/pmg/SIGPRO/>

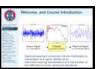
The latest edition was printed in 2002, but any edition will do. An alternative presentation of roughly the same material is provided by the following book:

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

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

In particular, the appendix on complex numbers may prove useful.

8.2 Signal Classification



Topic Summary 50 Deterministic time-series signal classification

New slide

Topic Objectives:

- Distinguish periodic and non-periodic, discrete-time and continuous signals.
- Ability to distinguish different signal types.

Topic Activities:

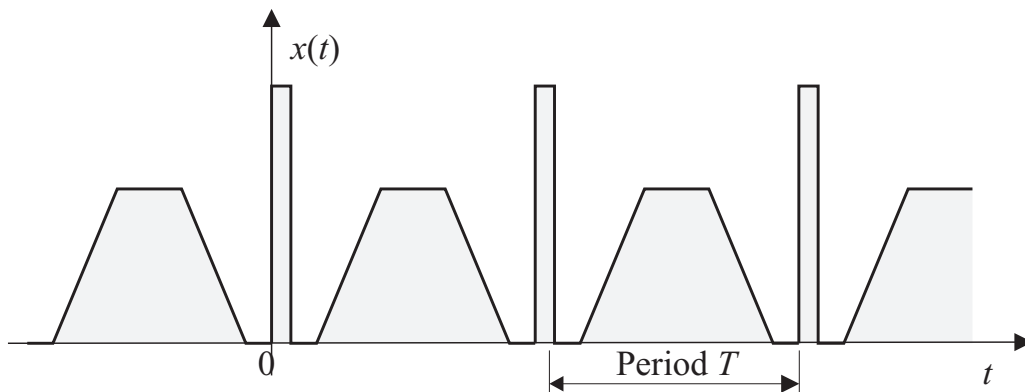
Type	Details	Duration	Progress
Read Handout	Read page 277 to page 281	8 mins/page	

Before considering the analysis of signals and systems, it is necessary to be aware of the general classifications to which signals can belong, and to be aware of the significance of some subtle characteristics that determine how a signal can be analysed. Not all signals can be analysed using a particular technique.

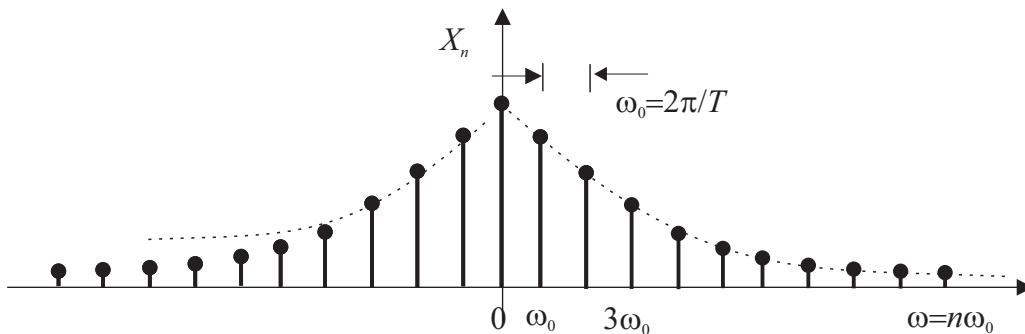
Different types of **deterministic** signals include:

- continuous-time periodic signals;
- continuous-time non-periodic (or aperiodic) signals;
- discrete-time periodic signals;
- discrete-time aperiodic signals.

The variety of signal classes rapidly changes when the notion of **random** or **stochastic** signals are introduced (not until the fourth-year!).



(a) An example of a periodic signal with period T .



(b) The Fourier series of the periodic signal in Figure 8.1a with fundamental frequency $\omega_0 = 2\pi/T$.

Figure 8.1: Example of a periodic signal and its spectral representation, found using the Fourier series.

8.2.1 Types of signal

In general, there are four distinct types of deterministic signals that must be analysed:

Continuous-time periodic Such signals repeat themselves after a fixed length of time known as the period, usually denoted by T . This repetition continues ad-infinitum (i.e. forever). Formally, a signal, $x(t)$, is periodic if

$$x(t) = x(t + mT), \forall m \in \mathbb{Z} \quad (8.1)$$

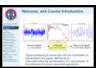
where the notation $\forall m \in \mathbb{Z}$ means that m takes on *all* integer values: in other-words, $m = -\infty, \dots, -2, -1, 0, 1, 2, \dots, \infty$. The smallest positive value of T which satisfies this condition is the defined as the **fundamental period**.

These signals will be analysed using the **Fourier Series**, and are used to represent real-world waveforms that are near to being periodic over a sufficiently long period of time.

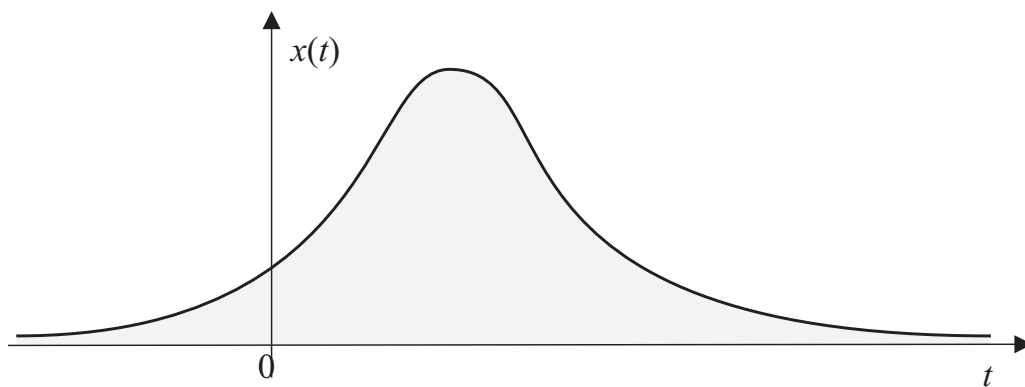
An example of a periodic signal is shown in Figure 8.1a. This kind of signal vaguely represents a line signal in analogue television, where the rectangular pulses represent line synchronisation signals.

Continuous-time aperiodic Continuous-time aperiodic signals are not periodic over all time, although they might be locally periodic over a short time-scale.

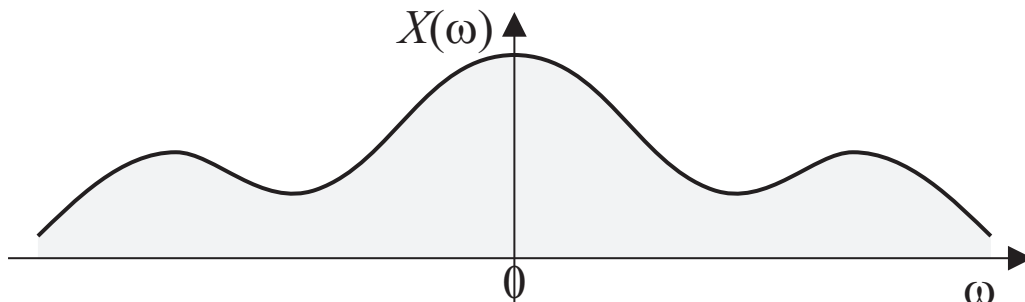
These signals can be analysed using the **Fourier transform** for most cases, and more often using the **Laplace transform**. Aperiodic signals are more representative of many real-world signals.



New slide



(a) An example of an aperiodic signal.



(b) The Fourier transform of the aperiodic signal in Figure 8.2a, where every possible frequency exists.

Figure 8.2: Example of an aperiodic signal and its spectral representation, found using the Fourier transform.

Again, real signals don't last for all time, although can last for a considerably long time. An example of an aperiodic signal is shown in Figure 8.2a.

Discrete-time periodic A discrete-time periodic signal is shown in Figure 8.3, which is essentially a *sampled* version of the signal shown in Figure 8.1a. Note in this case, the period is often denoted by N , primarily to reflect the fact the time-index is now n ; in other words, $x[n] = x(nT_s)$, $n \in \{0, 1, 2, \dots\}$, where T_s is the sampling interval.

A discrete-time signal, $x[n]$, is periodic if:

$$x[n] = x[n + mN], \forall m \in \mathbb{Z} \quad (8.2)$$

This is, of course, similar to Equation 8.1. Discrete-time periodic signals can be analysed using the discrete-time Fourier series or discrete Fourier transform (DFT) depending on whether the period is a multiple of the number of samples.

Discrete-time aperiodic Analogous to the continuous-time aperiodic signal in Figure 8.2a, a discrete-time aperiodic signal is shown in Figure 8.4.

Aperiodic discrete-time signals will be analysed using the discrete-time Fourier transform (DTFT). It can also be analysed using the so-called z -transform, which is the discrete-time version of the **Laplace transform**, although this will not be covered in complete detail until the third and fourth year courses, **Signals and Communications 3, Discrete-Time Signal Analysis**.

Finite-length discrete-time signals Discrete-time signals can also be classified as being finite in length. In other words, they are not assumed to exist for all-time, and what happens outside the **window** of data is assumed unknown. These signals can be modelled

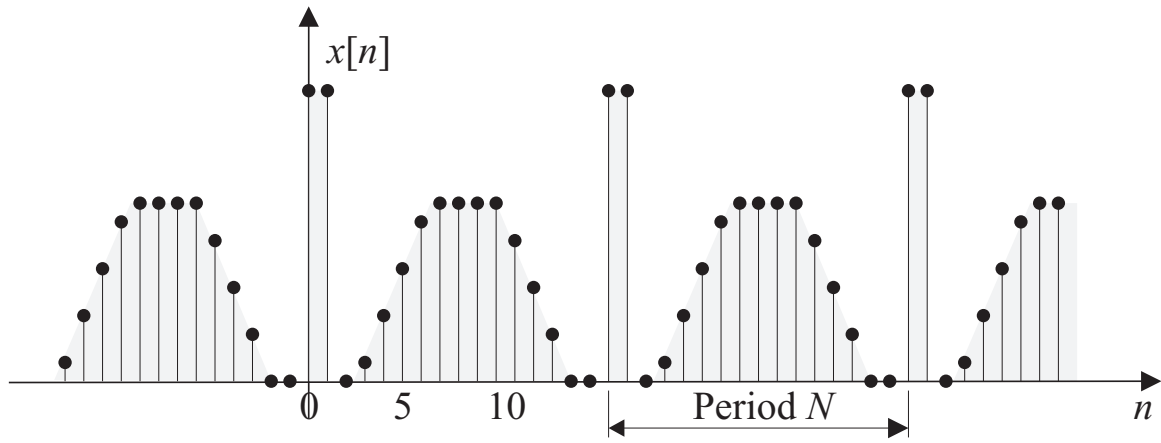


Figure 8.3: A discrete-time periodic signal.

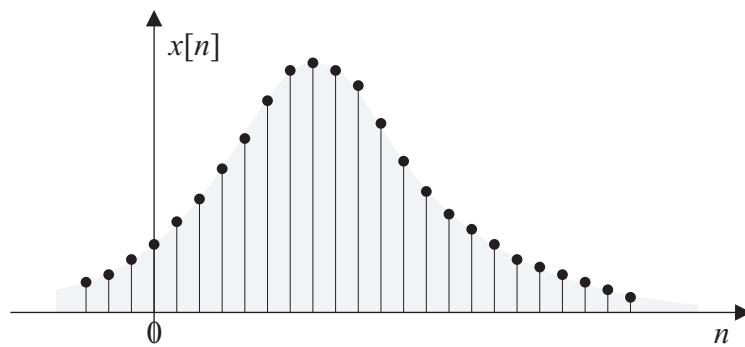


Figure 8.4: An example of a discrete-time aperiodic signal.

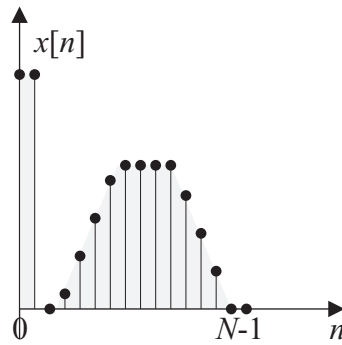


Figure 8.5: An example of a finite-duration signal.

using the so-called **DFT**, but again this is not covered until the fourth year course, **Discrete-Time Signal Analysis**. The Fast Fourier transform (FFT) is the well-known fast (low complexity) version of the DFT.

– End-of-Topic 50: **Summary of Different Types of Signals?** –



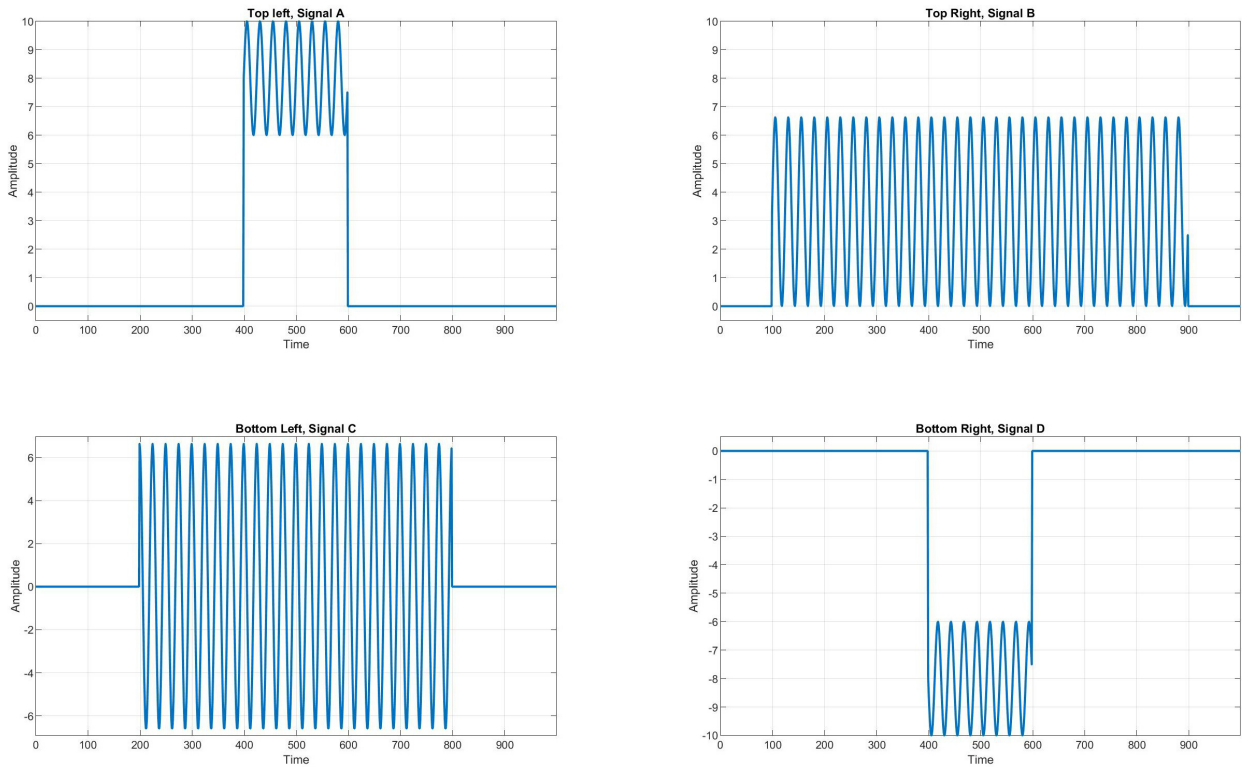
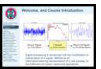


Figure 8.6: Which signal is the largest?

8.2.2 Energy and Power Signals



Topic Summary 51 Measuring the size of a signal (and introduction to signal norms)

New slide

Topic Objectives:

- Understanding how to measure the size (or norm) of a signal.
- Motivation for Energy and Power.

Topic Activities:

Type	Details	Duration	Progress
Read Handout	Read page 282 to page 284	8 mins/page	
Try Example	Try Example 8.1	15 mins	

There are many applications, such as signal detection, where knowing the *size* of a signal is important. A large signal such as aircraft noise as it flies over a particular town might be more or less significant than a longer signal of lower amplitude, but it all depends on the application.

Example 8.1 (Multi-choice Question). Which of the signals shown in Figure 8.6 is the *largest*!?

Moreover, as stated in Section 8.2.1, signals can be analysed using a variety of frequency-domain transform methods, such as the **Fourier series**, **Fourier transform**, **Laplace transform**, and for discrete-time, the ***z*-transform**. For example, the Fourier transform is used to analyse aperiodic continuous-time signals.



Figure 8.7: What is the size of an object?

Sidebar 19 Size of a Human Being

Suppose we wish to devise a signal number V as a measure of the size of a human being. Then clearly, the width (or girth) must also be taken into account as well as the height. One could make the simplifying assumption that the shape of a person is a cylinder of variable radius r (which varies with the height h). Then one possible measure of the size of a person of height H is the person's volume, given by:

$$V = \pi \int_0^H r^2(h) dh \quad (8.3)$$

This can be found by dividing the object into circular discs (which is an approximation), where each disc has a volume $\delta V \approx \pi r^2(h) \delta h$. Then the total volume is given by $V = \int dV$.

However, not all aperiodic signals can be analysed using the Fourier transform, and the reason for this can be directly related to a fundamental property of a signal: a measure of *how much signal there is*.

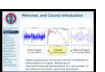
Therefore it is relevant to consider the **energy** or **power** as a means for characterising a signal. The concepts of **power** and **energy** intuitively follow from their use in other aspects of the physical sciences. However, the concept of signals which exist for all time requires careful definitions, in order to determine when it has **energy** and when it has **power**.

Intuitively, energy and power measure *how big* a signal is. A motivating example of measuring the size of something is given in Sidebar 19, and in Figure 8.7. However, there are other possible signal measures, as discussed in Sidebar 20.

8.2.2.1 Motivation for Energy and Power Expressions

Considering power from an electrical perspective, if a voltage $x(t)$ is connected across a resistance R , the dissipated power at time τ is given by:

$$P(\tau) = \frac{x^2(\tau)}{R} \propto x^2(\tau) \quad (8.4)$$



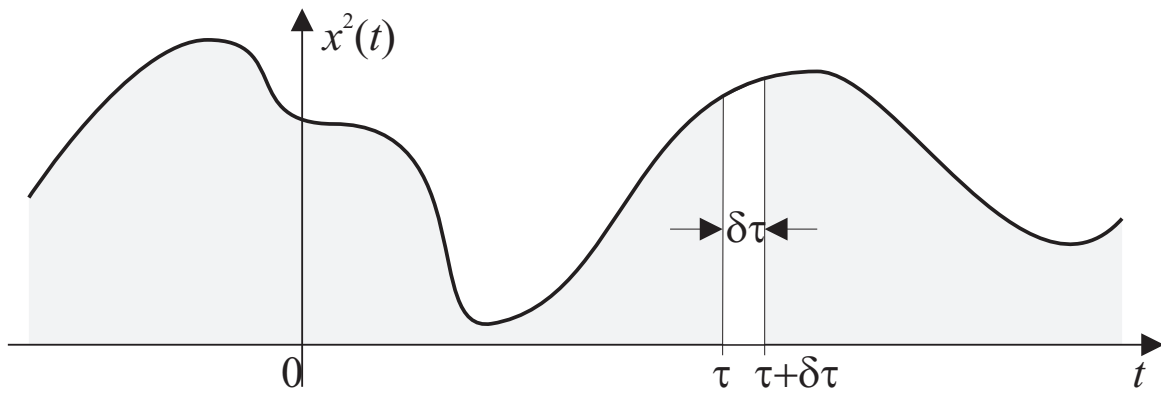


Figure 8.8: Energy Density.

where \propto denotes *proportional to*. In this case, the constant of proportionality is the inverse resistance. Since energy and power are related through the expression

$$\text{Energy} = \text{Power} \times \text{Time}, \quad (8.5)$$

the energy dissipated between times τ and $\tau + \delta\tau$, as indicated in Figure 8.8, is:

$$\delta E(\tau) \propto P(\tau) \delta\tau \propto x^2(\tau) \delta\tau \quad (8.6)$$

The total energy over all time can thus be found by integrating over all time:

$$E \propto \int_{-\infty}^{\infty} x^2(\tau) d\tau \quad (8.7)$$

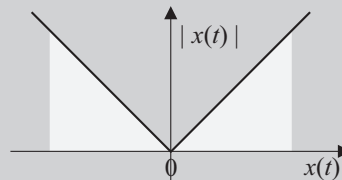
This leads to the formal definitions of energy and power.

– End-of-Topic 51: **Introduction to Energy and Power Signals** –



Sidebar 20 Other signal measures

1. While the area under a signal $x(t)$ is a possible measure of its size, because it takes account not only of the amplitude but also of the duration, is defective since even for a very large signal, the positive and negative areas could cancel each other out, indicating a signal of a small size.
2. Using the sum of square values can potentially give undue weighting to any outliers in the signal, where an outlier is defined as an unusual signal variation that is not characteristic of the rest of the signal; an example might be a high-energy shot burst of interference.
3. Therefore, taking the integral of the absolute value, $|x(t)| \equiv \text{abs } x(t)$, is a possible measure and in some circumstances can be used. The relationship between input and output for this signal measure is shown below.



Unfortunately, dealing with the absolute value of a function can be difficult to manipulate mathematically. However, using the area under the square of the function is not only more mathematically tractable but is also more meaningful when compared with the electrical examples and the volume in Sidebar 19.

4. These notions lead onto the more general subject of **signal norms**. The L_p -norm is defined by:

$$L_p(x) \triangleq \left(\int_{-\infty}^{\infty} |x(t)|^p dt \right)^{\frac{1}{p}}, \quad p \geq 1 \quad (8.8)$$

In particular, the expression for energy is related to the L_2 -norm, while using the absolute value of the signal gives rise to the L_1 -norm:

$$L_1(x) \triangleq \int_{-\infty}^{\infty} |x(t)| dt \quad (8.9)$$

which is the integral of the absolute value as described above in part 3.

5. While Parseval's theorem, described on later for the power of periodic signals, develops a relationship between the L_2 -norms in the time-domain and frequency-domain, in general no relation exists for other values of p .
6. Note that the L_p -norm generalises for discrete-time signals as follows:

$$L_p(x) \triangleq \left(\sum_{-\infty}^{\infty} |x[t]|^p \right)^{\frac{1}{p}}, \quad p \geq 1 \quad (8.10)$$

8.2.2.2 Formal Definitions for Energy and Power

Topic Summary 52 Energy and Power Definitions

Topic Objectives:

- Formal definitions for Energy and Power.
- Units of Energy and Power.

Topic Activities:

Type	Details	Duration	Progress
Read Handout	Read page 286 to page 288	8 mins/page	
Try Example	Try Examples 8.2, 8.3, and 8.4	15 mins	

Based on the justification in Section 8.2.2.1, the formal abstract definitions for energy and power that are independent of how the energy or power is dissipated are defined below.

Energy Signals A continuous-time signal $x(t)$ is said to be an **energy signal** if the total energy, E , dissipated by the signal over all time is both *nonzero* and *finite*. Thus:

$$0 < E < \infty \quad \text{where} \quad E = \int_{-\infty}^{\infty} |x(t)|^2 dt \quad (8.11)$$

where $|x(t)|$ means the magnitude of the signal $x(t)$. If $x(t)$ is a real-signal, this is just its amplitude. If $x(t)$ is a complex-signal, then $|x(t)|^2 = x(t) x^*(t)$ where $*$ denotes complex-conjugate. In this course, however, only real signals will be encountered.

A necessary condition for the energy to be finite is that the signal amplitude $|x(t)| \rightarrow 0$ as $|t| \rightarrow \infty$, otherwise the integral in Equation 8.11 will not exist. When the amplitude of $x(t)$ does not tend to zero as $|t| \rightarrow \infty$, the signal energy is likely to be infinite. A more meaningful measure of the signal size in such a case would be the time average of the energy if it exists. This measure is called the **power** of the signal.

Power signals If the average power delivered by the signal over all time is both *nonzero* and *finite*, the signal is classified as a power signal:

$$0 < P < \infty \quad \text{where} \quad P = \lim_{W \rightarrow \infty} \frac{1}{2W} \int_{-W}^W |x(t)|^2 dt \quad (8.12)$$

where the variable W can be considered as half of the width of a *window* that covers the signal and gets larger and larger.

Example 8.2. Name a type of signal which is not an example of an **energy signal**?

SOLUTION. A periodic signal has finite energy over one period, so consequently has infinite energy. However, as a result it has a finite average power and is therefore a power signal, and not an energy signal.

Example 8.3 (Rectangular Pulse). What is the energy of the rectangular pulse shown in Figure 8.9 as a function of τ , and for the particular case of $\tau = 4$?

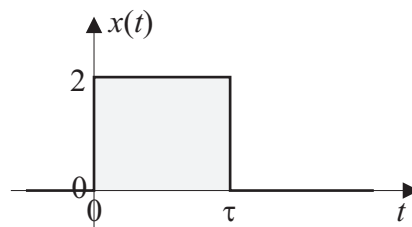
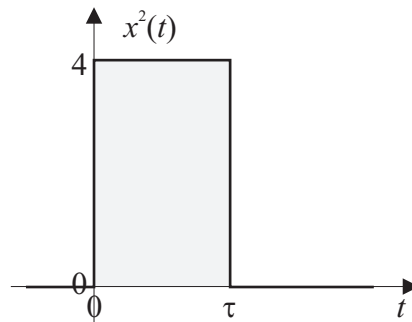
Figure 8.9: Rectangular pulse of length τ .

Figure 8.10: The total energy of the signal in Figure 8.9 can be found as the area under the curve representing the square of the rectangular pulse, as shown here.

SOLUTION. The signal can be represented by

$$x(t) = \begin{cases} 2 & 0 \leq t < \tau \\ 0 & \text{otherwise} \end{cases} \quad (8.13)$$

so that the square of the signal is also rectangular and given by

$$x^2(t) = \begin{cases} 4 & 0 \leq t < \tau \\ 0 & \text{otherwise} \end{cases} \quad (8.14)$$

Since an integral can be interpreted as the area under the curve (see Figure 8.10), the total energy is thus:

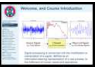
$$E = 4\tau \quad (8.15)$$

□

When $\tau = 4$, $E = 16$.

Example 8.4 (Multiple Choice). The signal $x(t) = \exp(-|t|)$ is:

1. an energy signal, but not a power signal;
2. a power signal, but not an energy signal;
3. both an energy and a power signal;
4. not an energy signal, nor a power signal?



8.2.2.3 Units of Energy and Power

It is important to consider the physical units associated with energy and power, and therefore to determine how the abstract definitions of E and P in Equation 8.11 and Equation 8.12 can be converted into real energy and power. New slide

Consider again power from an electrical perspective. When considering “direct current” (DC) signals, power is given by

$$P_{DC} = \frac{V^2}{R} = \frac{\text{Volts}^2}{\text{Ohms}} = \text{Watts} \quad (8.16)$$

where V is the signal voltage, and R is the resistance through which the power is dissipated. Consider now the units of the abstract definition of power, P in Equation 8.12:

$$P = \frac{1}{\text{time}} \times \text{Volts}^2 \times \text{time} = \text{Volts}^2 = \text{Watts} \times \text{Ohms} \quad (8.17)$$

where the second unit of *time* comes from the integral term dt , and in which the integral may be considered as a summation. Therefore, by comparing Equation 8.16 and Equation 8.12, the abstract definition of power, P , can be converted to real power by **Ohms** for the case of electrical circuits.

Similarly, the units of energy in Equation 8.11 is given by

$$E = \text{volts}^2 \times \text{time} \quad (8.18)$$

Therefore, to convert the abstract energy to Joules, it is again necessary to divide by **Ohms** by noting that energy is power multiplied by time.

8.2.2.4 Power for Periodic Signals

The expression for power in Equation 8.12 can be simplified for periodic signals. Consider the periodic signal in Figure 8.1a. Let $2W = T$ and define:

$$P_T = \frac{1}{2W} \int_{-W}^W |x(t)|^2 dt \quad (8.19)$$

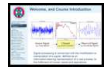
Thus, the average power over two periods is $2P_T$, and the average power over N periods is P_{NT} . Then, it should become clear that:

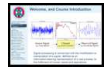
$$P_T = P_{NT}, \forall N \in \mathbb{Z} \quad (8.20)$$

since the average in each period is the same. Consequently, **power** for a periodic signal with period T may be defined as:

$$P = \frac{1}{T} \int_0^T |x(t)|^2 dt \quad (8.21)$$

Note that the limits in Equation 8.21 may be over any period and thus can be replaced by $(\tau, \tau + T)$ for any value of τ .





8.3 Fourier Series and Fourier Transforms

Topic Summary 53 Fourier Transform Theory

Topic Objectives:

- Understanding how to measure the size (or norm) of a signal.
- Motivation for Energy and Power.

Topic Activities:

Type	Details	Duration	Progress
Read Handout	Read page 289 to page 296	8 mins/page	
Try Example	Try Examples 8.5, 8.6, and 8.7	40 mins	

In this review of Fourier series and transforms, the topics covered are:

- Complex Fourier series
- **Fourier transform**
- The discrete-time Fourier transform
- Discrete Fourier transform

8.3.1 Complex Fourier series

The complex Fourier series is a frequency analysis tool for continuous time periodic signals. Examples of periodic signals encountered in practice include square waves, triangular waves, sawtooth waves, pulse waves and, of course, sinusoids and complex exponentials, as well as half-wave rectified, full-wave rectified and p -phased rectified sinusoids. The basic mathematical representation of periodic signals is the Fourier series, which is a linear weighted sum of harmonically related sinusoids or complex exponentials.

A **periodic continuous-time** deterministic signal, $x_c(t)$, with fundamental period T_p can be expressed as a linear combination of harmonically related complex exponentials:

$$x_c(t) = \sum_{k=-\infty}^{\infty} \check{X}_c(k) e^{jk\omega_0 t}, \quad t \in \mathbb{R}, \quad (\text{M:2.2.1})$$

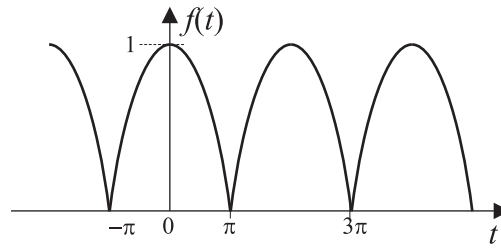
where $\omega_0 = 2\pi F_0 = \frac{2\pi}{T_p}$ is the **fundamental frequency**. Here, ω_0 is in radians per second, and the fundamental frequency, in Hertz, is given by $F_0 = \frac{1}{T_p}$. Moreover,

$$\check{X}_c(k) = \frac{1}{T_p} \int_0^{T_p} x_c(t) e^{-jk\omega_0 t} dt, \quad k \in \mathbb{Z} \quad (\text{M:2.2.2})$$

are termed the **Fourier coefficients**, or **spectrum** of $x_c(t)$. Note that although the region of integration in Equation M:2.2.2 is from 0 to T_p , it can actually be over any period of the waveform, since the signal, $x_c(t)$, is periodic with period T_p .

The k th frequency component corresponds to frequency $\omega_k = k\omega_0 = k\frac{2\pi}{T_p}$. The set of exponential functions

$$\mathcal{F}(t) = \{e^{j\omega_0 kt}, k \in \mathbb{Z}\} \quad (8.22)$$

Figure 8.11: Function $f(t)$ of Example 8.5

can be thought of as the basic *building blocks* from which periodic signals of various types can be constructed with the proper choice of fundamental frequency and Fourier coefficients.

Example 8.5 (Complex Fourier Series). Find the complex form of the Fourier series expansion of the periodic function $f(t)$ defined by:

$$\begin{aligned} f(t) &= \cos \frac{1}{2}t \quad (-\pi < t < \pi) \\ f(t + 2\pi) &= f(t) \end{aligned} \quad (8.23)$$

SOLUTION. A graph of the function $f(t)$ over the interval $-\pi \leq t \leq 3\pi$ is shown in Figure 8.11. The period $T_p = 2\pi$, so therefore the complex coefficients, denoted by F_n , are given by Equation M:2.2.2 as:

$$F_n = \frac{1}{T_p} \int_0^{T_p} f(t) e^{-jn\omega_0 t} dt, \quad n \in \mathbb{Z} \quad (8.24)$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos \frac{t}{2} e^{-jnt} dt = \frac{1}{4\pi} \int_{-\pi}^{\pi} \left(e^{j\frac{t}{2}} + e^{-j\frac{t}{2}} \right) e^{-jnt} dt \quad (8.25)$$

$$= \frac{1}{4\pi} \int_{-\pi}^{\pi} \left(e^{-j(n-\frac{1}{2})t} + e^{-j(n+\frac{1}{2})t} \right) dt \quad (8.26)$$

which, after some trivial integration, gives:

$$F_n = \frac{1}{4\pi} \left[\frac{-2e^{-j(2n-1)\frac{t}{2}}}{j(2n-1)} - \frac{2e^{-j(2n+1)\frac{t}{2}}}{j(2n+1)} \right]_{-\pi}^{\pi} \quad (8.27)$$

$$= \frac{j}{2\pi} \left[\left(\frac{e^{-jn\pi} e^{j\frac{\pi}{2}}}{2n-1} + \frac{e^{-jn\pi} e^{-j\frac{\pi}{2}}}{2n+1} \right) - \left(\frac{e^{jn\pi} e^{-j\frac{\pi}{2}}}{2n-1} + \frac{e^{jn\pi} e^{j\frac{\pi}{2}}}{2n+1} \right) \right] \quad (8.28)$$

Noting that $e^{\pm j\frac{\pi}{2}} = \pm j$, and $e^{\pm jn\pi} = \cos n\pi = (-1)^n$, then it follows that:

$$F_n = \frac{j}{2\pi} \left(\frac{j}{2n-1} - \frac{j}{2n+1} + \frac{j}{2n-1} - \frac{j}{2n+1} \right) (-1)^n \quad (8.29)$$

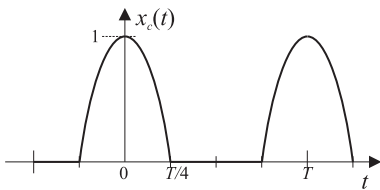
$$= \frac{(-1)^n}{\pi} \left(\frac{1}{2n+1} - \frac{1}{2n-1} \right) = \frac{2(-1)^{n+1}}{(4n^2-1)\pi} \quad (8.30)$$

Note that in this case, the coefficients F_n are real. This is expected, since the function $f(t)$ is an even function of t . The complex Fourier series expansion for $f(t)$ is therefore:

$$f(t) = \sum_{n=-\infty}^{\infty} \frac{2(-1)^{n+1}}{(4n^2-1)\pi} e^{jnt} \quad (8.31) \quad \square$$

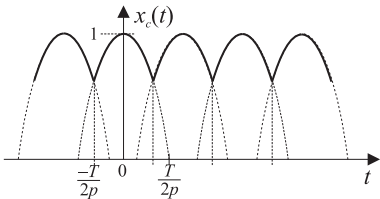
8.3.1.1 Common Fourier Series Expansions

In the following Fourier series expansions, $\omega_0 = \frac{2\pi}{T}$ is the fundamental frequency.



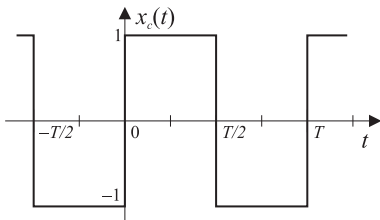
Half-wave rectified cosine-wave:

$$x_c(t) = \frac{1}{\pi} + \frac{1}{2} \cos \omega_0 t + \frac{2}{\pi} \sum_{n=1}^{\infty} (-1)^{n+1} \frac{\cos(2n\omega_0 t)}{4n^2 - 1}$$



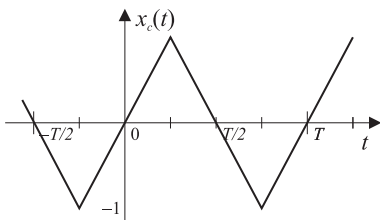
p-phase rectified cosine-wave ($p \geq 2$):

$$x_c(t) = \frac{p}{\pi} \sin \frac{\pi}{p} \left[1 + 2 \sum_{n=1}^{\infty} (-1)^{n+1} \frac{\cos(pn\omega_0 t)}{p^2 n^2 - 1} \right]$$



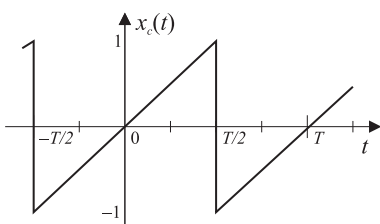
Square wave:

$$x_c(t) = \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\sin(2n-1)\omega_0 t}{2n-1}$$



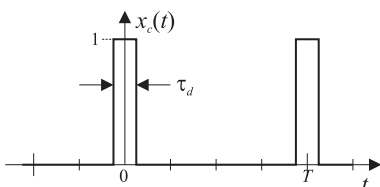
Triangular wave:

$$x_c(t) = \frac{8}{\pi^2} \sum_{n=1}^{\infty} (-1)^{n+1} \frac{\sin(2n-1)\omega_0 t}{(2n-1)^2}$$



Sawtooth wave:

$$x_c(t) = \frac{2}{\pi} \sum_{n=1}^{\infty} (-1)^{n+1} \frac{\sin n\omega_0 t}{n}$$



Pulse wave:

$$x_c(t) = \frac{\tau_d}{T} \left[1 + 2 \sum_{n=1}^{\infty} \frac{\sin(n\pi \frac{\tau_d}{T})}{(n\pi \frac{\tau_d}{T})} \cos(n\omega_0 t) \right]$$

8.3.1.2 Dirichlet Conditions

An important issue that arises in the representation of the continuous time periodic signal $x_c(t)$ by the Fourier series representation,

$$\bar{x}_c(t) = \sum_{k=-\infty}^{\infty} \check{X}_c(k) e^{jk\omega_0 t}, \quad (\text{P:4.1.5})$$

is whether or not the series converges for every value of $t \in \mathbb{R}$; i.e., is it true that

$$\bar{x}_c(t) \stackrel{?}{=} x_c(t), \quad \forall t \in \mathbb{R} \quad (8.32)$$

The so-called **Dirichlet conditions** guarantee that the Fourier series converges everywhere except at points of discontinuity. At these points, the Fourier series representation $\bar{x}_c(t)$ converges to the midpoint, or average value, of the discontinuity.

The **Dirichlet conditions** require that the signal $x_c(t)$:

1. has a finite number of discontinuities in any period;
2. contains a finite number of maxima and minima during any period;
3. is **absolutely integrable** in any period; that is:

$$\int_{T_p} |x_c(t)| dt < \infty \quad (\text{P:4.1.6})$$

where the integral is over one period. Many periodic signals of practical interest easily satisfy these conditions, and it is reasonable to assume that all practical periodic signals do. However, it is important to beware that pathological cases can make certain proofs or algorithms collapse.

8.3.1.3 Parseval's Theorem (for Fourier series)

It is sometimes relevant to consider the **energy** or **power** as a means for characterising a signal. These concepts of **power** and **energy** intuitively follow from their use in other aspects of the physical sciences. However, the concept of signals which exist for all time requires careful definitions for when it has **energy** and when it has **power**. Consider the following signal classifications:

Energy Signals A signal $x_c(t)$ is said to be an **energy signal** if the total energy, E , dissipated by the signal over all time is both *nonzero* and *finite*. Thus:

$$0 < E < \infty \quad \text{where} \quad E = \int_{-\infty}^{\infty} |x_c(t)|^2 dt \quad (8.33)$$

Power signals If the average power delivered by the signal over all time is both *nonzero* and *finite*, the signal is classified as a power signal:

$$0 < P < \infty \quad \text{where} \quad P = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T |x_c(t)|^2 dt \quad (8.34)$$

A periodic signal has infinite energy, but finite average power. The average power of $x_c(t)$ is given by **Parseval's theorem**:

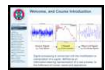
$$P_x = \frac{1}{T_p} \int_0^{T_p} |x_c(t)|^2 dt = \sum_{k=-\infty}^{\infty} |\check{X}_c(k)|^2 \quad (\text{M:2.2.3})$$

The term $|\check{X}_c(k)|^2$ represents the power in the k th frequency component, at frequency $\omega_k = k \frac{2\pi}{T_p}$. Hence,

$$\check{P}_x(k) = |\check{X}_c(k)|^2, \quad -\infty < k < \infty, k \in \mathbb{Z} \quad (8.35)$$

is called the **power spectrum** of $x_c(t)$. Consequently, it follows P_x may also be written as:

$$P_x = \sum_{k=-\infty}^{\infty} \check{P}_x(k) \quad (8.36)$$



New slide

PROOF. Starting with

$$P_x = \frac{1}{T_p} \int_0^{T_p} x_c(t) x_c^*(t) dt \quad (8.37)$$

then substituting for the Fourier series expansion of $x_c(t)$ gives:

$$P_x = \frac{1}{T_p} \int_0^{T_p} x_c(t) \left\{ \sum_{k=-\infty}^{\infty} \check{X}_c(k) e^{jk\omega_0 t} \right\}^* dt \quad (8.38)$$

Noting that the conjugate of a summation (multiplication) is the summation (multiplication) of the conjugates, then:

$$P_x = \frac{1}{T_p} \int_0^{T_p} x_c(t) \sum_{k=-\infty}^{\infty} \check{X}_c^*(k) e^{-jk\omega_0 t} dt \quad (8.39)$$

Rearranging the order of the integration and the summation gives:

$$P_x = \sum_{k=-\infty}^{\infty} \check{X}_c^*(k) \underbrace{\left\{ \frac{1}{T_p} \int_0^{T_p} x_c(t) e^{-jk\omega_0 t} dt \right\}}_{X_c(k)} \quad (8.40) \quad \square$$

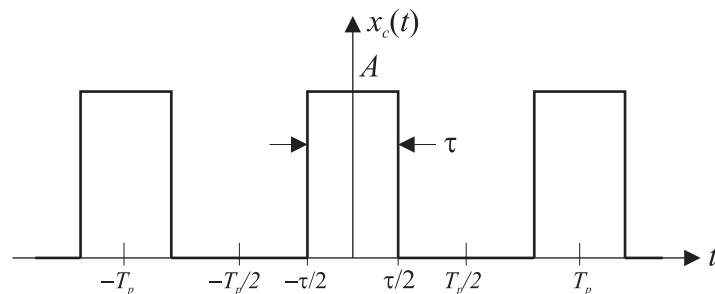
which is the desired result and completes the proof.

Later in this course, the notion of a **power spectrum** will be extended to *stochastic* signals.

Example 8.6 ([Proakis:1996, Example 4.1.1, Page 237]). Determine the Fourier series and the power density spectrum of a rectangular pulse train that is defined over *one* period as follows:

$$x_c(t) = \begin{cases} 0 & \text{if } -\frac{T_p}{2} \leq t < -\frac{\tau}{2} \\ A & \text{if } -\frac{\tau}{2} \leq t < \frac{\tau}{2} \\ 0 & \text{if } \frac{\tau}{2} \leq t < \frac{T_p}{2} \end{cases} \quad (8.41)$$

where $\tau < T_p$.



SOLUTION. The signal is periodic with fundamental period T_p and, clearly, satisfies the Dirichlet conditions. Consequently, this signal can be represented by the Fourier series. Hence, it follows that

$$\check{X}_c(k) = \frac{1}{T_p} \int_{-\frac{T_p}{2}}^{\frac{T_p}{2}} x_c(t) e^{-jk\omega_0 t} dt = \frac{A}{T_p} \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} e^{-jk\omega_0 t} dt \quad (8.42)$$

Two different integrals need to be performed depending on whether $k = 0$ or not. Considering the case when $k = 0$, then the average value of the signal is obtained and given by:

$$\check{X}_c(0) = \frac{1}{T_p} \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} x_c(t) dt = \frac{1}{T_p} \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} A dt = \frac{A\tau}{T_p} \quad (8.43)$$

For $k \neq 0$, then

$$\check{X}_c(k) = \frac{A}{T_p} \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} e^{-jk\omega_0 t} dt = \frac{A}{T_p} \left[\frac{e^{-jk\omega_0 t}}{-jk\omega_0} \right]_{-\frac{\tau}{2}}^{\frac{\tau}{2}} \quad (8.44)$$

$$= \frac{A}{jk\omega_0 T_p} (e^{jk\omega_0 \frac{\tau}{2}} - e^{-jk\omega_0 \frac{\tau}{2}}) = \frac{A\tau \sin \frac{\tau\omega_0 k}{2}}{T_p k\omega_0 \frac{\tau}{2}} \quad (8.45)$$

$$= \frac{A\tau}{T_p} \operatorname{sinc} \frac{\tau\omega_0 k}{2} \quad \text{where } \operatorname{sinc} x \triangleq \frac{\sin x}{x} \quad (8.46)$$

Hence, the power density spectrum for the rectangular pulse is:

$$|\check{X}_c(k)|^2 = \left(\frac{A\tau}{T_p} \right)^2 \operatorname{sinc}^2 \frac{\tau\omega_0 k}{2}, \quad k \in \mathbb{Z} \quad (\text{P:4.1.19}) \quad \square$$

where it is noted that $\operatorname{sinc}(0) = 1$.

8.3.2 Fourier transform

An **aperiodic continuous-time** deterministic signal, $x_c(t)$, can be expressed in the frequency domain using the **Fourier transform** pairs:

$$x_c(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X_c(\omega) e^{j\omega t} d\omega \quad (\text{M:2.2.5})$$

and

$$X_c(\omega) = \int_{-\infty}^{\infty} x_c(t) e^{-j\omega t} dt \quad (\text{M:2.2.4})$$

$X_c(\omega)$ is called the **spectrum** of $x_c(t)$. Again, note that [Manolakis:2000] uses the definition $F = \frac{\omega}{2\pi}$. Continuous-time aperiodic signals have continuous aperiodic spectra.

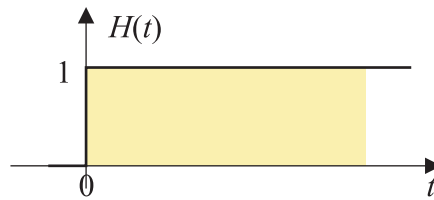
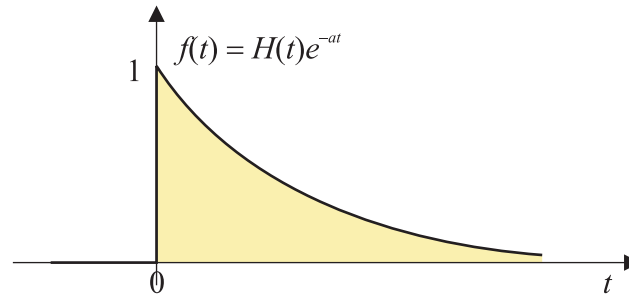
There are a few mathematical requirements that $x_c(t)$ must satisfy for $X_c(\omega)$ to exist; these can be summarised by the phrase that *the signal must be well-behaved*. More specifically, the set of conditions that guarantee the existence of the Fourier transform are the Dirichlet conditions which are the same as for Fourier series.

Example 8.7 (Fourier Transforms). Find the Fourier transform of the one-sided exponential function

$$f(t) = H(t) e^{-at} \quad \text{where } a > 0 \quad (8.47)$$

and where $H(t)$ is the Heaviside unit step function show in Figure 8.12 and given by:

$$H(t) = \begin{cases} 1 & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (8.48)$$

Figure 8.12: The Heaviside step function $H(t)$.Figure 8.13: Exponential decaying function, $f(t) = H(t)e^{-at}$ for $a > 0$.

SOLUTION. Since $f(t) \rightarrow 0$ as $t \rightarrow \infty$, then the signal energy is bounded, as indicated by plotting the graph of $f(t)$ as shown in Figure 8.13.

A Fourier transform therefore exists, and is given by:

$$X_c(\omega) = \int_{-\infty}^{\infty} H(t) e^{-at} e^{-j\omega t} dt \quad (8.49)$$

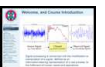
$$= \int_0^{\infty} e^{-(a+j\omega)t} dt = \left[-\frac{e^{-(a+j\omega)t}}{a+j\omega} \right]_0^{\infty} \quad (8.50)$$

giving

$$X_c(\omega) = \frac{1}{a+j\omega}, \quad \text{for } -\infty < \omega < \infty \quad (8.51)$$

□

8.3.2.1 Parseval's theorem (for Fourier transforms)



The *energy* of $x_c(t)$ is, as for **Fourier series**, computed in either the time or frequency domain by *New slide*
Parseval's theorem:

$$E_x = \int_{-\infty}^{\infty} |x_c(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |X_c(\omega)|^2 d\omega \quad (\text{M:2.2.7})$$

The function $|X_c(\omega)|^2 \geq 0$ shows the distribution of energy of $x_c(t)$ as a function of frequency, ω , and is called the **energy spectrum** of $x_c(t)$.

PROOF. The derivation of Parseval's theorem for Fourier transforms follows a similar line to the derivation of Parseval's theorem for Fourier series; it proceeds as follows:

$$\begin{aligned} E_x &= \int_{-\infty}^{\infty} x_c(t) x_c^*(t) dt = \int_{-\infty}^{\infty} x_c(t) \frac{1}{2\pi} \int_{-\infty}^{\infty} X_c^*(\omega) e^{-j\omega t} d\omega dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} X_c^*(\omega) \int_{-\infty}^{\infty} x_c(t) e^{-j\omega t} dt d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} X_c^*(\omega) X_c(\omega) d\omega \end{aligned} \quad (8.52)$$

□

– End-of-Topic 53: **Revision of the Continuous-Time Fourier Analysis:
the Complex Fourier Series and the Fourier Transform** –



8.3.3 The discrete-time Fourier transform

Topic Summary 54 Fourier Transform Theory

Topic Objectives:

- Understanding how to measure the size (or norm) of a signal.
- Motivation for Energy and Power.

Topic Activities:

Type	Details	Duration	Progress
Read Handout	Read page 297 to page 301	8 mins/page	

Turning to discrete-time deterministic signals, the natural starting point is to consider aperiodic signals that exist for all discrete-time; i.e. $\{x[n]\}_{-\infty}^{\infty}$. It is interesting to note that there are fewer convergence issues with transforms for discrete-time signals than there are in the continuous-time case.

An **aperiodic discrete-time** deterministic signal, $\{x[n]\}_{-\infty}^{\infty}$, can be synthesised from its **spectrum** using the inverse-discrete-time Fourier transform, given by:

$$x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega T}) e^{j\omega n} d\omega, \quad n \in \mathbb{Z} \quad (\text{M:2.2.13})$$

and the discrete-time Fourier transform (DTFT):

$$X(e^{j\omega T}) = \sum_{\text{all } n} x[n] e^{-j\omega n}, \quad \omega \in \mathbb{R} \quad (\text{M:2.2.12})$$

$X(e^{j\omega T})$ is the **spectrum** of $x[n]$.

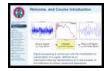
Since $X(e^{j\omega T}) = X(e^{j(\omega+2\pi k)T})$, discrete-time aperiodic signals have continuous periodic spectra with **fundamental period** 2π . However, this property is just a consequence of the fact that the frequency range of any discrete-time signal is limited to $[-\pi, \pi)$ or $[0, 2\pi)$; any frequency outside this interval is equivalent to some frequency within this interval.

There are two basic differences between the Fourier transform of a discrete-time finite-energy aperiodic signal, as represented by the discrete-time Fourier transform, and the Fourier transform of a finite-energy continuous-time aperiodic signal:

1. For continuous-time signals, the Fourier transform, and hence the spectrum of the signal, have a frequency range of $(-\infty, \infty)$. In contrast, the frequency range for a discrete-time signal is unique over the frequency range $[-\pi, \pi)$ or, equivalently, $[0, 2\pi)$.
2. Since $X(e^{j\omega T})$ in the DTFT is a periodic function of frequency, it has a Fourier series expansion, provided that the conditions for the existence of the Fourier series are satisfied. In fact, from the fact that $X(e^{j\omega T})$ is given by the summation of exponentially weighted versions of $x[n]$ it is clear that the DTFT already has the form of a Fourier series. This is not true for the Fourier transform.

In order for $X(e^{j\omega T})$ to exist, $x[n]$ must be absolutely summable:

$$\sum_{\text{all } n} |x[n]| < \infty \quad (\text{M:2.2.11})$$



Finally, as for the Fourier series, and the Fourier transform, discrete-time aperiodic signals have energy which satisfies Parseval's theorem:

$$E_x = \sum_{n=-\infty}^{\infty} |x[n]|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |X(e^{j\omega T})|^2 d\omega \quad (\text{P:4.2.41})$$

8.3.4 Discrete Fourier transform

Any finite-length or **periodic discrete-time** deterministic signal, $\{x[n]\}_0^{N-1}$, can be written by the Fourier series, or inverse-DFT (IDFT):

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{j\frac{2\pi}{N}nk}, \quad n \in \mathcal{N} \quad (\text{M:2.2.8})$$

where $\mathcal{N} = \{0, 1, \dots, N-1\} \subset \mathbb{Z}^+$, and where the DFT:

$$X_k = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}nk}, \quad k \in \mathcal{N} \quad (\text{M:2.2.9})$$

are the corresponding Fourier coefficients. The sequence X_k , $k \in \mathbb{R}$ is the **spectrum** of $x[n]$. X_k is discrete and periodic with the same period as $x[n]$.

Note that a finite-length discrete-time signal of length N has the same Fourier transform, through the DFT, as an infinite-length discrete-time periodic signal of period N . Hence, these equivalent perspectives will be considered synonymous.

PROOF (DERIVATION OF DISCRETE FOURIER TRANSFORM). If the **discrete-time** signal $x[n]$ is **periodic** over N samples, then it can be written over one period in continuous time as:

$$x_c(t) = T_p \sum_{n \in \mathcal{N}} x[n] \delta(t - nT_s), \quad 0 \leq t < T_p \quad (8.53)$$

where $\mathcal{N} = \{0, \dots, N-1\}$, T_s is the sampling period, and $T_p = NT_s$ is the period of the process. Substituting into the expression for the **Fourier series**, using the **sifting property** and noting that $\omega_0 = \frac{2\pi}{T_p} = \frac{2\pi}{NT_s}$, gives:

$$X_k = \frac{1}{T_p} \int_0^{T_p} x_c(t) e^{-jk\omega_0 t} dt \quad (8.54)$$

$$= \frac{1}{T_p} \int_0^{T_p} \left\{ T_p \sum_{n \in \mathcal{N}} x[n] \delta(t - nT_s) \right\} e^{-jk\omega_0 t} dt \quad (8.55)$$

$$= \sum_{n \in \mathcal{N}} x[n] \int_0^{T_p} \delta(t - nT_s) e^{-jk\omega_0 t} dt \quad (8.56)$$

$$= \sum_{n \in \mathcal{N}} x[n] e^{-j\frac{2\pi}{N}nk} \quad (8.57)$$

□

The IDFT can be obtained using the appropriate identities to ensure a unique inverse.

8.3.4.1 Parseval's Theorem for Finite Length Discrete-Time Signals

The average power of a finite length or periodic discrete-time signal with period N is given by the sum of squared sample values:

$$P_x = \sum_{n=0}^{N-1} |x[n]|^2 \quad (\text{P:4.2.10})$$

Through the same manipulations as for Parseval's theorems in the cases presented above, which are left as an exercise for the reader, it is straightforward to show that:

$$P_x = \sum_{n=0}^{N-1} |x[n]|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |X_k|^2 \quad (\text{P:4.2.11})$$

8.3.4.2 The DFT as a Linear Transformation

The formulas for the DFT and IDFT may be expressed as:

$$X_k = \sum_{n=0}^{N-1} x[n] W_N^{nk}, \quad k \in \mathcal{N} \quad (\text{P:5.1.20})$$

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X_k W_N^{-nk}, \quad n \in \mathcal{N} \quad (\text{P:5.1.21})$$

where, by definition:

$$W_N = e^{-j\frac{2\pi}{N}} \quad (\text{P:5.1.22})$$

which is the N th root of unity. Note here that, if W_N has been pre-calculated, then the computation of each point of the DFT can be accomplished by N complex multiplications and $N - 1$ complex additions. Hence, the N -point *DFT* can be computed in a total of N^2 complex multiplications and $N(N - 1)$ complex additions.

It is instructive to view the DFT and IDFT as linear transformations on the sequences $\{x[n]\}_0^{N-1}$ and $\{X_k\}_0^{N-1}$. Defining the following vectors and matrices:

$$\mathbf{x}_N = \begin{bmatrix} x[0] \\ \vdots \\ x[N-1] \end{bmatrix}, \quad \mathbf{X}_N = \begin{bmatrix} X_0 \\ \vdots \\ X_{N-1} \end{bmatrix} \quad (8.58)$$

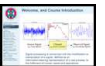
$$\mathbf{W}_N = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & W_N & W_N^2 & \cdots & W_N^{N-1} \\ 1 & W_N^2 & W_N^4 & \cdots & W_N^{2(N-1)} \\ \vdots & \vdots & \vdots & \cdot & \vdots \\ 1 & W_N^{N-1} & W_N^{2(N-1)} & \cdots & W_N^{(N-1)(N-1)} \end{bmatrix} \quad (8.59)$$

Observe that X_k can be obtained by the inner-product of the $(k - 1)$ th-order row with the column \mathbf{x}_N :

$$X_k = \begin{bmatrix} 1 & W_N^k & W_N^{2k} & \cdots & W_N^{(N-1)k} \end{bmatrix} \begin{bmatrix} x[0] \\ \vdots \\ x[N-1] \end{bmatrix} \quad (8.60)$$

Then the N -point DFT may be expressed in vector-matrix form as:

$$\mathbf{X}_N = \mathbf{W}_N \mathbf{x}_N \quad (\text{P:5.1.24})$$



New slide

where \mathbf{W}_N is the matrix of the linear transformation. Observe that \mathbf{W}_N is a symmetric matrix. Assuming that the inverse of \mathbf{W}_N exists, then Equation P:5.1.24 can be inverted by pre-multiplying both sides by \mathbf{W}_N^{-1} , to obtain:

$$\mathbf{x}_N = \mathbf{W}_N^{-1} \mathbf{X}_N \quad (\text{P:5.1.25})$$

This is the expression for the IDFT, which can also be expressed in matrix form as:

$$\mathbf{x}_N = \frac{1}{N} \mathbf{W}_N^* \mathbf{X}_N \quad (\text{P:5.1.26})$$

where \mathbf{W}_N^* denotes the complex conjugate of the matrix \mathbf{W}_N . Hence, it follows that:

$$\mathbf{W}_N^{-1} = \frac{1}{N} \mathbf{W}_N^* \quad \text{or} \quad \mathbf{W}_N \mathbf{W}_N^* = N \mathbf{I}_N \quad (\text{P:5.1.27})$$

where \mathbf{I}_N is the $N \times N$ identity matrix. Hence, \mathbf{W}_N is an orthogonal or unity matrix.

8.3.4.3 Properties of the discrete Fourier transforms

There are some important basic properties of the DFT that should be noted. The notation used to denote the N -point DFT pair $x[n]$ and X_k is

$$x[n] \stackrel{\text{DFT}}{\rightleftharpoons} X_k \quad (8.61)$$

Periodicity If $x[n] \stackrel{\text{DFT}}{\rightleftharpoons} X_k$, then:

$$x[n + N] = x[n] \quad \text{for all } n \quad (\text{P:5.2.4})$$

$$X_{k+N} = X_k \quad \text{for all } k \quad (\text{P:5.2.5})$$

These periodicities in $x[n]$ and X_k follow immediately from the definitions of the DFT and IDFT.

Linearity If $x[n] \stackrel{\text{DFT}}{\rightleftharpoons} X_k$ and $y[n] \stackrel{\text{DFT}}{\rightleftharpoons} Y_k$, then

$$\alpha_1 x[n] + \alpha_2 y[n] \stackrel{\text{DFT}}{\rightleftharpoons} \alpha_1 X_k + \alpha_2 Y_k \quad (\text{P:5.2.6})$$

for any real or complex-valued constants α_1 and α_2 .

Symmetry of real-valued sequences If the sequence $x[n] \stackrel{\text{DFT}}{\rightleftharpoons} X_k$ is real, then

$$X_{N-k} = X_k^* = X_{-k} \quad (\text{P:5.2.24})$$

Complex-conjugate properties If $x[n] \stackrel{\text{DFT}}{\rightleftharpoons} X_k$ then

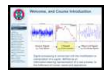
$$x^*[n] \stackrel{\text{DFT}}{\rightleftharpoons} X_{N-k}^* \quad (\text{P:5.2.45})$$

PROOF. The DFT of the sequence $x[n]$ is given by:

$$X_k = \sum_{n=0}^{N-1} x[n] e^{-j \frac{2\pi}{N} nk}, \quad k \in \mathcal{N} \quad (\text{M:2.2.9})$$

and the DFT of $y[n] = x^*[n]$ is given by:

$$Y_k = \sum_{n=0}^{N-1} x^*[n] e^{-j \frac{2\pi}{N} nk} \quad (8.62)$$



New slide

Taking complex conjugates, and noting that $e^{j\frac{2\pi}{N}mk} = e^{-j\frac{2\pi}{N}m(N-k)}$, then:

$$Y_k^* = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}m(N-k)} = X_{N-k} \quad (8.63) \quad \square$$

Hence, giving $x^*[n] \stackrel{\text{DFT}}{\rightleftharpoons} X_{N-k}^*$ as required.

Time reversal of a sequence If $x[n] \stackrel{\text{DFT}}{\rightleftharpoons} X_k$ then

$$x[N-n] \stackrel{\text{DFT}}{\rightleftharpoons} X_{N-k} \quad (\text{P:5.2.42})$$

Hence, reversing the N -point sequence in time is equivalent to reversing the DFT values in frequency.

PROOF. From the definition of the DFT, if $y[n] = x[N-n]$, then:

$$Y_k = \sum_{n=0}^{N-1} x[N-n] e^{-j\frac{2\pi}{N}nk} = \sum_{m=1}^N x[m] e^{-j\frac{2\pi}{N}(N-m)k} \quad (8.64)$$

where the second summation comes from changing the index from n to $m = N - n$. Noting then, that $x[N] = x[0]$, then this may be written as

$$Y_k = \sum_{m=0}^{N-1} x[m] e^{-j\frac{2\pi}{N}(N-m)k} = \sum_{m=0}^{N-1} x[m] e^{j\frac{2\pi}{N}mk} \quad (8.65)$$

$$= \sum_{m=0}^{N-1} x[m] e^{-j\frac{2\pi}{N}m(N-k)} = X_{N-k} \quad (8.66) \quad \square$$

as required.

Circular Convolution As with many linear transforms, convolution in the time-domain becomes multiplication in the frequency domain, and vice-versa. Since the signals are periodic, it is necessary to introduce the idea of circular convolution. Details of this are discussed in depth in [Proakis:1996, Section 5.2.2, Page 415] and are currently omitted here. However, assuming that convolution is interpreted in the circular sense (i.e. taking advantage of the periodicity of the time-domain signals), then if $x[n] \stackrel{\text{DFT}}{\rightleftharpoons} X_k$ and $y[n] \stackrel{\text{DFT}}{\rightleftharpoons} Y_k$, then:

$$x[n] * y[n] \stackrel{\text{DFT}}{\rightleftharpoons} X_k Y_k \quad (\text{P:5.2.41})$$



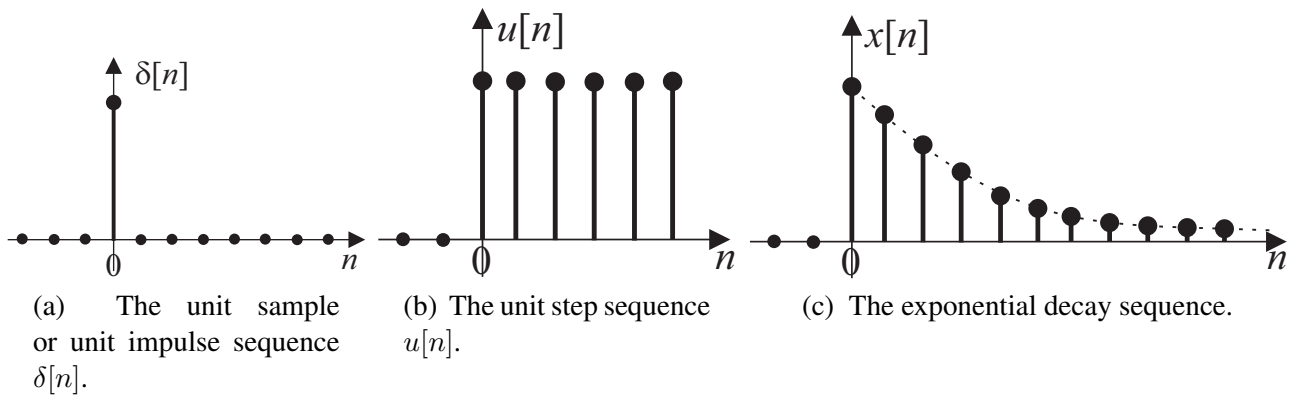
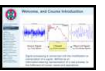


Figure 8.14: Basic discrete-time signals.

8.4 Review of discrete-time systems



Topic Summary 55 Fourier Transform Theory

New slide

Topic Objectives:

- Understanding how to measure the size (or norm) of a signal.
- Motivation for Energy and Power.

Topic Activities:

Type	Details	Duration	Progress
Read Handout	Read page 302 to page 311	8 mins/page	
Try Example	Try Examples 8.8 and 8.9	40 mins	

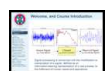
The following aspects of **discrete-time systems** are reviewed:

- Basic discrete-time signals
- The z -transform
- Review of **linear time-invariant** systems
- Rational **transfer functions**

8.4.1 Basic discrete-time signals

In general, the notation $x[n]$ is used to denote a sequence of numbers that represent a discrete-time signal. The n th sample refers to the value of this sequence for a specific value of n . In a strict sense, this terminology is only correct if the discrete-time signal has been obtained by sampling a continuous-time signal $x_c(t)$. In the case of periodic sampling with sampling period T , then $x[n] = x_c(nT)$, $n \in \mathbb{Z}$; that is, $x[n]$ is the n th sample of $x_c(t)$.

There are some basic discrete-time signals that will be used repeatedly throughout the course, and these are shown in Figure 8.14:



New slide

1. The **unit sample** or **unit impulse** sequence $\delta[n]$ is defined as:

$$\delta[n] = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases} \quad (\text{M:2.1.1})$$

2. The **unit step** sequence, $u[n]$ is defined as:

$$u[n] = \begin{cases} 1 & n \geq 0 \\ 0 & n < 0 \end{cases} \quad (\text{M:2.1.2})$$

3. The **exponential sequence** is of the form

$$x[n] = a^n, \quad -\infty < n < \infty, n \in \mathbb{Z} \quad (\text{M:2.1.3})$$

If a is a complex number, such that $a = r e^{j\omega_0}$ for $r > 0$, $\omega_0 \neq 0, \pi$, then $x[n]$ is complex valued and given by:

$$\begin{aligned} x[n] &= r^n e^{j\omega_0 n} = x_R[n] + jx_I[n] & (\text{M:2.1.4}) \\ &= r^n \cos \omega_0 n + jr^n \sin \omega_0 n & (8.67) \end{aligned}$$

where $x_R[n]$ and $x_I[n]$ are real sequences given by:

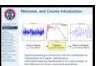
$$x_R[n] = r^n \cos \omega_0 n \quad \text{and} \quad x_I[n] = r^n \sin \omega_0 n \quad (\text{M:2.1.5})$$

4. The **critical decay sequence** is of the form

$$x[n] = a n r^n u[n], \quad n \in \mathbb{Z} \quad (8.68)$$

which is discussed further in Sidebar 21.

8.4.2 The z -transform



New slide

The z -transform of a sequence is a very powerful tool for the analysis of discrete linear and time-invariant systems; it plays the same role in the analysis of discrete-time signals and linear time-invariant (LTI) systems as the Laplace transform does in the analysis of continuous-time signals and LTI systems. For example, as will be seen, in the z -domain, also known as the complex z -plane, the convolution of two time-domain signals is equivalent to multiplication of their corresponding z -transforms. This property greatly simplifies the analysis of the response of an LTI system to various inputs.

Although the Fourier transform also satisfies the property that convolution in the time domain becomes multiplication in the frequency domain, it is not always possible to calculate the Fourier transform of a signal, $x[n]$, even for some elementary signals that are important for the analysis of systems. For example, if $x[n]$ is a **power signal** (having finite power), rather than an **energy signal**, the discrete-time Fourier transform (DTFT) does not exist.

One such signal, of practical importance, is the unit step function, $u[t]$, as can be illustrated by attempting to evaluate the DTFT:

$$U(e^{j\omega T}) = \sum_{n=-\infty}^{\infty} u[n] e^{-j\omega n} = \sum_{n=0}^{\infty} e^{-j\omega n} \quad (8.72)$$

Sidebar 21 The signal $n r^n$

The discrete-time signal

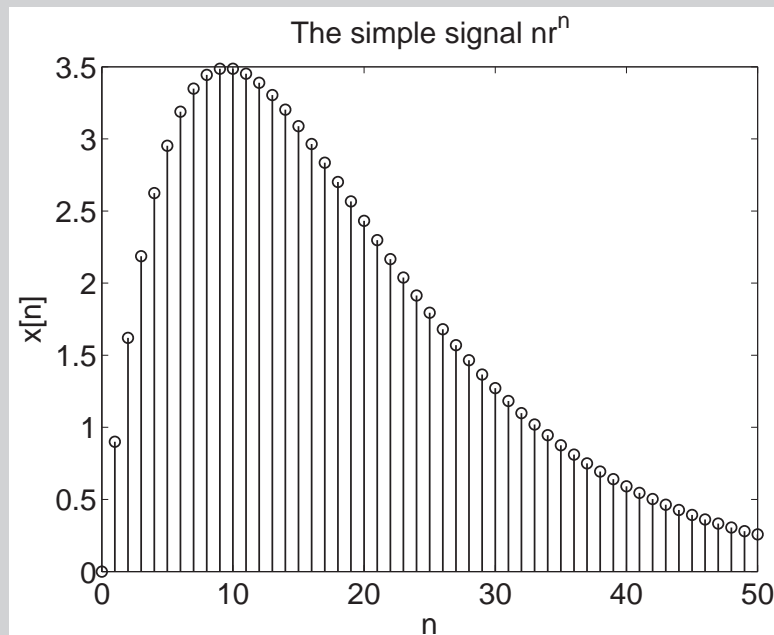
$$x[n] = a n r^n \quad (8.69)$$

is equivalent to the continuous-time signal $x[t] = t e^{-\alpha t}$, and both are important, as they represent the response of a **critically damped system**, as will be seen later. Note in both cases that:

$$\lim_{n \rightarrow \infty} n r^n \rightarrow 0 \quad (8.70)$$

The shape of $x[n]$ is shown below for $r = 0.9$, and note the relationship derived in Sidebar 22 that:

$$n r^n \stackrel{z^+}{\rightleftharpoons} \frac{r}{(1-r)^2} \quad \text{if } |r| < 1 \quad (8.71)$$

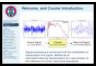


This is a geometric series, of the form $\sum_{n=0}^{\infty} r^n$ where $r = e^{-j\omega}$; however, this series *diverges* since $|r| = 1$. Therefore, the DTFT does not exist; this could also have been deduced from the fact that $u[n]$ is not absolutely summable, which a necessary condition for a Fourier transform to exist:

$$\sum_{\text{all } n} |u[n]| = \sum_{n=0}^{\infty} 1 \not< \infty \quad (8.73)$$

The solution is to multiply the signal by a convergence factor, which leads to the z -transform. Details of the derivation can be found in some text books.

8.4.2.1 Bilateral z -transform



New slide

The bilateral z -transform is defined by the following pairs of equations:

$$X(z) \triangleq \mathcal{Z}[x[n]] = \sum_{n=-\infty}^{\infty} x[n] z^{-n} \quad (\text{M:2.2.29})$$

$$x[n] = \frac{1}{2\pi j} \oint_C X(z) z^{n-1} dz \quad (\text{M:2.2.30})$$

where z is a complex variable. This is usually denoted as:

$$x[n] \stackrel{z}{\rightleftharpoons} X(z) \quad \text{or} \quad X(z) = \mathcal{Z}[x[n]] \quad (8.74)$$

The set of values of z for which the power series in the (direct) z -transform converges is called the region of convergence (ROC) of $X(z)$. A sufficient condition for convergence is:

$$\sum_{n=-\infty}^{\infty} |x[n]| |z^{-n}| < \infty \quad (\text{M:2.2.31})$$

The unilateral or one-sided z -transform, which is more commonly encountered in undergraduate Engineering courses, is discussed below in Section 8.4.2.3. For the moment, it suffices to mention that the difference between them usually comes down to the initial conditions, and therefore a discussion of the bilateral transform is not too restrictive at this point.

By evaluating the z -transform on the unit circle of the z -plane, such that $z = e^{j\omega}$, then:

$$X(z)|_{z=e^{j\omega}} = X(e^{j\omega T}) = \sum_{n=-\infty}^{\infty} x[n] e^{-j\omega n} \quad (\text{M:2.2.32})$$

$$x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega T}) e^{j\omega n} d\omega \quad (\text{M:2.2.33})$$

which are the DTFT and inverse-DTFT relating the signals $x[n]$ and $X(e^{j\omega T})$. This relation holds only if the unit circle is inside the ROC.

Example 8.8 ([Proakis:1996, Example 3.1.3, Page 154]). Determine the z -transform of the signal:

$$x[n] = \alpha^n u[n] \equiv \begin{cases} \alpha^n & n \geq 0 \\ 0 & n < 0 \end{cases} \quad (8.80)$$

Sidebar 22 The Ubiquitous Geometric Progression

The **geometric progression** occurs frequently in discrete-time analysis due to the existence of the summation operator and the commonality of exponential decay functions. It is essentially the discrete-time equivalent of integrating an exponential function. The geometric progression is given by

$$\sum_{n=0}^N a r^n = a \frac{1 - r^{N+1}}{1 - r} \quad (8.75)$$

$$\sum_{n=0}^{\infty} a r^n = a \frac{1}{1 - r} \quad \text{if } |r| < 1 \quad (8.76)$$

More interesting are variants of the geometric progression that can be obtained by simple manipulations, such as differentiating both sides of Equation 8.76 with respect to (w. r. t.) r :

$$\frac{d}{dr} \left[\sum_{n=0}^{\infty} a r^n \right] = \frac{d}{dr} \left[a \frac{1}{1 - r} \right] \quad (8.77)$$

$$\sum_{n=0}^{\infty} a n r^{n-1} = a \frac{1}{(1 - r)^2} \quad (8.78)$$

or, multiplying both sides by r , gives:

$$\sum_{n=0}^{\infty} a n r^n = a \frac{r}{(1 - r)^2} \quad \text{if } |r| < 1 \quad (8.79)$$

which is also a useful identity. The signal $x[n] = n r^n$ is an important one and discussed further in Sidebar 21. Differentiating repeated times gives a general expression for $\sum n^p r^n$ which can often be useful.

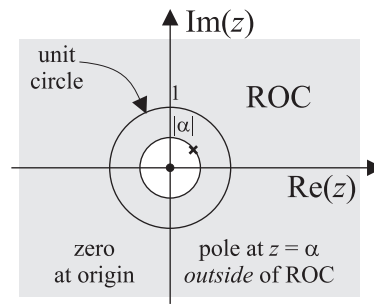


Figure 8.15: The region of convergence (ROC) for the transfer function in Equation P:3.1.7.

SOLUTION. From the definition of the z -transform, it follows that:

$$X(z) = \sum_{k=0}^{\infty} \alpha^n z^{-n} = \sum_{n=0}^{\infty} (\alpha z^{-1})^n \quad (8.81)$$

The summation on the right is a geometric progression, and converges to $\frac{1}{1-\alpha z^{-1}}$ if, and only if, (iff) $|\alpha z^{-1}| < 1$ or, equivalently, $|z| > |\alpha|$. Further details on the geometric progression are given in Sidebar 22. Thus, this gives the z -transform pair:

$$x[n] = \alpha^n u[n] \stackrel{z}{\rightleftharpoons} X(z) = \frac{1}{1-\alpha z^{-1}} \quad \text{ROC: } |z| > |\alpha| \quad (\text{P:3.1.7})$$

Note that, in general, α need not be real. The ROC is the exterior of a circle having radius $|\alpha|$. The ROC is shown in Figure 8.15. The z -transform in Equation P:3.1.7 may be written as:

$$X(z) = \frac{z}{z-\alpha} \quad \text{ROC: } |z| > |\alpha| \quad (8.82)$$

□

and therefore it has a pole at $z = \alpha$ and a zero at $z = 0$. The position of the pole is outside the ROC, which is as expected since the z -transform does not converge at a pole; it tends to infinity instead. However, simply because there is a zero at the origin does not mean the z -transform converges at that point – it doesn't, since it is outside of the ROC. However, the concept of the zero is still important and is thus still drawn on the pole-zero diagram.

Example 8.9 (Two-sided exponential (Laplacian exponential)). What is the bilateral z -transform of the sequence $x[n] = a^{|n|}$ for all n and some real constant a , where $|a| < 1$?

SOLUTION. The bilateral z -transform of a sequence $x[n] = a^{|n|}$, shown in Figure 8.16, is given by:

$$X(z) = \sum_{n=-\infty}^{\infty} x[n] z^{-n} = \sum_{n=-\infty}^{\infty} a^{|n|} z^{-n} \quad (8.83)$$

$$= \sum_{n=-\infty}^{-1} a^{-n} z^{-n} + \sum_{n=0}^{\infty} a^n z^{-n} \quad (8.84)$$

Setting $m = -n$ in the first summation, noting that when $n = -\infty$ then $m = \infty$, and when $n = 0$ then $m = 0$, gives:

$$X(z) = \sum_{n=1}^{\infty} (az)^n + \sum_{n=0}^{\infty} \left(\frac{a}{z}\right)^n \quad (8.85)$$

$$= \sum_{n=0}^{\infty} (az)^n - 1 + \sum_{n=0}^{\infty} \left(\frac{a}{z}\right)^n \quad (8.86)$$

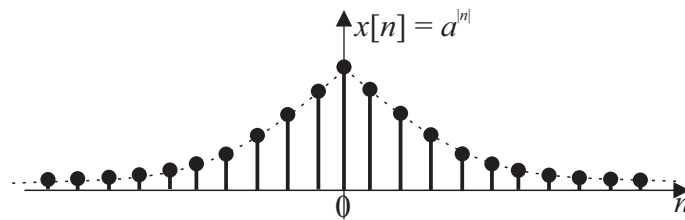


Figure 8.16: The sequence $x[n] = a^{|n|}$.

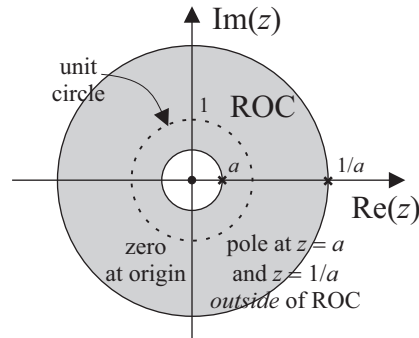


Figure 8.17: The region of convergence (ROC) for the transfer function in Equation 8.88.

giving:

$$X(z) = \frac{1}{1-az} - 1 + \frac{1}{1-\frac{a}{z}} \quad (8.87)$$

where the expression for an infinite geometric progression has been used. Note, however, that each summation has different convergence constraints. Thus, note that the first summation only exists for $|az| < 1$, while the second summation only exists for $|\frac{a}{z}| < 1$. This means that the ROC for this transform is the ring $|a| < z < \frac{1}{|a|}$. The ROC is thus shown in Figure 8.17.

Combining the various terms and a slight rearrangement gives the expression:

$$X(z) = \frac{1-a^2}{(1-az)(1-az^{-1})} \quad (8.88)$$

which has a zero at $z = 0$ and poles at $z = a$ and $z = \frac{1}{a}$.

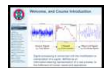
8.4.2.2 Properties of the z -transform

The power of the z -transform is a consequence of some very important properties that the transform possesses. Some of these properties are listed below, as a re-cap. Note that the proof of many of these properties follows immediately from the definition of the property itself and the z -transform, and is left as an exercise for the reader. Alternatively, cheat and look in, for example, [Proakis:1996].

Linearity If $x_1[n] \stackrel{z}{\rightleftharpoons} X_1(z)$ and $x_2[n] \stackrel{z}{\rightleftharpoons} X_2(z)$, then by linearity

$$x[n] = \alpha_1 x_1[n] + \alpha_2 x_2[n] \stackrel{z}{\rightleftharpoons} X(z) = \alpha_1 X_1(z) + \alpha_2 X_2(z) \quad (\text{P:3.2.1})$$

for any constants α_1 and α_2 .



New slide

Obviously, this property can be generalised for an arbitrary number of signals, and therefore if $x_m[n] \stackrel{z}{\rightleftharpoons} X_m(z)$ for $m = \{1, \dots, M\}$

$$x[n] = \sum_{m=1}^M \alpha_m x_m[n] \stackrel{z}{\rightleftharpoons} X(z) = \sum_{m=1}^M \alpha_m X_m(z) \quad (8.89)$$

for any constants $\{\alpha_m\}_1^M$.

Time shifting If $x[n] \stackrel{z}{\rightleftharpoons} X(z)$ then:

$$x[n-k] \stackrel{z}{\rightleftharpoons} z^{-k} X(z) \quad (8.90)$$

The ROC of $z^{-k} X(z)$ is the same as that of $X(z)$ except for $z = 0$ if $k > 0$ and $z = \infty$ if $k < 0$.

Scaling in the z -domain If $x[n] \stackrel{z}{\rightleftharpoons} X(z)$ with ROC $r_1 < |z| < r_2$, then

$$a^n x[n] \stackrel{z}{\rightleftharpoons} X(a^{-1}z) \quad \text{ROC: } |a|r_1 < |z| < |a|r_2 \quad (\text{P:3.2.9})$$

for any constant a .

Time reversal If $x[n] \stackrel{z}{\rightleftharpoons} X(z)$ with ROC $r_1 < |z| < r_2$, then

$$x[-n] \stackrel{z}{\rightleftharpoons} X(z^{-1}) \quad \text{ROC: } \frac{1}{r_1} < |z| < \frac{1}{r_2} \quad (\text{P:3.2.12})$$

Differentiation in the z -domain If $x[n] \stackrel{z}{\rightleftharpoons} X(z)$ then

$$nx[n] \stackrel{z}{\rightleftharpoons} -z \frac{dX(z)}{dz} \quad (\text{P:3.2.14})$$

PROOF. Since

$$X(z) = \sum_{n=-\infty}^{\infty} x[n] z^{-n} \quad (8.91)$$

then differentiating both sides gives:

$$\frac{dX(z)}{dz} = -z^{-1} \sum_{n=-\infty}^{\infty} [nx[n]] z^{-n} = -z^{-1} \mathcal{Z}[nx[n]] \quad (8.92) \quad \square$$

Both transforms have the same ROC.

Convolution If $x_1[n] \stackrel{z}{\rightleftharpoons} X_1(z)$ and $x_2[n] \stackrel{z}{\rightleftharpoons} X_2(z)$, then

$$x[n] = x_1[n] * x_2[n] \stackrel{z}{\rightleftharpoons} X(z) = X_1(z)X_2(z) \quad (3.2.17)$$

The ROC of $X(z)$ is, at least, the intersection of that for $X_1(z)$ and $X_2(z)$.

PROOF. The convolution of $x_1[n]$ and $x_2[n]$ is defined as:

$$x[n] = \sum_{k=-\infty}^{\infty} x_1[k] x_2[n-k] \quad (8.93)$$

The z -transform of $x[n]$ is:

$$X(z) = \sum_{n=-\infty}^{\infty} x[n] z^{-n} = \sum_{n=-\infty}^{\infty} \left[\sum_{k=-\infty}^{\infty} x_1[k] x_2[n-k] \right] z^{-n} \quad (8.94)$$

Upon changing the order of the summations, then:

$$X(z) = \sum_{k=-\infty}^{\infty} x_1[k] \underbrace{\left[\sum_{n=-\infty}^{\infty} x_2[n-k] z^{-n} \right]}_{X_2(z) z^{-k}} = X_2(z) \underbrace{\sum_{k=-\infty}^{\infty} x_1[k] z^{-k}}_{X_1(z)} \quad (8.95) \quad \square$$

giving the desired result.

The Initial Value Theorem If $x[n] = 0$, $n < 0$ is a **causal** sequence, then

$$x[0] = \lim_{z \rightarrow \infty} X(z) \quad (\text{P:3.2.23})$$

PROOF. Since $x[n]$ is causal, then:

$$X(z) = x[0] + x[1] z^{-1} + x[2] z^{-2} + \dots \quad (8.96) \quad \square$$

Hence, as $z \rightarrow \infty$, $z^{-n} \rightarrow 0$ since $n > 0$, and thus the desired result is obtained.

8.4.2.3 The Unilateral z -transform

The two-sided z -transform requires that the corresponding signals be specified for the entire time range $n \in \mathbb{Z}$. This requirement prevents its use for systems that are described by difference equations with nonzero initial conditions. Since the input is applied at a finite time, say n_0 , both input and output signals are specified for $n \geq n_0$, but are not necessarily zero for $n < 0$. Thus the two-sided z -transform cannot be used.

The one-sided **unilateral z -transform** of a signal $x[n]$ is defined by:

$$X^+(z) \equiv \sum_{n=0}^{\infty} x[n] z^{-n} \quad (\text{P:3.5.1})$$

This is usually denoted as:

$$x[n] \stackrel{z^+}{\rightleftharpoons} X^+(z) \quad \text{or} \quad X^+(z) = \mathcal{Z}^+[x[n]] \quad (8.97)$$

The unilateral z -transform differs from the bilateral transform in the lower limit of the summation, which is always zero, whether or not the signal $x[n]$ is zero for $n < 0$ (i.e., causal). Therefore, the unilateral z -transform contains no information about the signal $x[n]$ for negative values of time, and is therefore *unique* only for causal signals. The unilateral and bilateral z -transforms are, consequentially, identical for the signal $x[n] u[n]$ where $u[n]$ is the step function. Since $x[n] u[n]$ is causal, the ROC of its transform, and hence the ROC of $X^+(z)$, is always the exterior of a circle. Thus, when discussing the unilateral z -transform, it is not necessary to refer to their ROC - which perhaps explains why this is the more commonly discussed transform in undergraduate courses.

Almost all the properties for the bilateral z -transform carry over to the unilateral transform with the exception of the shifting property.

Shifting property: Time Delay If $x[n] \stackrel{z^+}{\rightleftharpoons} X^+(z)$ then:

$$x[n-k] \stackrel{z^+}{\rightleftharpoons} z^{-k} X^+(z) + \underbrace{\sum_{n=-k}^{-1} x[n] z^{-(n+k)}}_{\text{initial conditions}}, \quad k > 0 \quad (8.98)$$

PROOF. Since

$$X^+(z) \equiv \sum_{n=0}^{\infty} x[n] z^{-n} \quad (\text{P:3.5.1})$$

then it follows that

$$\mathcal{Z}^+[x[n-k]] = \sum_{n=0}^{\infty} x[n-k] z^{-n} = \sum_{m=-k}^{\infty} x[m] z^{-(m+k)} \quad (8.99)$$

by the change of index $m = n - k$,

$$= z^{-k} \sum_{m=-k}^{-1} x[m] z^{-m} + z^{-k} \underbrace{\sum_{m=0}^{\infty} x[m] z^{-m}}_{X^+(z)} \quad (8.100) \quad \square$$

This is the desired result.

Shifting property: Time Advance If $x[n] \stackrel{z^+}{\rightleftharpoons} X^+(z)$ then:

$$x[n+k] \stackrel{z^+}{\rightleftharpoons} z^k X^+(z) - \sum_{n=0}^{k-1} x[n] z^{k-n}, \quad k > 0 \quad (8.101)$$

PROOF. From the definition of the unilateral transform, it follows

$$\mathcal{Z}^+[x[n+k]] = \sum_{n=0}^{\infty} x[n+k] z^{-n} = \sum_{m=k}^{\infty} x[m] z^{-(m-k)} \quad (8.102)$$

by the change of index $m = n + k$. Thus,

$$= z^k \underbrace{\sum_{m=0}^{\infty} x[m] z^{-m}}_{X^+(z)} - z^k \sum_{m=1}^{k-1} x[m] z^{-m} \quad (8.103) \quad \square$$

This is the desired result.

Final Value Theorem If $x[n] \stackrel{z^+}{\rightleftharpoons} X^+(z)$ then:

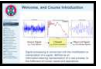
$$\lim_{n \rightarrow \infty} x[n] = \lim_{z \rightarrow 1} (z-1)X^+(z) \quad (\text{P:3.5.6})$$

The limit on the right hand side (RHS) exists if the ROC of $(z-1)X^+(z)$ includes the unit circle.

Further information can be found in books on discrete-time systems, for example [Proakis:1996, Section 3.5, Page 197].



8.4.3 Review of linear time-invariant systems



Topic Summary 56 Fourier Transform Theory

New slide

Topic Objectives:

- Understanding how to measure the size (or norm) of a signal.
- Motivation for Energy and Power.

Topic Activities:

Type	Details	Duration	Progress
Read Handout	Read page 312 to page 314	8 mins/page	
Try Example	Try Examples 8.8 and 8.9	40 mins	

- Systems which are **LTI** can be elegantly analysed in both the time and frequency domain: **convolution** in time, multiplication in frequency.
- For signals and sequences, it is common to write $\{y[n]\}_{n=-\infty}^{\infty}$, or even $\{y[n]\}_{n \in \mathbb{Z}}$ rather than simply $y[n]$: the latter is sufficient for these notes.
- Output, $y[n]$, of a **LTI** system is the **convolution** of the input, $x[n]$, and the **impulse response** of the system, $h[n]$:

$$y[n] = x[n] * h[n] \triangleq \sum_{k \in \mathbb{Z}} x[k] h[n - k] \tag{M:2.3.2}$$

- By making the substitution $\hat{k} = n - k$, it follows:

$$y[n] = \sum_{k \in \mathbb{Z}} h[k] x[n - k] = h[n] * x[n] \tag{M:2.3.3}$$

8.4.3.1 Matrix-vector formulation for convolution

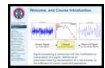
If $x[n]$ and $h[n]$ are sequences of finite duration, the **convolution** operation can be written in matrix-vector form. Let $x[n]$, $0 \leq n \leq N - 1$ and $h[n]$, $0 \leq n \leq M - 1$ be finite-duration sequences, then $y[n]$, $0 \leq n \leq L - 1$, where $L = N + M - 1$, can be written as:

$$\begin{bmatrix} y[0] \\ y[1] \\ \vdots \\ y[M-1] \\ \vdots \\ y[N-1] \\ \vdots \\ y[L-2] \\ y[L-1] \end{bmatrix} = \begin{bmatrix} x[0] & 0 & \cdots & 0 \\ x[1] & x[0] & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ x[M-1] & \cdots & \cdots & x[0] \\ \vdots & \ddots & \ddots & \vdots \\ x[N-1] & \cdots & \cdots & x[N-M] \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & x[N-1] & x[N-2] \\ 0 & \cdots & 0 & x[N-1] \end{bmatrix} \begin{bmatrix} h[0] \\ h[1] \\ \vdots \\ h[M-1] \end{bmatrix} \tag{M:2.3.4}$$

or

$$\mathbf{y} = \mathbf{X} \mathbf{h} \tag{M:2.3.5}$$

- Here, $\mathbf{y} \in \mathbb{R}^L$, $\mathbf{X} \in \mathbb{R}^{L \times M}$, and $\mathbf{h} \in \mathbb{R}^M$.



New slide

- The matrix \mathbf{X} is termed an **input data matrix**, and has the property that it is **toeplitz**.¹
- The observation or output vector \mathbf{y} can also be written in a similar way as:

$$\mathbf{y} = \mathbf{H} \mathbf{x} \quad (\text{M:2.3.6})$$

in which \mathbf{H} is also **toeplitz**.

- A system is **causal** if the present output sample depends only on past and/or present input samples.
- Assume system is asymptotically stable.

8.4.3.2 Transform-domain analysis of LTI systems

Time-domain **convolution**:

$$y[n] = \sum_{k \in \mathbb{Z}} x[k] h[n - k] \quad (\text{M:2.3.2})$$

or

$$y[n] = \sum_{k \in \mathbb{Z}} h[k] x[n - k] \quad (\text{M:2.3.3})$$

Taking z -**transforms** gives:

$$Y(z) = H(z) X(z) \quad (\text{M:2.3.8})$$

where $X(z)$, $Y(z)$ and $H(z)$ are the z -**transforms** of the input, output, and impulse response sequences respectively. $H(z) = \mathcal{Z}[h[n]]$ is the **system function** or **transfer function**.

8.4.3.3 Frequency response of LTI systems

The **frequency response** of the system is found by evaluating the z -transform on the unit circle, so $z = e^{j\omega}$:

$$Y(e^{j\omega T}) = H(e^{j\omega T}) X(e^{j\omega T}) \quad (\text{M:2.3.9})$$

- $|H(e^{j\omega})|$ is the **magnitude response** of the system, and $\arg H(e^{j\omega})$ is the **phase response**.
- The **group delay** of the system is a measure of the average delay of the system as a function of frequency:

$$\tau(e^{j\omega}) = -\frac{d}{d\omega} \arg H(e^{j\omega}) \quad (\text{M:2.3.11})$$

8.4.3.4 Frequency response to Periodic Inputs

Although the convolution summation formula can be used to compute the response of a stable system to any input, the frequency-domain input-output relationship for a **LTI** cannot be used with periodic inputs, since periodic signals do not strictly possess a z -transform. However, it is possible to develop an expression for the frequency response of LTI from first principles. Let $x[n]$ be a periodic signal with fundamental period N . This signal can be expanded using an IDFT as:

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{j\frac{2\pi}{N}kn}, \quad n \in \{0, \dots, N-1\} \quad (\text{M:2.3.19})$$

¹A **Toeplitz** matrix is one in which the elements along each diagonal, parallel to the main diagonal each descending from left to right, are constant. Note that the anti-diagonals are not necessarily equal.

where X_k are the Fourier components.

Hence, it follows that on substitution into the convolution equation:

$$y[n] = \sum_{m=-\infty}^{\infty} h[m] x[n-m] = \frac{1}{N} \sum_{m=-\infty}^{\infty} h[m] \sum_{k=0}^{N-1} X_k e^{j\frac{2\pi}{N}k(n-m)} \quad (\text{M:2.3.20})$$

which, by interchanging the order of summation (noting that the limits are over a rectangular region of summation), gives;

$$y[n] = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{j\frac{2\pi}{N}kn} \underbrace{\sum_{m=-\infty}^{\infty} h[m] e^{-j\frac{2\pi}{N}km}}_{H(e^{j\frac{2\pi}{N}k})} \quad (8.104)$$

where $H(e^{j\frac{2\pi}{N}k})$ are samples of $H(e^{j\omega})$. Hence,

$$y[n] = \frac{1}{N} \sum_{k=0}^{N-1} \left\{ H(e^{j\frac{2\pi}{N}k}) X_k \right\} e^{j\frac{2\pi}{N}kn} \quad (8.105)$$

However, this is just the inverse-DFT expansion of $y[n]$, and therefore:

$$Y_k = H(e^{j\frac{2\pi}{N}k}) X_k \quad k \in \{0, \dots, N-1\} \quad (\text{M:2.3.21})$$

Thus, the response of a LTI system to a periodic input is also periodic with the same period. The magnitude of the input components is modified by $|H(e^{j\frac{2\pi}{N}k})|$, and the phase is modified by $\arg H(e^{j\frac{2\pi}{N}k})$.

8.4.4 Rational transfer functions

Many systems can be expressed in the z -domain by a **rational transfer function**. They are described in the time domain by:

$$y[n] = - \sum_{k=1}^P a_k y[n-k] + \sum_{k=0}^Q d_k x[n-k] \quad (\text{M:2.3.12})$$

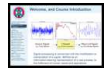
Taking z -transforms gives:

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{k=0}^Q d_k z^{-k}}{1 + \sum_{k=1}^P a_k z^{-k}} \triangleq \frac{D(z)}{A(z)} \quad (\text{M:2.3.13})$$

This can be described in the complex z -plane as:

$$H(z) = \frac{D(z)}{A(z)} = G \frac{\prod_{k=1}^Q (1 - z_k z^{-1})}{\prod_{k=1}^P (1 - p_k z^{-1})} \quad (\text{M:2.3.14})$$

where p_k are the poles of the system, and z_k are the zeros.



9

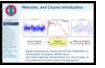
Discrete-Time Stochastic Processes

Introduces the notion of time-series or random processes. Gives an interpretation using ensembles, and covers second-order statistics including correlation sequences. Discusses types of stationary processes, ergodicity, joint-signal statistics, and correlation matrices.

9.1 A Note on Notation

Note that, unfortunately, for this module, a slightly different (and abusive use of) notation for random quantities is used than what was presented in the first four handouts of the *Probability, Random Variables, and Estimation Theory (PET)* module. In the literature, most time series are described using lower-case letters, primarily since once the notation for the representation of a random process in the frequency domain is discussed, 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. Hence, random variables and vectors in this handout will not always be denoted using upper-case letters.

9.2 Definition of a Stochastic Process



Topic Summary 57 Introduction to Stochastic Processes

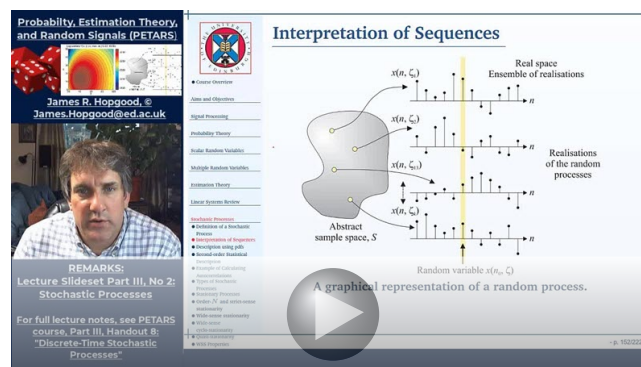
New slide

Topic Objectives:

- Definition of a Stochastic Process.
- Concept of an ensemble of realisations.
- Example of drawing an ensemble and the corresponding sample space.
- Interpretation of the random sequences.

Topic Activities:

Type	Details	Duration	Progress
Watch video	13 : 22 min video	3 × length	
Read Handout	Read page 316 to page 318	8 mins/page	
Try Example	Try Example 9.1	10 mins	



http://media.ed.ac.uk/media/1_f7d1ldvi

Video Summary: In this first video of the Statistical Signal Processing part of the PETARS course, the notion of random signals, or stochastic processes is introduced. It is defined as a natural extension to the conceptual development of random variables and random vectors, but where a deterministic signal is associated with each outcome of the experiment. After a formal definition of the random process, the notion of an ensemble of realisations is considered. The video then gives an example of plotting an ensemble for a particular problem. The random process is also considered as a sequence of random variables, where the random variables have dependences. Finally, this video discusses the general concepts regarding analysing random processes that will be considered in the remainder of the course.

After studying random variables and vectors, these concepts can now (easily) be extended to discrete-time signals or sequences.

- Natural discrete-time signals can be characterised as random signals, since their values cannot be determined precisely; that is, they are **unpredictable**. A natural mathematical framework for the description of these discrete-time random signals is provided by discrete-time stochastic processes.

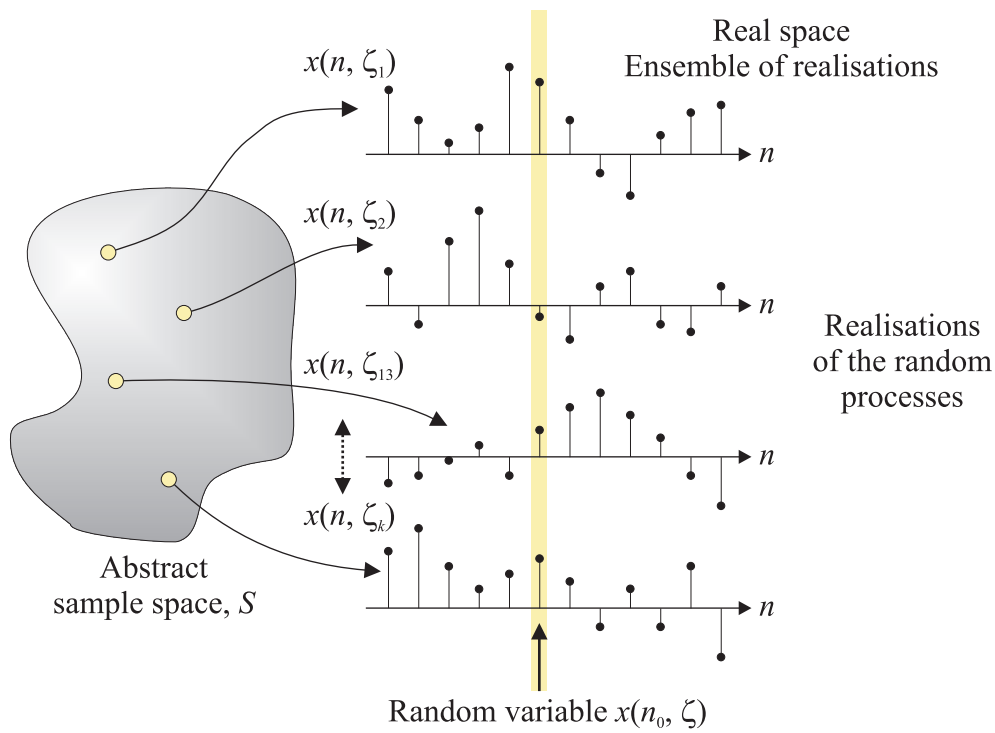
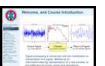


Figure 9.1: A graphical representation of a random process.

- To obtain a formal definition, consider an experiment with a finite or infinite number of unpredictable outcomes from a sample space $\mathcal{S} = \{\zeta_k, k \in \mathbb{Z}^+\}$, each occurring with probability $\Pr(\zeta_k)$. Assign by some rule to each $\zeta_k \in \mathcal{S}$ a deterministic sequence $x[n, \zeta_k], n \in \mathbb{Z}$.
- The sample space \mathcal{S} , probabilities $\Pr(\zeta_k)$, and the sequences $x[n, \zeta_k], n \in \mathbb{Z}$ constitute a **discrete-time stochastic process**, or **random sequence**.
- Formally, $x[n, \zeta_k], n \in \mathbb{Z}$ is a random sequence or **stochastic process** if, for a fixed value $n_0 \in \mathbb{Z}^+$ of n , $x[n_0, \zeta], n \in \mathbb{Z}$ is a random variable.
- A random or stochastic process is also known as a **time series** in the statistics literature.
- It is an infinite sequence of random variables, so could be thought of as an infinite-dimensional random vector. Indeed, finite-length random signals and sequences can specifically be represented by the concept of a random vector.

9.2.1 Interpretation of Sequences



Example 9.1. Consider a continuous-time random process, $x(t, \zeta)$, defined by a finite sized ensemble *New slide* consisting of four equally probable functions given by:

$$\begin{aligned} x(t, 1) &= -3 u(t) & x(t, 2) &= \cos(5\pi t) u(t) \\ x(t, 3) &= 10 t u(t) & x(t, 4) &= 2 \sin(6\pi t + 0.2) \end{aligned}$$

1. Draw the ensemble.
2. For $t = 0.2$, determine the sample space.

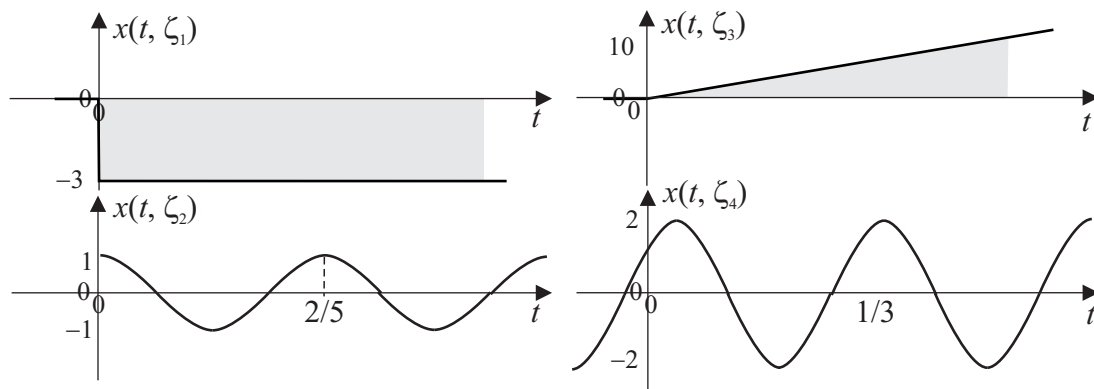


Figure 9.2: Ensemble of waveforms.

SOLUTION. 1. To plot the ensemble, draw all the realisations. The ensemble is therefore shown in Figure 9.2.

2. The sample space is thus $\{-3, -1, 2, -1.4736\}$.

The set of all possible sequences $\{x[n, \zeta]\}$ is called an **ensemble**, and each individual sequence $x[n, \zeta_k]$, corresponding to a specific value of $\zeta = \zeta_k$, is called a **realisation** or a **sample sequence** of the ensemble. Hence, when a random process is observed through the outcome of a single experiment, one member of the ensemble is selected randomly and presented. A graphical representation of a random process is shown in Figure 9.7.

There are four possible interpretations of $x[n, \zeta]$:

	ζ Fixed	ζ Variable
n Fixed	Number	Random variable
n Variable	Sample sequence	Stochastic process

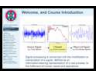
Use simplified notation $x[n] \equiv x[n, \zeta]$ to denote both a stochastic process, and a single realisation. The word *stochastic* is derived from the Greek word *stochasticos*, which means skillful in aiming or guessing. Use the terms **random process** and **stochastic process** interchangeably throughout this course.

Building on these interpretations of sequences, this course will therefore investigate:

- The statistical properties of random signals, the statistical dependence of samples at different points in time.
- Interpreting stochastic signals in the frequency domain, the notion of a random spectrum, and the concept of the power spectral density.
- What happens to a stochastic process and signals as it passes through systems?
- The notion of signal modelling for signal analysis and prediction.



9.2.2 Description using probability density functions (pdfs)



Topic Summary 58 Statistical Description of Random Processes

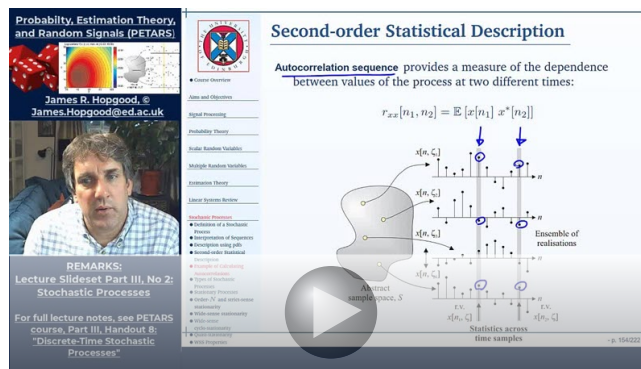
New slide

Topic Objectives:

- Concept of second-order statistical descriptions.
- Calculating autocorrelation sequence (ACS) from a signal model.
- Calculating ACS of a linear function of random processes.

Topic Activities:

Type	Details	Duration	Progress
Watch video	25 : 34 min video	3 × length	
Read Handout	Read page 319 to page 323	8 mins/page	
Try Example	Try 9.2 and 9.3	10 mins	



http://media.ed.ac.uk/media/1_fpiz8i47

Video Summary: As with random vectors, other than in certain special cases, it can be difficult to describe and manipulate random processes through the use of joint-pdfs, although the definitions for the joint-pdf is provided. Instead, the video discusses that second-order statistics including the mean sequence, the ACS (second-moment), and the autocovariance sequence (central moment) are often adequate for capturing key salient features of the random processes. After extending the definitions for the mean and correlations previously seen for random vectors to random processes, two examples are given. The first example derives the ACS for a process which is based on an *a priori* defined physics-based model (namely, an harmonic process). The second example considers finding the ACS of a linear function of random processes (in this case, a non-causal delay).

For fixed $n = n_0$, it is clear from Figure 9.7 that $x[n_0, \zeta]$ is a random variable. Moreover, the random vector formed from the k random variables $\{x[n_j], j \in \{1, \dots, k\}\}$ is characterised by the joint-cumulative distribution function (cdf) and pdfs:

$$F_X(x_1 \dots x_k | n_1 \dots n_k) = \Pr(x[n_1] \leq x_1, \dots, x[n_k] \leq x_k) \tag{9.1}$$

$$f_X(x_1 \dots x_k | n_1 \dots n_k) = \frac{\partial^k F_X(x_1 \dots x_k | n_1 \dots n_k)}{\partial x_1 \dots \partial x_k} \tag{9.2}$$

In exactly the same way as with random variables and random vectors, it is:

- difficult to estimate these probability functions without considerable additional information or assumptions;
- possible to frequently characterise stochastic processes usefully with much less information.

Thus, the density and distribution functions are characterised using moments and, in particular, second-order moments.

9.3 Second-order Statistical Description

Random variables can be characterised, upto second-order statistics, using the mean and variance; random vectors are characterised by the mean vector, auto-correlation and auto-covariance matrices. Random processes, however, are characterised by sequences, where a particular sample, n_0 , of this sequence characterises the random variable $x[n_0, \zeta]$. These sequences are the mean and variance sequence, the autocorrelation and autocovariance sequences, as outlined below.

Mean and Variance Sequence At time n , the **ensemble** mean and variance are given by:

$$\mu_x[n] = \mathbb{E} [x[n]] \quad (\text{M:3.3.3})$$

$$\sigma_x^2[n] = \mathbb{E} [|x[n] - \mu_x[n]|^2] = \mathbb{E} [|x[n]|^2] - |\mu_x[n]|^2 \quad (\text{M:3.3.4})$$

Both $\mu_x[n]$ and $\sigma_x^2[n]$ are deterministic sequences.

Autocorrelation sequence The second-order statistic $r_{xx}[n_1, n_2]$ provides a measure of the dependence between values of the process at two different times; it can provide information about the time variation of the process:

$$r_{xx}[n_1, n_2] = \mathbb{E} [x[n_1] x^*[n_2]] \quad (\text{M:3.3.5})$$

Note this definition is not consistent across all text book, or indeed University courses!

Autocovariance sequence The autocovariance sequence provides a measure of how similar the deviation from the mean of a process is at two different time instances:

$$\begin{aligned} \gamma_{xx}[n_1, n_2] &= \mathbb{E} [(x[n_1] - \mu_x[n_1])(x[n_2] - \mu_x[n_2])^*] \\ &= r_{xx}[n_1, n_2] - \mu_x[n_1] \mu_x^*[n_2] \end{aligned} \quad (\text{M:3.3.6})$$

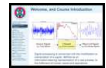
To show how these deterministic sequences of a stochastic process can be calculated, several examples are considered in detail below.

9.3.1 Example of Calculating Autocorrelations

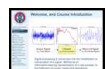
Example 9.2 ([Manolakis:2000, Ex 3.9, page 144]). The harmonic process $x[n]$ is defined by:

$$x[n] = \sum_{k=1}^M A_k \cos(\omega_k n + \phi_k), \quad \omega_k \neq 0 \quad (\text{M:3.3.50})$$

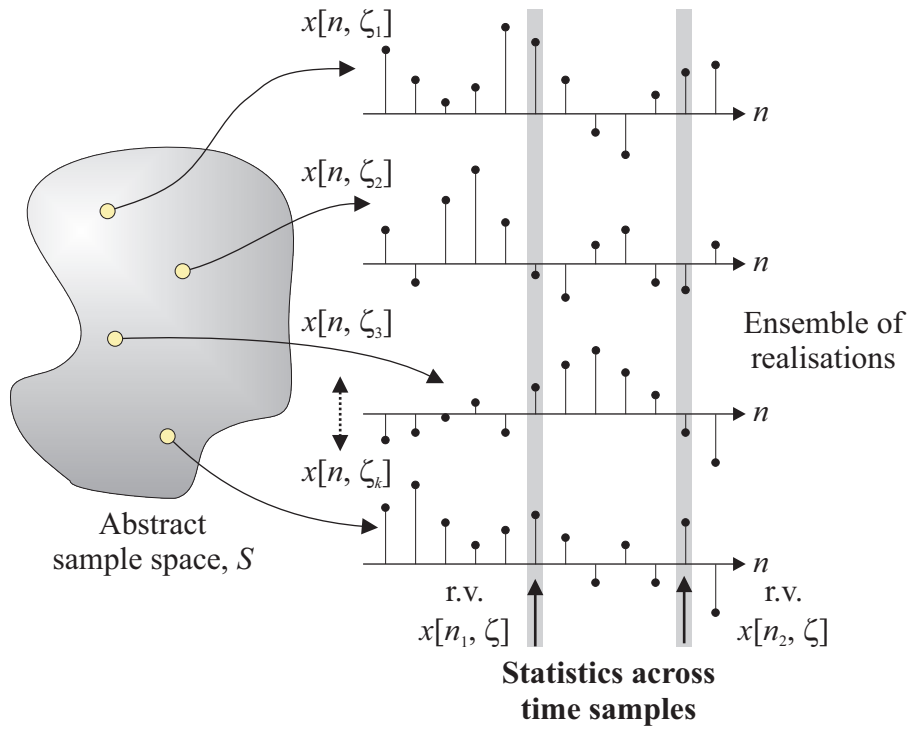
where M , $\{A_k\}_1^M$ and $\{\omega_k\}_1^M$ are constants, and $\{\phi_k\}_1^M$ are pairwise independent random variables uniformly distributed in the interval $[0, 2\pi]$.



New slide



New slide



1. Determine the mean of $x[n]$.
2. Show the autocorrelation sequence is given by

$$r_{xx}[\ell] = \frac{1}{2} \sum_{k=1}^M |A_k|^2 \cos \omega_k \ell, \quad -\infty < \ell < \infty \tag{9.3}$$

where $\ell \triangleq n_1 - n_2$, and $r_{xx}[\ell] \triangleq r_{xx}[n_1, n_1 + \ell]$ for any n_1 .

SOLUTION. 1. The expected value of the process is straightforwardly given by:

$$\mathbb{E} [x[n]] = \mathbb{E} \left[\sum_{k=1}^M A_k \cos(\omega_k n + \phi_k) \right] = \sum_{k=1}^M A_k \mathbb{E} [\cos(\omega_k n + \phi_k)] \tag{9.4}$$

Recall from results derived earlier in the course that if $x[n, \zeta] = g(n, \phi(\zeta))$ is a random variable obtained by transforming $\phi(\zeta)$ through a known function, g , the expectation of $x[n] = x[n, \zeta]$ is:

$$\mathbb{E} [x[n]] = \int x[n] p(x[n]) dx[n] \tag{9.5}$$

$$= \int_{-\infty}^{\infty} g(n, \phi) p_{\Phi}(\phi) d\phi \tag{9.6}$$

This property results from the invariance of the expectation operator, and helps for problems like the present one; this invariance was covered back in the handout on Scalar random variables. It is important to consider n as a constant.

Since a co-sinusoid is zero-mean, then:

$$\mathbb{E} [\cos(\omega_k n + \phi_k)] = \int \cos(\omega_k n + \phi_k) f_{\Phi_k}(\phi_k) d\phi_k \tag{9.7}$$

$$= \int_0^{2\pi} \cos(\omega_k n + \phi_k) \times \frac{1}{2\pi} \times d\phi_k = 0 \tag{9.8}$$

Hence, it follows:

$$\mathbb{E}[x[n]] = 0, \quad \forall n \quad (9.9)$$

2. The autocorrelation $r_{xx}[n_1, n_2] = \mathbb{E}[x[n_1] x^*[n_2]]$ follows similarly:

$$r_{xx}[n_1, n_2] = \mathbb{E} \left[\sum_{k=1}^M A_k \cos(\omega_k n_1 + \phi_k) \sum_{j=1}^M A_j^* \cos(\omega_j n_2 + \phi_j) \right] \quad (9.10)$$

$$= \sum_{k=1}^M \sum_{j=1}^M A_k A_j^* \underbrace{\mathbb{E}[\cos(\omega_k n_1 + \phi_k) \cos(\omega_j n_2 + \phi_j)]}_{r(\phi_k, \phi_j)} \quad (9.11)$$

After some algebra, it can be shown that the term $r(\phi_k, \phi_j)$:

$$\mathbb{E}[\cos(\omega_k n_1 + \phi_k) \cos(\omega_j n_2 + \phi_j)] = \begin{cases} \frac{1}{2} \cos \omega_k (n_1 - n_2) & k = j \\ 0 & \text{otherwise} \end{cases} \quad (9.12)$$

The proof of this statement is obtained by considering the term

$$r(\phi_k, \phi_j) = \mathbb{E}[\cos(\omega_k n_1 + \phi_k) \cos(\omega_j n_2 + \phi_j)] \quad (9.13)$$

for the cases when $k \neq j$, and when $k = j$. Considering the former case first, $k \neq j$, then

$$r(\phi_k, \phi_j) \iint \cos(\omega_k n_1 + \phi_k) \cos(\omega_j n_2 + \phi_j) f_{\Phi_j \Phi_k}(\phi_j, \phi_k) d\phi_j d\phi_k \quad (9.14)$$

Using the fact that $\{\phi_k\}_1^M$ come from the uniform density, then $f_{\Phi_j \Phi_k}(\phi_j, \phi_k) = (\frac{1}{2\pi})^2 \mathbb{I}_{[0, 2\pi]}(\phi_j) \mathbb{I}_{[0, 2\pi]}(\phi_k)$, then:

$$r(\phi_k, \phi_j) = \frac{1}{4\pi^2} \int_0^{2\pi} \int_0^{2\pi} \cos(\omega_k n_1 + \phi_k) \cos(\omega_j n_2 + \phi_j) d\phi_j d\phi_k \quad (9.15)$$

$$= \frac{1}{4\pi^2} \int_0^{2\pi} \cos(\omega_k n_1 + \phi_k) d\phi_k \int_0^{2\pi} \cos(\omega_j n_2 + \phi_j) d\phi_j \quad (9.16)$$

$$= 0 \quad (9.17)$$

An alternative derivation for this case when $k \neq j$, which might be considered more straightforward, is to observe that Equation 9.13 might also be written as:

$$r(\phi_k, \phi_j) = \mathbb{E}[g(\phi_k) h(\phi_j)] = \mathbb{E}[g(\phi_k)] \mathbb{E}[h(\phi_j)] \quad (9.18)$$

where $g(\phi_k) = \cos(\omega_k n_1 + \phi_k)$ and $h(\phi_k) = \cos(\omega_j n_2 + \phi_j)$, and the fact that ϕ_k and ϕ_j are independent implies the expectation function may be factorised.

For the case when $k = j$ such that $\phi = \phi_k = \phi_j$ and $\omega = \omega_k = \omega_j$, then:

$$r(\phi, \phi) = \int \cos(\omega n_1 + \phi) \cos(\omega n_2 + \phi) f_{\Phi}(\phi) d\phi \quad (9.19)$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \cos(\omega n_1 + \phi) \cos(\omega n_2 + \phi) d\phi \quad (9.20)$$

Using the trigonometric identity $\cos A \cos B = \frac{1}{2} (\cos(A + B) + \cos(A - B))$, then:

$$r(\phi_k, \phi_j) = \frac{1}{4\pi} \int_0^{2\pi} \{\cos \omega (n_1 - n_2) + \cos(\omega (n_1 + n_2) + 2\phi)\} d\phi \quad (9.21)$$

$$= \frac{1}{2} \cos \omega (n_1 - n_2) \quad (9.22)$$

giving the result above; namely:

$$\mathbb{E} [\cos(\omega_k n_1 + \phi_k) \cos(\omega_j n_2 + \phi_j)] = \frac{1}{2} \cos \omega_k (n_1 - n_2) \delta(k - j) \quad (9.23)$$

Substituting this expression into

$$r_{xx}[n_1, n_2] = \sum_{k=1}^M \sum_{j=1}^M A_k A_j^* \mathbb{E} [\cos(\omega_k n_1 + \phi_k) \cos(\omega_j n_2 + \phi_j)] \quad (9.24)$$

thus leads to the desired result, where $\ell = n_1 - n_2$. It can be seen that the process $x[n]$ must be a stationary process, as it is only a function of the lag ℓ :

$$r_{xx}[\ell] = \frac{1}{2} \sum_{k=1}^M |A_k|^2 \cos \omega_k \ell, \quad -\infty < \ell < \infty \quad (9.25) \quad \square$$

Note finally that these are ensemble statistics, meaning that they are expected values across the different realisations (i.e. across the ensemble).

Example 9.3 (Functions of Random Process). A random variable $y[n]$ is defined to be:

$$y[n] = x[n] + x[n + m] \quad (9.26)$$

where m is some integer, and $x[n]$ is a stochastic process whose ACS is given by:

$$r_{xx}[n_1, n_2] = e^{-(n_1 - n_2)^2} \quad (9.27)$$

Derive an expression for the ACS of the stochastic process $y[n]$, denoted $r_{yy}[n_1, n_2]$.

SOLUTION. In this example, it is simplest to form the product:

$$y[n_1] y^*[n_2] = [x[n_1] + x[n_1 + m]] [x^*[n_2] + x^*[n_2 + m]] \quad (9.28)$$

$$\begin{aligned} &= x[n_1] x^*[n_2] + x[n_1] x^*[n_2 + m] \\ &\quad + x[n_1 + m] x^*[n_2] + x[n_1 + m] x^*[n_2] \end{aligned} \quad (9.29)$$

Then, taking expectations, it follows:

$$r_{yy}[n_1, n_2] = r_{xx}[n_1, n_2] + r_{xx}[n_1, n_2 + m] \quad (9.30)$$

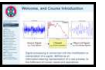
$$+ r_{xx}[n_1 + m, n_2] + r_{xx}[n_1 + m, n_2 + m] \quad (9.31)$$

Using the result $r_{xx}[n_1, n_2] = e^{-(n_1 - n_2)^2}$ gives, in this particular case:

$$r_{yy}[r_1, r_2] = 2 e^{-(n_1 - n_2)^2} + e^{-(n_1 - n_2 + m)^2} + e^{-(n_1 - n_2 - m)^2} \quad (9.32) \quad \square$$



9.4 Types of Stochastic Processes



Topic Summary 59 Important Types of Stochastic Processes

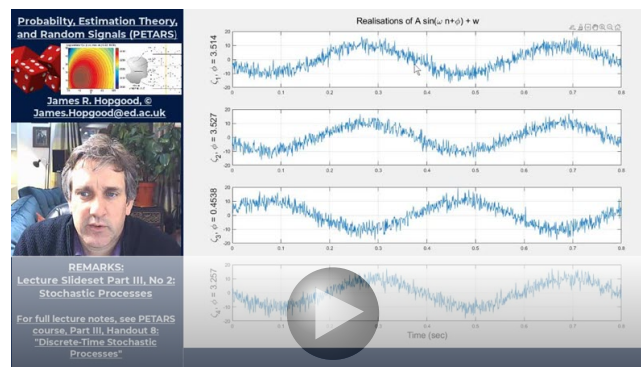
New slide

Topic Objectives:

- Concepts and definitions of fundamental types of stochastic processes.
- Understanding predictable and unpredictable processes and signal decompositions.

Topic Activities:

Type	Details	Duration	Progress
Watch video	16 : 37 min video	3× length	
Read Handout	Read page 324 to page 327	8 mins/page	
Try Code	Use the MATLAB code	10 mins	
Try Example	Try Example 9.4	20 mins	



http://media.ed.ac.uk/media/1_rnwrdpim

Video Summary: This video discusses some fundamental types of stochastic processes, including predictable processes, unpredictable processes, independent and independent and identically distributed processes, uncorrelated, and orthogonal processes, and an introduction to stationary processes. These fundamental processes are introduced mainly to define terminology for the rest of the course, but also to discuss the importance of some of these processes in signal modelling, and processes that can be dealt with in a mathematically convenient manner. There are no examples associated with this topic, but there is a MATLAB example for generating a linear combination of a predictable and unpredictable process (and Wold's decomposition theorem is mentioned in passing).

Some useful types of stochastic properties, based on their statistical properties, are now introduced:

Predictable Processes A deterministic signal is by definition exactly predictable; it assumes there exists a certain functional relationship that completely describes the signal, even if that functional relationship is not available or is extremely difficult to describe. The unpredictability of a random process is, in general, the combined result of the following two characteristics:

1. The selection of a single realisation of a stochastic process is based on the outcome of a random experiment; in other-words, it depends on ζ .

2. No functional description is available for *all* realisations of the *ensemble*. In other-words, even if a functional relationship is available for a subset of the ensemble, it might not be available for all members of the ensemble.

In some special cases, however, a functional relationship is available. This means that after the occurrence of all samples of a particular realisation up to a particular point, n , all future values can be predicted exactly from the past ones.

If this is the case for a random process, then it is called **predictable**, otherwise it is said to be **unpredictable** or a **regular process**.

KEYPOINT! (Predictable Process). As an example of a predictable process, consider the signal:

$$x[n, \zeta] = A \sin(\omega n + \phi) \quad (9.33)$$

□

where A is a known amplitude, ω is a known normalised angular frequency, and ϕ is a random phase, where $\phi \sim f_{\Phi}(\phi)$ is its pdf.

As an outline of this idea, suppose that all the samples of a stochastic process $x[n, \zeta]$ upto sample $n - 1$ are known; thus, $\{x[k, \zeta]\}_{k=-\infty}^{n-1}$ are known. Then the predicted value of $x[n]$ might, for example, be expressed as:

$$\hat{x}[n] = - \sum_{k=1}^{\infty} a_k^* x[n - k] \quad (\text{T:7.189})$$

The error in this prediction is given by

$$\epsilon[n] = x[n] - \hat{x}[n] = \sum_{k=0}^{\infty} a_k^* x[n - k] \quad (\text{T:7.190})$$

where $a_0 = 1$. The process is said to be **predictable** if the $\{a_k\}$'s can be chosen such that:

$$\sigma_{\epsilon}^2 = \mathbb{E} [|\epsilon[n]|^2] = 0 \quad (\text{T:7.191})$$

Otherwise the process is not predictable. The phrase *not predictable* is somewhat misleading, since the **linear prediction** in Equation T:7.189 can be applied to any process, whether predictable or not, with satisfactory results. If a process is not predictable, it just means that the prediction error variance is not zero.

An example of **predictable process** is the process $x[n, \zeta] = c$, where c is a random variable, since every realisation of the discrete-time signal has a constant amplitude, and once $x[n_0, \zeta_k]$ is known for a particular realisation, all other samples of that process have also been determined.

The notion of predictable and regular processes is formally presented through the **Wold decomposition**, and further details of this very important theorem can be found in [Therrien:1992, Section 7.6, Page 390] and [Papoulis:1991, Page 420].

Independence A stochastic process is independent if, and only if, (iff)

$$f_X(x_1, \dots, x_N | n_1, \dots, n_N) = \prod_{k=1}^N f_{X_k}(x_k | n_k) \quad (\text{M:3.3.10})$$

$\forall N, n_k, k \in \{1, \dots, N\}$. Here, therefore, $x[n]$ is a sequence of independent random variables.

An independent and identically distributed (i. i. d.) process is one where all the random variables $\{x[n_k, \zeta], n_k \in \mathbb{Z}\}$ have the same pdf, and $x[n]$ will be called an **i. i. d.** random process.

Example 9.4 (Independence: i. i. d. processes). I am selling my house, and have decided to accept the first offer exceeding K pounds. Assuming that the offers are i. i. d. random variables, with common cumulative distribution function $F_X(x)$, where x is the offer price, find the expected number of offers received before I sell the house.

SOLUTION. Suppose that I sell the house after N offers. Then there are $N - 1$ offers that are less than K , which occur with probability $F_X(K)$. Thus, the probability of selling the house after N offers is:

$$\Pr(N = n) = F_X(K)^{n-1} [1 - F_X(K)] \quad n \geq 1 \quad (9.34)$$

This is a **geometric distribution**, and its mean can either be looked up in tables, or calculated:

$$\mu_N = \sum_{n=1}^{\infty} n \Pr(N = n) = \sum_{n=1}^{\infty} n F_X(K)^{n-1} [1 - F_X(K)] \quad (9.35)$$

$$= \left[\frac{1-r}{r} \right] \sum_{n=0}^{\infty} n r^n \quad (9.36)$$

where $r = F_X(K)$. There is a general result which can be found in mathematical tables that [Gradshteyn:1994]:

$$\sum_{n=0}^{N-1} (a + nb)r^n = \frac{a - [a + (N-1)b]r^N}{1-r} + \frac{br(1-r^{N-1})}{(1-r)^2}, \quad r \neq 0, N > 1 \quad (9.37)$$

Therefore, in the case when $a = 0$, $r = 1$, and $N \rightarrow \infty$, and $0 < r < 1$ then:

$$\sum_{n=0}^{\infty} n r^n = \frac{r}{(1-r)^2}, \quad 0 < r < 1 \quad (9.38)$$

Hence, this gives the mean of the geometric distribution as:

$$\mu_N = \left[\frac{1-r}{r} \right] \frac{r}{(1-r)^2} = \frac{1}{1-r} = [1 - F_X(K)]^{-1} \quad (9.39) \quad \square$$

An uncorrelated processes is a sequence of uncorrelated random variables:

$$\gamma_{xx}[n_1, n_2] = \sigma_x^2[n_1] \delta[n_1 - n_2] \quad (\text{M:3.3.11})$$

Alternatively, the ACS can be written as:

$$r_{xx}[n_1, n_2] = \begin{cases} \sigma_x^2[n_1] + |\mu_x[n_1]|^2 & n_1 = n_2 \\ \mu_x[n_1] \mu_x^*[n_2] & n_1 \neq n_2 \end{cases} \quad (\text{M:3.3.12})$$

An **orthogonal process** is a sequence of **orthogonal random variables**, and is given by:

$$r_{xx}[n_1, n_2] = \mathbb{E} [|x[n_1]|^2] \delta[n_1 - n_2] \quad (\text{M:3.3.13})$$

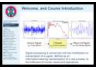
If a process is zero-mean, then it is both **orthogonal** and **uncorrelated** since $\gamma_{xx}[n_1, n_2] = r_{xx}[n_1, n_2]$. More often than not, in this course, we shall consider zero-mean processes.

A **stationary process** is a random process where its statistical properties do not vary with time. Put another way, it would be impossible to distinguish the statistical characteristics of a process at time t from those at some other time, t' . Processes whose statistical properties **do** change with time are referred to as **nonstationary**.

– End-of-Topic 59: **Types of Random Signals** –



9.5 Stationary Processes



Topic Summary 60 Stationary and wide-sense stationary (WSS) processes

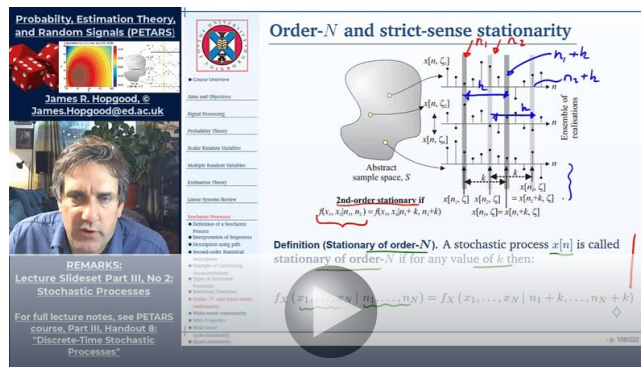
New slide

Topic Objectives:

- Awareness of common types and definitions of stationary random processes.
- Understand order- N , strict-sense stationary, and wide-sense stationary processes.
- Examples of manipulating means and autocorrelation sequences for stationary processes.

Topic Activities:

Type	Details	Duration	Progress
Watch video	18 : 07 min video	3× length	
Read Handout	Read page 328 to page 334	8 mins/page	
Try Example	Try Examples 9.5, 9.6, and 9.7.	15 mins	
Practice Exercises	Exercise ??	15 mins	



http://media.ed.ac.uk/media/1_c2z10igx

Video Summary: This video starts by considering the common types and definitions of stationary processes used in time-series analysis. This topic then considers meanings and relationships of order- N , strict-sense stationarity, and wide-sense stationarity. The second half of the video focusses on an example of showing that the sum of a co-sinusoid and sinusoid with independent random amplitudes but fixed phase and frequency is a stationary process (and although not mentioned, will of course be a predictable processes). Other examples are included in the handout associated with this video.

A random process $x[n]$ has been called **stationary** if its statistics determined for $x[n]$ are equal to those for $x[n + k]$, for every k . There are various formal definitions of **stationarity**, along with **quasi-stationary** processes, which are discussed below.

- **Order- N and strict-sense stationarity**
- **Wide-sense stationarity**
- Autocorrelation properties for WSS processes
- **Wide-sense periodicity and cyclo-stationarity**

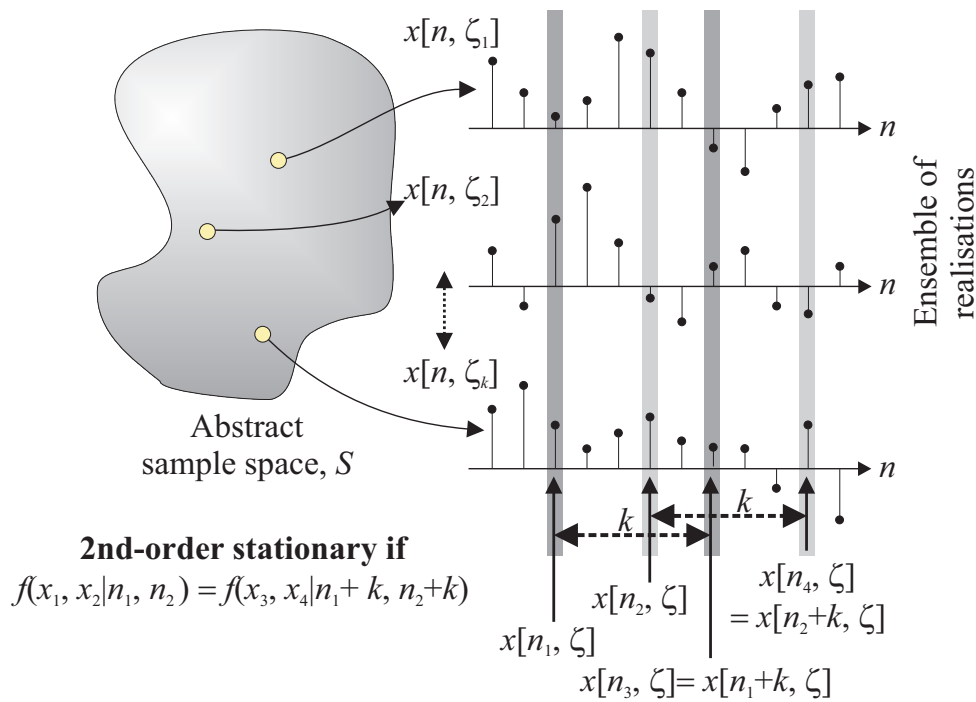


Figure 9.3: Demonstrating 2nd-order stationarity.

- Local- or **quasi-stationary** processes

After this, some examples of various stationary processes will be given.

9.5.1 Order- N and strict-sense stationarity

Definition 9.1 (Stationary of order- N). A stochastic process $x[n]$ is called **stationary of order- N** if for any value of k then:

$$f_X(x_1, \dots, x_N | n_1, \dots, n_N) = f_X(x_1, \dots, x_N | n_1 + k, \dots, n_N + k) \tag{M:3.3.21}$$



Definition 9.2 (Strict-sense stationary). If $x[n]$ is stationary for all orders $N \in \mathbb{Z}^+$, it is said to be **strict-sense stationary (SSS)**.

Clearly, any stochastic process that is stationary of order- N is also stationary of order- M , where $M \leq N$.

An independent and identically distributed process is SSS since, in this case, $f_{X_k}(x_k | n_k) = f_X(x_k)$ is independent of n , and therefore also of $n + k$. However, SSS is more restrictive than necessary in practical applications, and is a rarely required property.

9.5.2 Wide-sense stationarity

A more relaxed form of stationarity, which is sufficient for practical problems, occurs when a random process is stationary order-2; such a process is **wide-sense stationary (WSS)**.

Definition 9.3 (Wide-sense stationarity). A random signal $x[n]$ is called wide-sense stationary if:

- the mean and variance is constant and independent of n :

$$\mathbb{E}[x[n]] = \mu_x \quad (\text{M:3.3.22})$$

$$\text{var}[x[n]] = \sigma_x^2 \quad (\text{M:3.3.23})$$

- the autocorrelation depends only on the time difference $\ell = n_1 - n_2$, called the lag:

$$\begin{aligned} r_{xx}[n_1, n_2] &= r_{xx}^*[n_2, n_1] = \mathbb{E}[x[n_1] x^*[n_2]] \\ &= r_{xx}[\ell] = r_{xx}[n_1 - n_2] = \mathbb{E}[x[n_1] x^*[n_1 - \ell]] \\ &= \mathbb{E}[x[n_2 + \ell] x^*[n_2]] \end{aligned} \quad (\text{M:3.3.24}) \quad \diamond$$

KEYPOINT! (Inconsistency of definition of lag). The definition of the **lag** is not consistent across textbooks, or indeed courses on this MSc! Elsewhere, the following definition is used for a stationary process:

$$\begin{aligned} r_{xx}[n_1, n_2] &\triangleq \mathbb{E}[x[n_1] x^*[n_1 + \hat{\ell}]] \\ r_{xx}[\hat{\ell}] &\triangleq \mathbb{E}[x[n - \hat{\ell}] x^*[n]] \end{aligned} \quad (9.40) \quad \square$$

Although a minor change in sign, this does have implications when considering results that are functions of random processes, such as a signal passing through a linear system, or frequency-domain analysis. It is simply something to become used to, and to understand the equations and use the appropriate subsequent results carefully.

Additionally:

- The autocovariance sequence is given by:

$$\gamma_{xx}[\ell] = r_{xx}[\ell] - |\mu_x|^2 \quad (9.41)$$

- Since 2nd-order moments are defined in terms of 2nd-order pdf, then strict-sense stationary are always WSS, but not necessarily *vice-versa*, except if the signal is Gaussian.
- In practice, however, it is very rare to encounter a signal that is stationary in the wide-sense, but not stationary in the strict sense.

Example 9.5 ([Manolakis:2000, Example 3.3.1, Page 102]). Let $w[n]$ be a zero-mean, uncorrelated Gaussian random sequence with variance $\sigma_w^2[n] = 1$.

1. Characterise the random sequence $w[n]$.
2. Define $x[n] = w[n] + w[n - 1]$, $n \in \mathbb{Z}$. Determine the mean and autocorrelation of $x[n]$. Also, characterise $x[n]$.

SOLUTION. Note that the variance of $w[n]$ is a constant.

1. Since uncorrelatedness implies independence for Gaussian random variables, then $w[n]$ is an independent random sequence. Since its mean and variance are constants, it is at least stationary of first-order. Furthermore, from Equation M:3.3.12 or from Equation M:3.3.13, then:

$$r_{ww}[n_1, n_2] = \sigma_w^2 \delta[n_1 - n_2] = \delta[n_1 - n_2] \quad (9.42)$$

Since the autocorrelation sequence depends only on the lag $n_1 - n_2$, then by definition it is WSS process.

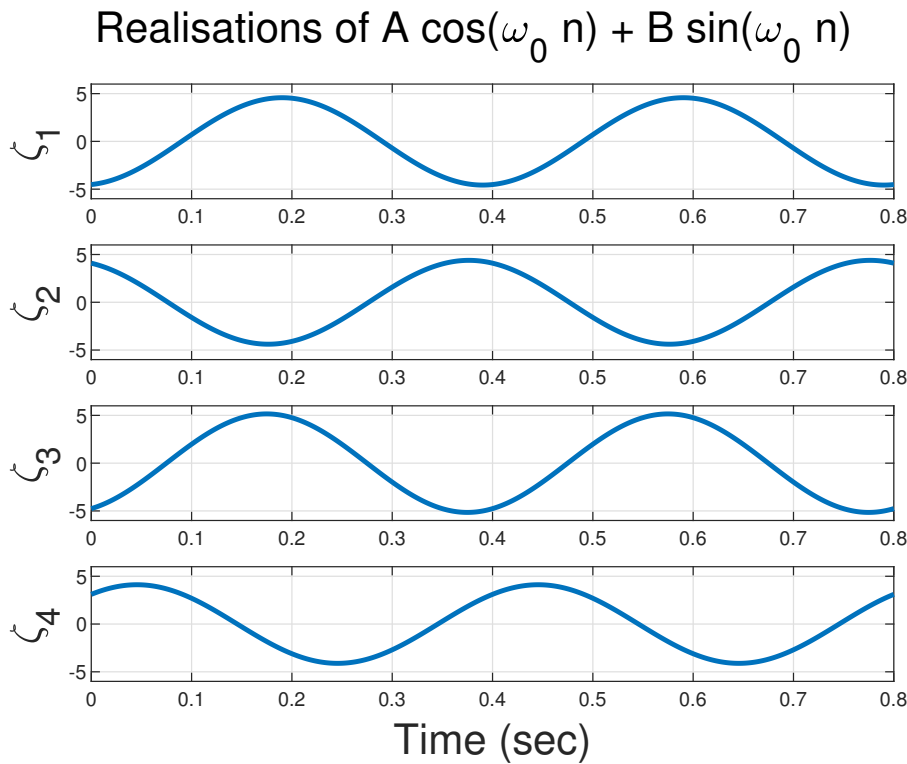


Figure 9.4: Ensemble of waveforms for the problem in Example 9.6.

2. The mean of $x[n]$ is zero for all n since $w[n]$ is a zero-mean process. Next, consider:

$$r_{xx}[n_1, n_2] = \mathbb{E} [x[n_1] x^*[n_2]] \quad (9.43)$$

$$= \mathbb{E} [[w(n_1) + w(n_1 - 1)][w^*(n_2) + w^*(n_2 - 1)]] \quad (9.44)$$

$$= r_{ww}(n_1, n_2) + r_{ww}(n_1, n_2 - 1) + r_{ww}(n_1 - 1, n_2) + r_{ww}(n_1 - 1, n_2 - 1) \quad (9.45)$$

$$= 2\delta(n_1 - n_2) + \delta(n_1 - n_2 + 1) + \delta(n_1 - n_2 - 1) \quad (9.46)$$

$$= 2\delta(l) + \delta(l + 1) + \delta(l - 1), \quad l = n_1 - n_2 \quad (9.47)$$

□

Hence, since $r_{xx}(n_1, n_2) \equiv r_{xx}(l)$ is a function of the difference between n_1 and n_2 only, then $x(n)$ is a WSS sequence. However, it is not an independent process since both $x(n)$ and $x(n+1)$ both depend on $w(n)$.

Example 9.6 (Sum of sinusoids). A discrete-time random process, $g[n]$, is defined as

$$g[n] = A \sin(\omega_0 n) + B \cos(\omega_0 n)$$

where A and B are independent random variables each having zero mean and variance σ^2 , ω_0 is a fixed frequency, and n is the time-index. An example of realisations from this random process are shown in Figure 9.4.

- Determine the mean and autocovariance function of $g[n]$.
- Determine whether or not $g[n]$ is a WSS process. Explain your answer.

SOLUTION. • Noting that the expectation operator is linear:

$$\mu_g[n] = \mathbb{E}[g[n]] = \mathbb{E}[A \sin \omega_0 n] + \mathbb{E}[B \cos \omega_0 n] \quad (9.48)$$

Since $\sin \omega_0 n$ and $\cos \omega_0 n$ are deterministic functions, and $\mathbb{E}[A] = \mathbb{E}[B] = 0$, the expectation simplifies to:

$$\mu_g[n] = \mathbb{E}[A] \sin \omega_0 n + \mathbb{E}[B] \cos \omega_0 n = 0 \quad (9.49)$$

The autocovariance function is given by:

$$\gamma_{gg}[n_1, n_2] = \mathbb{E}[(g[n_1] - \mu_g[n_1])(g[n_2] - \mu_g[n_2])] \quad (9.50)$$

Hence, since $\mu_g[n_i] = 0$, it follows:

$$\gamma_{gg}[n_1, n_2] = \mathbb{E}[(A \sin \omega_0 n_1 + B \cos \omega_0 n_1)(A \sin \omega_0 n_2 + B \cos \omega_0 n_2)] \quad (9.51)$$

$$\begin{aligned} &= \mathbb{E}[A^2] \sin \omega_0 n_1 \sin \omega_0 n_2 + \mathbb{E}[AB] \sin \omega_0 n_1 \cos \omega_0 n_2 \\ &\quad + \mathbb{E}[BA] \cos \omega_0 n_1 \sin \omega_0 n_2 + \mathbb{E}[B^2] \cos \omega_0 n_1 \cos \omega_0 n_2 \end{aligned} \quad (9.52)$$

Since A and B are independent random variables (RVs), $\mathbb{E}[AB] = \mathbb{E}[BA] = \mathbb{E}[A] \mathbb{E}[B] = 0 \times 0 = 0$. Noting $\text{var}[A] = \text{var}[B] = \sigma^2$ and that

$$\text{var}[A] = \mathbb{E}[A^2] - \mathbb{E}^2[A] \quad (9.53)$$

means that $\mathbb{E}[A^2] = \mathbb{E}[B^2] = \sigma^2$. Thus,

$$\gamma_{gg}[n_1, n_2] = \sigma^2 (\sin \omega_0 n_1 \sin \omega_0 n_2 + \cos \omega_0 n_1 \cos \omega_0 n_2) \quad (9.54)$$

Using the supplied trigonometric identity, it follows that:

$$\gamma_{gg}[n_1, n_2] = \sigma^2 \cos \omega_0 (n_1 - n_2) \quad (9.55)$$

- For a process to be WSS, the mean and variance must be constant, and the ACS a function of the time difference or lag $\ell = n_1 - n_2$. The ACS is thus also given by:

$$r_{gg}[n_1, n_2] = \gamma_{gg}[n_1, n_2] + \mu_g[n_1] \mu_g[n_2] \quad (9.56)$$

$$= \sigma^2 \cos \omega_0 (n_1 - n_2) \quad (9.57)$$

□

Thus, it can be seen the mean is constant, and the ACS is a function of the time difference $n_1 - n_2$ only. Therefore it is WSS.

Example 9.7 ([Manolakis:2000, Example 3.3.2, Page 103]: Wiener Process). A coin is tossed at each $n \in \mathbb{Z}$. Let:

$$w[n] = \begin{cases} +S & \text{if heads is the outcome, with probability } \Pr(H) = p \\ -S & \text{if tails is the outcome, with probability } \Pr(T) = 1 - p \end{cases} \quad (9.58)$$

where S is some arbitrary increment or step size in the process $w[n]$. Since $w[n]$, for a given n , is a discrete-random variable taking on two possible values (either S or $-S$), then $w[n]$ is an independent random process with mean:

$$\mathbb{E}[w[n]] = S \Pr(H) + (-S) \Pr(T) \quad (9.59)$$

$$\mu_w = Sp + (-S)(1 - p) = S(2p - 1) \quad (9.60)$$

and second moment:

$$\mathbb{E} [w^2[n]] = \sigma_w^2 + \mu_w^2 \quad (9.61)$$

$$= S^2 \Pr(H) + (-S)^2 \Pr(T) \quad (9.62)$$

$$= S^2 p + S^2 (1 - p) = S^2 \quad (9.63)$$

This in turn means that the autocorrelation function for $w[n]$ is given by:

$$r_{ww}[n, m] = \begin{cases} S^2 & \text{if } n = m \\ \mu_w^2 = S^2 (2p - 1)^2 & \text{if } n \neq m \end{cases} \quad (9.64)$$

Not only is the process $w[n]$ an i. i. d. process, it is also SSS, and therefore, it is also WSS.

Now, define a new random process $x[n]$, $n \geq 1$, as:

$$x[1] = w[1] \quad (9.65)$$

$$x[2] = x[1] + w[2] = w[1] + w[2] \quad (9.66)$$

$$\vdots \quad (9.67)$$

$$x[n] = x[n - 1] + w[n] \quad (9.68)$$

$$= \sum_{k=1}^n w[k] \quad (9.69)$$

Note that $x[n]$ is a running or cumulative sum of independent increments; this is known as an **independent increment process**. Such a sequence is called a **discrete Wiener process** or **random walk**. It can easily be seen that the mean is given by:

$$\mu_x[n] = \mathbb{E} [x[n]] = \mathbb{E} \left[\sum_{k=1}^n w[k] \right] \quad (9.70)$$

$$= n S (2p - 1) \quad (9.71)$$

The variance of $x[n]$ is given by:

$$\sigma_x^2[n] = \mathbb{E} [x^2[n]] - \mu_x^2[n] = \mathbb{E} \left[\sum_{k=1}^n w[k] \sum_{\ell=1}^n w[\ell] \right] - \mu_x^2[n] \quad (9.72)$$

$$= \mathbb{E} \left[\sum_{k=1}^n \sum_{\ell=1}^n w[k] w[\ell] \right] - \mu_x^2[n] = \sum_{k=1}^n \sum_{\ell=1}^n r_{ww}[k - \ell] - \mu_x^2[n] \quad (9.73)$$

$$= \sum_{k=1}^n [S^2 + (n - 1) S^2 (2p - 1)^2] - (n S (2p - 1))^2 \quad (9.74)$$

$$= n S^2 + (n(n - 1) - n^2) S^2 (2p - 1)^2 = [1 - (2p - 1)^2] n S^2 \quad (9.75)$$

$$= 4p(1 - p) n S^2 \quad (9.76)$$

Therefore, the random walk is a nonstationary (or evolutionary) process with a mean and variance that grows linearly with n , the number of steps taken.

It is worth noting that finding the autocorrelation the process $x[n]$ is somewhat more involved, as it involves a calculation involving different limits in each summation:

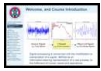
$$\mathbb{E} [x[n] x[m]] = \sum_{k=1}^n \sum_{\ell=1}^m \mathbb{E} [w[k] w[\ell]] \quad (9.77)$$

✘

Substituting the expression for $r_{ww}[k, \ell]$, and rearranging will give the desired answer. This is left as an exercise to the reader, but note that you will need to consider the cases when $m < n$ and $n \geq m$.

– End-of-Topic 60: **Overview of types of stationary processes, and examples of WSS processes** –





New slide

9.5.3 Autocorrelation properties for WSS processes

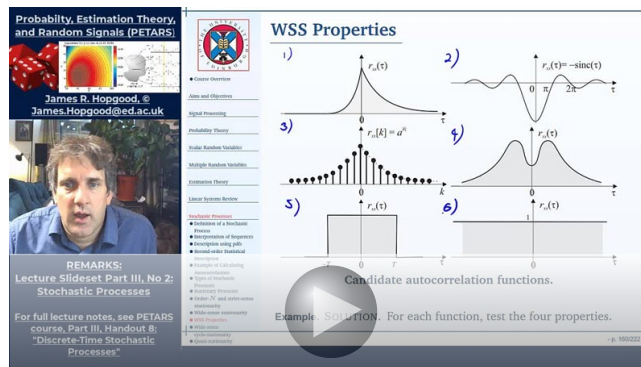
Topic Summary 61 Properties of autocorrelation sequences of WSS processes

Topic Objectives:

- Properties of autocorrelation sequence for wide-sense stationary processes.
- Testing the positive semi-definite property in the lag domain but also frequency domain.
- Examples of testing these properties on sequences and functions.

Topic Activities:

Type	Details	Duration	Progress
Watch video	21 : 10 min video	3 × length	
Read Handout	Read page 335 to page 339	8 mins/page	
Try Example	Try Examples 9.8 and Example 9.9.	15 mins	
Practice Exercises	Exercises ??, ??, and ??	75 mins	



http://media.ed.ac.uk/media/1_6n3mjxwo

Video Summary: Second-order statistics are fundamental to the definition of WSS processes, and this video considers the properties that a WSS ACS must satisfy. Some of properties primarily derive from the key property that the ACS must be positive semi-definite, but other basic ones include the symmetrical property of the ACS, that a random variable cannot be more correlated with another random variable than itself, and that the second moment must always be positive. The video then considers a couple of examples, which tests whether a particular sequence or function is indeed valid.

The average power of a WSS process $x[n]$ satisfies:

$$r_{xx}[0] = \sigma_x^2 + |\mu_x|^2 \geq 0 \tag{M:3.3.27}$$

$$r_{xx}[0] \geq |r_{xx}[\ell]|, \quad \text{for all } \ell \tag{M:3.3.28}$$

The expression for power can be broken down as follows:

Average DC Power: $|\mu_x|^2$

Average AC Power: σ_x^2

Total average power: $r_{xx}[0] \geq 0$

In other words,

$$\text{Total average power} = \text{Average DC power} + \text{Average AC power} \quad (\text{M:3.3.27})$$

To prove $r_{xx}[0] \geq |r_{xx}[\ell]|$, observe that $\mathbb{E}[|x[n+\ell] \pm x[n]|^2] \geq 0$. On expansion, this yields the desired result; this is left as an exercise to the reader, see [Manolakis:2000, Exercise 3.21, Page 145].

Moreover, it follows that $\gamma_{xx}[0] \geq |\gamma_{xx}[\ell]|$.

It is also intuitively obvious, since the autocorrelation of a function should be maximum when it is “self-aligned” with itself. This property is also useful for **template-matching** time-series; i.e. to find which of a particular set of realisations is *most like* a given separate realisation.

It is left as an exercise to show that the ACS $r_{xx}[\ell]$ satisfies two more properties, namely it is:

- a conjugate symmetric function of the lag ℓ :

$$r_{xx}^*[-\ell] = r_{xx}[\ell] \quad (\text{M:3.3.29})$$

- a **nonnegative-definite** or **positive semi-definite** function, such that for any sequence $\alpha[n]$:

$$\sum_{n=1}^M \sum_{m=1}^M \alpha^*[n] r_{xx}[n-m] \alpha[m] \geq 0 \quad (\text{M:3.3.30})$$

Note that, more generally, even a correlation function for a nonstationary random process is **positive semi-definite**:

$$\sum_{n=1}^M \sum_{m=1}^M \alpha^*[n] r_{xx}[n, m] \alpha[m] \geq 0 \quad \text{for any sequence } \alpha[n] \quad (9.78)$$

When dealing with stationary processes, this course will exclusively consider wide-sense stationary (WSS) rather than strict-sense stationary (SSS) processes. Therefore, the term *stationary* will be used to mean WSS from here onwards.

Example 9.8 (Cosinusoid). The function $r[\ell] = \cos \omega_0 \ell$ is claimed to be a valid ACS. Test the properties of this function to determine if this claim is true or not.

SOLUTION. The function $r[\ell] = \cos \omega_0 \ell$ satisfies: the symmetric property, $r[\ell] = r[-\ell]$; the equality $r[0] \geq |r[\ell]|$ for all ℓ ; and $r[0] \geq 0$.

The final property of positive semi-definiteness is a little more tedious to verify. Let:

$$I = \sum_{n=1}^M \sum_{m=1}^M \alpha^*[n] r_{xx}[n-m] \alpha[m] \quad (9.79)$$

$$= \sum_{n=1}^M \sum_{m=1}^M \alpha[n] \alpha[m] \cos \omega_0 (n-m) \quad (9.80)$$

Using the trigonometric identity: $\cos \omega_0 (n-m) = \cos \omega_0 n \cos \omega_0 m + \sin \omega_0 n \sin \omega_0 m$, then consider the resulting first term and using the fact $r[\ell]$ is real:

$$I_1 = \sum_{n=1}^M \sum_{m=1}^M \alpha[n] \alpha[m] \cos \omega_0 n \cos \omega_0 m \quad (9.81)$$

$$= \left(\sum_{n=1}^M \alpha[n] \cos \omega_0 n \right) \left(\sum_{m=1}^M \alpha[m] \cos \omega_0 m \right) \quad (9.82)$$

$$= \left(\sum_{n=1}^M \alpha[n] \cos \omega_0 n \right)^2 \geq 0 \quad (9.83)$$

□

A similar argument can be made for the second term as well, showing that $I \geq 0$. This proof is a little tedious, and can often be more easily shown using the following equivalent result.

KEYPOINT! (Equivalent condition for positive semi-definiteness). The Fourier transform of an autocorrelation sequence (ACS) or autocorrelation function (ACF) is an extremely important concept, called the power spectral density (PSD) which will be discussed in the next handout. It will be proved that the PSD should always be positive. It is easy to prove that an ACS or ACF has a positive Fourier transform if, and only if, it is positive semi-definite.

To prove this result, then writing the inverse discrete-time Fourier transform (DTFT) for $r_{xx}[\ell]$:

$$r_{xx}[\ell] = \frac{1}{2\pi} \int_{-\pi}^{\pi} S(e^{j\omega}) e^{j\omega\ell} d\omega \quad (9.84)$$

Substituting into Equation M:3.3.30 (but not assuming the inequality) gives:

$$I = \sum_{n=1}^M \sum_{m=1}^M \alpha^*[n] \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} S(e^{j\omega}) e^{j\omega(n-m)} d\omega \right\} \alpha[m] \quad (9.85)$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} S(e^{j\omega}) \left\{ \sum_{n=1}^M \sum_{m=1}^M \alpha^*[n] e^{j\omega n} e^{-j\omega m} \alpha[m] \right\} d\omega \quad (9.86)$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} S(e^{j\omega}) \left| \sum_{m=1}^M \alpha[m] e^{-j\omega m} \right|^2 d\omega \quad (9.87)$$

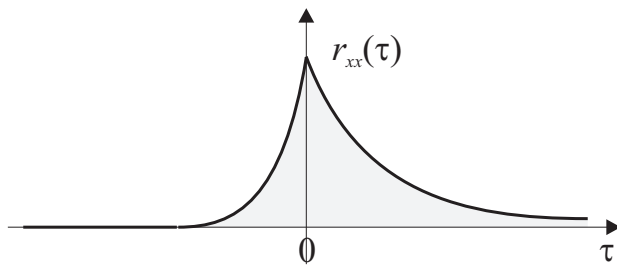
$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} S(e^{j\omega}) |A(e^{j\omega})|^2 d\omega \quad (9.88)$$

where $\alpha[n] \stackrel{\text{DTFT}}{\rightleftharpoons} A(e^{j\omega})$ are DTFT pairs. Since $S(e^{j\omega}) \geq 0$, then so is $I \geq 0$.

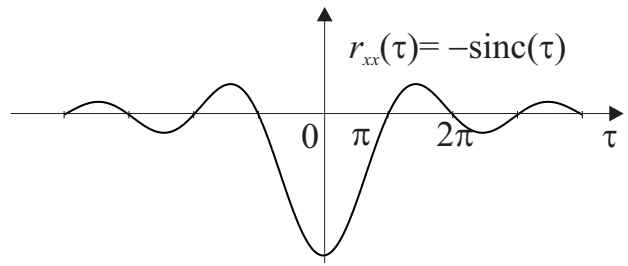
Example 9.9. Consider the functions shown in Figure 9.5. For each function, state whether it is a valid autocorrelation function or autocorrelation sequence or not. Explain carefully the reasoning for your answers, but no detailed calculations are required.

SOLUTION. Consider each function or sequence in turn. For each function or sequence, test the four properties.

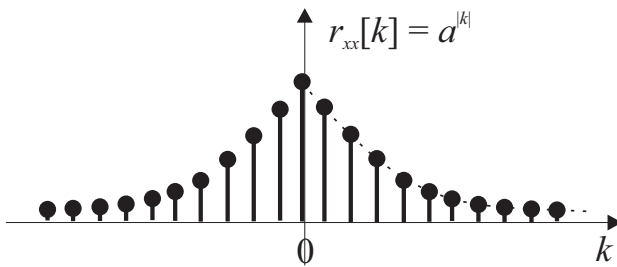
1. The first function violates the symmetry rule.
2. The second function violates for property that $r_{xx}[0] \geq 0$ in order to have positive power.
3. This function is valid, as it is symmetric, satisfies the positive power condition, that the largest ACS value occurs at zero-lag. The final condition of testing the positive semi-definiteness is most easily done by noting that the Fourier transform of the ACS is positive, but this is left as an exercise to the reader.
4. This violates the property that $r_{xx}[0] \geq |r_{xx}[\ell]|$ for all ℓ .
5. The Fourier transform of this function is not always positive.
6. This function satisfies all the properties and is therefore valid.



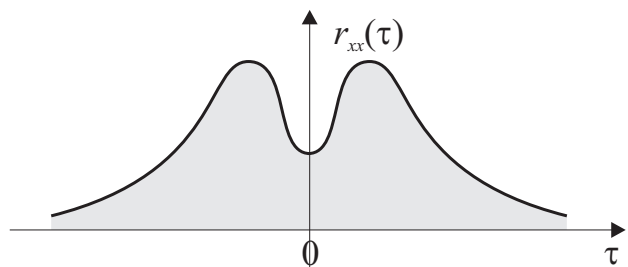
(a) Two-sided exponential function, where $r_{xx}(\tau) = e^{2\tau}$ if $\tau < 0$ and $r_{xx}(\tau) = e^{-\tau}$ if $\tau \geq 0$.



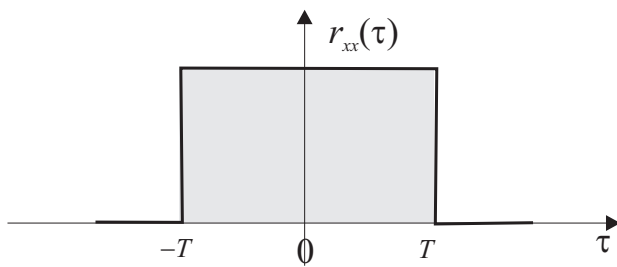
(b) Inverted sinc function ($\text{sinc } \tau = \frac{\sin \tau}{\tau}$).



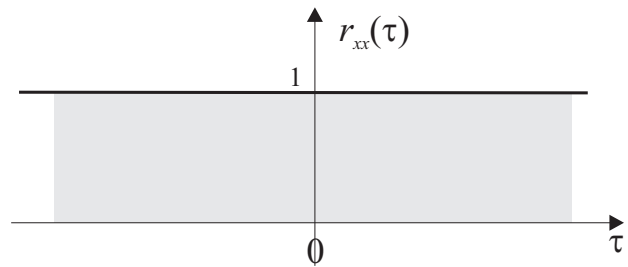
(c) Discrete exponential decay.



(d) Double-peaked function.



(e) Rectangular function.



(f) Constant value.

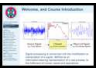
Figure 9.5: Candidate autocorrelation functions.

Thus, in summary, the claims are satisfied as follows: 1)-2) and 4)-5), No; 3) and 6) Yes!

– End-of-Topic 61: **Properties of the ACS for WSS** –



9.5.4 Wide-sense periodicity and cyclo-stationarity



Topic Summary 62 Wide-sense periodic, wide-sense cyclo-stationary, and quasi-stationary processes New slide

Topic Objectives:

- Concept of nonstationary process that have structured second-order statistics.
- Definition of wide-sense periodic (WSP) and wide-sense cyclo-stationary processes.
- Example of wide-sense cyclostationary process in a communications system.
- Notion of quasi-stationary processes.

Topic Activities:

Type	Details	Duration	Progress
Watch video	17 : 33 min video	3 × length	
Read Handout	Read page 340 to page 344	8 mins/page	
Try Example	Try Example 9.10	10 mins	

Wide-sense cyclo-stationarity

An example pulse and typical transmit signal.

Example (Pulse-Amplitude Modulation). An important example of a cyclo-stationary process is the random signal:

$$\text{Transmitted waveform } \rightarrow x[n] = \sum_{m=-\infty}^{\infty} c_m h[n - mT]$$

for some period T , where c_m is a stationary sequence with ACS $r_{c_m}(n_1, n_2) = E\{c_{n_1} c_{n_2}^*\} = r_{c_m}(n_1 - n_2)$, and $h[n]$ is a given deterministic sequence, usually an impulse response.

http://media.ed.ac.uk/media/1_tql3v66m

Video Summary: This video considers a wider class of nonstationary processes that share some similarities with WSS processes. Such nonstationary processes occur in systems where, for example, there is some aspect of upsampling, or a random process generates a new process that is a function of some deterministic signal that has temporal extent. This video looks at wide-sense periodic and wide-sense cyclo-stationary processes. An example of pulse-amplitude modulation is presented. Finally, globally non-stationary but locally-stationary processes are discussed, called quasi-stationary processes. The application of speech modelling is considered as an example of a quasi-stationary process.

A signal whose statistical properties vary *cyclically* with time is called a cyclostationary process. A cyclostationary process can be viewed as several interleaved stationary processes. For example, the maximum daily temperature in Edinburgh can be modeled as a cyclostationary process: the maximum temperature on July 21 is statistically different from the temperature on December 18; however, the temperature on December 18 of different years has (arguably) identical statistics (although unfortunately there seems to be a growing trend).

Two classes of **cyclostationary signals** that are actually **nonstationary process** which, in part, have

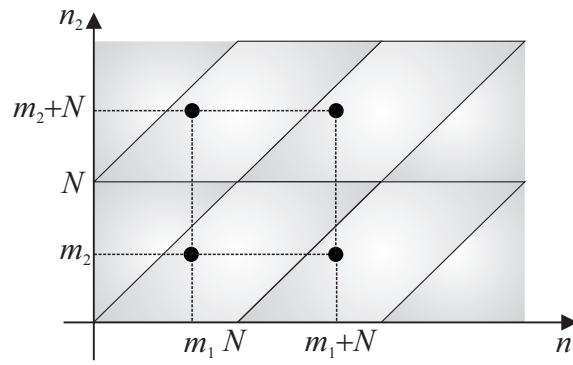


Figure 9.6: The periodicity of the ACS for a WSP signal.

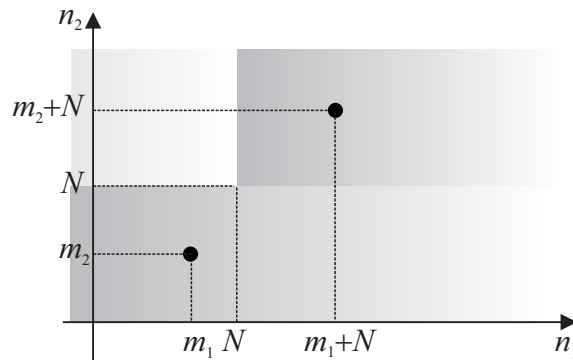


Figure 9.7: The periodicity of the ACS for a wide-sense cyclo-stationary process.

properties resembling stationary signals are:

1. A **WSP process** is classified as signals whose mean is periodic, and whose ACS is periodic in both dimensions:

$$\mu_x[n] = \mu_x[n + N] \tag{M:3.3.14}$$

$$\begin{aligned} r_{xx}[n_1, n_2] &= r_{xx}[n_1 + N, n_2] = r_{xx}[n_1, n_2 + N] \\ &= r_{xx}[n_1 + N, n_2 + N] \end{aligned} \tag{M:3.3.15}$$

for all n, n_1 and n_2 . These are quite tight constraints for practical signals.

2. A **wide-sense cyclo-stationary process** has similar but less restrictive properties than a WSP process, in that the mean is periodic, but the ACS is now just invariant to a shift by N in both of its arguments:

$$\mu_x[n] = \mu_x[n + N] \tag{M:3.3.16}$$

$$r_{xx}[n_1, n_2] = r_{xx}[n_1 + N, n_2 + N] \tag{M:3.3.17}$$

for all n, n_1 and n_2 .

Example 9.10 (Pulse-Amplitude Modulation). An important example of a cyclo-stationary process is the random signal:

$$x[n] = \sum_{m=-\infty}^{\infty} c_m h[n - mT] \tag{9.89}$$

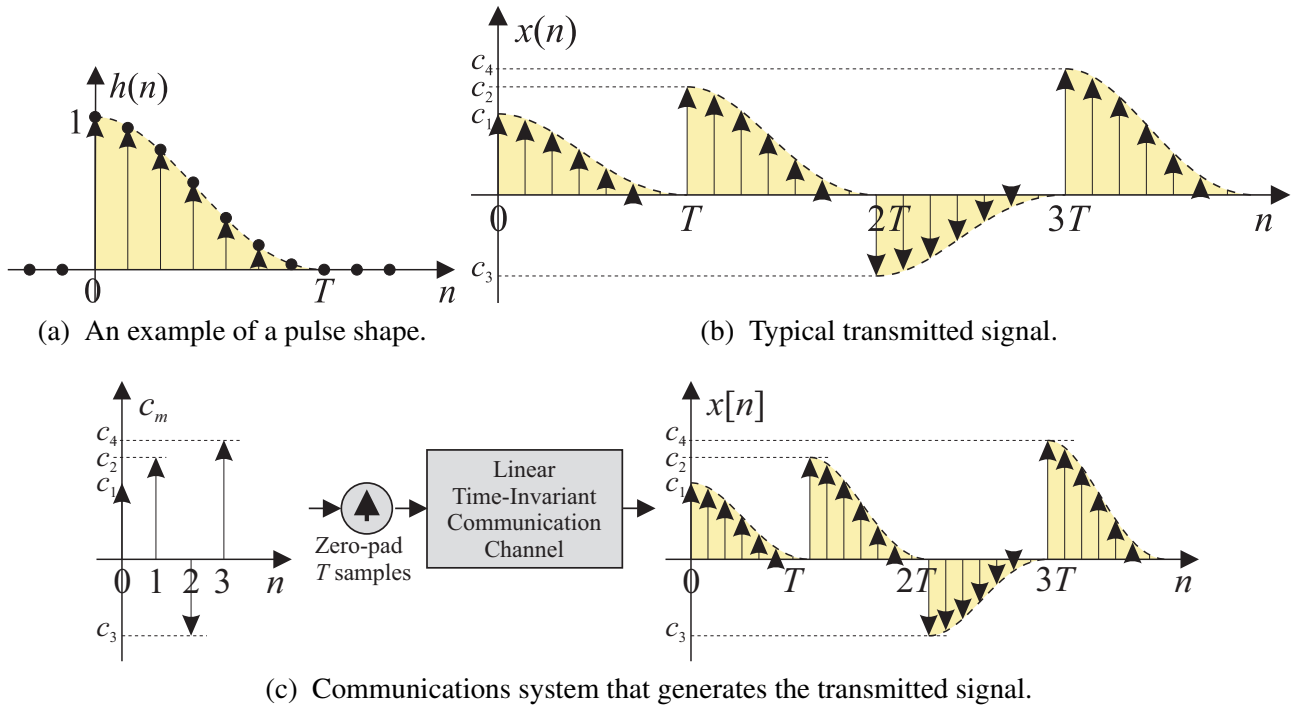


Figure 9.8: An example pulse shape and typical transmit signal in a communications system.

for some period T , and where c_m is a stationary sequence of RVs with ACS $r_{cc}[n_1, n_2] = \mathbb{E}[c_{n_1} c_{n_2}^*] = r_{cc}[n_1 - n_2]$, and $h[n]$ is a given deterministic sequence, usually an impulse response. An example of a particular pulse shape for $h[n]$ and a typical sequence $x[n]$ is shown in Figure 9.8.

Show that $x[n]$ satisfies the properties of a wide-sense cyclo-stationary process.

SOLUTION. The stochastic process $x[n]$ represents the signal for several different types of linear modulation techniques used in digital communication systems. The sequence $\{c_m\}$ represents the digital information (of symbols) that is transmitted over the communication channel, and $\frac{1}{T}$ represents the rate of transmission of the information symbols.

Note that this example demonstrates why notation can become an issue: how is it possible to determine that c_n is a RV, while $h[n]$ is not?

To see that this is a wide-sense cyclo-stationary process, first begin by writing:

$$\mu_x[n] = \mathbb{E}[x[n]] = \sum_{m=-\infty}^{\infty} \mathbb{E}[c_m] h[n - mT] = \mu_c \sum_{m=-\infty}^{\infty} h[n - mT] \quad (9.90)$$

where $\mu_c[n] = \mu_c$ since it is a stationary process. Thus, observe that:

$$\mu_x[n + kT] = \mu_c \sum_{m=-\infty}^{\infty} h[n + kT - mT] = \mu_c \sum_{r=-\infty}^{\infty} h[n - Tr] = \mu_x[n] \quad (9.91)$$

by a change of variables $r = m - k$.

Next consider the autocorrelation function given by:

$$\begin{aligned} r_{xx}[n_1, n_2] &= \mathbb{E}[x[n_1] x^*[n_2]] \\ &= \sum_{m=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} h[n_1 - mT] h^*[n_2 - T\ell] r_{cc}[m - \ell] \end{aligned} \quad (9.92)$$

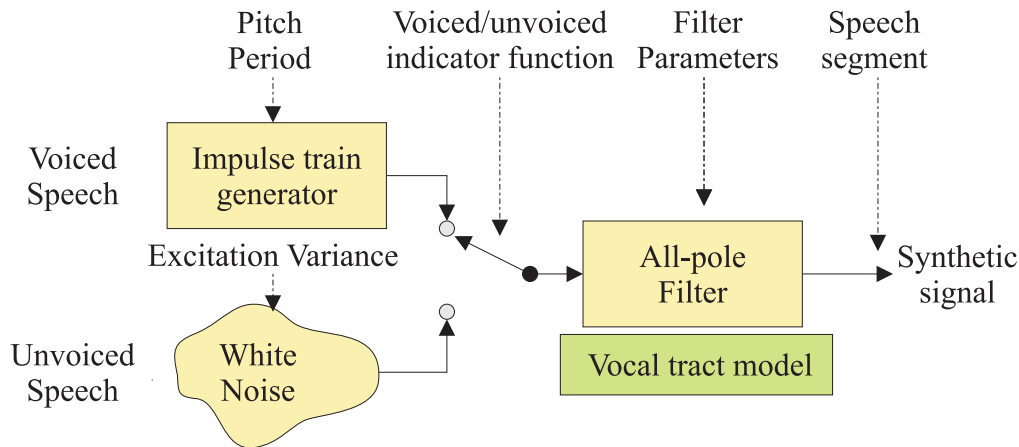


Figure 9.9: The speech synthesis model (repeated from Introduction handout).

where it has been noted that $r_{cc}[m, \ell] = \mathbb{E} [c_m c_\ell^*] = r_{cc}[m - \ell]$ since it is a stationary process. Similar to the approach with the mean above, then set $n_1 \rightarrow n_1 + pT$ and $n_2 \rightarrow n_2 + qT$.

Therefore, it follows:

$$\begin{aligned}
 & r_{xx}[n_1 + pT, n_2 + qT] \\
 &= \sum_{m=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} h[n_1 - T(m - p)] h[n_2 - T(\ell - q)] r_{cc}[m - \ell]
 \end{aligned} \tag{9.93}$$

Again, by the change of variables $r = m - p$ and $s = \ell - q$, it can be seen that:

$$\begin{aligned}
 & r_{xx}[n_1 + pT, n_2 + qT] \\
 &= \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} h[n_1 - Tr] h[n_2 - Ts] r_{cc}[r - s + p - q]
 \end{aligned} \tag{9.94}$$

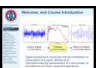
In the case that $p = q$, then comparing Equation 9.92 and Equation 9.94, it finally follows that:

$$r_{xx}[n_1 + pT, n_2 + pT] = r_{xx}[n_1, n_2] \tag{9.95}$$

□

By definition, $x[n]$ is therefore a cyclo-stationary process.

9.5.5 Local- or quasi-stationary processes



At the introduction of this lecture course, it was noted that in the analysis of speech signals, the speech waveform is broken up into short segments whose duration is typically 10 to 20 milliseconds. New slide

This is because speech can be modelled as a **locally stationary** or **quasi-stationary** process. Such processes possess statistical properties that change *slowly* over short periods of time. They are *globally* nonstationary, but are approximately *locally* stationary, and are modelled as if the statistics *actually are* stationary over a short segment of time.

Quasi-stationary models are, in fact, just a special case of nonstationary processes, but are distinguished since their characterisation closely resemble stationary processes.

– End-of-Topic 62: **Wide-sense periodic and cyclostationary signals, and other forms of nonstationary signals** –



9.6 Estimating statistical properties

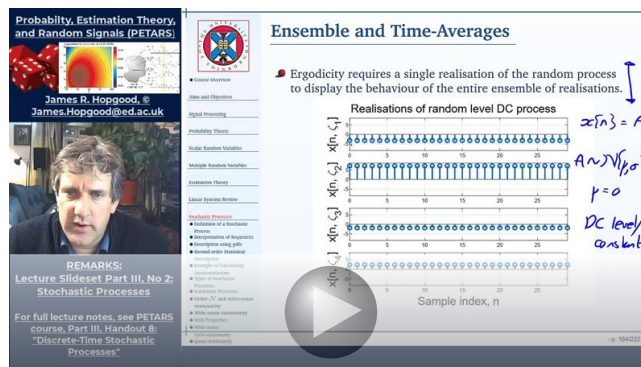
Topic Summary 63 Time-Averages and Ergodicity

Topic Objectives:

- Concept of estimating statistical averages from a single realisation of a stochastic process.
- Introduction to the notion of ergodicity and estimating ensemble averages from time-averages.
- Examples of testing if a process is ergodic or not.

Topic Activities:

Type	Details	Duration	Progress
Watch video	26 : 59 min video	3 × length	
Read Handout	Read page 345 to page 350	8 mins/page	
Try Example	Try Example 9.11.	15 mins	



http://media.ed.ac.uk/media/1_1nebqv6a

Video Summary: This Topic introduces the notion of estimating statistical averages from a single realisation of a stochastic process. This concept is most easily developed for estimating first and second moments of stationary random processes using time-averages. This requires the process to be Ergodic and WSS. The video first introduces ergodicity from an intuitive perspective, and then further expands the definition in terms of using the properties of a consistent estimator. This is expressed through the two definitions of ergodic in the mean, or ergodic in correlation. Examples of non-ergodic and ergodic processes are presented. One very detailed example proves a process is ergodic in the mean through calculating the bias and variance of the time-average.

- A stochastic process consists of the ensemble, $x[n, \zeta]$, and a probability law, $f_X(\{x\} | \{n\})$. If this information is available $\forall n$, the statistical properties are easily determined.
- In practice, only a limited number of realisations of a process is available, and often only one: i.e. $\{x[n, \zeta_k], k \in \{1, \dots, K\}\}$ is known for some K , but $f_X(x | n)$ is unknown.
- Is it possible to infer the statistical characteristics of a process from a single realisation? Yes, for the following class of signals:

– **ergodic** processes;

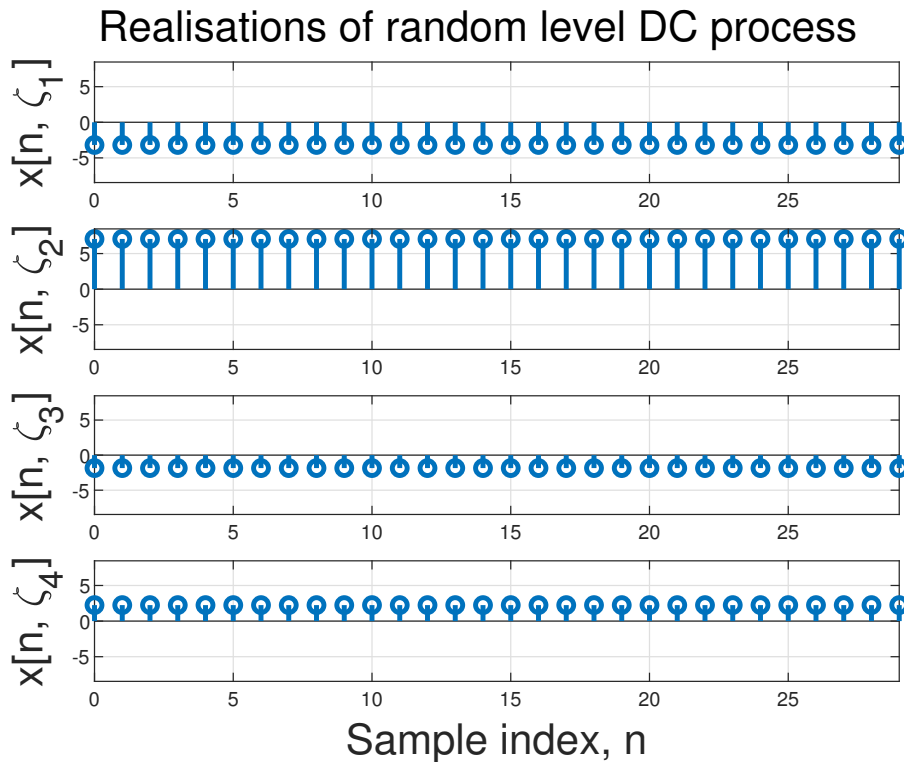


Figure 9.10: The temporal-variability of the DC level does not capture the ensemble statistics.

- nonstationary processes where additional structure about the autocorrelation function is known (beyond the scope of this course).

9.6.1 Ensemble and Time-Averages

Ensemble averaging, as considered so far in the course, is not frequently used in practice since it is impractical to obtain the number of realisations needed for an accurate estimate.

A statistical average that can be obtained from a **single** realisation of a process is a **time-average**, defined by:

$$\langle g(x[n]) \rangle \triangleq \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N g(x[n]) \quad (\text{M:3.3.32})$$

For every ensemble average, a corresponding time-average can be defined; the time-average above corresponds to: $\mathbb{E}[g(x[n])]$.

Time-averages are random variables since they implicitly depend on the particular realisation, given by ζ . Averages of deterministic signals are fixed numbers or sequences, even though they are given by the same expression.

It should be intuitive that ergodicity requires a single realisation of the random process to display the behaviour of the entire ensemble of realisations. If not, ergodicity will not hold.

9.6.2 Ergodicity

A stochastic process, $x[n]$, is **ergodic** if its ensemble averages can be estimated from a single realisation of a process using time averages.

The two most important degrees of ergodicity are:

Mean-Ergodic (or ergodic in the mean) processes have identical expected values and sample-means:

$$\langle x[n] \rangle = \mathbb{E} [x[n]] \quad (\text{M:3.3.34})$$

Covariance-Ergodic Processes (or ergodic in correlation) have the property that:

$$\langle x[n] x^*[n-l] \rangle = \mathbb{E} [x[n] x^*[n-l]] \quad (\text{M:3.3.35})$$

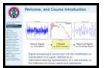
Another form of ergodicity is a **distribution-ergodic** process, but this will not be discussed here.

- It should be intuitiveness obvious that ergodic processes must be stationary and, moreover, that a process which is ergodic both in the mean and correlation is WSS.
- WSS processes are not necessarily ergodic.
- Ergodic is often used to mean both ergodic in the mean and correlation.
- In practice, only finite records of data are available, and therefore an estimate of the time-average will be given by

$$\langle g(x[n]) \rangle = \frac{1}{N} \sum_{n \in \mathcal{N}} g(x[n]) \quad (\text{M:3.3.37})$$

where N is the number of data-points available.

9.6.3 More Details on Mean-Ergodicity



Returning to the definition of mean-ergodicity, a little more detail of conditions on the random process *New slide* is given.

The **time-average** over $2N + 1$ samples, $\{x[n]\}_{-N}^N$ is given by:

$$\mu_x|_N = \langle x[n] \rangle = \frac{1}{2N + 1} \sum_{n=-N}^N x[n] \quad (9.96)$$

Clearly, $\mu_x|_N$ is a random variable with mean:

$$\mathbb{E} [\mu_x|_N] = \frac{1}{2N + 1} \sum_{n=-N}^N \mathbb{E} [x[n]] = \mu_x \quad (9.97)$$

since $x[n]$ is a stationary stochastic process. As is seen elsewhere in these lectures, this is known as an **unbiased estimate** since the **sample mean** is equal to the **ensemble mean**.

Since $\mu_x|_N$ is a random variable, then it must have a variance as well:

$$\text{var} [\mu_x|_N] = \text{var} \left[\frac{1}{2N + 1} \sum_{n=-N}^N x[n] \right] \quad (9.98)$$

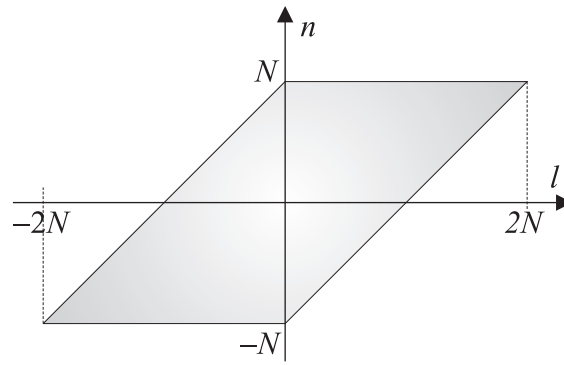


Figure 9.11: Region of summation for deriving the variance of the time-average.

Theorem 9.1 (Variance of estimator). Suppose the sample mean of a sequence of a WSS process, $x[n]$, is given by:

$$\mu_x|_N = \frac{1}{2N+1} \sum_{n=-N}^N x[n] \quad (9.99)$$

If the process $x[n]$ has ACS $\gamma_{xx}[\ell]$, then the variance of the sample mean can be expressed as:

$$\text{var} [\mu_x|_N] = \frac{1}{2N+1} \sum_{\ell=-2N}^{2N} \left(1 - \frac{|\ell|}{2N+1}\right) \gamma_{xx}[\ell] \quad (9.100)$$

PROOF. Noting the mean of the expression in the square brackets on the right hand side (RHS) of Equation 9.98 is equal to μ_x , then:

$$\text{var} [\mu_x|_N] = \frac{1}{(2N+1)^2} \mathbb{E} \left[\sum_{n=-N}^N \sum_{m=-N}^N x[n] x^*[m] \right] - \mu_x^2 \quad (9.101)$$

$$= \frac{1}{(2N+1)^2} \left\{ \sum_{n=-N}^N \sum_{m=-N}^N r_{xx}[n-m] \right\} - \mu_x^2 \quad (9.102)$$

since $x[n]$ is a stationary process, and therefore its ACS only depends on the time difference. With a little manipulation, then noting that the autocovariance is given by $\gamma_{xx}[\ell] = r_{xx}[\ell] - \mu_x^2$, it follows that:

$$\text{var} [\mu_x|_N] = \frac{1}{(2N+1)^2} \sum_{n=-N}^N \sum_{m=-N}^N \gamma_{xx}[n-m] \quad (9.103)$$

A change of variable can now be performed by setting $\ell = n - m$. Hence:

$$\text{var} [\mu_x|_N] = \frac{1}{(2N+1)^2} \sum_{n=-N}^N \sum_{\ell=n-N}^{n+N} \gamma_{xx}[\ell] \quad (9.104)$$

The region of summation is shown in Figure 9.11.

Thus, the next step is to change the order of summation (as this is the usual trick), and so considering

the region of summation, then summing l first:

$$\text{var} [\mu_x|_N] = \frac{1}{(2N+1)^2} \sum_{\ell=-2N}^{2N} \sum_{n=\max\{-N, \ell-N\}}^{\min\{N, \ell+N\}} \gamma_{xx}[\ell] \quad (9.105)$$

$$= \frac{1}{(2N+1)^2} \sum_{\ell=-2N}^{2N} (2N+1-|\ell|) \gamma_{xx}[\ell] \quad (9.106)$$

$$= \frac{1}{2N+1} \sum_{\ell=-2N}^{2N} \left(1 - \frac{|\ell|}{2N+1}\right) \gamma_{xx}[\ell] \quad (9.107)$$

□

as required.

KEYPOINT! (Mean-ergodic). If the variance $\lim_{N \rightarrow \infty} \text{var} [\mu_x|_N] = 0$, then $\mu_x|_N \rightarrow \mu_x$ in the mean-square sense. In this case, it is said that the time average $\mu_x|_N$ computed from a single realisation of $x[n]$ is close to μ_x with probability close to 1. If this is true, then the technical definition is that the process $x[n]$ is **mean-ergodic**.

The result presented above leads to the following conclusion:

Theorem 9.2 (Mean-ergodic processes). A discrete-random process $x[n]$ with autocovariance $\gamma_{xx}[\ell]$ is mean-ergodic iff:

$$\lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{\ell=-2N}^{2N} \left(1 - \frac{|\ell|}{2N+1}\right) \gamma_{xx}[\ell] = 0 \quad (9.108)$$

PROOF. See discussion above.

Example 9.11 ([Papoulis:1991, Example 13.3, Page 429]). A stationary stochastic process $x[n]$ has an ACS given by $\gamma_{xx}[\ell] = q e^{-c|\ell|}$ for some constants q and c . Is the process $x[n]$ ergodic in the mean?

SOLUTION. Writing:

$$\text{var} [\mu_x|_N] = \frac{1}{2N+1} \sum_{\ell=-2N}^{2N} \left(1 - \frac{|\ell|}{2N+1}\right) \gamma_{xx}[\ell] \quad (9.109)$$

$$= \frac{q}{2N+1} \sum_{\ell=-2N}^{2N} \left(1 - \frac{|\ell|}{2N+1}\right) e^{-c|\ell|} \quad (9.110)$$

which can be rearranged to give as:

$$\text{var} [\mu_x|_N] = \frac{q}{2N+1} \left\{ 2 \sum_{\ell=0}^{2N} \left(1 - \frac{\ell}{2N+1}\right) e^{-c\ell} - 1 \right\} \quad (9.111)$$

Now, noting the general result which can be found in mathematical tables [Gradshteyn:1994]:

$$\sum_{n=0}^{N-1} (a+nb)r^n = \frac{a - [a + (N-1)b]r^N}{1-r} + \frac{br(1-r^{N-1})}{(1-r)^2}, \quad r \neq 0, N > 1 \quad (9.112)$$

then by setting $a = 1$, $b = -\frac{1}{2N+1}$ and $r = e^{-c}$, with $n = \ell$ and $N \rightarrow 2N + 1$:

$$\text{var} [\mu_x|_N] = 2q \left[\frac{\frac{1}{M} - \frac{1}{M^2}e^{-Mc}}{1 - e^{-c}} + \frac{\frac{1}{M^2}e^{-c} - \frac{1}{M^2}e^{-Mc}}{(1 - e^{-c})^2} - \frac{1}{2M} \right] \quad (9.113)$$

where $M = 2N + 1$. Now, by setting $N \rightarrow \infty$, which is equivalent to $M \rightarrow \infty$, and noting the relationship that:

$$\lim_{n \rightarrow \infty} n^s x^n \rightarrow 0 \quad \text{if } |x| < 1 \text{ for any real value of } s \quad (9.114)$$

it can easily be seen that

$$\lim_{N \rightarrow \infty} \text{var} [\mu_x|_N] = 0 \quad (9.115)$$

□

and therefore $x[n]$ is mean-ergodic.

– End-of-Topic 63: **Ergodicity and time-average estimates of statistics of WSS processes** –



9.7 Joint Signal Statistics

Topic Summary 64 Joint Signal Statistics and Correlation Matrices

Topic Objectives:

- Extending definitions presented previously to Joint signal statistics.
- Understanding notion of cross-correlation and cross-covariance.
- Application of these techniques to Blind Source Separation.
- Definition and use of Correlation Matrices.

Topic Activities:

Type	Details	Duration	Progress
Watch video	19 : 21 min video	3× length	
Read Handout	Read page 351 to page 354	8 mins/page	
Try Example	Try Example 9.12.	15 mins	
Practice Exercises	Exercises ?? to ??	75 mins	

http://media.ed.ac.uk/media/1_smlhq601

Video Summary: This video starts to wrap up the Chapter on Stochastic processes by looking at joint signal statistics, such as cross-correlation and cross-covariance, uncorrelated pairs of random processes, and an extension of the various concepts previously developed for analysing random processes. An example is presented of using cross-covariance as a surrogate for measuring independence of signals in the classic signal processing problem of blind source separation. Finally, the Topic introduces the use of correlation matrices for analysing a finite-block or window of samples. Correlation matrices are a convenient way of representing signal statistics when it comes to creating real signal processing algorithms.

Next, it is important to consider the dependence between two different random processes, and these follow similar definitions to those introduced for random vectors. In this section, consider the interaction between two random processes $x[n]$ and $y[n]$.

Cross-correlation and cross-covariance A measure of the dependence between values of two *different* stochastic processes is given by the **cross-correlation** and **cross-covariance**

functions:

$$r_{xy}[n_1, n_2] = \mathbb{E} [x[n_1] y^*[n_2]] \quad (\text{M:3.3.7})$$

$$\gamma_{xy}[n_1, n_2] = r_{xy}[n_1, n_2] - \mu_x[n_1] \mu_y^*[n_2] \quad (\text{M:3.3.8})$$

Normalised cross-correlation (or cross-covariance) The cross-covariance provides a measure of similarity of the deviation from the respective means of two processes. It makes sense to consider this deviation relative to their **standard deviations**; thus, **normalised cross-correlations**:

$$\rho_{xy}[n_1, n_2] = \frac{\gamma_{xy}[n_1, n_2]}{\sigma_x[n_1] \sigma_y[n_2]} \quad (\text{M:3.3.9})$$

9.7.1 Types of Joint Stochastic Processes

The definitions introduced earlier for a single stochastic process can be extended to the case of two joint stochastic processes:

Statistically independence of two stochastic processes occurs when, for every n_x and n_y ,

$$f_{XY}(x, y | n_x, n_y) = f_X(x | n_x) f_Y(y | n_y) \quad (\text{M:3.3.18})$$

Uncorrelated stochastic processes have, for all n_x & $n_y \neq n_x$:

$$\begin{aligned} \gamma_{xy}[n_x, n_y] &= 0 \\ r_{xy}[n_x, n_y] &= \mu_x[n_x] \mu_y[n_y] \end{aligned} \quad (\text{M:3.3.19})$$

Joint stochastic processes that are statistically independent are uncorrelated, but not necessarily vice-versa, except for Gaussian processes. Nevertheless, a measure of uncorrelatedness is often used as a measure of independence. More on this later.

Further definitions include:

Orthogonal joint processes have, for every n_1 and $n_2 \neq n_1$:

$$r_{xy}[n_1, n_2] = 0 \quad (\text{M:3.3.20})$$

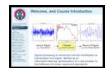
Joint WSS is similar to WSS for a single stochastic process, and is useful since it facilitates a spectral description, as discussed later in this course:

$$r_{xy}[\ell] = r_{xy}[n_1 - n_2] = r_{yx}^*[-\ell] = \mathbb{E} [x[n] y^*[n - \ell]] \quad (9.116)$$

$$\gamma_{xy}[\ell] = \gamma_{xy}[n_1 - n_2] = \gamma_{yx}^*[-\ell] = r_{xy}[\ell] - \mu_x \mu_y^* \quad (9.117)$$

Joint-Ergodicity applies to two ergodic processes, $x[n]$ and $y[n]$, whose ensemble cross-correlation can be estimated from a time-average:

$$\langle x[n] y^*[n - \ell] \rangle = \mathbb{E} [x[n] y^*[n - \ell]] \quad (\text{M:3.3.36})$$



New slide

9.8 Correlation Matrices for Random Processes

A stochastic process can also be represented as a random vector, and its second-order statistics given by the mean vector and the correlation matrix. Obviously these quantities are functions of the index n .

Let an M -dimensional random vector $\mathbf{X}[n, \zeta] \equiv \mathbf{X}[n]$ be derived from the random process $x[n]$ as follows:

$$\mathbf{X}[n] \triangleq [x[n] \quad x[n-1] \quad \cdots \quad x[n-M+1]]^T \quad (\text{M:3.4.56})$$

Then its mean is given by an M -vector

$$\boldsymbol{\mu}_{\mathbf{X}}[n] \triangleq [\mu_x[n] \quad \mu_x[n-1] \quad \cdots \quad \mu_x[n-M+1]]^T \quad (\text{M:3.4.57})$$

and the $M \times M$ correlation matrix is given by:

$$\mathbf{R}_{\mathbf{X}}[n] = \mathbb{E} [\mathbf{X}[n] \mathbf{X}^H[n]] \quad (\text{T:4.23})$$

which can explicitly be written as:

$$\mathbf{R}_{\mathbf{X}}[n] \triangleq \begin{bmatrix} r_{xx}[n, n] & \cdots & r_{xx}[n, n-M+1] \\ \vdots & \ddots & \vdots \\ r_{xx}[n-M+1, n] & \cdots & r_{xx}[n-M+1, n-M+1] \end{bmatrix} \quad (\text{M:3.4.58})$$

Clearly $\mathbf{R}_{\mathbf{X}}[n]$ is Hermitian, since $r_{xx}[n-i, n-j] = \mathbb{E} [x[n-i] x^*[n-j]] = r_{xx}^*[n-j, n-i]$, $0 \leq i, j \leq M-1$. This vector representation can be useful in discussion of optimum filters.

For WSS processes, the correlation matrix has an interesting additional structure. Note that:

1. $\mathbf{R}_{\mathbf{X}}[n]$ is a constant matrix $\mathbf{R}_{\mathbf{X}}$;
2. $r_{xx}[n-i, n-j] = r_{xx}[j-i] = r_{xx}[\ell]$, $\ell = j-i$;
3. conjugate symmetry gives $r_{xx}[\ell] = r_{xx}^*[-\ell]$.

Hence, the matrix \mathbf{R}_{xx} is given by:

$$\mathbf{R}_{\mathbf{X}} \triangleq \begin{bmatrix} r_{xx}[0] & r_{xx}[1] & r_{xx}[2] & \cdots & r_{xx}[M-1] \\ r_{xx}^*[1] & r_{xx}[0] & r_{xx}[1] & \cdots & r_{xx}[M-2] \\ r_{xx}^*[2] & r_{xx}^*[1] & r_{xx}[0] & \cdots & r_{xx}[M-3] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{xx}^*[M-1] & r_{xx}^*[M-2] & r_{xx}^*[M-3] & \cdots & r_{xx}[0] \end{bmatrix} \quad (\text{M:3.4.60})$$

It can easily be seen that $\mathbf{R}_{\mathbf{X}}$ is Hermitian and **Toeplitz**; a **Toeplitz** matrix is one in which the elements along each diagonal, parallel to the main diagonal, are equal. Note that the anti-diagonals are not necessarily equal. Thus, the autocorrelation matrix of a stationary process is Hermitian, nonnegative definite, and Toeplitz.

Example 9.12 (Correlation matrices). The correlation function for a certain random process $x[n]$ has the exponential form:

$$r_{xx}[\ell] = 4(-0.5)^{|\ell|} \quad (9.118)$$

Hence, the correlation matrix for $N = 3$ is given by:

$$\mathbf{R}_{\mathbf{X}} = \begin{bmatrix} r_{xx}[0] & r_{xx}[1] & r_{xx}[2] \\ r_{xx}^*[1] & r_{xx}[0] & r_{xx}[1] \\ r_{xx}^*[2] & r_{xx}^*[1] & r_{xx}^*[0] \end{bmatrix} \quad (9.119)$$

$$= \begin{bmatrix} 4(-0.5)^0 & 4(-0.5)^1 & 4(-0.5)^2 \\ 4(-0.5)^1 & 4(-0.5)^0 & 4(-0.5)^1 \\ 4(-0.5)^2 & 4(-0.5)^1 & 4(-0.5)^0 \end{bmatrix} = \begin{bmatrix} 4 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 4 \end{bmatrix} \quad (9.120)$$

⊗

which is clearly Toeplitz.

Note that the definition of a covariance matrix for a random process follows an almost identical form, except with the elements of the autocorrelation functions replaced by the autocovariance functions. Finally, note that it is possible to define a correlation or covariance matrix for a random vector that consists of non-consecutive samples from a random process. Hence, if

$$\mathbf{X}(\{n\}) \triangleq [x(n_1) \quad x(n_2) \quad \cdots \quad x(n_M)]^T \quad (9.121)$$

where $\{n_k\}_1^M$ are unique arbitrary indices to samples from the random process, then the correlation matrix is still defined as:

$$\mathbf{R}_{\mathbf{X}}(\{n\}) = \mathbb{E} [\mathbf{X}(\{n\}) \mathbf{X}^H(\{n\})] \quad (\text{T:4.23})$$

– End-of-Topic 64: **Joint Statistics and Correlation Matrices** –



9.9 Markov Processes

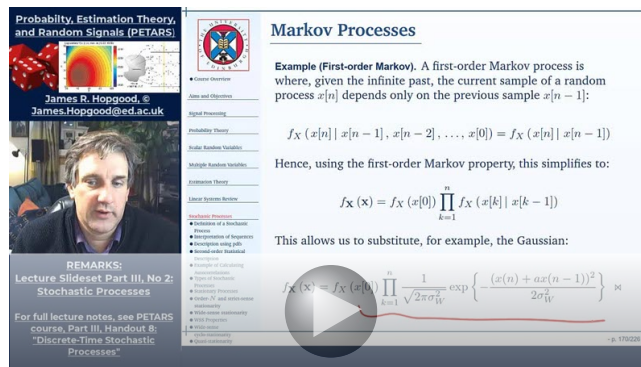
Topic Summary 65 Brief Introduction to Markov Processes

Topic Objectives:

- Introduction to advantages of the Markov model.
- Definitions of a Markov Process.
- Calculating the joint-pdf for first-order Markov process.

Topic Activities:

Type	Details	Duration	Progress
Watch video	8 : 32 min video	3 × length	
Read Handout	Read page 355 to page 356	8 mins/page	
Try Example	Try Example 9.13.	15 mins	



http://media.ed.ac.uk/media/1_y9zrkrsk

Video Summary: This video gives a very brief introduction to the powerful Markov model for random processes. It considers in detail the first-order Markov process, deriving the joint-pdf for a Gaussian-excited process. This powerful model allows certain problems to be analysed in a comprehensive manner. The video mentions higher-order Markov processes, as well as Markov Chains.

Finally, in this handout, a powerful model for a stochastic process known as a **Markov model** is introduced; such a process that satisfies this model is known as a **Markov process**. Quite simply, a Markov process is one in which the probability of any particular value in a sequence is dependent upon the preceding sample values. The simplest kind of dependence arises when the probability of any sample depends only upon the value of the *immediately preceding* sample, and this is known as a **first-order Markov process**. This simple process is a surprisingly good model for a number of practical signal processing, communications and control problems.

As an example of a Markov process, consider the process generated by the difference equation

$$x[n] = -a x[n - 1] + w[n] \tag{T:3.17}$$

where a is a known constant; and $w[n]$ is a sequence of zero-mean i. i. d. Gaussian random variables

with variance σ_W^2 density:

$$f_W(w[n]) = \frac{1}{\sqrt{2\pi\sigma_W^2}} \exp\left\{-\frac{w^2[n]}{2\sigma_W^2}\right\} \quad (\text{T:3.18})$$

The conditional density of $x[n]$ given $x[n-1]$ is also Gaussian, and using the probability transformation rule for which the Jacobian evaluates to one, it can be shown that

$$f_X(x[n] | x[n-1]) = \frac{1}{\sqrt{2\pi\sigma_W^2}} \exp\left\{-\frac{(x[n] + ax[n-1])^2}{2\sigma_W^2}\right\} \quad (\text{T:3.19})$$

In fact, if $w[n]$ is independent with *any* density $f_W(w[n])$, the conditional density of $x[n]$ given $x[n-1]$ is $f_W(x[n] + ax[n-1])$. Note that $x[n-1]$ completely determines the distribution for $x[n]$, and $x[n]$ completely determines the distribution for $x[n+1]$ and so forth. Thus, the value of the sequence at any time n_0 completely determines the distribution of $x[n]$ for any $n > n_0$. The following serves as a formal definition of a Markov process.

Definition 9.4 (Markov Process). A random process is a P th-order Markov process if the distribution of $x[n]$, given the infinite past, depends only on the previous P samples $\{x[n-1], \dots, x[n-P]\}$; that is, if:

$$f_X(x[n] | x[n-1], x[n-2], \dots) = f_X(x[n] | x[n-1], \dots, x[n-P]) \quad (\text{T:3.20})$$

◇

Example 9.13 (First-order Markov). A first-order Markov process is where, given the infinite past, the current sample of a random process $x[n]$ depends only on the previous sample $x[n-1]$; that is, if:

$$f_X(x[n] | x[n-1], x[n-2], \dots, x[0]) = f_X(x[n] | x[n-1]) \quad (9.122)$$

Note that using the probability chain rule, and defining $\mathbf{x} = \{x[n], x[n-1], \dots, x[0]\}$, the general joint-pdf of all samples can be written as:

$$f_{\mathbf{X}}(\mathbf{x}) = f_X(x[n] | x[n-1], x[n-2], \dots, x[0]) \times f_X(x[n-1] | x[n-2], x[n-3], \dots, x[0]) \cdots f_X(x[0]) \quad (9.123)$$

This can be written in the form:

$$f_{\mathbf{X}}(\mathbf{x}) = f_X(x[0]) \prod_{k=1}^n f_X(x[k] | x[k-1], \dots, x[0]) \quad (9.124)$$

Hence, using the first-order Markov property, this simplifies to:

$$f_{\mathbf{X}}(\mathbf{x}) = f_X(x[0]) \prod_{k=1}^n f_X(x[k] | x[k-1]) \quad (9.125)$$

This allows us to substitute, for example, the Gaussian expression in Equation T:3.19:

$$f_{\mathbf{X}}(\mathbf{x}) = f_X(x[0]) \prod_{k=1}^n \frac{1}{\sqrt{2\pi\sigma_W^2}} \exp\left\{-\frac{(x[k] + ax[k-1])^2}{2\sigma_W^2}\right\} \quad (9.126)$$

⋈

Finally, it is noted that if $x[n]$ takes on a countable (discrete) set of values, a Markov random process is called a **Markov chain**. This will always be the case in digital signal processing since the values of the random sequence are represented with a finite number of bits. There is a tremendous volume of results on Markov chains, but they will not presently be covered in this course.

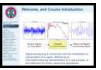


10

Frequency-Domain Description of Stationary Processes

Introduces the notion of a frequency-domain description of stationary random processes, defining the power spectral density (PSD) as the Fourier transform of the autocorrelation function. Considers the properties of the PSD including the PSD of harmonic processes. Defines the cross-PSD and the complex spectral density.

10.1 Introduction to the power spectral density



Topic Summary 66 Concept of the Power Spectral Definition and its Origins

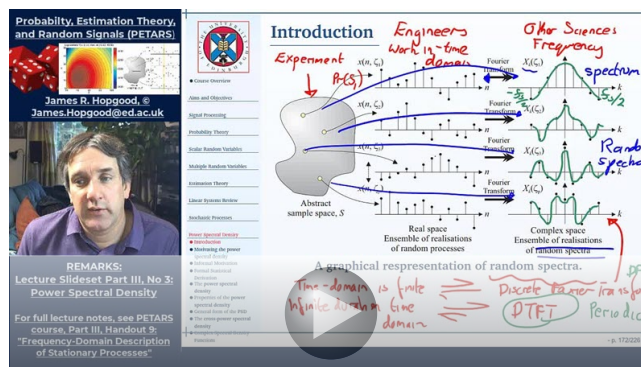
New slide

Topic Objectives:

- Notion of representing a random process in the frequency domain.
- Development of the power spectral density.
- Introduction to the Wiener-Khinchin(-Einstein-Kolmogorov) theorem.

Topic Activities:

Type	Details	Duration	Progress
Watch video	22 : 59 min video	3× length	
Read Handout	Read page 358 to page 363	8 mins/page	



http://media.ed.ac.uk/media/1_zk01rnwd

Video Summary: This video introduces the frequency-domain description of stationary processes, through the equivalent but conceptually different ideas of stochastic decompositions and Fourier transforms of moments (such as the autocorrelation or autocovariance). The video considers the conceptual equivalence of a random spectrum and random time-series. The power spectral density is developed in an informal method by calculating the second moment of the Fourier transforms of the realisations of the random signals. This is then formalised as a limiting process, to develop the infamous Wiener-Khinchin(-Einstein-Kolmogorov) theorem. The video considers the conceptual traps that you should be aware of, although ultimately the theory all leads to the definition that the power spectral density is the Fourier Transform of the autocorrelation sequence.

Frequency- and transform-domain methods including the Fourier-transform and z -transform are very powerful tools for the analysis of deterministic sequences. It seems natural to extend these techniques to analysis stationary **random processes**. In principle, it would make sense to extend the techniques to non-stationary processes, but this requires further insight and additional constraints to come up with a general theory.

So far in this course, **stationary stochastic processes** have been considered in the time-domain through the use of the **autocorrelation sequence (ACS)**. Since the ACS for a stationary process is a function of a single-discrete time process, then the question arises as to what the discrete-time

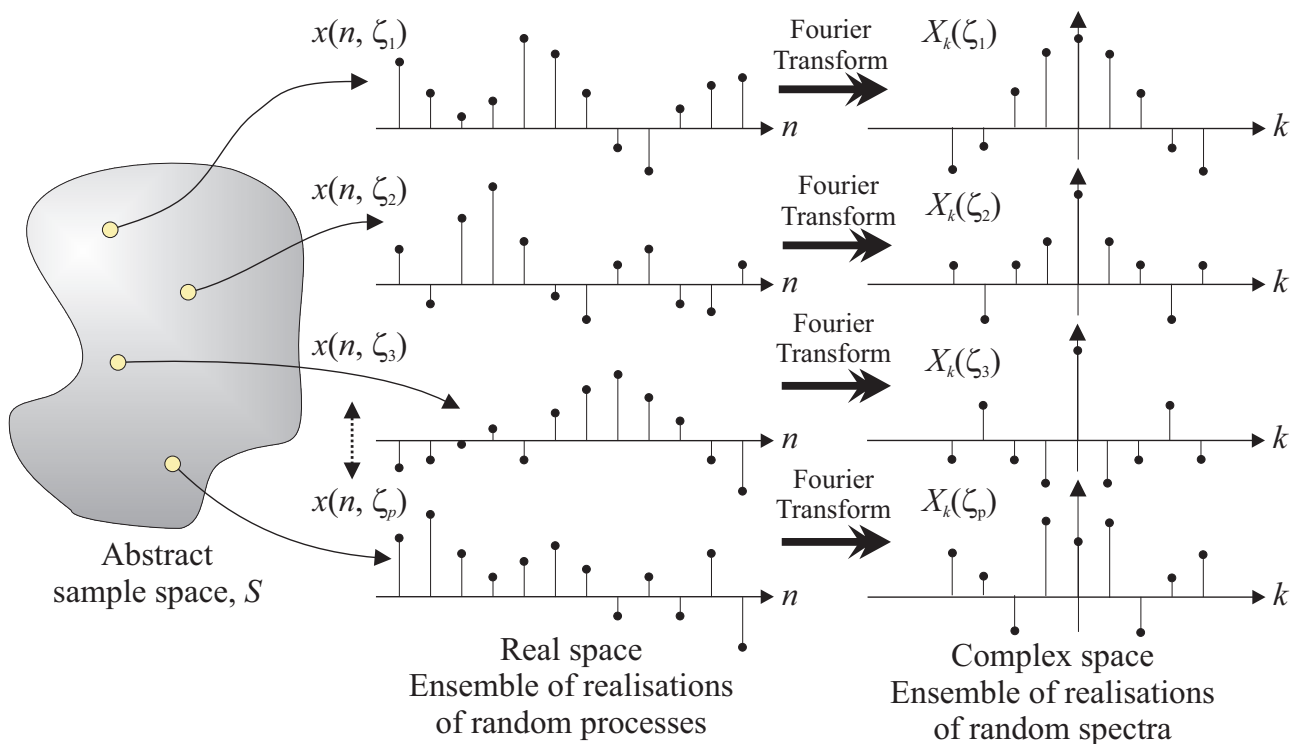


Figure 10.1: A graphical representation of random spectra.

Fourier transform (DTFT) of the ACS corresponds to. It turns out to be known as the **power spectral density (PSD)** of a stationary random process, and the PSD is an extremely powerful and conceptually appealing tool in statistical signal processing. This handout will study the PSD in some detail.

In signal theory for deterministic signals, spectra are used to represent a function as a superposition of exponential functions. For random signals, the notion of a spectrum has two interpretations:

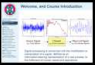
Transform of averages The first involves transform of averages (or moments). As will be seen, this will be the Fourier transform of the autocorrelation function.

Stochastic decomposition The second interpretation, and arguably more natural perspective, represents a stochastic process as a superposition of exponentials, where the coefficients are themselves random variables. Hence, a stochastic process $x[n]$ can be represented as:

$$x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega T}) e^{j\omega n} d\omega, \quad n \in \mathbb{R} \tag{10.1}$$

where $X(e^{j\omega})$ is a random variable for a given value of ω . Alternatively, $X(e^{j\omega})$ can be considered as a continuous random-process, as a function of ω . This interpretation is extremely powerful, and can in fact be extended to the superposition of any set of **basis functions**; the Karhunen-Loeve (KL) transform is an example of such a decomposition. Unfortunately, there is not time in this course to consider this spectral representation in detail, extremely interesting as it is, although it will be used below to motivate the PSD for stationary signals.

10.2 Motivating the power spectral density



It is important to appreciate that most realisations of stationary random signals, $x[n, \zeta]$, do not have finite energy, as they usually don't decay away as $n \rightarrow \pm\infty$. This is because the statistics as $n \rightarrow \pm\infty$ are the same as the statistics at any other time. Therefore, technically, these realisations do not possess a corresponding DTFT, and hence it is not possible simply to take the DTFT of the random signal without further addressing these technicalities.

Moreover, noting that a random signal is actually an ensemble of realisations, each realisation occurring with a different probability, it raises the question of what does it mean to take the DTFT of a random process directly? It should also be remembered that the DTFT of a particular observed realisation, even if it existed, is itself a realisation of a **random process**, albeit as a function of frequency rather than time. Therefore, it is necessary to take an alternative perspective, as discussed in Section 10.2.2. However, in order to motivate the PSD, first an informal and imprecise, yet insightful analysis is given in the next section.

10.2.1 Informal Motivation

This section contains an informal but insightful derivation of the PSD. Assume for the moment that the DTFT of a realisation from a stationary random process does in fact exist, by ignoring any issues with convergence of the sequence. If a particular realisation is denoted by $x[n, \zeta]$, then suppose the corresponding DTFT is denoted by:

$$X_{\zeta}(e^{j\omega T}) = \sum_{n=-\infty}^{\infty} x[n, \zeta] e^{-j\omega n} \quad (10.2)$$

where $|\omega| < \pi$ is the normalised frequency (with respect to the sampling frequency). The collection of different DTFTs forms an ensemble of frequency-domain realisations, as shown in Figure 10.1.

As this spectrum is continuous, the second-order autocorrelation function (ACF) is a seemingly important statistic to consider, representing the correlation between two frequencies at ω_1 and ω_2 , say. Hence, consider forming:

$$R_{XX}(\omega_1, \omega_2) = \mathbb{E} [X_{\zeta}(e^{j\omega_1}) X_{\zeta}^*(e^{j\omega_2})] \quad (10.3)$$

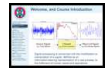
Substituting the DTFT expression from Equation 10.2 into this expression, and reorganising where possible:

$$R_{XX}(\omega_1, \omega_2) = \mathbb{E} \left[\sum_{n=-\infty}^{\infty} x[n, \zeta] e^{-j\omega_1 n} \sum_{m=-\infty}^{\infty} x^*[m, \zeta] e^{j\omega_2 m} \right] \quad (10.4)$$

$$= \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \mathbb{E} [x[n, \zeta] x^*[m, \zeta]] e^{-j(\omega_1 n - \omega_2 m)} \quad (10.5)$$

$$= \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} r_{xx}[n, m] e^{-j(\omega_1 n - \omega_2 m)} \quad (10.6)$$

At this stage, this is quite a generic expression; note further, that a very similar result can be obtained if the random process in the time-domain were continuous, where the DTFT would be replaced by the continuous-time Fourier transform (CTFT) which amounts to replacing summations by integrals. However, it can be seen though that it is indicative of a frequency domain correlation being some kind of Fourier transform of the corresponding time-domain correlation.



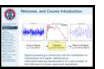
New slide

Indeed, as it has been assumed $x[n, \zeta]$ is stationary, then let $r_{xx}[n, m] = r_{xx}[n - m]$. Consider finding the second-moment or power at a given frequency, so setting $\omega = \omega_1 = \omega_2$, and then undertaking a change in variable of summation such that $\ell = n - m$. Then, it follows that:

$$R_{XX}(\omega) = \sum_{n=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} r_{xx}[\ell] e^{-j\omega\ell} = \sum_{n=-\infty}^{\infty} \mathcal{F}(r_{xx}[\ell]) \quad (10.7)$$

The additional summation results from the fact the realisations of the process do not have finite-energy, and the mathematical treatment somewhat informal. However, it clearly indicates that the power at each frequency can be found from the Fourier transform of the ACS, and is therefore the PSD. This proof can be tidied up somewhat by using careful limiting operations, as described in the next section. It can also easily be extended to the continuous-time case, by effectively just replacing the summations by integrals.

10.2.2 Formal Statistical Derivation



New slide

Motivated by the stochastic decomposition in Equation 10.1, and restricting the analysis to wide-sense stationary (WSS) processes, consider the random variable, $X(e^{j\omega T})$, resulting from the DTFT of a random signal, $x[n]$:

$$X(e^{j\omega T}) = \sum_{n=-\infty}^{\infty} x[n] e^{-j\omega n} \quad (10.8)$$

It is of interest to consider the **total power** in the rv, $X(e^{j\omega T})$, which is given by the second moment:

$$P_{XX}(e^{j\omega T}) = \mathbb{E} \left[|X(e^{j\omega T})|^2 \right] \quad (10.9)$$

Since random signals are not finite energy, then this expression will diverge, so consider instead the definition:

$$P_{XX}(e^{j\omega T}) = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \mathbb{E} \left[|X_N(e^{j\omega})|^2 \right] \quad (10.10)$$

where $X_N(e^{j\omega})$ is the truncated Fourier transform of $x[n]$, or basically a **windowed** version of the sequence $x[n]$ between $-N$ and N , as given by:

$$X_N(e^{j\omega T}) \triangleq \sum_{n=-N}^N x[n] e^{-j\omega n} = \sum_{n=-\infty}^{\infty} w[n] x[n] e^{-j\omega n} \quad (10.11)$$

where $w[n]$ is the window function:

$$w[n] = \begin{cases} 1 & -N \leq n \leq N \\ 0 & \text{otherwise} \end{cases} \quad (10.12)$$

Then, substituting Equation 10.11 into Equation 10.10 and rearranging gives:

$$P_{XX}(e^{j\omega T}) = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \mathbb{E} \left[\sum_{n=-N}^N x[n] e^{-j\omega n} \sum_{m=-N}^N x^*[m] e^{j\omega m} \right] \quad (10.13)$$

$$= \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N \sum_{m=-N}^N \mathbb{E} [x[n] x^*[m]] e^{-j\omega(n-m)} \quad (10.14)$$

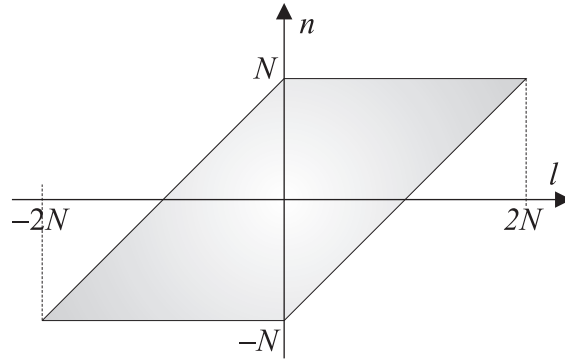


Figure 10.2: Region of summation for deriving the variance of the time-average.

It can be shown, through the following manipulations, that in the limit this expression does indeed simplify to DTFT of the ACS.

To show this, first substitute the variable $\ell = n - m$, such that when $m = \pm N$, then $\ell = n \mp N$. Since the summation is over integers, which means that $\sum_a^b(\cdot) = \sum_b^a(\cdot)$, and noting that for WSS processes, $\mathbb{E}[x[n]x^*[n-\ell]] = r_{xx}[\ell]$ this means Equation 10.14 becomes:

$$P_{XX}(e^{j\omega}) = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N \sum_{\ell=n-N}^{n+N} r_{xx}[\ell] e^{-j\omega\ell} \quad (10.15)$$

The region of summation is shown in Figure 10.2. Changing the order of summation (as this is the usual trick), to sum over ℓ first, then it can be seen that ℓ varies from $-2N$ to $2N$, while n will vary from $\max\{-N, \ell - N\}$ to $\min\{N, \ell + N\}$. Hence, Equation 10.15 becomes:

$$P_{XX}(e^{j\omega}) = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{\ell=-2N}^{2N} \sum_{n=\max\{-N, \ell-N\}}^{\min\{N, \ell+N\}} r_{xx}[\ell] e^{-j\omega\ell} \quad (10.16)$$

$$P_{XX}(e^{j\omega}) = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{\ell=-2N}^{2N} r_{xx}[\ell] e^{-j\omega\ell} \left[\sum_{n=\max\{-N, \ell-N\}}^{\min\{N, \ell+N\}} 1 \right] \quad (10.17)$$

The second summation in the square brackets can be shown by, simple counting, to simplify to $2N + 1 - |\ell|$, and therefore:

$$P_{XX}(e^{j\omega}) = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{\ell=-2N}^{2N} (2N+1-|\ell|) r_{xx}[\ell] e^{-j\omega\ell} \quad (10.18)$$

$$= \sum_{\ell=-\infty}^{\infty} r_{xx}[\ell] e^{-j\omega\ell} - \lim_{N \rightarrow \infty} \sum_{\ell=-2N}^{2N} \frac{|\ell|}{2N+1} r_{xx}[\ell] e^{-j\omega\ell} \quad (10.19)$$

Assuming the mild assumption that the autocorrelation sequence $r_{xx}[\ell]$ decays sufficiently rapidly such that:

$$\lim_{N \rightarrow \infty} \sum_{\ell=-2N}^{2N} |\ell| |r_{xx}[\ell]| = 0 \quad (10.20)$$

then Equation 10.19 simplifies to:

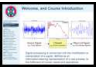
$$P_{XX}(e^{j\omega}) = \sum_{\ell=-\infty}^{\infty} r_{xx}[\ell] e^{-j\omega\ell} \quad (10.21)$$

Hence, $P_{XX}(e^{j\omega T})$ can be viewed as the average power, or energy, of the Fourier transform of a random process at frequency ω . Clearly, this gives an indication of whether, *on average*, there are dominant frequencies present in the realisations of $x[n]$.

– End-of-Topic 66: **Introduction to the concept of the PSD** –



10.3 The power spectral density



Topic Summary 67 Definition and Properties of the PSD

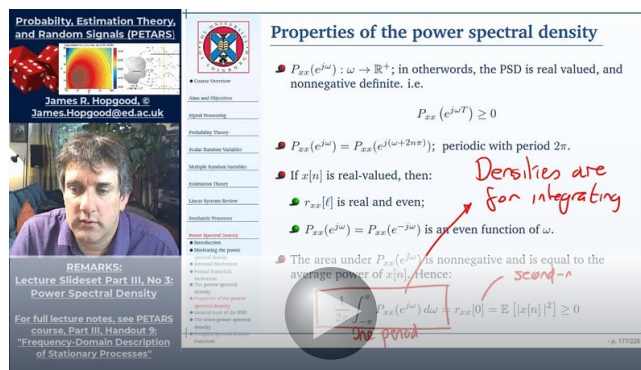
New slide

Topic Objectives:

- Definition and Properties of the PSD.
- Dealing with periodic and non-periodic components in an ACS.
- Examples of calculating PSDs.

Topic Activities:

Type	Details	Duration	Progress
Watch video	20 : 15 min video	3× length	
Read Handout	Read page 364 to page 367	8 mins/page	
Try Example	Try Examples 10.1 and Example 10.2.	25 mins	
Try Code	Use the MATLAB code	10 mins	
Practice Exercises	Exercise ??	20 mins	



http://media.ed.ac.uk/media/1_yoe37jow

Video Summary: This video presents the formal definition of the PSD of a WSS process, and its inverse relationship, both through DTFT pairs. The video presents the key properties of the PSD, many of which are related to properties of the Fourier transform, but also some key conceptual properties such as positivity, total power, and being a real function. Several examples worked examples for calculating PSDs are presented, including a detailed analysis of dealing with ACSs that have a periodic component, as well as a non-periodic component.

The discrete-time Fourier transform of the autocorrelation sequence of a stationary stochastic process $x[n, \zeta]$ is known as the **power spectral density (PSD)**, is denoted by $P_{xx}(e^{j\omega})$, and is given by:

$$P_{xx}(e^{j\omega}) = \sum_{\ell \in \mathbb{Z}} r_{xx}[\ell] e^{-j\omega\ell} \tag{M:3.3.39}$$

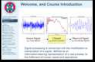
where ω is frequency in radians per sample.

The autocorrelation sequence, $r_{xx}[\ell]$, can be recovered from the **PSD** by using the inverse-DTFT:

$$r_{xx}[\ell] = \frac{1}{2\pi} \int_{-\pi}^{\pi} P_{xx}(e^{j\omega}) e^{j\omega\ell} d\omega, \quad \ell \in \mathbb{Z} \tag{M:3.3.41}$$

Sometimes the PSD is called the auto-PSD to distinguish it from the cross-PSD introduced in Section 10.4. In the case that $r_{xx}[\ell]$ is periodic, corresponding to a wide-sense periodic stochastic process, then the power spectral density is defined as the discrete Fourier transform of the autocorrelation sequence. This natural extension is easily obtained once the aperiodic-case is considered in depth.

10.3.1 Properties of the power spectral density



There are a number of properties of the power spectral density that follow from the corresponding *New slide* properties of the autocorrelation sequence, and the discrete-time Fourier transform.

- $P_{xx}(e^{j\omega}) : \omega \rightarrow \mathbb{R}^+$; in other words, the PSD is real valued, and nonnegative definite. i.e.

$$P_{xx}(e^{j\omega T}) \geq 0 \quad (\text{M:3.3.44})$$

This property follows from the positive semi-definiteness of the autocorrelation sequence.

- $P_{xx}(e^{j\omega}) = P_{xx}(e^{j(\omega+2n\pi)})$; in other words, the PSD is periodic with period 2π .
- If $x[n]$ is real-valued, then:
 - $r_{xx}[\ell]$ is real and even;
 - $P_{xx}(e^{j\omega}) = P_{xx}(e^{-j\omega})$ is an even function of ω .

- The area under $P_{xx}(e^{j\omega})$ is nonnegative and is equal to the average power of $x[n]$. Hence:

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} P_{xx}(e^{j\omega}) d\omega = r_{xx}[0] = \mathbb{E}[|x[n]|^2] \geq 0 \quad (\text{M:3.3.45})$$

Example 10.1 ([Manolakis:2001, Example 3.3.4, Page 109]). Determine the PSD of a zero-mean WSS process $x[n]$ with autocorrelation sequence $r_{xx}[\ell] = a^{|\ell|}$, $-1 < a < 1$.

SOLUTION. Using the definition of the PSD directly, then:

$$P_{xx}(e^{j\omega}) = \sum_{\ell \in \mathbb{Z}} r_{xx}[\ell] e^{-j\omega\ell} \quad (10.22)$$

$$= \sum_{\ell \in \mathbb{Z}} a^{|\ell|} e^{-j\omega\ell} \quad (10.23)$$

$$= \sum_{\ell=0}^{\infty} (a e^{-j\omega})^{\ell} + \sum_{\ell=0}^{\infty} (a e^{j\omega})^{\ell} - 1 \quad (10.24)$$

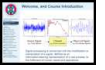
Hence, by using the expressions for geometric series, the PSD can be written as:

$$P_{xx}(e^{j\omega}) = \frac{1}{1 - a e^{-j\omega}} + \frac{1}{1 - a e^{j\omega}} - 1 \quad (\text{M:3.3.42})$$

$$= \frac{1 - a^2}{1 - 2a \cos \omega + a^2} \quad (10.25)$$

□

which is a real-valued, even, and nonnegative function of ω .



10.3.2 General form of the PSD

A process, $x[n]$, and its corresponding autocorrelation sequence (ACS), $r_{xx}[\ell]$, can be decomposed into a zero-mean aperiodic component, $r_{xx}^{(a)}[\ell]$, and a non-zero-mean periodic component, $r_{xx}^{(p)}[\ell]$: New slide

$$r_{xx}[\ell] = r_{xx}^{(a)}[\ell] + r_{xx}^{(p)}[\ell] \quad (10.26)$$

Theorem 10.1 (PSD of a non-zero-mean process with periodic component). The most general definition of the PSD for a non-zero-mean stochastic process with a periodic component is given by:

$$P_{xx}(e^{j\omega}) = P_{xx}^{(a)}(e^{j\omega}) + \frac{2\pi}{K} \sum_{k \in \mathcal{K}} P_{xx}^{(p)}(k) \delta(\omega - \omega_k) \quad (\text{T:4.41})$$

The term $P_{xx}^{(a)}(e^{j\omega})$ is the DTFT of the aperiodic component $r_{xx}^{(a)}[\ell]$, while $P_{xx}^{(p)}(k)$ are the discrete Fourier transform (DFT) coefficients for the periodic component $r_{xx}^{(p)}[\ell]$ assuming a periodicity of length K , and where $\omega_k = \frac{2\pi k}{K}$.

Moreover, it can be seen that $P_{xx}^{(a)}(e^{j\omega})$ represents the continuous part of the spectrum, while the sum of weighted impulses represent the discrete part or *lines* of the spectrum.

PROOF. The non-zero-mean periodic component, $r_{xx}^{(p)}(l)$ can itself be decomposed using a **discrete Fourier transform**:

$$r_{xx}^{(p)}(l) = \frac{1}{K} \sum_{k \in \mathcal{K}} P_{xx}^{(p)}(k) e^{j\omega_k l} \quad (10.27)$$

where $\mathcal{K} = \{0, \dots, K-1\}$, and $\omega_k = \frac{2\pi}{K}k$. Thus, the **PSD** of $X(\zeta)$, becomes:

$$P_{xx}(e^{j\omega}) = P_{xx}^{(a)}(e^{j\omega}) + \frac{1}{K} \sum_{\ell \in \mathbb{Z}} \sum_{k \in \mathcal{K}} P_{xx}^{(p)}(k) e^{j\omega_k \ell} e^{-j\omega \ell} \quad (10.28)$$

As usual, change the order of summation:

$$= P_{xx}^{(a)}(e^{j\omega}) + \frac{1}{K} \sum_{k \in \mathcal{K}} P_{xx}^{(p)}(k) \sum_{\ell \in \mathbb{Z}} e^{-j\ell(\omega - \omega_k)} \quad (10.29)$$

$$= P_{xx}^{(a)}(e^{j\omega}) + \frac{2\pi}{K} \sum_{k \in \mathcal{K}} P_{xx}^{(p)}(k) \delta(\omega - \omega_k) \quad (10.30)$$

where **Poisson's formula**, which can be derived by writing down the Fourier series for an impulse train, is used:

$$\sum_{n=-\infty}^{\infty} \delta(t - nT) = \frac{1}{T} \sum_{\ell=-\infty}^{\infty} e^{-j\ell\omega_0 t} \quad (10.31)$$

where $\omega_0 = \frac{2\pi}{T}$. Thus, by letting $T = 2\pi$, and $t = \omega - \omega_k$, then:

$$2\pi \sum_{n=-\infty}^{\infty} \delta(\omega - \omega_k - 2\pi n) = \sum_{\ell=-\infty}^{\infty} e^{-j\ell(\omega - \omega_k)} \quad (10.32)$$

Since $-2\pi < \omega_k \leq 2\pi$, and $P_{xx}(e^{j\omega})$ is periodic in ω with period 2π , then it is sufficient to write for $|\omega| \leq 2\pi$, that:

$$2\pi \delta(\omega - \omega_k) = \sum_{\ell=-\infty}^{\infty} e^{-j\ell(\omega - \omega_k)} \quad (10.33) \quad \square$$

which can be substituted to give the desired result.

Example 10.2 ([Manolakis:2001, Harmonic Processes, Page 110-111]). Determine the PSD of the **harmonic process** introduced in the previous handout and defined by:

$$x[n] = \sum_{k=1}^M A_k \cos(\omega_k n + \phi_k), \quad \omega_k \neq 0 \quad (\text{M:3.3.50})$$

where M , $\{A_k\}_1^M$ and $\{\omega_k\}_1^M$ are constants, and $\{\phi_k\}_1^M$ are pairwise independent and identically distributed (i. i. d.) random variables (RVs) uniformly distributed in the interval $[0, 2\pi]$.

SOLUTION. As shown in the previous handout, $x[n]$ is a zero-mean stationary process, and ACS:

$$r_{xx}[\ell] = \frac{1}{2} \sum_{k=1}^M |A_k|^2 \cos \omega_k \ell, \quad -\infty < \ell < \infty \quad (\text{M:3.3.52})$$

Note that $r_{xx}[\ell]$ consists of a sum of *in-phase* cosines with the same frequencies as in $x[n]$. By writing

$$\cos \omega_k \ell = \frac{e^{j\omega_k \ell} + e^{-j\omega_k \ell}}{2} \quad (10.34)$$

then Equation M:3.3.52 may be written as:

$$\begin{aligned} r_{xx}[\ell] &= \frac{1}{4} \sum_{k=1}^M |A_k|^2 (e^{j\omega_k \ell} + e^{-j\omega_k \ell}) \\ &= \sum_{k=1}^M \frac{|A_k|^2}{4} e^{j\omega_k \ell} + \sum_{k=1}^M \frac{|A_k|^2}{4} e^{-j\omega_k \ell} \\ &= \sum_{k=1}^M \frac{|A_k|^2}{4} e^{j\omega_k \ell} + \sum_{\hat{k}=-1}^{-M} \frac{|A_{-\hat{k}}|^2}{4} e^{-j\omega_{-\hat{k}} \ell} \end{aligned} \quad (10.35)$$

Hence, the ACS can be written as:

$$r_{xx}[\ell] = \sum_{k=-M}^M \frac{|A_k|^2}{4} e^{j\omega_k \ell}, \quad -\infty < \ell < \infty \quad (10.36)$$

where the following are defined: $A_0 = 0$, $A_k = A_{-k}$, and $\omega_{-k} = -\omega_k$.

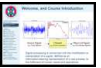
Hence, it directly follows using the results above that:

$$P_{xx}(e^{j\omega}) = 2\pi \sum_{k=-M}^M \frac{|A_k|^2}{4} \delta(\omega - \omega_k) = \frac{\pi}{2} \sum_{k=-M}^M |A_k|^2 \delta(\omega - \omega_k) \quad (10.37) \quad \square$$

The harmonic process is predictable because any given realisation is a sinusoidal sequence with fixed amplitude, frequency and phase. The independence and uniform distribution of the phase, however, is strictly required to ensure the stationarity of the process $x[n]$.



10.4 The cross-power spectral density



Topic Summary 68 Cross-Power and Complex Spectral Densities

New slide

Topic Objectives:

- Definition and Properties of the cross-power spectral density (CPSD).
- Introducing the Complex- and Cross-Spectral Density Functions and their properties.
- Examples of calculating the complex-spectral density of a challenging ACS.
- Using tables of z -transforms.

Topic Activities:

Type	Details	Duration	Progress
Watch video	23 : 18 min video	3× length	
Read Handout	Read page 368 to page ??	8 mins/page	
Try Example	Try Example 10.3	25 mins	
Practice Exercises	Exercise ?? to ??	80 mins	

http://media.ed.ac.uk/media/1_ocukvbyi

Video Summary: This Topic extends the definition of the PSD in two ways. First, it considers the CPSD for considering the spectral characteristics of two-jointly stationary processes. It takes the natural definition of being the DTFT of the cross-correlation function. The video considers some relevant properties of the CPSD. The Topic then considers that, due to technical limitations of the DTFT, taking the bilateral z -transform of the auto- or cross-correlation sequences is a more powerful technique. This is defined as the complex- and cross-complex spectral densities. An example of the complex-spectral density is calculated. Finally, a discussion of using z -transform tables for taking inverse transforms is provided.

The cross-power spectral density (CPSD) of two jointly stationary stochastic processes, $x[n]$ and $y[n]$, provides a description of their statistical relations in the frequency domain. It is defined, naturally, as the DTFT of the cross-correlation, $r_{xy}[\ell] \triangleq \mathbb{E} [x[n] y^*[n - \ell]]$:

$$P_{xy} (e^{j\omega T}) = \mathcal{F}\{r_{xy}[\ell]\} = \sum_{\ell \in \mathbb{Z}} r_{xy}[\ell] e^{-j\omega \ell} \tag{M:3.3.56}$$

The cross-correlation $r_{xy}[\ell]$ can be recovered by using the inverse-DTFT:

$$r_{xy}[\ell] = \frac{1}{2\pi} \int_{-\pi}^{\pi} P_{xy}(e^{j\omega T}) e^{j\omega\ell} d\omega, \quad \ell \in \mathbb{R} \quad (\text{M:3.3.57})$$

Since this integral is essentially a summation, then an interpretation that can be given to the cross-spectrum is that $P_{xy}(e^{j\omega T})$ measures the correlation between two RVs at a given frequency ω_0 .

The cross-spectrum $P_{xy}(e^{j\omega T})$ is, in general, a complex function of ω .

Some properties of the CPSD and related definitions include:

1. $P_{xy}(e^{j\omega T})$ is periodic in ω with period 2π .
2. Since $r_{xy}[\ell] = r_{yx}^*[-\ell]$, then it follows:

$$P_{xy}(e^{j\omega T}) = P_{yx}^*(e^{j\omega T}) \quad (\text{M:3.3.58})$$

Thus, $P_{xy}(e^{j\omega})$ and $P_{yx}(e^{j\omega})$ have the same magnitude, but opposite phase.

3. If the process $x[n]$ is real, then $r_{xy}[\ell]$ is real, and:

$$P_{xy}(e^{j\omega}) = P_{xy}^*(e^{-j\omega}) \quad (10.38)$$

4. The normalised cross-correlation, or **coherence function**, is given by:

$$\Gamma_{xy}(e^{j\omega}) \triangleq \frac{P_{xy}(e^{j\omega})}{\sqrt{P_{xx}(e^{j\omega})} \sqrt{P_{yy}(e^{j\omega})}} \quad (\text{M:3.3.59})$$

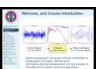
Its squared magnitude is known as the magnitude square coherence (MSC) function.

$$|\Gamma_{xy}(e^{j\omega})|^2 = \frac{|P_{xy}(e^{j\omega})|^2}{P_{xx}(e^{j\omega}) P_{yy}(e^{j\omega})} \quad (10.39)$$

If $y[n] = x[n]$, then $\Gamma_{xy}(e^{j\omega})$, corresponding to maximum correlation, whereas if $x[n]$ and $y[n]$ are uncorrelated, then $r_{xy}[\ell] = 0$, and therefore $\Gamma_{xy}(e^{j\omega}) = 0$. Hence:

$$0 \leq |\Gamma_{xy}(e^{j\omega})|^2 \leq 1 \quad (\text{M:3.3.60})$$

10.5 Complex Spectral Density Functions



New slide

The analysis of discrete-deterministic signals is also performed through the the z -transform and, therefore, in addition to using the Fourier transform, it is also very important to analyse stationary random processes using this transform; it is a perfectly natural extension.

The second moment quantities that described a random process in the z -transform domain are known as the **complex spectral density** and **complex cross-spectral density** functions. The PSD and CPSD functions discussed previously can be considered as special cases of the complex spectral density functions when the latter are evaluated on the unit circle.

If the sequences $r_{xx}[\ell]$ and $r_{xy}[\ell]$ are absolutely summable within a certain ring of the complex z -plane, then their z -transforms exist. Hence, $r_{xx}[\ell] \stackrel{z}{\rightleftharpoons} P_{xx}(z)$ and $r_{xy}[\ell] \stackrel{z}{\rightleftharpoons} P_{xy}(z)$, where:

$$P_{xx}(z) = \sum_{\ell \in \mathbb{Z}} r_{xx}[\ell] z^{-\ell} \quad (\text{M:3.3.61})$$

$$P_{xy}(z) = \sum_{\ell \in \mathbb{Z}} r_{xy}[\ell] z^{-\ell} \quad (\text{M:3.3.62})$$

Note that these are bilateral z -transforms. If the unit circle, defined by $z = e^{j\omega}$ is within the region of convergence of these summations, then:

$$P_{xx}(e^{j\omega}) = P_{xx}(z)|_{z=e^{j\omega}} \quad (\text{M:3.3.63})$$

$$P_{xy}(e^{j\omega}) = P_{xy}(z)|_{z=e^{j\omega}} \quad (\text{M:3.3.64})$$

Example 10.3 (Interleaved Example). Find the complex spectral-density of the sequence:

$$r[n] = \begin{cases} a^{|\frac{n}{2}|} & n \in \{0, \text{even}\} \\ 0 & \text{for } n \text{ odd} \end{cases} \quad (10.40)$$

SOLUTION. Writing the z -transform, noting that the all odd-values are zero:

$$P(z) = \sum_{\ell=-\infty}^{\infty} r[\ell] z^{-\ell} \quad (10.41)$$

$$= \underbrace{\sum_{\ell_o=-\infty}^{\infty} r[2\ell_o + 1] z^{-(2\ell_o+1)}}_{\text{Odd terms}} + \underbrace{\sum_{\ell_e=-\infty}^{\infty} r[2\ell_e] z^{-2\ell_e}}_{\text{Even terms}} \quad (10.42)$$

$$= \sum_{\ell_e=-\infty}^{\infty} a^{|\frac{2\ell_e}{2}|} z^{-2\ell_e} = \sum_{\ell_e=-\infty}^{\infty} a^{|\ell_e|} z^{-2\ell_e} \quad (10.43)$$

Splitting this into two further summations, as previous done with an earlier example:

$$P(z) = \sum_{\ell_e=-\infty}^0 a^{-\ell_e} z^{-2\ell_e} + \sum_{\ell_e=0}^{\infty} a^{\ell_e} z^{-2\ell_e} - 1 \quad (10.44)$$

$$= \sum_{\ell_e=0}^{\infty} (a z^2)^{\ell_e} + \sum_{\ell_e=0}^{\infty} \left(\frac{a}{z^2}\right)^{\ell_e} - 1 \quad (10.45)$$

Finally, applying the geometric progression formula $\sum_{\ell=0}^{\infty} r^{\ell} = \frac{1}{1-r}$ gives the desired result:

$$P(z) = \frac{1}{1 - a z^2} + \frac{1}{1 - a z^{-2}} - 1 \quad (10.46)$$

$$= \frac{1}{1 - a z^2} + \frac{a z^{-2}}{1 - a z^{-2}} \quad (10.47)$$

Note that this could have, equivalently, been written as:

$$P(z) = \frac{a z^2}{1 - a z^2} + \frac{1}{1 - a z^{-2}} \quad (10.48)$$

□

The inverse of the complex spectral and cross-spectral densities are given by the contour integral:

$$r_{xx}[\ell] = \frac{1}{2\pi j} \oint_C P_{xx}(z) z^{\ell-1} dz \quad (10.49)$$

$$r_{xy}[\ell] = \frac{1}{2\pi j} \oint_C P_{xy}(z) z^{\ell-1} dz \quad (10.50)$$

where the contour of integration C is to be taken counterclockwise and in the region of convergence. In practice, these integrals are usually never performed, and tables, instead, are used.

Some properties of the complex spectral densities include:

1. Conjugate-symmetry:

$$P_{xx}(z) = P_{xx}^*(1/z^*) \quad \text{and} \quad P_{xy}(z) = P_{yx}^*(1/z^*) \quad (10.51)$$

2. For the case when $x(n)$ is real, then:

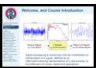
$$P_{xx}(z) = P_{xx}(z^{-1}) \quad (10.52)$$

The possible existence of lines in the PSD function due to a periodic component of the random process, as discussed in Section 10.3.2, poses some mathematical problems in defining the complex spectral density function since the z -transform does not exist. A similar approach to that in Equation T:4.41 is used here, and the complex spectral density function is written as:

$$P_{xx}(z) = P_{xx}^{(a)}(z) + 2\pi \sum_{k \in \mathcal{K}} P_{xx}^{(p)}(k) \delta(z - e^{j\omega_k}) \quad (10.53)$$

where $P_{xx}^{(a)}(z)$ corresponds to the aperiodic component of the autocorrelation function, and the second summation term denotes the line spectra.

10.6 Table of bilateral z -transforms



New slide

The **bilateral z -transform** is defined by the following pairs of equations:

$$X(z) \triangleq \mathcal{Z}[x[n]] = \sum_{n=-\infty}^{\infty} x[n] z^{-n} \quad (\text{M:2.2.29})$$

$$x[n] = \frac{1}{2\pi j} \oint_C X(z) z^{n-1} dz \quad (\text{M:2.2.30})$$

In the following table, it is assumed that $|a| \leq 1$. It is important to note that this is a crucial condition, as it will distinguish signals that exist only for $n \geq 0$ and those for $n < 0$. To use these tables, it is crucial to match an expression with an identity exactly, otherwise the incorrect inverse transform might accidentally be used.

For the purposes of the table, recall that $u[n]$ is the **discrete-time step function** given by:

$$u[n] = \begin{cases} 1 & n \geq 0 \\ 0 & n < 0 \end{cases} \quad (10.54)$$

The region of convergence (ROC) is also shown for completeness, although it is usual to assume that z is only considered within the ROC. Note that if the signal $x[n] = 0$ for $n < 0$, it is known as a **causal sequence**, and if $x[n] = 0$ for $n > 0$, it is known as an **anticausal sequence**.

Notes	$x[n]$	$X(z)$	ROC
$x[n] = 0, n < 0$	$u[n]$	$\frac{1}{1 - z^{-1}} \equiv \frac{z}{z - 1}$	$ z > 1$
$x[n] = 0, n > 0$	$u[-n]$	$\frac{1}{1 - z}$	$ z < 1$
$x[n] = 0, n < 0$	$a^n u[n]$	$\frac{1}{1 - az^{-1}} \equiv \frac{z}{z - a}$	$ z > a $
$x[n] = 0, n \leq 0$	$a^n u[n - 1]$	$\frac{a}{z - a} \equiv \frac{az^{-1}}{1 - az^{-1}}$	$ z > a $
$x[n] = 0, n > 0$	$a^{-n} u[-n]$	$\frac{1}{1 - az} \equiv \frac{z^{-1}}{z^{-1} - a}$	$ z < \frac{1}{ a }$
$x[n] = 0, n \geq 0$	$a^{-n} u[-n - 1]$	$\frac{az}{1 - az} \equiv \frac{a}{z^{-1} - a}$	$ z < \frac{1}{ a }$
$x[n] = 0, n < 0$	$na^n u[n]$	$\frac{az^{-1}}{(1 - az^{-1})^2}$	$ z > a $
$x[n] = 0, n \geq 0$	$-na^{-n} u[-n - 1]$	$\frac{az}{(1 - az)^2}$	$ z < \frac{1}{ a }$
See note 3	$\begin{cases} a^{\lfloor \frac{n}{2} \rfloor} & n \in \{0, \text{even}\} \\ 0 & \text{for } n \text{ odd} \end{cases}$	$\frac{1}{1 - az^2} + \frac{az^{-2}}{1 - az^{-2}}$ or $\frac{1 - a^2}{(1 - az^2)(1 - az^{-2})}$	$ a ^{\frac{1}{2}} < z < \frac{1}{ a ^{\frac{1}{2}}}$
	$\begin{cases} a^{\lfloor \frac{n}{2} \rfloor + \frac{1}{2}} & \text{for } n \text{ odd} \\ 0 & \text{otherwise} \end{cases}$	$\frac{az}{1 - az^2} + \frac{az^{-1}}{1 - az^{-2}}$ or $\frac{a(1 - a)(z + z^{-1})}{(1 - az^2)(1 - az^{-2})}$	$ a ^{\frac{1}{2}} < z < \frac{1}{ a ^{\frac{1}{2}}}$
See notes 1, 3	$a^{ n }$	$\frac{1}{1 - az^{-1}} + \frac{az}{1 - az}$ or $\frac{1 - a^2}{(1 - az)(1 - az^{-1})}$	$ a < z < \frac{1}{ a }$
See note 2	$ n a^{ n }$	$\frac{az^{-1}}{(1 - az^{-1})^2} + \frac{az}{(1 - az)^2}$	$ a < z < \frac{1}{ a }$

- Notes:**
1. This identity follows since $a^{|n|} \equiv a^n u[n] + a^{-n} u[-n - 1]$.
 2. Similarly, note that $|n|a^{|n|} = na^n u[n] - na^{-n} u[-n - 1]$.
 3. Note other similar expressions result, as shown below.

A variety of equivalent expressions can result from some simple manipulations; thus, other tables of

z -transforms may appear to list different results, but are actually equivalent. Some examples include:

$$\begin{aligned}
 x[n] &= \begin{cases} a^{\lfloor \frac{n}{2} \rfloor} & n \in \{0, \text{even}\} \\ 0 & \text{for } n \text{ odd} \end{cases} \\
 &\stackrel{z}{\Leftrightarrow} \frac{1}{1-az^2} + \frac{az^{-2}}{1-az^{-2}} = \left\{ \frac{az^2}{1-az^2} + 1 \right\} + \left\{ \frac{1}{1-az^{-2}} - 1 \right\} \\
 &= \frac{az^2}{1-az^2} + \frac{1}{1-az^{-2}}
 \end{aligned}$$

and

$$\begin{aligned}
 x[n] = a^{|n|} &\stackrel{z}{\Leftrightarrow} \frac{1}{1-az^{-1}} + \frac{az}{1-az} = \left\{ \frac{1}{1-az^{-1}} - 1 \right\} + \left\{ \frac{az}{1-az} + 1 \right\} \\
 &= \frac{az^{-1}}{1-az^{-1}} + \frac{1}{1-az}
 \end{aligned}$$

The fact that there are so many equivalent expressions means that sometimes it can be difficult to find the exact transform relation in tables. The particular form of the z -transform that needs to be inverted can vary depending on how it is calculated.

11

Linear Systems with Stationary Random Inputs

Considers the concept of applying a stochastic signal to the input of a system and determining the resulting output. Looks at the special case of linear time-invariant (LTI) systems with stationary inputs. Analysis by looking at the input and output statistics, as well as the input-output joint-statistics. Discusses system identification using cross-correlation. Provides examples for systems with rational transfer functions (using time domain analysis by solving difference equations and frequency domain analysis).

11.1 Systems with Stochastic Inputs

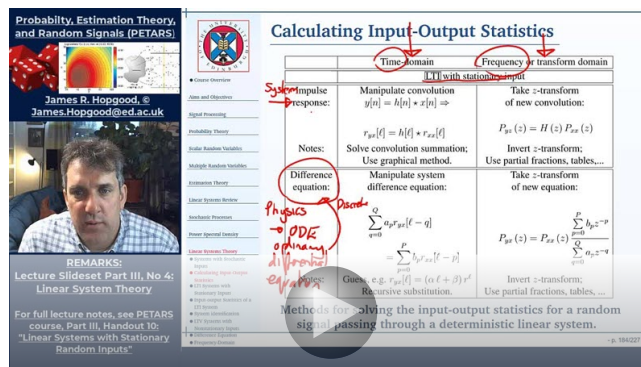
Topic Summary 69 Introduction to System Response to Random Signals

Topic Objectives:

- Concept of the output of a system to stochastic input.
- Overview of methods for Calculating Input-Output Statistics.
- Introduction of Monte Carlo calculation for Input-Output Statistics.

Topic Activities:

Type	Details	Duration	Progress
Watch video	18 : 19 min video	3× length	
Read Handout	Read page 375 to page 377	8 mins/page	
Try Example	Try Example 11.1 using MATLAB	25 mins	



http://media.ed.ac.uk/media/1_dak8253r

Video Summary: This Topic introduces the concept of calculating the stochastic process at the output of a known deterministic system, given a stochastic process at the input of the system. This concept is approached by considering the operation of the system on each realisation of the input stochastic process, and calculating the statistics over the resulting ensemble at the output. The video discusses why it is necessary, in this course, to restrict the analysis to known linear time-invariant (LTI) systems with wide-sense stationary (WSS) inputs. An overview is provided for the four different methods for calculating the input-output statistics, namely in the time-domain or frequency-domain, and either using the system impulse-response or the system-difference equation. Finally, an example of simulating the ensemble statistics through a Monte Carlo experiment is shown.

Signal processing involves the transformation of signals to enhance certain characteristics; for example, to suppress noise, or to extract meaningful information. This handout considers the processing of random processes by systems, and in particular linear systems.

What does it mean to apply a stochastic signal to the input of a system? This question is an interesting one since a stochastic process is not just a single sequence but an ensemble of sequences.

If the system is a general nonlinear possibly time-varying system, then one approach of expressing

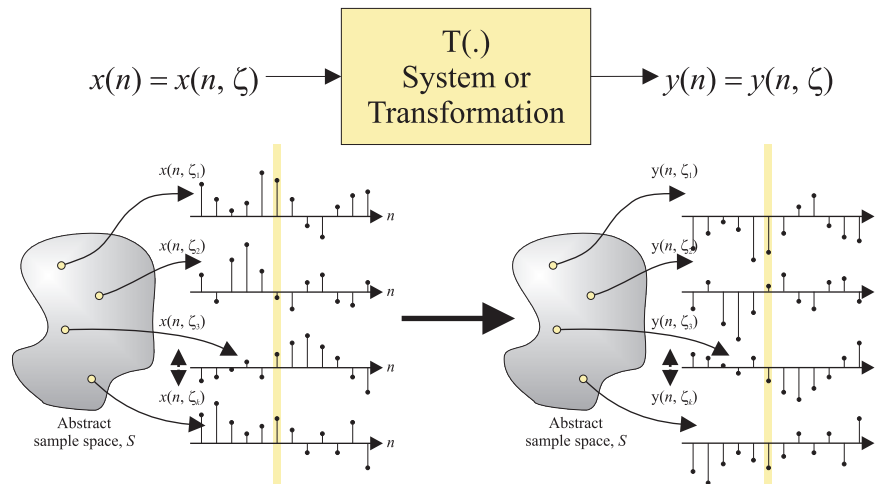


Figure 11.1: A graphical representation of a random process at the output of a system in relation to a random process at the input of the system.

the relationship is as follows: Given a stochastic process $x[n, \zeta]$, assign according to some rule to each of its realisations $x[n, \zeta_k]$ a function $y[n, \zeta_k]$. Thus, another process has been created in which:

$$y[n] = T[x[n]] \quad (11.1)$$

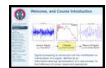
whose realisations are the functions $\{y[n, \zeta_k]\}$. This process $y[n]$ so formed can be considered as the output of a **system** or transformation with, as its input, the process $x[n]$. The system is completely specified in terms of the transformation function (or operator) T ; that is, the rule of correspondence between the samples of the input $x[n]$ and the output $y[n]$.

In principle, the statistics of the output of any system can be expressed in terms of the statistics of the input. However, in general this is a complicated problem except in special cases of particular types of signals or particular types of systems. A special case is that of known-deterministic *linear systems*, and this is considered in the next section. In particular, if the input is a stationary stochastic process, and the system is linear time-invariant (LTI), then the statistics are even simpler. Moreover, it leads to a slightly simpler and intuitive explanation for the response of the system to the input. There are other systems that can be analysed, but due to time constraints, they are not considered in this course. For more information see, for example, [Papoulis:1991, Chapter 10]. The case of random signals going through random systems is of great interest, but also beyond the scope of this course.

11.2 Methods for Calculating Input-Output Statistics

There are four different methods for calculating the input-output statistics for a WSS stochastic process passing through a known deterministic linear system. The techniques build on the theory that is already well understood in signals and systems theory, and therefore it should be familiar. The techniques involve either a time-domain solution, or a frequency-domain solution. In the time-domain, the problem can be solved either using convolution, if the system impulse response is known, or by solving difference equations if that description of the linear system is available.

Similarly, in the frequency domain, the transfer function approach can be used in which either the transfer function of the impulse response is known, or the rational transfer function of the difference equation describing the system is available. These four different methods are summarised in the table below.



New slide

	Time-domain	Frequency or transform domain
LTI with stationary input		
Impulse response:	Manipulate convolution $y[n] = h[n] \star x[n] \Rightarrow$	Take z -transform of new convolution:
Notes:	$r_{yx}[\ell] = h[\ell] \star r_{xx}[\ell]$ Solve convolution summation; Use graphical method.	$P_{yz}(z) = H(z) P_{xx}(z)$ Invert z -transform; Use partial fractions, tables,...
Difference equation:	Manipulate system difference equation:	Take z -transform of new equation:
Notes:	$\sum_{q=0}^Q a_p r_{yx}[\ell - q]$ $= \sum_{p=0}^P b_p r_{xx}[\ell - p]$ Guess, e.g. $r_{yx}[\ell] = (\alpha \ell + \beta) r^\ell$ Recursive substitution.	$P_{yx}(z) = P_{xx}(z) \frac{\sum_{p=0}^P b_p z^{-p}}{\sum_{q=0}^Q a_p z^{-q}}$ Invert z -transform; Use partial fractions, tables, ...

Example 11.1 (Typical Question). A real-valued discrete-time random process $x[n]$ consists of independent and identically distributed (i. i. d.) random variables each with uniform density on the interval $[0, 6]$.

The process $x[n]$ is applied to a linear time-invariant (LTI) system with impulse response:

$$h[n] = \begin{cases} \left(\frac{2}{3}\right)^n, & n \geq 0 \\ 0, & n < 0 \end{cases}$$

The output of this linear system is denoted as $y[n]$.

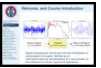
1. Calculate the output autocorrelation function $r_{yy}[\ell]$.
2. Suppose the i. i. d. process $x[n]$ now has a Weibull distribution with unit mean and variance of
3. Explain how your previous result might change, justifying your answer.

SOLUTION. You can try and answer this question after studying the rest of the handout!

– End-of-Topic 68: **Summary of methods for calculating input-output statistics** –



11.3 LTI Systems with Stationary Inputs



Topic Summary 70 Time-Domain Analysis of Response to Random Signals using the System Impulse Response New slide

Topic Objectives:

- Time-domain relationships for input-output statistics in terms of impulse response.
- Example of calculations for a typical problem.
- Observation of how WSS processes of arbitrary autocorrelation sequence (ACS) can be obtained by driving a LTI system by white Gaussian noise (WGN).

Topic Activities:

Type	Details	Duration	Progress
Watch video	32 : 23 min video	3 × length	
Read Handout	Read page 378 to page 384	8 mins/page	
Try Example	Try Example 11.2	30 mins	
Practice Exercises	Exercises ?? and ??	40 mins	

http://media.ed.ac.uk/media/1_8i50x9zo

Video Summary: This video looks at the method for calculating the output statistics for a LTI system in response to a WSS input using a time-domain method given the system impulse response. The Topic begins by highlighting the conceptual idea that the expectation of a linear operator or system is equivalent to the linear operator applied to expectations. This leads to the general idea that the output statistics are the convolution of the impulse response of the system with the input statistics. The specific details are presented, including calculating the mean at the output, the output-input cross-correlation, the output cross-correlation, and the equivalent covariance results. A detailed and typical example is presented, demonstrating the different stages of the calculations. Finally, the relationship of these results to the application of stochastic signal modelling is mentioned, and this will be addressed in detail in a later topic.

The notation:

$$y[n] = \mathcal{L}[x[n]]$$

(P:10-76)

will indicate that $y[n]$ is the output of a **linear system** with input $x[n]$. This means that for K random processes $\{x_k[n]\}_{k=1}^K$ and K scalar values $\{\alpha_k\}_{k=1}^K$, then

$$y[n] = \mathcal{L} \left[\sum_{k=1}^K \alpha_k x_k[n] \right] = \sum_{k=1}^K \alpha_k \mathcal{L}[x_k[n]] \quad (\text{P:10-77})$$

Since each sequence (realisation) of a stochastic process is a deterministic signal, there is a well-defined input signal producing a well-defined output signal corresponding to a single realisation of the output stochastic process:

$$y[n, \zeta] = \sum_{k=-\infty}^{\infty} h[k] x[n-k, \zeta] \quad (\text{M:3.4.1})$$

This is the familiar convolution integral for LTI systems, and the impulse response of this system is given by:

$$h[n] = \mathcal{L}[\delta[n]] \quad (\text{P:10-78})$$

If the sum in the right hand side (RHS) of Equation M:3.4.1 exists for all ζ such that $\Pr(\zeta) = 1$, then it is said that this sum has *almost-everywhere convergence* with probability of 1.

Theorem 11.1 (Input-output realisations for a LTI). If the process $x[n, \zeta]$ is stationary with $\mathbb{E}[|x[n, \zeta]|] < \infty$ and if the system is bounded-input, bounded-output (BIBO) stable, such that $\sum_{-\infty}^{\infty} |h[k]| < \infty$, then the output $y[n, \zeta]$ of the system in Equation M:3.4.1 converges absolutely with probability 1, or:

$$y[n, \zeta] = \sum_{k=-\infty}^{\infty} h[k] x[n-k, \zeta] \quad \text{for all } \zeta \in \mathcal{A}, \Pr(\mathcal{A}) = 1 \quad (\text{M:3.4.2})$$

- A complete description of $y[n, \zeta]$ requires the computation of an infinite number of convolutions, corresponding to each value of ζ .
- Thus, a better description would be to consider the statistical properties of $y[n, \zeta]$ in terms of the statistical properties of the input and the characteristics of the system. For Gaussian signals, which are used very often in practice, first- and second- order statistics are sufficient, since higher-order statistics are completely specified by these first two moments.

To investigate the statistical input-output properties of a linear system, note the following fundamental theorem:

Theorem 11.2 (Expectation in Linear Systems). For any linear system,

$$\mathbb{E}[\mathcal{L}[x[n]]] = \mathcal{L}[\mathbb{E}[x[n]]] \quad (11.2)$$

In other words, for example, the mean $\mu_y[n]$ of the output $y[n]$ equals the response of the system to the mean $\mu_x[n]$ of the input:

$$\mu_y[n] = \mathcal{L}[\mu_x[n]] \quad (11.3)$$

However, the definition extends to other statistics as well.

PROOF. This is a simple extension of the linearity of expected values to arbitrary linear operators.

This result will be used throughout the next section, where possible. Note, however, that while this result is very useful, it is often more practical to derive most equations from first principals.

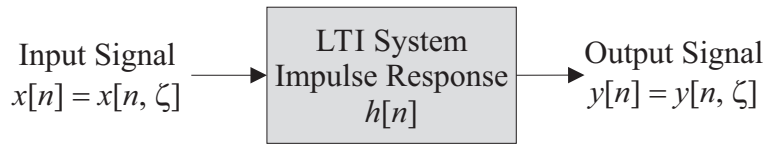
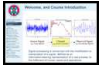


Figure 11.2: A linear time-invariant (LTI) system.

11.3.1 Input-output Statistics of a LTI System



New slide

If a stationary stochastic process $x[n]$ with mean value μ_x and correlation $r_{xx}[\ell]$ is applied to the input of a LTI system with impulse response $h[n]$ and transfer function $H(e^{j\omega})$, then the:

Output mean value is given by:

$$\mu_y = \mu_x \sum_{k=-\infty}^{\infty} h[k] = \mu_x H(e^{j0}) \quad (\text{M:3.4.4})$$

This is easily shown by using the linearity property of the expectation operator:

$$\mu_y[n] = \mathbb{E} \left[\sum_{k=-\infty}^{\infty} h[k] x[n-k] \right] = \sum_{k=-\infty}^{\infty} h[k] \mathbb{E} [x[n-k]] \quad (\text{M:3.4.4})$$

and since the process $x[n]$ is stationary, then $\mathbb{E} [x[n-k]] = \mu_x$, giving the desired result. Since μ_x and $H(e^{j0})$ are constant, μ_y is also constant. Note that $H(e^{j0})$ is the “direct current” (DC) gain of the spectrum.

Input-output cross-correlation is given by:

$$r_{xy}[\ell] = h^*[-\ell] * r_{xx}[\ell] = \sum_{k=-\infty}^{\infty} h^*[-k] r_{xx}[\ell - k] \quad (\text{M:3.4.5})$$

This can be shown by writing:

$$r_{xy}[\ell] = \mathbb{E} [x[n] y^*[n-\ell]] = \mathbb{E} [x[n+\ell] y^*[n]] \quad (11.4)$$

$$= \mathbb{E} \left[x[n+\ell] \sum_{k=-\infty}^{\infty} h^*[k] x^*[n-k] \right] \quad (11.5)$$

$$= \sum_{k=-\infty}^{\infty} h^*[k] \mathbb{E} [x[n+\ell] x^*[n-k]] \quad (11.6)$$

$$= \sum_{k=-\infty}^{\infty} h^*[k] r_{xx}[\ell+k] \quad (11.7)$$

which by making the substitution $m = -k$, gives:

$$r_{xy}[\ell] = \sum_{m=-\infty}^{\infty} h^*[-m] r_{xx}[\ell-m] = h^*[-\ell] * r_{xx}[\ell] \quad (11.8)$$

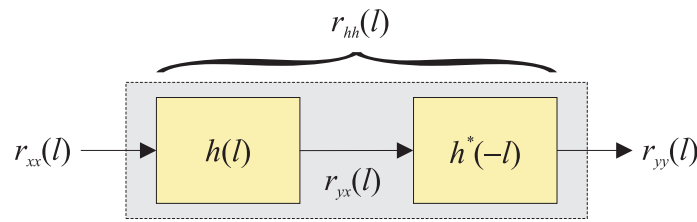


Figure 11.3: An equivalent LTI system for autocorrelation filtration.

Similarly, it follows that $r_{yx}[\ell] = h[\ell] * r_{xx}[\ell]$, and is arguably easier to prove:

$$r_{yx}[\ell] = \mathbb{E} [y[n] x^*[n - \ell]] \quad (11.9)$$

$$= \mathbb{E} \left[\sum_{k=-\infty}^{\infty} h[k] x[n - k] x^*[n - \ell] \right] \quad (11.10)$$

$$= \sum_{k=-\infty}^{\infty} h[k] \mathbb{E} [x[n - k] x^*[n - \ell]] \quad (11.11)$$

$$= \sum_{k=-\infty}^{\infty} h[k] r_{xx}[\ell - k] = h[\ell] * r_{xx}[\ell] \quad (11.12)$$

Since $r_{xy}[\ell]$ depends only on the lag ℓ , then the input and output processes of a BIBO stable linear time-invariant system, when driven by a WSS input, are jointly WSS.

Output autocorrelation is obtained by post-multiplying the system-output by $y^*[n - \ell]$ and taking expectations:

$$r_{yy}[\ell] = \mathbb{E} [y[n] y^*[n - \ell]] \quad (11.13)$$

$$= \mathbb{E} \left[\sum_{k=-\infty}^{\infty} h[k] x[n - k] y^*[n - \ell] \right] \quad (11.14)$$

and applying the linearity of the expectation operator, it follows:

$$r_{yy}[\ell] = \sum_{k=-\infty}^{\infty} h[k] \mathbb{E} [x[n - k] y^*[n - \ell]] = h[\ell] * r_{xy}[\ell] \quad (\text{M:3.4.8})$$

Substituting the expression for $r_{xy}[\ell] = h^*[-\ell] * r_{xx}[\ell]$ gives:

$$r_{yy}[\ell] = h[\ell] * h^*[-\ell] * r_{xx}[\ell] = r_{hh}[\ell] * r_{xx}[\ell] \quad (\text{M:3.4.10})$$

where $r_{hh}[\ell] = r_{hh}^*[-\ell]$ is the *autocorrelation*, for want of a better phrase, of the system impulse response:

$$r_{hh}[\ell] \triangleq h[\ell] * h^*[-\ell] = \sum_{n=-\infty}^{\infty} h[n] h^*[n - \ell] \quad (\text{M:3.4.11})$$

where \triangleq means *defined as*. If the relationship in Equation M:3.4.11 is not apparent, it can be proven by writing $g[\ell] = h^*[-\ell]$, such that the standard convolution formula gives:

$$r_{hh}[\ell] \triangleq h[\ell] * g[\ell] = \sum_{n=-\infty}^{\infty} h[n] g[\ell - n] \quad (11.15)$$

and, since $g[\ell - n] = h^*[-(\ell - n)] = h^*[n - \ell]$, Equation M:3.4.11 follows. However, this equation can also be written in an alternative form by making the substitution $m = n - \ell$ such that when $n \rightarrow \pm\infty$, $m \rightarrow \pm\infty$, and Equation M:3.4.11 becomes:

$$r_{hh}[\ell] \triangleq h[\ell] * h^*[-\ell] = \sum_{m=-\infty}^{\infty} h[m + \ell] h^*[m] \quad (\text{M:3.4.11})$$

Both of these forms of the convolution $r_{hh}[\ell] \triangleq h[\ell] * h^*[-\ell]$ are equally valid. It is straightforward to show that $r_{hh}[\ell] = r_{hh}^*[-\ell]$ by writing:

$$r_{hh}^*[-\ell] = (h[-\ell] * h^*[+\ell])^* = h[-\ell]^* * h[+\ell] = r_{hh}[\ell] \quad (11.16)$$

Since μ_y , as given by Equation M:3.4.4 is constant, and $r_{yy}[\ell]$ depends only on the lag ℓ , the response of a BIBO stable linear time-invariant to a stationary input is also a stationary process. A careful examination of Equation M:3.4.10 shows that when a signal $x[n]$ is filtered by a LTI system with impulse response $h[n]$, its autocorrelation sequence is *filtered* by a system with impulse response equal to the *autocorrelation* of its impulse response. This idea is illustrated in Figure 11.3.

Output-power of the process $y[n]$ is given by $r_{yy}[0] = \mathbb{E}[|y[n]|^2]$, and therefore since $r_{yy}[\ell] = r_{hh}[\ell] * r_{xx}[\ell]$,

$$P_{yy} = r_{hh}[\ell] * r_{xx}[\ell]|_{\ell=0} = \sum_{k=-\infty}^{\infty} r_{hh}[k] r_{xx}[-k] \quad (11.17)$$

Noting power, P_{yy} , is real, then taking complex-conjugates using $r_{xx}^*[-\ell] = r_{xx}[\ell]$:

$$P_{yy} = \sum_{k=-\infty}^{\infty} r_{hh}^*[k] r_{xx}[k] = \sum_{n=-\infty}^{\infty} h^*[n] \sum_{k=-\infty}^{\infty} r_{xx}[n+k] h[k] \quad (11.18)$$

This last step can be shown as follows:

$$P_{yy} = \sum_{k=-\infty}^{\infty} r_{hh}^*(k) r_{xx}(k) = \sum_{k=-\infty}^{\infty} \left\{ \sum_{n=-\infty}^{\infty} h^*(n) h(n-k) \right\} r_{xx}(n) \quad (11.19)$$

Hence, by rearranging the order of summation, and bringing the $h^*[n]$ forward, this gives:

$$= \sum_{n=-\infty}^{\infty} h^*(n) \sum_{k=-\infty}^{\infty} h(n-k) r_{xx}(n) \quad (11.20)$$

Then, by letting $m = n - k$, the desired result is obtained.

Output probability density function (pdf) It, in general, it is very difficult to calculate the pdf of the output of a LTI system, except in special cases, namely Gaussian processes.

If $x[n]$ is a Gaussian process, then the output is also a Gaussian process with mean and autocorrelation sequence given by Equation M:3.4.4 and Equation M:3.4.10 above. Also, if $x[n]$ is i. i. d., the pdf of the output is obtained by noting that $y[n]$ is a weighted sum of independent random variables (RVs). Indeed, as shown in earlier handouts, the pdf of the sum of independent RVs is the convolution of their pdfs or the product of their characteristic functions.

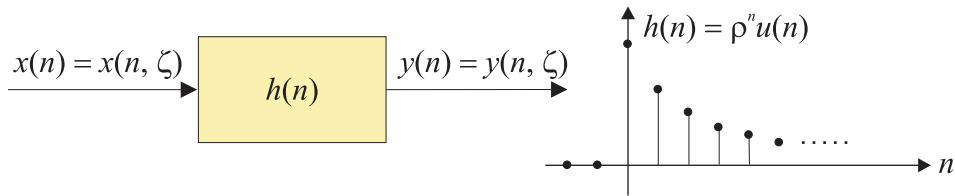


Figure 11.4: A LTI system for [Therrien:1991, Example 5.1, Page 229].

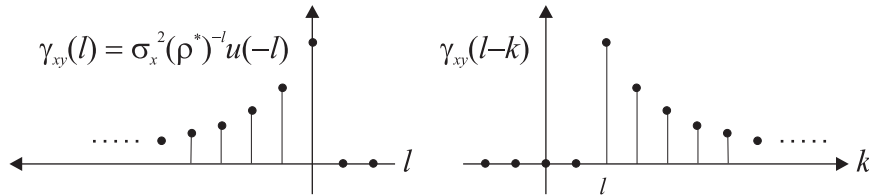


Figure 11.5: The input-output cross-covariance sequences for [Therrien:1991, Example 5.1, Page 229].

Finally, before concluding this section, note that the covariance sequences (or functions) is just the correlation sequences for the random process with the mean removed. As a result, the covariance functions satisfy a set of equations analogous to those derived above. For completeness, they are listed below:

$$\gamma_{yx}[\ell] = h[\ell] * \gamma_{xx}[\ell] \tag{T:5.18}$$

$$\gamma_{xy}[\ell] = h^*[-\ell] * \gamma_{xx}[\ell] \tag{T:5.19}$$

$$\gamma_{yy}[\ell] = h[\ell] * \gamma_{xy}[\ell] \tag{T:5.20}$$

$$= h[\ell] * h^*[-\ell] * \gamma_{xx}[\ell] \tag{T:5.21}$$

The following example illustrates the application of these results.

Example 11.2 (Simple example [Therrien:1991, Example 5.1, Page 229]). The LTI system shown in Figure 11.4 is driven by a process with mean μ_x and covariance sequence $\gamma_{xx}[\ell] = \sigma_x^2\delta[\ell]$; note that this input process is white noise with an added nonzero mean.

Calculate the mean, autocorrelation and autocovariance sequences of the output, $y[n]$, as well as the cross-correlation and cross-covariance functions between the input and the output.

SOLUTION. Each of these functions may be calculated using the equations listed in this section. Hence:

Output mean value First, calculate the mean. Using Equation M:3.4.4, then:

$$\mu_y = \mu_x \sum_{k=-\infty}^{\infty} h[k] = \mu_x \sum_{k=0}^{\infty} \rho^k = \frac{\mu_x}{1 - \rho} \tag{11.21}$$

Input-output cross-covariance Since the input and the output both have nonzero mean, then it is easiest to first calculate the auto- and cross-covariance functions, and then use these to find the auto- and cross-correlation functions.

Thus, the output-input cross-covariance is given by Equation T:5.18:

$$\gamma_{yx}[\ell] = h[\ell] * \gamma_{xx}[\ell] = (\rho^\ell u[\ell]) * (\sigma_x^2\delta[\ell]) = \sigma_x^2\rho^\ell u[\ell] \tag{11.22}$$

and therefore the input-output cross-covariance is

$$\gamma_{xy}[\ell] = \gamma_{yx}^*[-\ell] = \sigma_x^2(\rho^*)^{-\ell}u[-\ell] \quad (11.23)$$

Output autocovariance Next, using Equation T:5.20, then:

$$\gamma_{yy}[\ell] = h[\ell] * \gamma_{xy}[\ell] = \sum_{k=-\infty}^{\infty} h[k] \gamma_{xy}[\ell - k] \quad (11.24)$$

The input-output cross-covariance sequence, $\gamma_{xy}[\ell]$, is plotted in Figure 11.5, along with $\gamma_{xy}[\ell - k]$ as a function of k .

Hence, if $\ell > 0$ it follows

$$\gamma_{yy}[\ell] = \sum_{k=\ell}^{\infty} h[k] \gamma_{xy}[\ell - k] = \sum_{k=\ell}^{\infty} \rho^k \sigma_x^2(\rho^*)^{-(\ell-k)} \quad (11.25)$$

Substituting $m = k - \ell$, such that when $k = \{\ell, \infty\}$, then $m = \{0, \infty\}$, and so:

$$\gamma_{yy}[\ell] = \sigma_x^2 \sum_{m=0}^{\infty} \rho^\ell \rho^m (\rho^*)^m \quad (11.26)$$

$$= \sigma_x^2 \rho^\ell \sum_{m=0}^{\infty} (|\rho|^2)^m = \frac{\sigma_x^2 \rho^\ell}{1 - |\rho|^2}, \ell > 0 \quad (11.27)$$

If $\ell \leq 0$, then the summation is slightly different:

$$\gamma_{yy}[\ell] = \sum_{k=0}^{\infty} \rho^k \sigma_x^2(\rho^*)^{-(\ell-k)} \quad (11.28)$$

$$= \sigma_x^2(\rho^*)^{-\ell} \sum_{k=0}^{\infty} (|\rho|^2)^k = \frac{\sigma_x^2(\rho^*)^{-\ell}}{1 - |\rho|^2}, \ell \leq 0 \quad (11.29)$$

Input-output cross-correlation This can now be calculated using the relationship:

$$r_{xy}[\ell] = \gamma_{xy}[\ell] + \mu_x \mu_y^* \quad (11.30)$$

$$= \sigma_x^2(\rho^*)^{-\ell}u[-\ell] + \mu_x \frac{\mu_x^*}{1 - \rho^*} \quad (11.31)$$

$$= \sigma_x^2(\rho^*)^{-\ell}u[-\ell] + \frac{|\mu_x|^2}{1 - \rho^*} \quad (11.32)$$

Output autocorrelation In a similar manner, the autocorrelation of the output is given by:

$$r_{yy}[\ell] = \gamma_{yy}[\ell] + |\mu_y|^2 = \begin{cases} \frac{\sigma_x^2 \rho^\ell}{1 - |\rho|^2} + \left| \frac{\mu_x}{1 - \rho} \right|^2 & \ell > 0 \\ \frac{\sigma_x^2(\rho^*)^{-\ell}}{1 - |\rho|^2} + \left| \frac{\mu_x}{1 - \rho} \right|^2 & \ell \leq 0 \end{cases} \quad (11.33) \quad \square$$

Note that these results show that a process with the exponential correlation function can always be generated by applying white noise to a stable first-order system. More generally, in the next handout, it will be seen that wide-sense stationary of arbitrary autocorrelation sequence can be obtained by driving a LTI system by WGN.



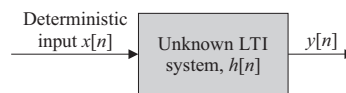
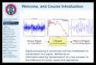


Figure 11.6: What signals might be used for System Identification?

11.3.2 System identification by cross-correlation



Topic Summary 71 Application of Cross-Correlation to System Identification

New slide

Topic Objectives:

- Concept of the output of a system to stochastic input.
- Overview of methods for Calculating Input-Output Statistics.
- Introduction of Monte Carlo calculation for Input-Output Statistics.

Topic Activities:

Type	Details	Duration	Progress
Watch video	14 : 24 min video	3 × length	
Read Handout	Read page 385 to page 387	8 mins/page	
Try Example	Try Example 11.3 using MATLAB	10 mins	



http://media.ed.ac.uk/media/1_e6662yx1

Video Summary: This video introduces the important signal processing application of system identification; identifying the system impulse response or transfer function through measurements. The video highlights the advantages and disadvantages of the three key deterministic approaches, using as the input an impulse, or step function, or harmonic input. A fourth method which relies on a stochastic input is then presented, namely driving a system with WGN. It is then shown, using the theory presented earlier in the course, that the cross-correlation between the input and output is the impulse response. The sample cross-correlation is highlighted as a way of estimating the cross-correlation from a single realisation of the random process, where ergodicity of the output has been assumed. Finally, as simple exam is implemented in MATLAB.

There are three key methods from our deterministic signal analysis for system identification:

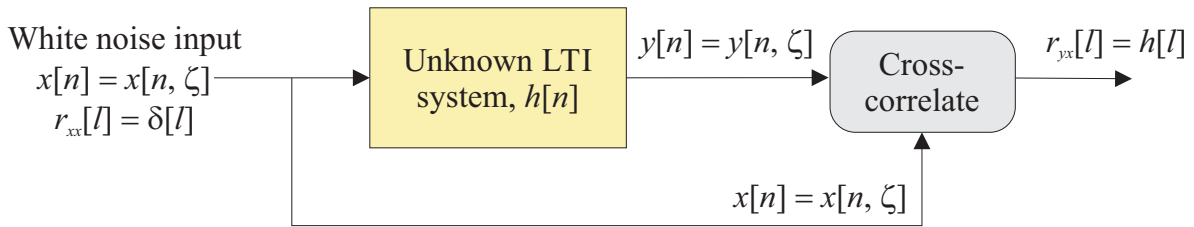


Figure 11.7: System identification by cross-correlation.

Impulse A simple input, but **difficult to generate**. The output is $y[n] = h[n]$, the system impulse response.

Step input A simple to generate signal, with the output $y[n] = \sum_{k=0}^n h[k]$ being the step response. The impulse response is obtained by taking the difference sequence at the output (equivalent to differentiating).

This is problematic, as the difference signal can lead to errors when there is a small amount of noise in the signals.

Harmonic input A simple to generate signal, $x[n] = \cos \omega_0 n$, leading to the output:

$$y[n] = |H(e^{j\omega_0})| \cos(\omega_0 n + \arg H(e^{j\omega_0})) \quad (11.34)$$

By sweeping across frequencies, the magnitude and phase response of $H(e^{j\omega})$ can be calculated. The inverse-discrete-time Fourier transform (DTFT) can then be used to reconstruct the impulse response, $h[n]$.

This method is potentially very accurate, but equally it is very slow as a result.

The input-output cross-correlation of a LTI system is the basis for a classical method of identification of an unknown linear system.

The system is excited with a WGN input with ACS:

$$r_{xx}[\ell] = \delta[\ell] \quad (11.35)$$

Since the output-input cross-correlation can be written as:

$$r_{yx}[\ell] = h[\ell] * r_{xx}[\ell] \quad (\text{M:3.4.6})$$

then, with $r_{xx}[\ell] = \delta[\ell]$, it follows:

$$r_{yx}[\ell] = h[\ell] * \delta[\ell] = h[\ell] \quad (11.36)$$

Hence, the impulse response of an unknown LTI system can be estimated by exciting the system with WGN and evaluating the input-output cross-correlation.

If the discrete system represents a sampled continuous system, this method of estimating the impulse response out-performs an estimation based on simply driving the system by an impulse since:

1. it is easier to generate an approximation to white noise than to generate an approximation to an impulse, since the latter must have finite energy in an almost zero-width pulse;
2. application of an impulse to a physical system requires driving it *very* hard, albeit for a very short time, and may cause damage. Driving a system with white noise is less traumatic. As an example, consider estimating the acoustic impulse response (AIR) of a concert hall or office; one method of generating an impulse is to fire a gun and, obviously, this will damage the concert hall, which is less than desirable.

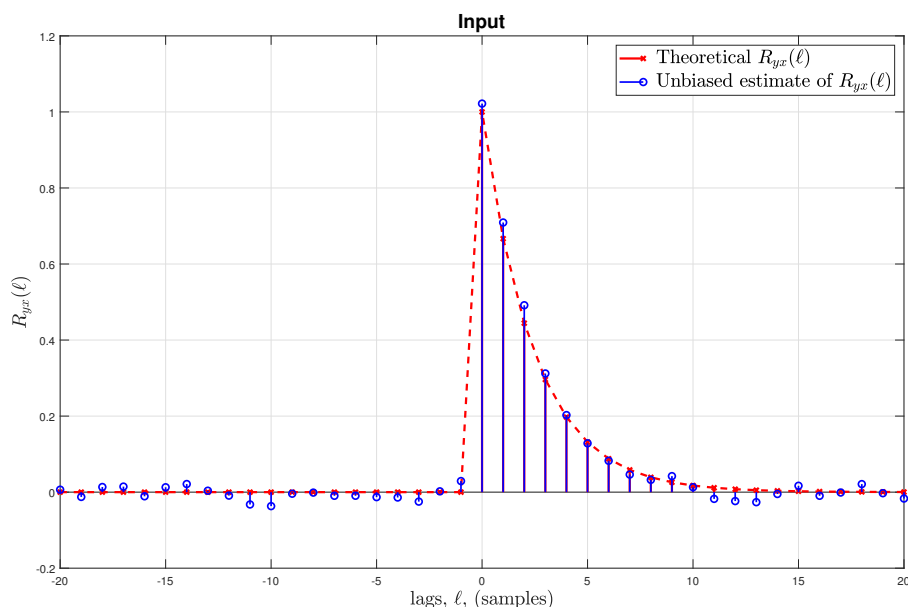


Figure 11.8: The theoretical impulse response $h[n] = \left(\frac{2}{3}\right)^n u[n]$ and the time-averaged estimate of the cross-correlation sequence $\hat{R}_{yx}[\ell]$.

As the input or excitation process is WGN, then the output is WSS, and in many cases will be ergodic. Hence, the cross-correlation (and therefore system impulse response) can be estimated from a single realisation using the *sample cross-correlation function*:

$$\hat{r}_{yx}[\ell] = \frac{1}{N} \sum_{n=0}^{N-1-|\ell|} y[n+|\ell|] x[n], \quad |\ell| < N \quad (11.37)$$

$$\hat{r}'_{yx}[\ell] = \frac{1}{N-|\ell|} \sum_{n=0}^{N-1-|\ell|} y[n+|\ell|] x[n], \quad |\ell| < N \quad (11.38)$$

It is simple to generate an example in MATLAB.

Example 11.3 (Low-pass filter). A system is described by $y[n] = \frac{2}{3}y[n-1] + x[n]$, although this is not known to the observer initially. By driving the system with WGN, calculate the impulse response of the system through numerical simulation.

SOLUTION. See the MATLAB code on LEARN, to obtain the numerical result shown in Figure 11.8.

– End-of-Topic 70: **Application of Cross-Correlation to System Identification** –



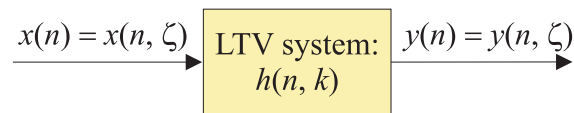
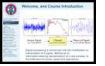


Figure 11.9: General linear time-varying (LTV) system with nonstationary input; the impulse response $h[n, k]$ is the response at index n to an impulse occurring at time index k .

11.4 LTV Systems with Nonstationary Inputs



Topic Summary 72 Analysis of linear time-varying (LTV) systems and other special cases

New slide

Topic Objectives:

- This topic is not currently examinable.

It is also possible to analyse a general linear system that is not necessarily time-invariant, as shown in Figure 11.9; such a system is called linear time-varying (LTV).

The input and output are related by the generalised convolution:

$$y(n) = \sum_{k=-\infty}^{\infty} h(n, k) x(k) \quad (\text{T:5.1})$$

where $h(n, k)$ is the response at time-index n to an impulse occurring at the system input at time-index k . The mean, autocorrelation and autocovariance sequences of the output, $y(n)$, as well as the cross-correlation and cross-covariance functions between the input and the output, can be calculated in a similar way as for LTI systems with stationary inputs. It is left as an exercise to the reader to derive these, but the results are summarised in the next section.

11.4.1 Input-output Statistics of a LTV System

It is important to note that the input-output statistics of a LTI system with a stationary input are simply special cases of the following results. Thus, it is perhaps preferable to remember these more general results and simplify them as necessary.

Output mean value is given by

$$\mu_y(n) = \sum_{k=-\infty}^{\infty} h(n, k) \mu_x(k) \quad (\text{T:5.2})$$

This can be written as:

$$\mu_y(n) = L[\mu_x(n)] \quad (\text{P:10-80})$$

Output-input cross-correlation is given by

$$r_{yx}(n, m) = \sum_{k=-\infty}^{\infty} h(n, k) r_{xx}(k, m) \quad (\text{T:5.5})$$

and the input-output cross-correlation is:

$$r_{xy}(n, m) = r_{yx}^*(m, n) \quad (\text{T:5.4})$$

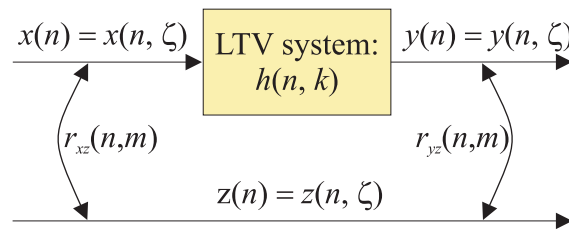


Figure 11.10: Cross-correlation with respect to a third random process.

Output autocorrelation is a similar form, given by:

$$r_{yy}(n, m) = \sum_{k=-\infty}^{\infty} h(n, k) r_{xy}(k, m) \quad (\text{T:5.3})$$

Output-input cross-covariance has an identical form to that for the input-output cross-correlation functions:

$$\gamma_{yx}(n, m) = r_{yx}(n, m) - \mu_y(n) \mu_x^*(m) \quad (11.39)$$

$$= \sum_{k=-\infty}^{\infty} h(n, k) \gamma_{xx}(k, m) \quad (\text{T:5.9})$$

and

$$\gamma_{yx}(n, m) = \gamma_{xy}^*(m, n) \quad (\text{T:5.8})$$

Output autocovariance is given by:

$$\gamma_{yy}(n, m) = r_{yy}(n, m) - \mu_y(n) \mu_y^*(m) \quad (\text{T:5.6})$$

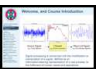
$$= \sum_{k=-\infty}^{\infty} h(n, k) \gamma_{xy}(k, m) \quad (\text{T:5.7})$$

Note that if the impulse response of the system has *finite support*, in the sense the region over which it has non-zero values is a well-defined finite region, then it is possible to represent the correlation functions and the impulse response function in matrix form:

$$\mathbf{R}_{yy} = \mathbf{H} \mathbf{R}_{xx} \mathbf{H}^H \quad (11.40)$$

Correlation matrices were introduced in an earlier handout.

11.4.2 Effect of Linear Transformations on Cross-correlation



Another situation worth considering is the cross-correlation with respect to a third random process, as shown in Figure 11.10. *New slide*

A random process $x[n]$ is transformed by a LTV system to produce another signal $y[n]$. The process $x[n]$ is related to a third process $z[n]$, and $r_{xz}[n_1, n_2]$ is known. It is desirable to find $r_{yz}[n_1, n_2]$. The response of the LTV system to $x[n]$ is:

$$y[n] = \sum_{k \in \mathbb{Z}} h[n, k] x[k] \quad (\text{T:5.22})$$

Hence, multiplying both sides by $z^*[m]$ and taking expectations:

$$r_{yz}[n, m] = \sum_{k \in \mathbb{Z}} h[n, k] r_{xz}[k, m] = h[n, k] * r_{xz}[k, m] \quad (\text{T:5.24})$$

If the system is LTI, then this simplifies to:

$$r_{yz}[\ell] = \sum_{k \in \mathbb{Z}} h[k] r_{xz}[\ell - k] = h[\ell] * r_{xz}[\ell] \quad (11.41)$$

– End-of-Topic 71: **Analysis of LTV systems and other special cases** –



11.5 Time-Domain Analysis with Difference Equations

Topic Summary 73 Difference Equation Analysis of Input-Output Time-Domain Statistics

Topic Objectives:

- Revising the difference-equation formulation of linear systems.
- Deriving the input-output statistics in terms of the difference equations.
- A worked example of solving the difference equations for a first-order system.

Topic Activities:

Type	Details	Duration	Progress
Watch video	20 : 14 min video	3× length	
Read Handout	Read page 391 to page 394	8 mins/page	
Try Example	Try Example 11.4	25 mins	
Practice Exercises	Exercise ??	60 mins	

The screenshot shows a video player interface. On the left is a video thumbnail of James R. Hoggood. The main content is a slide titled "Analysis with Difference Equations". The slide contains a block diagram of an LTI system with feedback and feedforward paths, and a text box stating: "A mathematically elegant analysis of stochastic systems comes when a LTI system can be represented by difference equations." The slide also includes a table of contents on the left and a play button in the center.

http://media.ed.ac.uk/media/1_wmwxloel

Video Summary: This topic considers extending previous topics on calculating the input-output statistics of a LTI system in response to a WSS process at the input, when the LTI system is described by a difference equation. The video begins by reviewing the difference-equations description of linear filters, and different possibilities for manipulating the system. The video proposes a single approach by showing that the input-output statistics satisfy the same difference equation that describes the system. Therefore, through solving this difference equation, the desired statistics can be obtained. A detailed example is then provided for a first-order linear system.

A mathematically elegant analysis of stochastic systems comes about when a LTI system can be represented by difference equations. This will be particularly useful in the next handout on linear signal models. Although the results of the preceding sections apply to these systems, the difference equation approach offers an alternative representation of the results obtained with the impulse response function, that can sometimes be quite useful and important. It is possible to use a combination of methods, such as taking the transfer function of a difference to find the impulse response, and then use convolution. The purpose of the difference equation approach is to do the calculations in a single approach.

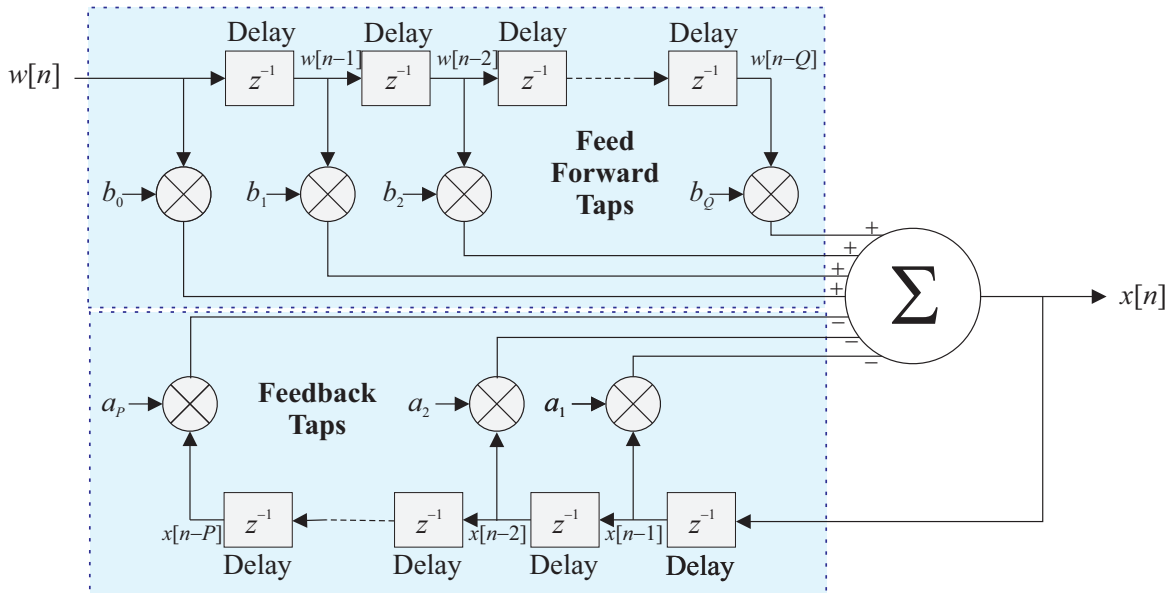


Figure 11.11: Difference-equation description of a LTI system.

Consider a LTI system that can be represented by a difference equation:

$$y[n] = - \sum_{p=1}^P a_p y[n-p] + \sum_{q=0}^Q b_q x[n-q] \quad (11.42)$$

which is often written in the more compact form:

$$\sum_{p=0}^P a_p y[n-p] = \sum_{q=0}^Q b_q x[n-q] \quad (11.43)$$

where $a_0 \triangleq 1$. Assuming that both $x[n]$ and $y[n]$ are stationary processes, such that $\mathbb{E}[x[n-p]] = \mu_x$ and $\mathbb{E}[y[n-q]] = \mu_y$, then taking expectations of both sides gives, after a little rearrangement:

$$\mu_y = \frac{\sum_{q=0}^Q b_q}{1 + \sum_{p=1}^P a_p} \mu_x \quad (11.44)$$

Without a priori assuming stationarity, then multiplying the system equation throughout by $y^*[m]$ and taking expectations gives:

$$\sum_{p=0}^P a_p r_{yy}[n-p, m] = \sum_{q=0}^Q b_q r_{xy}[n-q, m] \quad (11.45)$$

Assuming stationarity, and setting $\ell = n - m$, this simplifies to:

$$\sum_{p=0}^P a_p r_{yy}[\ell-p] = \sum_{q=0}^Q b_q r_{xy}[\ell-q] \quad (11.46)$$

Similarly, rather than multiplying throughout the system equation by $y^*[m]$, instead multiply through by $x^*[m]$ to give:

$$\sum_{p=0}^P a_p r_{yx}[n-p, m] = \sum_{q=0}^Q b_q r_{xx}[n-q, m] \quad (11.47)$$

and again assuming stationarity, this simplifies to:

$$\sum_{p=0}^P a_p r_{yx}[\ell - p] = \sum_{q=0}^Q b_q r_{xx}[\ell - q] \quad (11.48)$$

These two sets of difference equations may be used to solve for $r_{yy}[n_1, n_2]$ and $r_{xy}[n_1, n_2]$ in the nonstationary case, or in the stationary case. Note the statistics auto- and cross-correlation statistics satisfy the original difference equations. Similar expressions can be obtained for the covariance sequences. They are given by:

$$\sum_{p=0}^P a_p \gamma_{yy}[n - p, m] = \sum_{q=0}^Q b_q \gamma_{xy}[n - q, m] \quad (11.49)$$

and

$$\sum_{p=0}^P a_p \gamma_{yx}[n - p, m] = \sum_{q=0}^Q b_q \gamma_{xx}[n - q, m] \quad (11.50)$$

or, if the signals are stationary, then:

$$\sum_{p=0}^P a_p \gamma_{yy}[\ell - p] = \sum_{q=0}^Q b_q \gamma_{xy}[\ell - q] \quad (11.51)$$

and

$$\sum_{p=0}^P a_p \gamma_{yx}[\ell - p] = \sum_{q=0}^Q b_q \gamma_{xx}[\ell - q] \quad (11.52)$$

Example 11.4 ([Manolakis:2000, Example 3.6.2, Page 141]). Let $x[n]$ be a random process generated by the first order difference equation given by:

$$x[n] = \alpha x[n - 1] + w[n], \quad |\alpha| \leq 1, n \in \mathbb{Z} \quad (11.53)$$

where $w[n] \sim \mathcal{N}(\mu_w, \sigma_w^2)$ is an i. i. d. WGN process.

- Demonstrate that the process $x[n]$ is stationary, and calculate the mean μ_x .
- Determine the autocovariance and autocorrelation sequences, $\gamma_{xx}[\ell]$ and $r_{xx}[\ell]$.

SOLUTION. Note that this is a first-order autoregressive (AR) process, which will be discussed in more detail later in the lecture course. The case written above is, in fact, the stationary case, and [Manolakis, Exercise 3.23, Page 145] poses the case where there is an initial transient, resulting in a nonstationary autocorrelation function. This exercise is left for those interested, although be forewarned that this is not an easy exercise. This example uses the theory described above.

- The output of a LTI system with a stationary input is always stationary, although this can also be proved explicitly. It follows directly from the results above that:

$$\mu_x = \frac{\mu_w}{1 - \alpha} \quad (11.54)$$

- Using the results for the input-output covariance of a LTI system represented by difference equation:

$$\gamma_{xx}[n, m] - \alpha \gamma_{xx}[n-1, m] = \gamma_{wx}[n, m] \quad (11.55)$$

$$\gamma_{xw}[n, m] - \alpha \gamma_{xw}[n-1, m] = \gamma_{ww}[n, m] \quad (11.56)$$

which, since the system is stationary, can be written as:

$$\gamma_{xx}[\ell] - \alpha \gamma_{xx}[\ell-1] = \gamma_{wx}[\ell] \quad (11.57)$$

$$\gamma_{xw}[\ell] - \alpha \gamma_{xw}[\ell-1] = \gamma_{ww}[\ell] \quad (11.58)$$

Noting $x[n]$ cannot depend on future values of $w[n]$, then $\gamma_{xw}[n+\ell, n] = \gamma_{xw}[\ell] = 0$, $\ell < 0$. This can be demonstrated by explicitly evaluating $r_{xw}[n, m]$, $m < n$ or $r_{xw}[\ell] = \mathbb{E}[x[n] w^*[n-\ell]]$, and noting that $x[n]$ and $w[n]$ are independent. If $\ell < 0$, then $w[n-\ell]$ is a sample with time-index greater than that of $x[n]$, or in other words a future value.

Since $\gamma_{ww}[\ell] = \sigma_w^2 \delta[\ell]$, the second of the difference equations above becomes:

$$\gamma_{xw}[\ell] = \begin{cases} \alpha \gamma_{xw}[\ell-1] & \ell > 0 \\ \sigma_w^2 & \ell = 0 \\ 0 & \ell < 0 \end{cases} \quad (11.59)$$

Solving for $\ell \geq 0$ gives by repeated substitution, $\gamma_{xw}[\ell] = \alpha^\ell \sigma_w^2$, and zero for $\ell < 0$.

Since $\gamma_{wx}[\ell] = \gamma_{xw}^*[-\ell]$, then the difference equation for the autocovariance function of $x[n]$ simplifies to:

$$\gamma_{xx}[\ell] - \alpha \gamma_{xx}[\ell-1] = \begin{cases} 0 & \ell > 0 \\ \alpha^{-\ell} \sigma_w^2 & \ell \leq 0 \end{cases} \quad (11.60)$$

Note the solution for $\ell > 0$ is the solution of the homogeneous equation. Hence, since $\gamma_{xx}[\ell] = \gamma_{xx}[-\ell]$ for a real process, then this equation is solved by assuming the solution:

Assuming the solution:

$$\gamma_{xx}[\ell] = a \alpha^{|\ell|} + b \quad (11.61)$$

The values of a and b can be found by directly substituting the proposed solution for $\ell \leq 0$ into the difference equation:

$$a \alpha^{-\ell} + b - \alpha (a \alpha^{-(\ell-1)} + b) = \alpha^{-\ell} \sigma_w^2 \quad (11.62)$$

$$\alpha^{-\ell} (1 - \alpha^2) a + (1 - \alpha) b = \alpha^{-\ell} \sigma_w^2 \quad (11.63)$$

from which it directly follows that $b = 0$ and $a = \sigma_w^2 = \frac{\sigma_w^2}{1-\alpha^2}$, corresponding to the case when $\ell = 0$.

Hence, in conclusion

$$\gamma_{xx}[\ell] = \frac{\sigma_w^2}{1-\alpha^2} \alpha^{|\ell|} \quad (11.64)$$

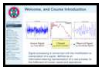
Using the relationship that $r_{xx}[\ell] = \gamma_{xx}[\ell] + \mu_x^2$, it follows that the output auto-correlation is given by:

$$r_{xx}[\ell] = \frac{\sigma_w^2}{1-\alpha^2} \alpha^{|\ell|} + \frac{\mu_w^2}{(1-\alpha)^2} \quad (11.65)$$

As usual, if $\mu_w = 0$, then $r_{xx}[\ell] = \gamma_{xx}[\ell]$. □



11.6 Frequency-Domain Analysis of LTI systems



Topic Summary 74 Frequency-domain analysis of input-output statistics

New slide

- Topic Objectives:**
- Introduction to frequency and transform domain analysis for input-output statistics.
 - Derivation and property of the complex spectral relationships between the system input and system output.
 - Several worked examples of calculations in the transform domain.

Topic Activities:

Type	Details	Duration	Progress
Watch video	28 : 04 min video	3× length	
Read Handout	Read page 395 to page 400	8 mins/page	
Try Example	Try Examples 11.5 and 11.6	25 mins	
Practice Exercises	Exercises ?? to ??	40 mins	

http://media.ed.ac.uk/media/1_xzqslifj

Video Summary: This Topic gives a comprehensive overview of using a frequency-domain analysis technique for evaluating the input-output statistics of a LTI system with a WSS input. By taking the DTFT or z -transforms of the time-domain relationships introduced in earlier topics, the transform domain relationships are obtained. The video then covers two detailed examples showing the various steps in the analysis technique; namely, first, find the system transfer function and complex-spectral density of the input statistics; second, simplify the transform domain using, for example, partial fraction expansion; and third, take inverse-transforms using, for example, z -transform tables. The video briefly discusses the trade-off between using the transform vs time-domain analysis techniques.

Now consider how a LTI transformation affects the power spectra and complex power density spectra of a stationary random process. Recall that the power spectral density (PSD) is the Fourier transform of the **autocorrelation** functions. Alternatively, it is possible to note that the frequency response of a system is the z -transform evaluated on the unit circle.

Taking the DTFT of the time-domain relationships for the input-output statistics in terms of the system

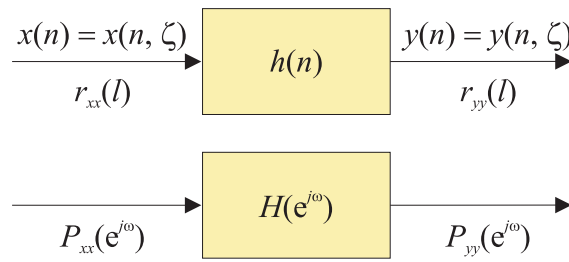


Figure 11.12: The PSD at the input and output of a LTI system with WSS input.

impulse response leads to the following spectral densities:

$$r_{xy}[\ell] = h^*[-\ell] * r_{xx}[\ell] \Rightarrow P_{xy}(e^{j\omega}) = H^*(e^{j\omega}) P_{xx}(e^{j\omega}) \quad (\text{M:3.4.19})$$

$$r_{yx}[\ell] = h[\ell] * r_{xx}[\ell] \Rightarrow P_{yx}(e^{j\omega}) = H(e^{j\omega}) P_{xx}(e^{j\omega}) \quad (\text{M:3.4.20})$$

$$r_{yy}[\ell] = h^*[-\ell] * h[\ell] * r_{xx}[\ell] \Rightarrow P_{yy}(e^{j\omega}) = |H(e^{j\omega})|^2 P_{xx}(e^{j\omega}) \quad (\text{M:3.4.21})$$

These results are derived very easily from the results in Section 11.3.1 and the properties of the Fourier transform, especially that convolution becomes multiplication. It is important to stress the similarity of these results with those for the frequency analysis of linear time-invariant systems with deterministic signal inputs. The system is depicted in Figure 11.12.

- If the input and output autocorrelations or autospectral densities are known, the magnitude response of a system $|H(e^{j\omega})|$ can be determined, but not the phase response.
- Only cross-correlation or cross-spectral information can help determine the phase response.

A set of similar relations to Equation M:3.4.19, Equation M:3.4.20 and Equation M:3.4.21 can also be derived for the complex spectral density function. Specifically, if: $h[\ell] \stackrel{z}{\rightleftharpoons} H(z)$, then:

$$h^*[-\ell] \stackrel{z}{\rightleftharpoons} H^*(1/z^*) \quad (11.66)$$

Therefore, the input output relationships:

$$r_{xy}[\ell] = h^*[-\ell] * r_{xx}[\ell] \quad (11.67)$$

$$r_{yx}[\ell] = h[\ell] * r_{xx}[\ell] \quad (11.68)$$

$$r_{yy}[\ell] = h[\ell] * r_{xy}[\ell] \quad (11.69)$$

$$= h[\ell] * h^*[-\ell] * r_{xx}[\ell] \quad (11.70)$$

transform to the spectral relationships:

$$P_{xy}(z) = H^*(1/z^*) P_{xx}(z) \quad (\text{T:5.41})$$

$$P_{yx}(z) = H(z) P_{xx}(z) \quad (\text{T:5.40})$$

$$P_{yy}(z) = H(z) P_{xy}(z) \quad (\text{T:5.42})$$

$$P_{yy}(z) = H(z) H^*(1/z^*) P_{xx}(z) \quad (\text{T:5.44})$$

Note that $P_{yy}(z)$ satisfies the required property for a complex spectral density function, namely that $P_{yy}(z) = P_{yy}^*(1/z^*)$. Also, note the following result for real filters that make the above equations simplify accordingly.

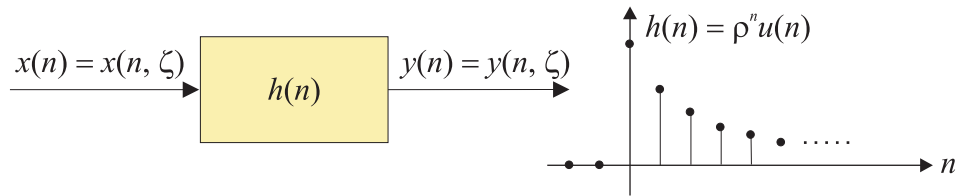


Figure 11.13: Equivalent figure to Figure 11.13: a LTI system for [Therrien:1991, Example 5.1, Page 229].

Theorem 11.3 (Transfer function for a real filter). For a real filter:

$$h[-\ell] \stackrel{z}{\rightleftharpoons} H^* \left(\frac{1}{z^*} \right) = H(z^{-1}) \tag{11.71}$$

PROOF. Writing:

$$H(z) = \sum_{n=-\infty}^{\infty} h[n] z^{-n} \tag{11.72}$$

then setting $z \rightarrow \frac{1}{z^*}$ gives:

$$H \left(\frac{1}{z^*} \right) = \sum_{n=-\infty}^{\infty} h[n] \left[\frac{1}{z^*} \right]^{-n} \tag{11.73}$$

Now, taking complex-conjugates, using the following facts:

- the conjugate of a sum/product of complex numbers is the sum/product of the conjugates of the complex numbers, or in otherwords $(a + b)^* = a^* + b^*$ and $(ab)^* = a^*b^*$,
- the filter coefficients are real, such that $h^*[n] = h[n]$,

then

$$H^* \left(\frac{1}{z^*} \right) = \sum_{n=-\infty}^{\infty} h(n) z^n \equiv \sum_{m=-\infty}^{\infty} h(-m) z^{-m} \tag{11.74}$$

where in the last step, the substitution $m = -n$ has been made. Hence, this gives the desired result. It is straightforward to adapt the final stage of this proof to show that $h^*[-\ell] \stackrel{z}{\rightleftharpoons} H^* \left(\frac{1}{z^*} \right)$ in general.

Consider again the earlier example based on [Therrien:1991, Example 5.1, Page 229].

Example 11.5 (Simple Example: [Therrien:1991, Example 5.3, Page 237]). Again, the LTI system shown in Figure 11.4 is driven by a process with mean μ_x and covariance sequence $\gamma_{xx}[\ell] = \sigma_x^2 \delta[\ell]$. Calculate the PSD, cross-power spectral density (CPSD) and the complex spectral densities.

SOLUTION. The first-order system with impulse response $h[n] = \rho^n u[n]$ has system transfer function:

$$H(z) = \frac{1}{1 - \rho z^{-1}} \tag{11.75}$$

The complex spectral density function for the white noise with added mean is given by the z -transform of the autocorrelation sequence. Since $\gamma_{xx}[\ell] = \sigma_x^2 \delta[\ell]$, then:

$$r_{xx}[\ell] = \gamma_{xx}[\ell] + \mu_x^2 = \sigma_x^2 \delta[\ell] + |\mu_x|^2 \tag{11.76}$$

Taking z -transforms gives:

$$P_{xx}(z) = \sigma_x^2 + 2\pi|\mu_x|^2\delta(z - e^{j0}) \quad (11.77)$$

$$= \sigma_x^2 + 2\pi|\mu_x|^2\delta(z - 1) \quad (11.78)$$

where the complex spectral density result in Equation (T:4.59) at the end of the previous handout has been used. Hence, the complex cross-spectral density is given by:

$$P_{xy}(z) = H^*(1/z^*) P_{xx}(z) \quad (11.79)$$

$$= \left(\frac{1}{1 - \rho \left[\frac{1}{z^*} \right]^{-1}} \right)^* [\sigma_x^2 + 2\pi|\mu_x|^2\delta(z - 1)] \quad (11.80)$$

$$= \frac{\sigma_x^2}{1 - \rho^* z} + \frac{2\pi|\mu_x|^2}{1 - \rho^* z} \delta(z - 1) \quad (11.81)$$

Moreover, the complex spectral density is given by:

$$P_{yy}(z) = H(z) P_{xy}(z) \quad (11.82)$$

$$= \left(\frac{1}{1 - \rho z^{-1}} \right) \left(\frac{1}{1 - \rho^* z} \right) [\sigma_x^2 + 2\pi|\mu_x|^2\delta(z - 1)] \quad (11.83)$$

$$= \frac{\sigma_x^2}{1 - |\rho|^2} \frac{1 - |\rho|^2}{(1 - \rho z^{-1})(1 - \rho^* z)} + \frac{2\pi|\mu_x|^2}{|1 - \rho|^2} \delta(z - 1) \quad (11.84)$$

$$= \frac{\sigma_x^2}{1 + |\rho|^2 - \rho^* z - \rho z^{-1}} + \frac{2\pi|\mu_x|^2}{|1 - \rho|^2} \delta(z - 1) \quad (11.85)$$

The CPSD and the PSD are found by setting $z = e^{j\omega}$ to obtain:

$$P_{xy}(e^{j\omega}) = \frac{\sigma_x^2}{1 - \rho^* e^{j\omega}} + \frac{2\pi|\mu_x|^2}{1 - \rho^* e^{j\omega}} \delta(e^{j\omega} - 1) \quad (11.86)$$

Moreover, the PSD is given by:

$$P_{yy}(e^{j\omega}) = \frac{\sigma_x^2}{1 - |\rho|^2} \frac{1 - |\rho|^2}{1 + |\rho|^2 - 2|\rho| \cos(\omega - \arg \rho)} + \frac{2\pi|\mu_x|^2}{|1 - \rho|^2} \delta(e^{j\omega} - 1) \quad (11.87)$$

where the simplification that:

$$\rho^* e^{j\omega} + \rho e^{-j\omega} = |\rho| [e^{-j \arg \rho} e^{j\omega} + e^{j \arg \rho} e^{-j\omega}] = |\rho| [e^{j(\omega - \arg \rho)} + e^{-j(\omega - \arg \rho)}] \quad (11.88)$$

$$= 2|\rho| \cos(\omega - \arg \rho) \quad (11.89)$$

has been used.

Taking inverse z -transforms of Equation 11.84 gives the output ACS:

$$r_{yy}[\ell] = \frac{\sigma_x^2}{1 - |\rho|^2} \rho^{|\ell|} + \frac{|\mu_x|^2}{|1 - \rho|^2} \quad (11.90)$$

□

This matches the solutions found using: the impulse response approach, or the difference equation approach.

Example 11.6 (Partial Fractions Example). The signal $y[n]$ from Example 11.5 is applied to the input of a causal LTI system with output $s[n]$ which is characterised by the difference equation:

$$s[n] = \rho s[n - 1] + y[n] + y[n - 1] \quad (11.91)$$

- Show that the cross-power spectral density is given by:

$$P_{sy}(z) = \frac{\sigma_x^2}{1 - \rho z^{-1}} \left\{ \frac{1 + z^{-1}}{(1 - \rho z^{-1})(1 - \rho z)} \right\} \quad (11.92)$$

- Hence, find the cross-covariance sequence, $\gamma_{sy}[\ell]$, between the output, $s[n]$, and the input $y[n]$.

The following bilateral z -transform from the sample-domain, ℓ , to the z -domain might be useful:

$$\ell a^\ell u[\ell] \stackrel{z}{\rightleftharpoons} \frac{a z^{-1}}{(1 - a z^{-1})^2}, \quad |a| < 1 \quad (11.93)$$

where $u[\ell] = 1$ if $\ell \geq 0$ and zero otherwise.

SOLUTION. • The cross-complex spectral density at the output of the filter is given by:

$$P_{sy}(z) = G(z) P_{yy}(z) \quad (11.94)$$

where $G(z)$ is the transfer function of the system.

By taking z -transforms:

$$G(z) = \frac{1 + z^{-1}}{1 - \rho z^{-1}} \quad (11.95)$$

and therefore using the expression for $P_{yy}(z)$ from the previous example:

$$P_{sy}(z) = G(z) P_{yy}(z) = \frac{1 + z^{-1}}{1 - \rho z^{-1}} \frac{\sigma_x^2}{(1 - \rho z^{-1})(1 - \rho z)} \quad (11.96)$$

$$= \frac{\sigma_w^2}{1 - \rho z^{-1}} \left\{ \frac{1 + z^{-1}}{(1 - \rho z^{-1})(1 - \rho z)} \right\} \quad (11.97)$$

- The term in the curly brackets can be simplified as:

$$\frac{1 + z^{-1}}{(1 - \rho z^{-1})(1 - \rho z)} = \frac{z + 1}{(z - \rho)(1 - \rho z)} = \frac{A}{z - \rho} + \frac{B}{1 - \rho z} \quad (11.98)$$

Using the cover-up rule to find:

$$A: \times \text{ by } z - \rho \text{ \& set } z - \rho = 0; = \frac{z + 1}{(1 - \rho z)} = A + \underbrace{(z - \rho) \frac{B}{1 - \rho z}}_{=0}$$

$$B: \times \text{ by } 1 - \rho z \text{ \& set } 1 - \rho z = 0; = \frac{z + 1}{(z - \rho)} = \underbrace{(1 - \rho z) \frac{A}{z - \rho}}_{=0} + B$$

which may be rewritten as:

$$A = \frac{z + 1}{1 - \rho z} \Big|_{z=\rho} = \frac{1 + \rho}{1 - \rho^2} = \frac{1}{1 - \rho} \quad (11.99)$$

$$B = \frac{z + 1}{z - \rho} \Big|_{z=\frac{1}{\rho}} = \frac{1 + \rho}{1 - \rho^2} = \frac{1}{1 - \rho} = A \quad (11.100)$$

Hence, the cross-complex spectral density is given by:

$$P_{sy}(z) = \frac{\sigma_w^2}{1 - \rho z^{-1}} \frac{1}{1 - \rho} \left\{ \frac{1}{z - \rho} + \frac{1}{1 - \rho z} \right\} \quad (11.101)$$

$$= \frac{\sigma_w^2}{1 - \rho} \left\{ \frac{1}{1 - \rho z^{-1}} \right\} \left\{ \frac{z^{-1}}{1 - \rho z^{-1}} + \frac{1}{1 - \rho z} \right\} \quad (11.102)$$

$$= \frac{\sigma_w^2}{1 - \rho} \left\{ \frac{1}{\rho (1 - \rho z^{-1})^2} + \frac{1}{1 - \rho^2} \frac{1 - \rho^2}{(1 - \rho z)(1 - \rho z^{-1})} \right\} \quad (11.103)$$

Hence, taking inverse- z -transforms gives the cross-covariance:

$$\gamma_{sy}[\ell] = \frac{\sigma_w^2}{1 - \rho} \left\{ \frac{\ell}{\rho} \rho^\ell u[\ell] + \frac{1}{1 - \rho^2} \rho^{|\ell|} \right\} \quad (11.104)$$

□

To find the cross-correlation requires the addition of the mean components as before. To find the output auto-correlation requires substantially more work, and this is left as an exercise to the reader!

– End-of-Topic 73: **Frequency-domain analysis of input-output statistics** –

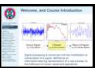


12

Linear Signal Models

This handout looks at the special class of stationary signals that are obtained by driving a linear time-invariant (LTI) system with white noise. A particular focus is placed on system functions that are rational; that is, they can be expressed at the ratio of two polynomials. Thus, the time-domain and frequency domain characteristics of pole-zero, all-pole, and all-zero models are investigated, including their time-series equivalents.

12.1 Abstract



- In the last lecture, the response of a linear-system when a stochastic process is applied at the input was considered. General linear systems were considered, and no focus on their interpretation or their practical applications was discussed. *New slide*
- This lecture looks at the special class of stationary signals that are obtained by driving a linear time-invariant (LTI) system with white noise. A particular focus is placed on **rational system functions**; that is, they can be expressed at the ratio of two polynomials. The power spectral density (PSD) of the resulting process is also rational, and its shape is completely determined by the filter coefficients. As a result, linear signal models provide a method for modelling the PSD of a process, and thus leads to **parametric PSD estimation**, also known as **modern spectral estimation**.
- The following models are considered in detail:
 - **All-pole** systems and autoregressive (AR) processes;
 - **All-zero** systems and moving average (MA) processes;
 - and **pole-zero** systems and autoregressive moving average (ARMA) processes.
- **Pole-zero** models are widely used for modelling stationary signals with short memory; the concepts will be extended, in overview at least, to nonstationary processes.

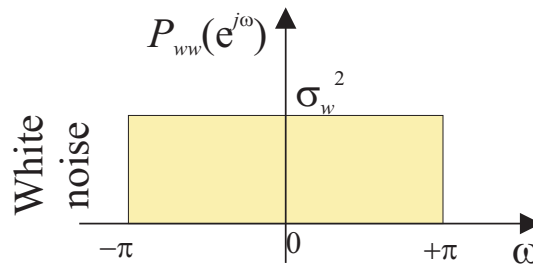


Figure 12.1: White noise PSD.

Linear signal models are developed first by assuming that the second order moments of the random process are known, and equations are developed whose solution provides the model parameters. In most practical applications of the theory, however, the fixed quantities in the equations, namely the correlation functions and the model orders, are not known *a priori* but need to be estimated from the data. This, as a result, introduces the issue of estimation of the model parameters and leads to the notion of, for example, maximum likelihood estimation and least squares estimates as discussed in the next handout.

12.2 The Ubiquitous WGN Sequence

The simplest random signal model is the wide-sense stationary (WSS) white Gaussian noise (WGN) sequence:

$$w[n] \sim \mathcal{N}(0, \sigma_w^2) \quad (12.1)$$

The sequence is independent and identically distributed (i. i. d.), and has a flat PSD: $P_{ww}(e^{j\omega T}) = \sigma_w^2$, $-\pi < \omega \leq \pi$. The PSD is shown below in Figure 12.1. It is also easy (as shown below) to generate samples using simple algorithms.

12.2.1 Generating 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 12.1 (Probability transformation rule (revised)). If $\{x_1, \dots, x_n\}$ are random variables with a joint-probability density function (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:

$$f_Y(y_1, \dots, y_n) = \frac{1}{|J(x_1, \dots, x_n)|} f_X(x_1, \dots, x_n) \quad (12.2)$$

where $J(x_1, \dots, x_n)$ is the **Jacobian** of the transformation given by:

$$J(x_1, \dots, x_n) = \frac{\partial(y_1, \dots, y_n)}{\partial(x_1, \dots, x_n)} \quad (12.3)$$

◇

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,

$$f_{X_k}(x_k) = \mathbb{I}_{0,1}(x_k), \quad k = 1, 2 \quad (12.4)$$

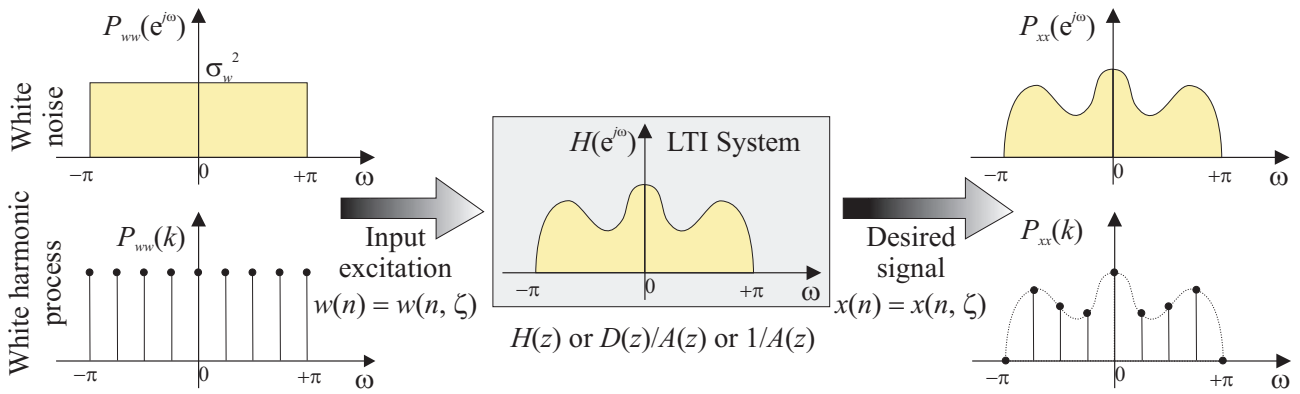


Figure 12.2: Signal models with continuous and discrete (line) power spectrum densities.

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:

$$y_1 = \sqrt{-2 \ln x_1} \cos 2\pi x_2 \tag{12.5}$$

$$y_2 = \sqrt{-2 \ln x_1} \sin 2\pi x_2 \tag{12.6}$$

It follows, by rearranging these equations, that:

$$x_1 = \exp \left[-\frac{1}{2}(y_1^2 + y_2^2) \right] \tag{12.7}$$

$$x_2 = \frac{1}{2\pi} \arctan \frac{y_2}{y_1} \tag{12.8}$$

The Jacobian determinant can be calculated as:

$$J(x_1, x_2) = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{vmatrix} = \begin{vmatrix} \frac{-1}{x_1 \sqrt{-2 \ln x_1}} \cos 2\pi x_2 & -2\pi \sqrt{-2 \ln x_1} \sin 2\pi x_2 \\ \frac{-1}{x_1 \sqrt{-2 \ln x_1}} \sin 2\pi x_2 & 2\pi \sqrt{-2 \ln x_1} \cos 2\pi x_2 \end{vmatrix} = \frac{2\pi}{x_1} \tag{12.9}$$

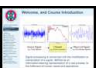
Hence, it follows:

$$f_Y(y_1, y_2) = \frac{x_1}{2\pi} = \left[\frac{1}{\sqrt{2\pi}} e^{-y_1^2/2} \right] \left[\frac{1}{\sqrt{2\pi}} e^{-y_2^2/2} \right] \tag{12.10}$$

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.

12.2.2 Filtration of WGN



New slide

By filtering a WGN through a stable LTI system, it is possible to obtain a stochastic signal at the output with almost any arbitrary aperiodic correlation function or continuous PSD. The PSD of the output is given by:

$$P_{xx}(e^{j\omega}) = \sigma_w^2 |H(e^{j\omega})|^2 = G^2 \frac{\prod_{k=1}^Q |1 - z_k e^{-j\omega}|^2}{\prod_{k=1}^P |1 - p_k e^{-j\omega}|^2} \tag{12.11}$$

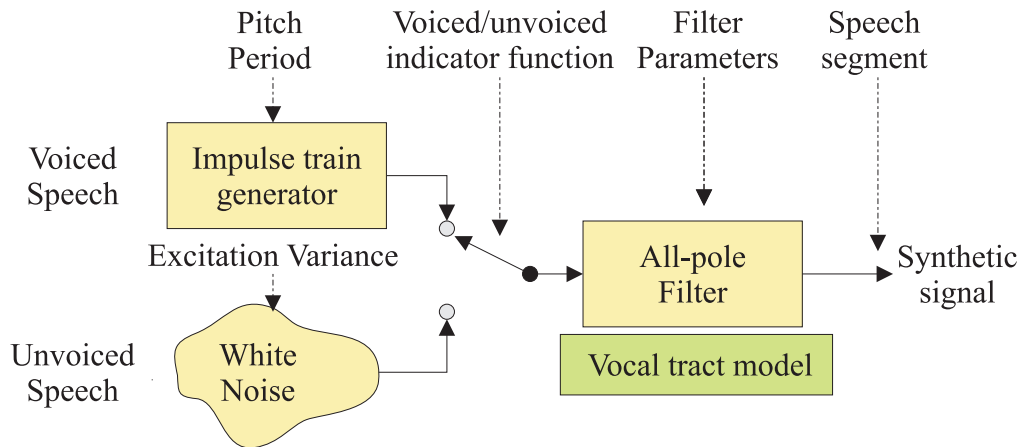


Figure 12.3: The speech synthesis model.

Note that the shape of the power spectrum depends only upon the magnitude of the filter's frequency response.

Random signals with line PSDs can be generated by using the **harmonic process** model, which is a linear combination of sinusoidal sequences with statistically independent random phases. Signal models with mixed PSDs can be obtained by combining these two models; a process justified by the **Wold decomposition**. This is highlighted in Figure 12.2; contrast this with the speech synthesis model shown in Figure 12.3, which was also shown in the introductory handout.

12.3 Nonparametric and parametric signal models

Nonparametric models have no restriction on its form, or the number of parameters characterising the model. For example, specifying a LTI filter by its impulse response is a nonparametric model.

If the input $w(n)$ is a zero-mean white noise process with variance σ_w^2 , autocorrelation $r_{ww}(l) = \sigma_w^2 \delta(l)$ and $P_{ww}(e^{j\omega}) = \sigma_w^2$, $-\pi < \omega \leq \pi$, then the autocorrelation, complex spectral density, and PSD of the output $x(n)$ are given by, respectively:

$$r_{xx}(l) = \sigma_w^2 \sum_{k=-\infty}^{\infty} h(k) h^*(k-l) = \sigma_w^2 r_{hh}(l) \quad (\text{M:4.1.2})$$

$$P_{xx}(z) = \sigma_w^2 H(z) H^* \left(\frac{1}{z^*} \right) \quad (\text{M:4.1.3})$$

$$P_{xx}(e^{j\omega}) = \sigma_w^2 |H(e^{j\omega})|^2 \quad (\text{M:4.1.4})$$

Notice that the shape of the autocorrelation and the power spectrum of the output signal are completely characterised by the system. This is known as a **system based signal model**, and in the case of linear systems, is also known as the **linear random signal model**, or the **general linear process model**.

Parametric models, on the other hand, describe a system with a finite number of parameters. For example, if a LTI filter is specified by a finite-order rational **system function**, it is a parametric model.

Two important analysis tools present themselves for parametric modelling:



New slide

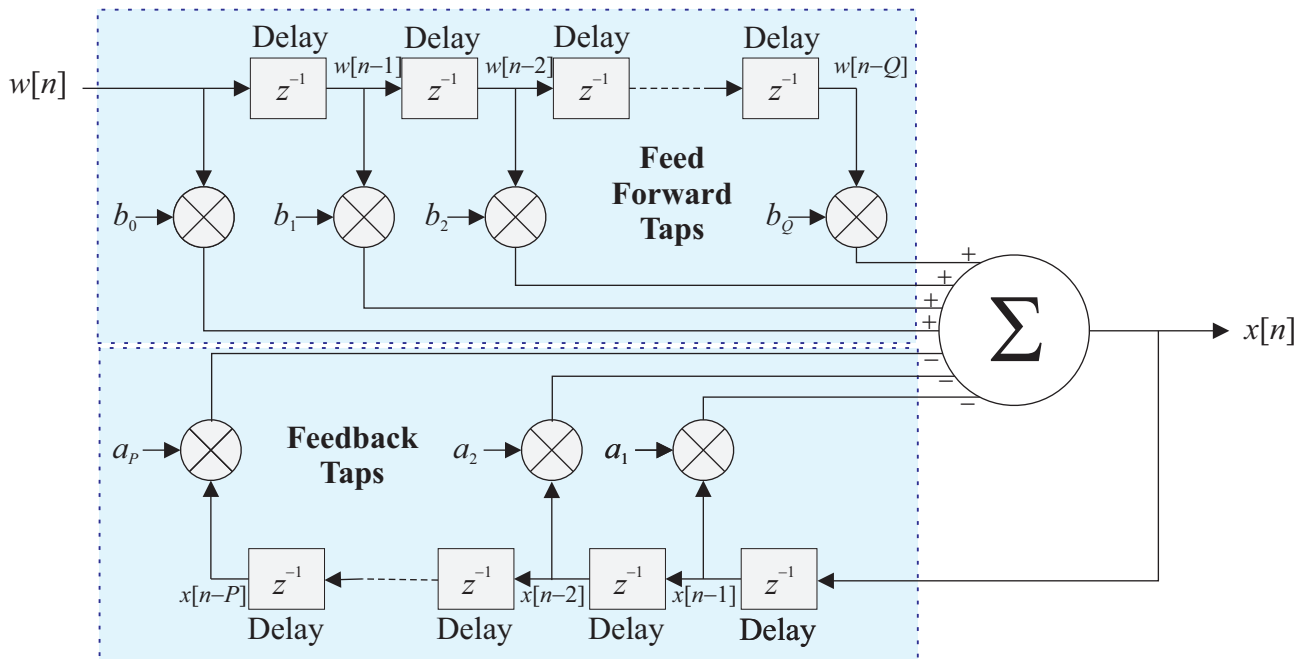
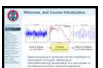


Figure 12.4: Filter block diagram for ARMA model.

1. given the parameters of the model, analyse the characteristics of that model (in terms of moments etc.);
2. design of a parametric system model to produce a random signal with a specified autocorrelation function or PSD. This problem is known as **signal modelling**.

12.4 Parametric Pole-Zero Signal Models



Parametric models describe a system with a finite number of parameters. Consider a system described *New slide* by the following linear constant-coefficient difference equation:

$$x[n] = - \sum_{k=1}^P a_k x[n - k] + \sum_{k=0}^Q d_k w[n - k] \tag{M:4.1.21}$$

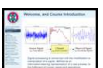
This rational transfer function was introduced in the first lecture, and the filter block diagram is shown in Figure 12.4. Taking z -transforms gives the system function:

$$H(z) = \frac{X(z)}{W(z)} = \frac{\sum_{k=0}^Q d_k z^{-k}}{1 + \sum_{k=1}^P a_k z^{-k}} \tag{M:4.1.22}$$

$$\triangleq \frac{D(z)}{A(z)} = G \frac{\prod_{k=1}^Q (1 - z_k z^{-1})}{\prod_{k=1}^P (1 - p_k z^{-1})} \tag{M:4.1.23}$$

This system has Q zeros, $\{z_k, k \in \mathcal{Q}\}$ where $\mathcal{Q} = \{1, \dots, Q\}$, and P poles, $\{p_k, k \in \mathcal{P}\}$. Note that poles and zeros at $z = 0$ are not considered here. The term G is the system gain. It is assumed that the polynomials $A(z)$ and $D(z)$ do not have any common roots.

12.4.1 Types of pole-zero models



There are three cases of interest as shown in Figure 12.5:

New slide

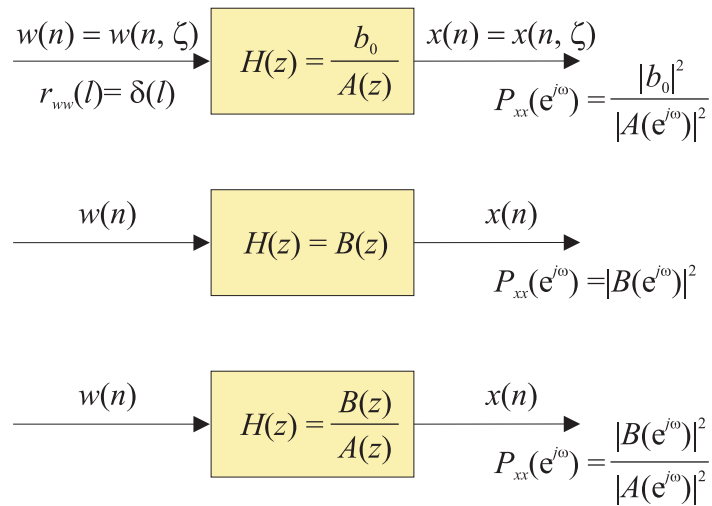


Figure 12.5: Types of linear model; top to bottom, these are the AR, MA and ARMA models.

All-pole model when $Q = 0$. The input-output difference equation is given by:

$$x[n] = - \sum_{k=1}^P a_k x[n-k] + d_0 w[n] \quad (\text{M:4.1.26})$$

This is commonly denoted as the $AP(P)$ model.

All-zero model when $P = 0$. The input-output relation is given by:

$$x[n] = \sum_{k=0}^Q d_k w[n-k] \quad (\text{M:4.1.25})$$

This is commonly denoted as the $AZ(Q)$ model.

Pole-zero model when $P > 0$ and $Q > 0$.

This is commonly denoted as the $PZ(P, Q)$ model, and if it is assumed to be causal, is given by Equation M:4.1.21.

If a parametric model is *excited* with WGN, the resulting output signal has second-order moments determined by the parameters of the model. These **stochastic processes** have special names in the literature, and are known as:

a moving average (MA) process when it is the output of an all-zero model;

an autoregressive (AR) process when it is the output of an all-pole model;

an autoregressive moving average (ARMA) process when it is the output of an pole-zero model;

each subject to a WGN process at the input.

The parametric signal model is usually specified by normalising $d_0 = 1$ and setting the variance of the input to σ_w^2 . The alternative is to specify $\sigma_w^2 = 1$ and leave d_0 arbitrary, but this isn't quite as elegant when it comes to deriving pdfs. It is also important to stress that these models assume the resulting processes are stationary, which is ensured if the corresponding systems are bounded-input, bounded-output (BIBO) stable.

12.4.2 All-pole Models

Assume an all-pole model of the form:

$$H(z) = \frac{d_0}{A(z)} = \frac{d_0}{1 + \sum_{k=1}^P a_k z^{-k}} = \frac{d_0}{\prod_{k=1}^P (1 - p_k z^{-1})} \quad (\text{M:4.2.1})$$

where d_0 is the system gain, and P is the order of the model.

All-pole models are frequently used in signal processing applications since they are:

- mathematically convenient since model parameters can be estimated by solving a set of linear equations, and
- they widely parsimoniously approximate rational transfer functions, especially resonant systems.

There are various model properties of the all-pole model that are useful; these include:

1. the systems impulse response;
2. the somewhat inappropriate term called the autocorrelation of the impulse response;
3. and minimum-phase conditions.

Although the autocorrelation of the impulse response is useful to gain additional insight into aspects of the all-pole filter, it is better to consider the autocorrelation function of an AR process (i.e. the autocorrelation function of the output of an all-pole filter). However, for completeness, the details of the autocorrelation of the impulse response is included in these notes.

12.4.2.1 Frequency Response of an All-Pole Filter

The all-pole model has form:

$$H(z) = \frac{d_0}{A(z)} = \frac{d_0}{1 + \sum_{k=1}^P a_k z^{-k}} = \frac{d_0}{\prod_{k=1}^P (1 - p_k z^{-1})} \quad (\text{M:4.2.1})$$

Therefore, its frequency response is given by:

$$H(e^{j\omega}) = \frac{d_0}{1 + \sum_{k=1}^P a_k e^{-jk\omega}} = \frac{d_0}{\prod_{k=1}^P (1 - p_k e^{-j\omega})} \quad (12.12)$$

When the poles are written in the form $p_k = r_k e^{j\omega_k}$, the frequency response can be written as:

$$H(e^{j\omega}) = \frac{d_0}{\prod_{k=1}^P (1 - r_k e^{-j(\omega - \omega_k)})} \quad (12.13)$$

Hence, it can be deduced that resonances occur near the frequencies corresponding to the phase position of the poles. When the system is real, the complex-poles occur in conjugate-pairs.

Hence, the PSD of the output of an all-pole filter is given by:

$$P_{xx}(e^{j\omega}) = \sigma_w^2 |H(e^{j\omega})|^2 = \frac{G^2}{\prod_{k=1}^P |1 - r_k e^{-j(\omega - \omega_k)}|^2} \quad (12.14)$$

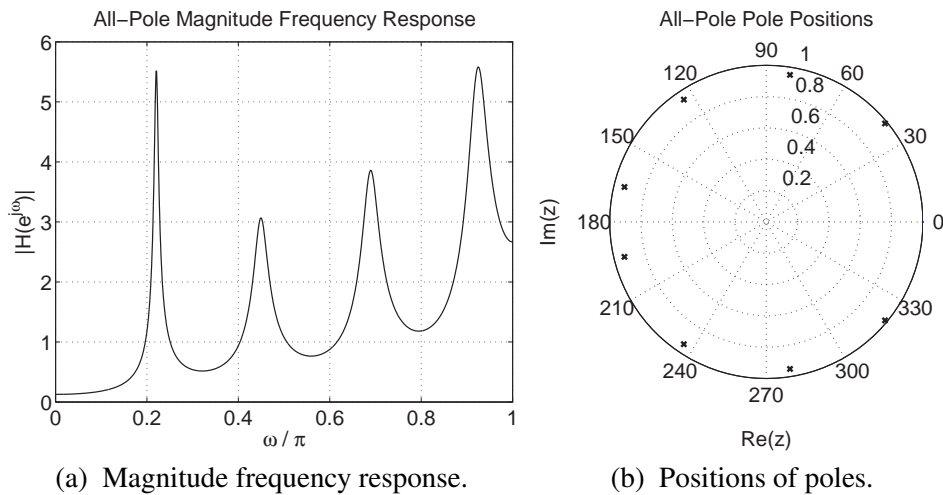


Figure 12.6: The frequency response and position of the poles in an all-pole system.

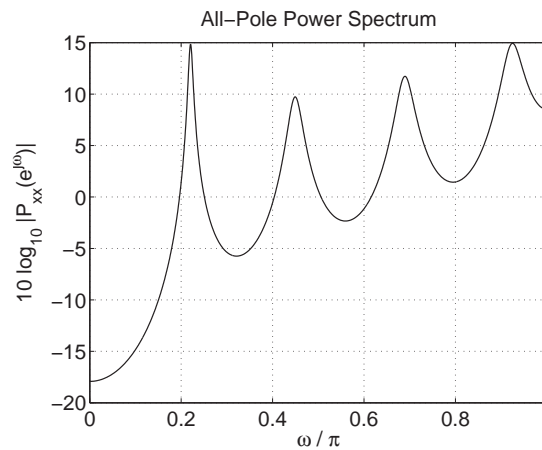


Figure 12.7: Power spectral response of an all-pole model.

where $G = \sigma_w d_0$ is the overall gain of the system.

Consider the all-pole model with poles at positions:

$$\{p_k\} = \{r_k e^{j\omega_k}\} \quad \text{where} \quad \begin{cases} \{r_k\} &= \{0.985, 0.951, 0.942, 0.933\} \\ \{\omega_k\} &= 2\pi \times \{270, 150, 120, 30\}/2450; \end{cases} \quad (12.15)$$

The pole positions and magnitude frequency response of this system is plotted in Figure 12.6. For comparison, the PSD of the output of the system is shown in Figure 12.7.

12.4.2.2 Impulse Response of an All-Pole Filter

Recalling that the input-output difference equation for an all-pole filter is given by:

$$x[n] = - \sum_{k=1}^P a_k x[n-k] + d_0 w[n] \quad (\text{M:4.1.26})$$

then the impulse response, $h[n]$, is the output when the input is a delta function, $w[n] = \delta[n]$.



New slide

The impulse response of the all-pole filter satisfies the equation:

$$h[n] = - \sum_{k=1}^P a_k h[n-k] + d_0 \delta[n] \quad (\text{M:4.2.3})$$

The derivation in [Manolakis:2000, page 157] is somewhat verbose; nevertheless, their approach is to re-write the system function of the all-pole filter as:

$$H(z) + \sum_{k=1}^P a_k H(z) z^{-k} = d_0 \quad (12.16)$$

and thus by taking the inverse z -transform gives the same result as above. If $H(z)$ has its poles inside the unit circle, then $h[n]$ is a causal, stable sequence, and the system is **minimum-phase**.

Assuming causality, such that $h[n] = 0$, $n < 0$ then it follows $h[-k] = 0$, $k > 0$, and therefore:

$$h[n] = \begin{cases} 0 & \text{if } n < 0 \\ d_0 & \text{if } n = 0 \\ - \sum_{k=1}^P a_k h[n-k] & \text{if } n > 0 \end{cases} \quad (\text{M:4.2.5})$$

Thus, except for the value at $n = 0$, $h[n]$ can be obtained recursively as a linearly weighted summation of its previous values, $\{h[n-p], p = \{1, \dots, P\}\}$. Thus, in this sense, $h[n]$ can be *predicted*, for $n \neq 0$, with zero error from the past P past values. Thus, the coefficients $\{a_k\}$ are often referred to as **predictor coefficients**.

Finally, note that a causal $H(z)$ can be written as a one-sided z -transform, or infinite polynomial, $H(z) = \sum_{n=0}^{\infty} h[n] z^{-n}$. This representation implies that any finite-order, all-pole model can be represented equivalently by an infinite number of zeros, and conversely a single zero can be represented by an infinite number of poles. If the poles are inside the unit circle, then so are the corresponding zeros, and vice-versa.

12.4.2.3 Autocorrelation of the Impulse Response

The autocorrelation of the system impulse response is given by:

$$r_{hh}(l) \triangleq h(l) * h^*(-l) = \sum_{n=-\infty}^{\infty} h(n) h^*(n-l) \quad (12.17)$$

Multiplying both side of Equation M:4.2.3 by $h^*[n-l]$ gives and summing over all n :

$$\sum_{n=-\infty}^{\infty} \sum_{k=0}^P a_k h(n-k) h^*(n-l) = d_0 \sum_{n=-\infty}^{\infty} h^*(n-l) \delta(n) \quad (\text{M:4.2.14})$$

where $a_0 = 1$. Interchanging the order of summations (as usual) in the left hand side (LHS), and setting $\hat{n} = n - k$ gives:

$$\sum_{k=0}^P a_k \sum_{\hat{n}=-\infty}^{\infty} h(\hat{n}) h^*(\hat{n} - (l-k)) = d_0 h^*(-l) \quad (12.18)$$

which can also be written as

$$\sum_{k=0}^P a_k r_{hh}(l-k) = d_0 h^*(-l) \quad (\text{M:4.2.15})$$

Since $h(n) = 0$, $n < 0$, then $h(-l) = 0$, $l > 0$, and $h(0) = d_0$, then:

$$r_{hh}(l) = \begin{cases} d_0 h^*(-l) - \sum_{k=1}^P a_k r_{hh}(l-k) & l < 0 \\ |d_0|^2 - \sum_{k=1}^P a_k r_{hh}(-k) & l = 0 \\ -\sum_{k=1}^P a_k r_{hh}(l-k) & l > 0 \end{cases} \quad (12.19)$$

These are recursive relationships for $r_{hh}[\ell]$ in terms of past values of the autocorrelation function.

It is also possible to write the autocorrelation in terms of the poles of the model, and to also investigate the response of the model to an impulse train (harmonic) excitation. These are not considered in this handout, but are detailed in [Manolakis:2000, Section 4.2].

12.4.2.4 All-Pole Modelling and Linear Prediction

A **linear predictor** forms an estimate, or *prediction*, $\hat{x}[n]$, of the present value of a stochastic process $x[n]$ from a linear combination of the past P samples; that is:

$$\hat{x}[n] = -\sum_{k=1}^P a_k x[n-k] \quad (\text{M:1.4.1})$$

The coefficients $\{a_k\}$ of the linear predictor are determined by attempting to minimise some function of the **prediction error** given by:

$$e(n) = x(n) - \hat{x}(n) \quad (\text{M:1.4.2})$$

Usually the objective function is equivalent to mean-squared error (MSE), given by $E = \sum_n e^2(n)$.

Hence, the prediction error can be written as:

$$e(n) = x(n) - \hat{x}(n) = x(n) + \sum_{k=1}^P a_k x(n-k) \quad (\text{M:4.2.50})$$

- Thus, the prediction error is equal to the excitation of the all-pole model; $e(n) = w(n)$. Clearly, **finite impulse response (FIR) linear prediction** and **all-pole modelling** are closely related.
- Many of the properties and algorithms developed for either **linear prediction** or **all-pole modelling** can be applied to the other.
- To all intents and purposes, linear prediction, **all-pole modelling**, and **AR processes** (discussed next) are equivalent terms for the same concept.

12.4.2.5 Autoregressive Processes

While **all-pole models** refer to the properties of a rational system containing only poles, **AR processes** refer to the resulting stochastic process that occurs as the result of **WGN** being applied to the input of an **all-pole filter**.

As such, the same input-output equations for all-pole models still apply although, in this case, the AR process refers to $x[n]$, whereas **all-pole modelling** would refer to the system itself, as defined by the linear difference equation and the parameters $\{a_k\}$.

Thus:

$$x[n] = - \sum_{k=1}^P a_k x[n-k] + w[n], \quad w[n] \sim \mathcal{N}(0, \sigma_w^2) \quad (\text{M:4.2.52})$$

The **AR process** is valid only if the corresponding **all-pole system** is stable. The autoregressive output, $x[n]$, is a stationary sequence with a mean value of zero, $\mu_x = 0$.

The autocorrelation sequence (ACS) can be calculated in a similar approach to finding the output autocorrelation and cross-correlation for linear systems.

Multiply the difference Equation M:4.2.52 through by $x^*(n-l)$ and take expectations to obtain:

$$r_{xx}(l) + \sum_{k=1}^P a_k r_{xx}(l-k) = r_{wx}(l) \quad (\text{M:4.2.54})$$

Observing that $x[n]$ cannot depend on future values of $w[n]$ since the system is causal, then $r_{wx}[\ell] = \mathbb{E}[w[n] x^*[n-\ell]]$ is zero if $\ell > 0$, and σ_w^2 if $\ell = 0$.

Thus, for $l = \{0, 1, \dots, P\}$ gives:

$$r_{xx}(0) + a_1 r_{xx}(-1) + \dots + a_P r_{xx}(-P) = \sigma_w^2 \quad (12.20)$$

$$r_{xx}(1) + a_1 r_{xx}(0) + \dots + a_P r_{xx}(-P+1) = 0 \quad (12.21)$$

$$\vdots \quad (12.22)$$

$$r_{xx}(P) + a_1 r_{xx}(P-1) + \dots + a_P r_{xx}(0) = 0 \quad (12.23)$$

This can be written in matrix-vector form (noting that $r_{xx}[\ell] = r_{xx}^*[-\ell]$ and that the parameters $\{a_k\}$ are real) as:

$$\begin{bmatrix} r_{xx}[0] & r_{xx}^*[1] & \dots & r_{xx}^*[P] \\ r_{xx}[1] & r_{xx}[0] & \dots & r_{xx}^*[P-1] \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx}[P] & r_{xx}[P-1] & \dots & r_{xx}^*[0] \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} \sigma_w^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (\text{M:4.2.56})$$

These **Yule-Walker equations** have an identical form to the **normal equations** which are a result of analysing linear prediction. The differences are minor, but the interested reader can find out more in [Therrien:1992, Chapter 8]. It is important to note that the Yule-Walker equations are linear in the parameters a_k , and there are several different efficient methods for solving them. Details, again, can be found in [Therrien:1992, Chapters 8 and 9].

12.4.2.6 Autocorrelation Function from AR parameters

In the previous section, an expression for calculating the AR coefficients given the autocorrelation values was given. But what if the AR coefficients are known, and it is desirable to calculate the autocorrelation function given these parameters. A formulation is given here. Assume that an AR process is real, such that the **Yule-Walker equations** become:

$$\begin{bmatrix} r_{xx}(0) & r_{xx}(1) & \dots & r_{xx}(P) \\ r_{xx}(1) & r_{xx}(0) & \dots & r_{xx}(P-1) \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx}(P) & r_{xx}(P-1) & \dots & r_{xx}(0) \end{bmatrix} \hat{\mathbf{a}} = \mathbf{b} \quad \text{where} \quad \hat{\mathbf{a}} = \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_P \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} \sigma_w^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (12.24)$$

To generate the autocorrelation values from the AR parameters, it is desirable to obtain an equation of the form $\mathbf{A} \mathbf{r} = \mathbf{b}$, where $[r_{xx}(0) \ \cdots \ r_{xx}(P)]^T$, and the matrix \mathbf{A} and vector \mathbf{b} are functions of the parameters $\{a_k\}$ and the input variance σ_w^2 . Write the Yule-Walker equations as:

$$r_{xx}(0) \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix} \hat{\mathbf{a}} + r_{xx}(1) \begin{bmatrix} 0 & 1 & \cdots & 0 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 1 & 0 \end{bmatrix} \hat{\mathbf{a}} + \cdots + r_{xx}(P) \begin{bmatrix} 0 & 0 & \cdots & 1 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 1 & \cdots & 0 & 0 \end{bmatrix} \hat{\mathbf{a}} = \mathbf{b} \quad (12.25)$$

By defining the $P \times P$ matrix $\mathbf{I}_{P,k}$ with ones on the k th diagonal away from the leading diagonal, and zero elsewhere, then it follows:

$$\sum_{k=0}^P (\mathbf{I}_{P+1,k} \hat{\mathbf{a}}) r_{xx}(k) = \mathbf{b} \quad (12.26)$$

Next defining the vector $\hat{\mathbf{a}}_k = \mathbf{I}_{P+1,k} \hat{\mathbf{a}}$ and the matrix $[\hat{\mathbf{a}}_0 \ \cdots \ \hat{\mathbf{a}}_P]$, then the matrix-vector equation

$$\mathbf{A} \mathbf{r} = \mathbf{b} \quad (12.27)$$

has been obtained. In low-order cases, it might be more straightforward to explicitly compute the autocorrelation functions by writing out the **Yule-Walker equations**.

All-pole models therefore have the unique property that the model parameters are completely specified by the first $P + 1$ autocorrelation coefficients via a set of linear equations, as given by the equation $\mathbf{A} \mathbf{r} = \mathbf{b}$. An alternative way of writing this is:

$$\begin{bmatrix} \sigma_w^2 \\ a_1 \\ \vdots \\ a_P \end{bmatrix} \leftrightarrow \begin{bmatrix} r_{xx}(0) \\ \vdots \\ r_{xx}(P) \end{bmatrix} \quad (12.28)$$

Thus, the mapping of the model parameters to the autocorrelation coefficients is reversible and unique. This *correlation matching* of **all-pole models** is quite remarkable, and is not shared by **all-zero models**, and is true for **pole-zero models** only under certain conditions.

Example 12.1 (Calculating Autocorrelation Functions of All-Pole Model). Given the parameters σ_w^2 , a_1 , and a_2 , of a second-order all-pole model, compute the autocorrelation values $r_{xx}(k)$ for $\{k = 0, 1, 2\}$.

SOLUTION. Using the results above, it follows that:

$$r_{xx}(0) \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ a_2 \end{bmatrix} + r_{xx}(1) \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ a_2 \end{bmatrix} + r_{xx}(2) \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sigma_w^2 \\ 0 \\ 0 \end{bmatrix} \quad (12.29)$$

or,

$$\begin{bmatrix} 1 & a_1 & a_2 \\ a_1 & 1 + a_2 & 0 \\ a_2 & a_1 & 1 \end{bmatrix} \begin{bmatrix} r_{xx}(0) \\ r_{xx}(1) \\ r_{xx}(2) \end{bmatrix} = \begin{bmatrix} \sigma_w^2 \\ 0 \\ 0 \end{bmatrix} \quad (12.30)$$

Although you could try a direct version to solve this, a slightly more ad-hoc approach quickly yields a solution in this case, and is related to Gaussian elimination. Multiplying the second row by a_1 and the last row by a_2 , and then subtracting them both from the first row gives:

$$\begin{bmatrix} 1 - a_1^2 - a_2^2 & -2a_1a_2 & 0 \\ a_1^2 & a_1(1 + a_2) & 0 \\ a_2^2 & a_1a_2 & a_2 \end{bmatrix} \begin{bmatrix} r_{xx}(0) \\ r_{xx}(1) \\ r_{xx}(2) \end{bmatrix} = \begin{bmatrix} \sigma_w^2 \\ 0 \\ 0 \end{bmatrix} \quad (12.31)$$

It can thus be seen that the first two equations for $r_{xx}(0)$ and $r_{xx}(1)$ do not depend on $r_{xx}(2)$ and therefore, by inverting the 2 by 2 matrix, this gives:

$$\begin{bmatrix} r_{xx}(0) \\ r_{xx}(1) \end{bmatrix} = \frac{1}{a_1(1+a_2)(1-a_1^2-a_2^2)+2a_1^3a_2} \begin{bmatrix} a_1(1+a_2) & 2a_1a_2 \\ -a_1^2 & 1-a_1^2-a_2^2 \end{bmatrix} \begin{bmatrix} \sigma_w^2 \\ 0 \end{bmatrix} \quad (12.32)$$

$$= \frac{\sigma_w^2}{(1-a_1^2-a_2^2)+\frac{2a_1^2a_2}{1+a_2}} \begin{bmatrix} 1 \\ -\frac{a_1}{1+a_2} \end{bmatrix} \quad (12.33)$$

Moreover,

$$r_{xx}(2) = -\frac{1}{a_2} \begin{bmatrix} a_2^2 & a_1a_2 \end{bmatrix} \begin{bmatrix} r_{xx}(0) \\ r_{xx}(1) \end{bmatrix} = \frac{\sigma_w^2}{(1-a_1^2-a_2^2)+\frac{2a_1^2a_2}{1+a_2}} \left(\frac{a_1^2}{1+a_2} - a_2 \right) \quad (12.34)$$

In summary,

$$\begin{bmatrix} r_{xx}(0) \\ r_{xx}(1) \\ r_{xx}(2) \end{bmatrix} = \frac{\sigma_w^2}{(1-a_1^2-a_2^2)+\frac{2a_1^2a_2}{1+a_2}} \begin{bmatrix} 1 \\ -\frac{a_1}{1+a_2} \\ \frac{a_1^2}{1+a_2} - a_2 \end{bmatrix} \quad (12.35) \quad \square$$

12.4.3 All-Zero models

Whereas **all-pole** models can capture resonant features of a particular PSD, it cannot capture *nulls* in the frequency response. These can only be modelled using a pole-zero or **all-zero** model. New slide

The output of an all-zero model is the weighted average of delayed versions of the input signal. Thus, assume an all-zero model of the form:

$$x[n] = \sum_{k=0}^Q d_k w[n-k] \quad (\text{M:4.3.1})$$

where Q is the order of the model, and the corresponding system function is given by:

$$H(z) = D(z) = \sum_{k=0}^Q d_k z^{-k} \quad (\text{M:4.3.2})$$

Similar to the relationship between **all-pole models** and **AR processes**, **all-zero models** refer to the properties of a rational system containing only zeros, while **MA processes** refer to the resulting stochastic process that occurs as the result of **WGN** being applied to the input of an **all-zero filter**.

All-zero models are difficult to deal with since, unlike the **Yule-Walker equations** for the **all-pole model**, the solution for model parameters given the autocorrelation functions involves solving nonlinear equations, which becomes quite a complicated task.

12.4.3.1 Frequency Response of an All-Zero Filter

The all-zero model has form:

$$H(z) = D(z) = \sum_{k=0}^Q d_k z^{-k} = d_0 \prod_{k=1}^Q (1 - z_k z^{-1}) \quad (12.36)$$

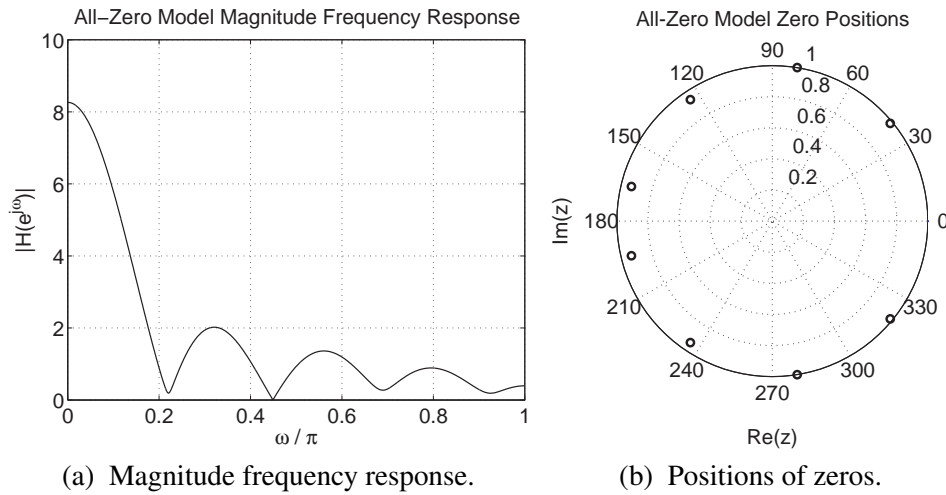


Figure 12.8: The frequency response and position of the zeros in an all-zero system.

where $\{z_k\}$ are the zeros of the all-zero model. Therefore, its frequency response is given by:

$$H(e^{j\omega}) = \sum_{k=0}^Q d_k e^{-jk\omega} = d_0 \prod_{k=1}^Q (1 - z_k e^{-j\omega}) \quad (12.37)$$

When the zeros are written in the form $z_k = r_k e^{j\omega_k}$, then the frequency response can be written as:

$$H(e^{j\omega}) = d_0 \prod_{k=1}^Q (1 - r_k e^{-j(\omega - \omega_k)}) \quad (12.38)$$

Hence, it can be deduced that troughs or nulls occur near frequencies corresponding to the phase position of the zeros. When the system is real, the complex-zeros occur in conjugate-pairs.

Hence, the PSD of the output of an all-zero filter is given by:

$$P_{xx}(e^{j\omega}) = \sigma_w^2 |H(e^{j\omega})|^2 = G^2 \prod_{k=1}^Q |1 - r_k e^{-j(\omega - \omega_k)}|^2 \quad (12.39)$$

where $G = \sigma_w d_0$ is the overall gain of the system. Consider the all-zero model with zeros at positions:

$$\{z_k\} = \{r_k e^{j\omega_k}\} \quad \text{where} \quad \begin{cases} \{r_k\} = \{0.985, 1, 0.942, 0.933\} \\ \{\omega_k\} = 2\pi \times \{270, 550, 844, 1131\}/2450; \end{cases} \quad (12.40)$$

The zero positions and magnitude frequency response of this system is plotted in Figure 12.8. For comparison, the power spectral density of the output of the system is shown in Figure 12.9. Note that one of the zeros is on the unit circle, and that the frequency response at this point is zero.

12.4.3.2 Impulse Response

The impulse response of an **all-zero** model is an **FIR system** with impulse response:

$$h(n) = \begin{cases} d_n & 0 \leq n \leq Q \\ 0 & \text{elsewhere} \end{cases} \quad (12.41)$$

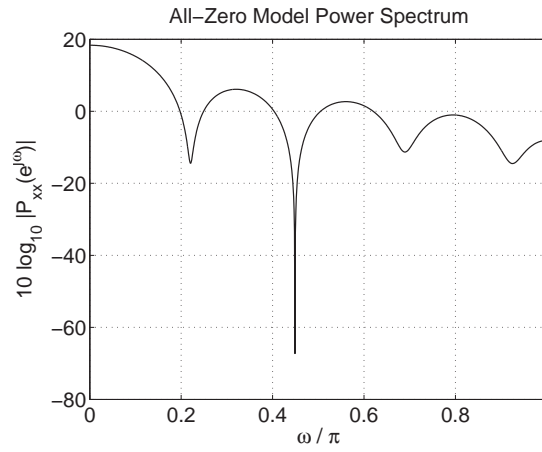


Figure 12.9: Power spectral response of an all-zero model.

12.4.3.3 Autocorrelation of the Impulse Response

Following a similar line to that shown for all-pole models, the autocorrelation of the impulse response of an all-zero system can be found.

Theorem 12.2. The autocorrelation sequence of the impulse response of an all-zero system is:

$$r_{hh}[\ell] = \sum_{n=-\infty}^{\infty} h[n] h^*[n - \ell] = \begin{cases} \sum_{k=0}^{Q-\ell} d_{k+\ell} d_k^* & 0 \leq \ell \leq Q \\ 0 & \ell > Q \end{cases} \quad (\text{M:4.3.4})$$

and $r_{hh}^*[-\ell] = r_{hh}[\ell]$ for all ℓ .

PROOF. The autocorrelation sequence of the impulse response of an all-zero system is given by the discrete-time convolution:

$$r_{hh}[\ell] = \sum_{n=-\infty}^{\infty} h[n] h^*[n - \ell] \quad (12.42)$$

Considering the term $h[n]$,

$$h[n] = \begin{cases} d_n & 0 \leq n \leq Q \\ 0 & \text{otherwise} \end{cases} \quad (12.43)$$

or, in otherwords, $h(n) = 0$ when $n < 0$ and $n > Q$. Hence Equation 12.42 becomes:

$$r_{hh}[\ell] = \sum_{n=0}^Q d_n h^*[n - \ell] \quad (12.44)$$

Moreover, the lower-limit is constrained since

$$h^*[n - \ell] = \begin{cases} d_{n-\ell}^* & 0 \leq n - \ell \leq Q \\ 0 & \text{otherwise} \end{cases} \quad (12.45)$$

or, in otherwords, $h^*[n - \ell] = 0$ if $n < \ell$ and when $n > Q + \ell$. Assuming that $\ell \geq 0$, the second condition is already met by the upper-limit in Equation 12.44. Therefore, Equation 12.44 becomes:

$$r_{hh}[\ell] = \sum_{n=\ell}^Q d_n d_{n-\ell}^* \quad (12.46)$$

By substituting $k = n - \ell$, such that when $n = \{\ell, Q\}$, $k = \{0, Q - \ell\}$, then:

$$r_{hh}[\ell] = \sum_{k=0}^{Q-\ell} d_{k+\ell} d_k^*, \quad \text{for } \ell \geq 0 \quad (12.47)$$

Clearly this expression is equal to zero if $\ell > Q$. Therefore, using the result from the previous handout that $r_{hh}[\ell] = r_{hh}^*[-\ell]$, it follows:

$$r_{hh}[\ell] = \sum_{n=-\infty}^{\infty} h[n] h^*[n - \ell] = \begin{cases} \sum_{k=0}^{Q-\ell} d_{k+\ell} d_k^* & 0 \leq \ell \leq Q \\ 0 & \ell > Q \end{cases} \quad (\text{M:4.3.4}) \quad \square$$

and $r_{hh}^*[-\ell] = r_{hh}[\ell]$ for all ℓ .

12.4.3.4 Moving-average processes

As an analogy with Section 12.4.2.5, a **MA process** refers to the stochastic process that is obtained at the output of an **all-zero filter** when a WGN sequence is applied to the input.

Thus, a MA process is an $AZ(Q)$ model with $d_0 = 1$ driven by WGN. That is,

$$x[n] = w[n] + \sum_{k=1}^Q d_k w[n - k], \quad w[n] \sim \mathcal{N}(0, \sigma_w^2) \quad (\text{M:4.3.9})$$

The output $x[n]$ has zero-mean, and variance of

$$\sigma_x^2 = \sigma_w^2 \left[1 + \sum_{k=1}^Q |d_k|^2 \right] \quad (12.48)$$

The autocorrelation sequence and PSD are given by:

$$r_{xx}[\ell] = \sigma_w^2 r_{hh}[\ell] = \sigma_w^2 \sum_{k=0}^{Q-\ell} d_{k+\ell} d_k^*, \quad \text{for } 0 \leq \ell \leq Q \quad (12.49)$$

and is zero for $\ell > Q$, with $r_{xx}[\ell] = r_{xx}^*[-\ell]$, where $d_0 = 1$, and also where $P_{xx}(e^{j\omega}) = \sigma_w^2 |D(e^{j\omega})|^2$.

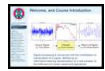
The fact that $r_{xx}[\ell] = 0$ if the samples are more than Q samples apart, means that they are therefore uncorrelated. An alternative derivation for the autocorrelation sequence for an MA process is given in the following section, Section 12.4.3.5.

12.4.3.5 Autocorrelation Function for MA Process

As stated in the previous section, using the results for the statistics of a stationary signal passed through a linear system, then the autocorrelation sequence for a MA process is given by $r_{xx}[\ell] = \sigma_w^2 r_{hh}[\ell]$, where $r_{hh}[\ell]$ is given by Equation M:4.3.4. For completeness, this section gives an alternative derivation from first principles.

Multiplying the difference equation, Equation M:4.3.1, through by $x^*[n - \ell]$ and taking expectations gives:

$$r_{xx}[\ell] = \sum_{k=0}^Q d_k r_{wx}[\ell - k] \quad (12.50)$$



New slide

Similarly, post-multiplying by $w^*[n - \ell]$ gives:

$$r_{xw}[\ell] = \sum_{k=0}^Q d_k r_{ww}[\ell - k] = \begin{cases} \sigma_w^2 d_\ell & 0 \leq \ell \leq Q \\ 0 & \text{otherwise} \end{cases} \quad (12.51)$$

since $r_{ww}[\ell] = \sigma_w^2 \delta(\ell)$. Recalling that $r_{wx}[\ell] = r_{xw}^*[-\ell]$, then:

$$r_{wx}[\ell] = \begin{cases} \sigma_w^2 d_{-\ell}^* & 0 \leq -\ell \leq Q \\ 0 & \text{otherwise} \end{cases} \quad (12.52)$$

with the limit $0 \leq -\ell \leq Q$ being equivalent to $-Q \leq \ell \leq 0$. Consequently,

$$r_{wx}[\ell - k] = \begin{cases} \sigma_w^2 d_{k-\ell}^* & 0 \leq k - \ell \leq Q \\ 0 & \text{otherwise} \end{cases} \quad (12.53)$$

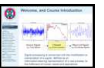
Considering $\ell > 0$, the autocorrelation sequence for an MA process is thus:

$$r_{xx}[\ell] = \sigma_w^2 \sum_{k=\ell}^Q d_k d_{k-\ell}^* = \sigma_w^2 \sum_{k=0}^{Q-\ell} d_{k+\ell} d_k^* \quad (12.54)$$

for $0 \leq \ell \leq Q$, and zero for $\ell > Q$. Using the relationship $r_{xx}[-\ell] = r_{xx}^*[\ell]$ gives the ACS for all values of ℓ .

Unlike AR models, it is not possible to solve for the model parameters using linear algebra techniques. It requires the solution of highly nonlinear equations, and is therefore more difficult than dealing with AR process. This, hence, is one reason why many algorithms in statistical signal processing prefer to use all-pole models over all-zero models.

12.4.4 Pole-Zero Models



Finally, the most general of LTI parametric signal models is the **pole-zero** model which, as the name suggests, is a combination of the all-pole and all-zero models, and can therefore model both resonances as well as nulls in a frequency response.

New slide

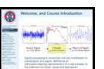
The output of a causal pole-zero model is given by the recursive input-output relationship:

$$x[n] = - \sum_{k=1}^P a_k x[n - k] + \sum_{k=0}^Q d_k w[n - k] \quad (\text{M:4.4.1})$$

where it is assumed that the model orders $P > 0$ and $Q \geq 1$. The corresponding system function is given by:

$$H(z) = \frac{D(z)}{A(z)} = \frac{\sum_{k=0}^Q d_k z^{-k}}{1 + \sum_{k=1}^P a_k z^{-k}} \quad (12.55)$$

12.4.4.1 Frequency Response of an Pole-Zero Model



The pole-zero model can be written as

New slide

$$H(z) = \frac{D(z)}{A(z)} = d_0 \frac{\prod_{k=1}^Q (1 - z_k z^{-1})}{\prod_{k=1}^P (1 - p_k z^{-1})} \quad (12.56)$$

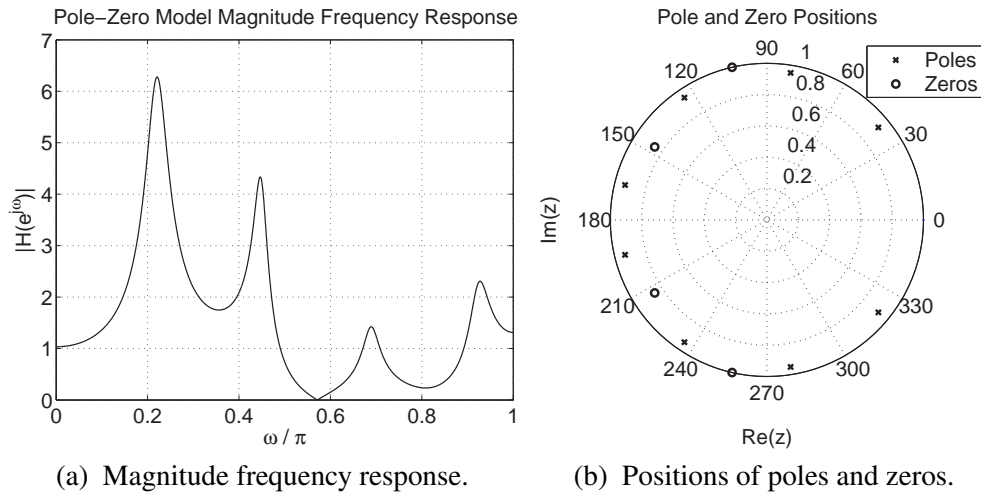


Figure 12.10: The frequency response and position of the poles and zeros in an pole-zero system.

where $\{p_k\}$ and $\{z_k\}$ are the poles and zeros of the model. Therefore, its frequency response is:

$$H(e^{j\omega}) = d_0 \frac{\prod_{k=1}^Q (1 - z_k e^{-j\omega})}{\prod_{k=1}^P (1 - p_k e^{-j\omega})} \quad (12.57)$$

As before, it can be deduced that troughs or nulls occur at frequencies corresponding to the phase position of the zeros, while resonances occur at frequencies corresponding to the phase of the poles. When the system is real, the complex-poles and complex-zeros occur in conjugate-pairs.

The PSD of the output of a pole-zero filter is given by:

$$P_{xx}(e^{j\omega}) = \sigma_w^2 |H(e^{j\omega})|^2 = G^2 \frac{\prod_{k=1}^Q |1 - z_k e^{-j\omega}|^2}{\prod_{k=1}^P |1 - p_k e^{-j\omega}|^2} \quad (12.58)$$

where $G = \sigma_w d_0$ is the overall gain of the system.

Consider the pole-zero model with poles at positions:

$$\{p_k\} = \{r_k e^{j\omega_k}\} \quad \text{where} \quad \begin{cases} \{r_k\} = \{0.925, 0.951, 0.942, 0.933\} \\ \{\omega_k\} = 2\pi \times \{270, 550, 844, 1131\}/2450; \end{cases} \quad (12.59)$$

and zeros at:

$$\{z_k\} = \{r_k e^{j\omega_k}\} \quad \text{where} \quad \begin{cases} \{r_k\} = \{1, 0.855\} \\ \{\omega_k\} = 2\pi \times \{700, 1000\}/2450; \end{cases} \quad (12.60)$$

The pole and zero positions, and the magnitude frequency response of this system is plotted in Figure 12.10, while the PSD of the output of the system is shown in Figure 12.11. Note again that one of the zeros lies on the unit-circle, and therefore at the corresponding frequency, the frequency response is zero.

12.4.4.2 Impulse Response

The impulse response of a causal pole-zero filter can be obtained from Equation M:4.4.1 by substituting $w(n) = \delta(n)$ and $x(n) = h(n)$, such that:

$$h(n) = - \sum_{k=1}^P a_k h(n-k) + d_n, \quad n \geq 0 \quad (\text{M:4.4.2})$$

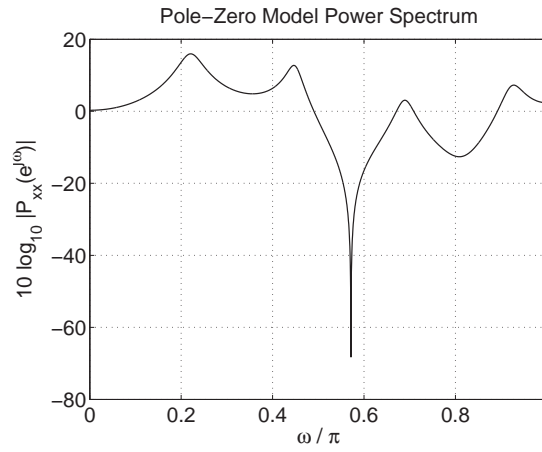


Figure 12.11: Power spectral response of an pole-zero model.

where $d_n = 0$ for $n > Q$ and $n < 0$, and $h(n) = 0$ for $n < 0$. Hence, writing this explicitly as:

$$h(n) = \begin{cases} 0 & n < 0 \\ -\sum_{k=1}^P a_k h(n-k) + d_n & 0 \leq n \leq Q \\ -\sum_{k=1}^P a_k h(n-k) & n > 0 \end{cases} \quad (12.61)$$

it can be seen that the impulse response obeys a *linear prediction* equation for $n > Q$. Thus, given $h(n)$ for $0 \leq n \leq P + Q$, the all-pole parameters $\{a_k\}$ can be calculated by using the P equations specified by $Q + 1 \leq n \leq P + Q$. Given the $\{a_k\}$'s, it is then possible to compute the all-zero parameters from Equation M:4.4.2 using the equations for $0 \leq n \leq Q$. Thus, it is clear that the first $P + Q + 1$ values of the impulse response completely specify the pole-zero model.

12.4.4.3 Autocorrelation of the Impulse Response

Multiplying both sides of Equation M:4.4.2 by $h^*(n-l)$ and summing over all n gives:

$$\sum_{n=-\infty}^{\infty} h(n)h^*(n-l) = -\sum_{k=1}^P a_k \sum_{n=-\infty}^{\infty} h(n-k)h^*(n-l) + \sum_{n=-\infty}^{\infty} d_n h^*(n-l) \quad (12.62)$$

Using the definition for $r_{hh}[\ell]$ and noting that $h^*[n-l] = 0$ for $n-l < 0$ then:

$$r_{hh}[\ell] = -\sum_{k=1}^P a_k r_{hh}[\ell-k] + \sum_{n=0}^Q d_n h^*[n-l] \quad (\text{M:4.4.6})$$

Since the impulse response $h[n]$ is a function of the parameters $\{a_k\}$'s and $\{d_k\}$'s, then this set of equations is nonlinear in terms of these parameters. However, noting that the right hand side (RHS) of this equation is zero for $l > Q$, then:

$$\sum_{k=1}^P a_k r_{hh}[\ell-k] = -r_{hh}[\ell], \quad \ell > Q \quad (12.63)$$

This equation, unlike Equation M:4.4.6, is linear in the all-pole parameters $\{a_k\}$'s. Therefore, given the autocorrelation of the impulse response, the all-pole parameters can be calculated by solving

Equation 12.63 for $l \in \{Q + 1, \dots, Q + P\}$ to give:

$$\begin{bmatrix} r_{hh}(Q) & r_{hh}(Q-1) & \cdots & r_{hh}(Q+1-P) \\ r_{hh}(Q+1) & r_{hh}(Q) & \cdots & r_{hh}(Q+2-P) \\ \vdots & \vdots & \ddots & \vdots \\ r_{hh}(Q+P-1) & r_{hh}(Q+P-2) & \cdots & r_{hh}(Q) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix} = - \begin{bmatrix} r_{hh}(Q+1) \\ r_{hh}(Q+2) \\ \vdots \\ r_{hh}(Q+P) \end{bmatrix} \quad (\text{M:4.4.8})$$

or, alternatively,

$$\mathbf{R}_{hh} \mathbf{a} = -\mathbf{r}_{hh} \quad (\text{M:4.4.9})$$

The matrix \mathbf{R}_{hh} in Equation M:4.4.9 is a non-Hermitian Toeplitz matrix; it can be solved using a variety of linear algebra techniques.

Given the all-pole parameters, it then falls to solve Equation M:4.4.6 for the all-zero parameters $\{d_k\}$'s. This is somewhat involved, but they can be found using spectral factorisation. The details are omitted from this handout, but can be found in [Therrien:1992, Section 9.1, page 509] or [Manolakis:2000, Page 178].

12.4.4.4 Autoregressive Moving-Average Processes

As with the all-pole and all-zero models, the corresponding random process associated with a pole-zero model is the **ARMA process**. This is the output of a pole-zero model, when the input of the system is driven by WGN. Hence, a causal ARMA model with model orders P and Q is defined by:

$$x[n] = - \sum_{k=1}^P a_k x[n-k] + w[n] + \sum_{k=1}^Q d_k w[n-k] \quad (\text{M:4.4.15})$$

where $w(n) \sim \mathcal{N}(0, \sigma_w^2)$, the model-orders are P and Q , and the full set of model parameters are $\{\sigma_w^2, a_1, \dots, a_P, d_1, \dots, d_Q\}$. The output has zero-mean and variance that can be shown to equal:

$$\sigma_x^2 = - \sum_{k=1}^P a_k r_{xx}(k) + \sigma_w^2 \left[1 + \sum_{k=1}^Q d_k h(k) \right] \quad (\text{M:4.4.16})$$

where $h[n]$ is the impulse response of the pole-zero filter.

Finally, the autocorrelation function for the output is given by:

$$r_{xx}(l) = - \sum_{k=1}^P a_k r_{xx}(l-k) + \sigma_w^2 \left[1 + \sum_{n=l}^Q d_n h^*(n-l) \right] \quad (12.64)$$

where it has been noted that $d_0 = 1$.

12.5 Estimation of AR Model Parameters from Data

The Yule-Walker equations introduced earlier in this handout provide an approach for finding the model parameters for an AR process. Although a valid technique, there are two implicit assumptions that limit its use for practical problems. These assumptions are:

- That the order, P , of the model is known.
- That the correlation function, $r_{xx}[\ell]$, is known.

If these two conditions are met then, using the Yule-Walker equations, the model parameters, a_k , can be found exactly. Unfortunately, in most practical situations, *neither* of these conditions is met.

From a theoretical perspective, the first assumption that the model order is known is less of an issue than the second assumption. This is since if a larger model order than the true model order is chosen, then the excess parameters will theoretically be zero. In practice, choosing the models order is not that straightforward, and there are numerous methods for model order estimation. Model order selection criteria include names such as 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). There is not time in this course to discuss these techniques, although there are plenty of tutorial papers in the literature, as well as being covered by many text books.

The second assumption leads to both theoretical and practical problems since, if the correlation function is not known, it must be estimated from the data. This brings up the following questions:

1. If the correlation function is estimated, how good is the resulting estimate for the model parameters, in a statistical sense?
2. Why estimate the correlation function at all when it is the model parameters that need to be estimated?
3. What is the best procedure for this problem?

12.5.1 LS AR parameter estimation

Suppose that a particular realisation of a process that is to be modelled as an AR process is given. It is possible to estimate the correlation function as a time-average from the realisation, assuming that the process is time-ergodic, and then use these estimates in the Yule-Walker equations. The method described in this chapter effectively estimates the AR parameters in this way, although the problem is not formulated as such. Two common data-oriented methods, known as the **autocorrelation method** and the **covariance method**, are presented in this section and the next section. A description of these methods begins with the **autocorrelation method**.

Suppose linear prediction is used to model a particular realisation of a random process as accurately as possible. Thus, suppose a **linear predictor** forms an estimate, or *prediction*, $\hat{x}[n]$, of the present value of a stochastic process $x[n]$ from a linear combination of the past P samples; that is:

$$\hat{x}[n] = - \sum_{k=1}^P a_k x[n-k] \quad (\text{M:1.4.1})$$

Then the **prediction error** is given by:

$$e[n] = x[n] - \hat{x}[n] = x[n] + \sum_{k=1}^P a_k x[n-k] \quad (\text{M:4.2.50})$$

Note that this is different to the WGN sequence that drives a linear system to generate an autoregressive random process; the difference is that here, the prediction error is the difference between the actual value and the predicted value of a *particular realisation* of a random process.

Writing Equation M:4.2.50 for $n \in \{n_I, \dots, n_F\}$, in matrix-vector form:

$$\underbrace{\begin{bmatrix} e[n_I] \\ e[n_I+1] \\ \vdots \\ e[n_F] \end{bmatrix}}_{\mathbf{e}} = \underbrace{\begin{bmatrix} x[n_I] \\ x[n_I+1] \\ \vdots \\ x[n_F] \end{bmatrix}}_{\mathbf{x}} + \underbrace{\begin{bmatrix} x[n_I-1] & x[n_I-2] & \cdots & x[n_I-P] \\ x[n_I] & x[n_I-1] & \cdots & x[n_I-P+1] \\ \vdots & \vdots & \cdots & \vdots \\ x[n_F-1] & x[n_F-2] & \cdots & x[n_F-P] \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix}}_{\mathbf{a}} \quad (12.65)$$

which can hence be written as:

$$\mathbf{e} = \mathbf{x} + \mathbf{X} \mathbf{a} \quad (12.66)$$

The parameters \mathbf{a} can now be estimated using any of the parameter estimation techniques discussed above. Here, the least-squares estimate (LSE) is used. Thus, noting that:

$$J(\mathbf{a}) = \sum_{n=n_I}^{n_F} e^2[n] = \mathbf{e}^T \mathbf{e} \quad (12.67)$$

$$= (\mathbf{x} + \mathbf{X} \mathbf{a})^T (\mathbf{x} + \mathbf{X} \mathbf{a}) \quad (12.68)$$

$$= \mathbf{x}^T \mathbf{x} + 2\mathbf{x}^T \mathbf{X} \mathbf{a} + \mathbf{a}^T \mathbf{X}^T \mathbf{X} \mathbf{a} \quad (12.69)$$

where it has been noted that $\mathbf{a}^T \mathbf{X}^T \mathbf{x} = \mathbf{x}^T \mathbf{X} \mathbf{a}$. Hence, differentiating with respect to (w. r. t.) \mathbf{a} and setting to zero gives the LSE, $\hat{\mathbf{a}}$,

$$\frac{\partial}{\partial \mathbf{a}} (\mathbf{b}^T \mathbf{a}) = \mathbf{b} \quad \text{and} \quad \frac{\partial}{\partial \mathbf{a}} (\mathbf{a}^T \mathbf{B} \mathbf{a}) = (\mathbf{B} + \mathbf{B}^T) \mathbf{a} \quad (12.70)$$

The reader is invited to derive these results. Hence,

$$\frac{\partial J(\mathbf{a})}{\partial \mathbf{a}} = 2\mathbf{X}^T \mathbf{x} + 2\mathbf{X}^T \mathbf{X} \mathbf{a} \quad (12.71)$$

where it has been noted that the matrix $\mathbf{X}^T \mathbf{X}$ is symmetric. Setting this to zero, and rearranging noting that $\mathbf{X}^T \mathbf{X}$ is of full rank, gives the LSE:

$$\mathbf{a}_{LSE} = -(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{x} \quad (12.72)$$

Defining $N_p = n_F - n_I + 1$, the least-squares (LS) error is then given by:

$$J(\mathbf{a}_{LSE}) = \mathbf{x}^T \left(\mathbf{I}_{N_p} - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right) \mathbf{x} \quad (12.73)$$

$$= \mathbf{x}^T (\mathbf{x} + \mathbf{X} \mathbf{a}_{LSE}) \quad (12.74)$$

Observe the similarity of these results with those of the linear LS formulation. In fact, this derivation is identical to the LS formulation with the matrix \mathbf{H} replaced by \mathbf{X} ! There are two different methods which result from different choices of n_I and n_F . These are called the **autocorrelation method** and the **covariance method**. However, as mentioned in [Therrien:1991], these terms do not bear any relation to the statistical meanings of these terms, and so they should not be confused with the statistical definitions. The names for these methods are unfortunate, but have found a niche in signal processing, and are unlikely to be changed.

12.5.2 Autocorrelation Method

In the **autocorrelation method**, the end points are chosen as $n_I = 0$ and $n_F = N + P - 1$. Thus, the AR filter model runs over the entire length of the data, predicting some of the early points from *zero valued samples*, and predicting P additional zero values at the end. Since this method uses zeros for the data outside of the given interval, it can be thought of as applying a rectangular window to the

data. For this method, the $(N + P) \times P$ data matrix \mathbf{X} has the specific structure:

$$\mathbf{X} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ x[0] & 0 & \cdots & 0 \\ x[1] & x[0] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x[P-1] & x[P-2] & \cdots & x[0] \\ x[P] & x[P-1] & \cdots & x[1] \\ \vdots & \vdots & & \vdots \\ x[N-1] & x[N-2] & \cdots & x[N-P] \\ 0 & x[N-1] & \cdots & x[N-P+1] \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x[N-1] \end{bmatrix} \quad (\text{T:9.112})$$

When formed into the product $\mathbf{X}^T \mathbf{X}$, this data matrix produces a Toeplitz correlation matrix; consequently, the **normal equations** may be solved very efficiently, for example using the **Levinson recursion**. Moreover, the matrix $\mathbf{X}^T \mathbf{X}$ is strictly positive definite, and thus a valid *correlation matrix*.

12.5.3 Covariance Method

An alternative method is to choose $n_I = P$ and $n_F = N - 1$. With this method, no zeros are either predicted, or used in the prediction. In other words, the limits are chosen so that the data that the AR filter operates on always remain within the measured data; no window is applied. For this method, the $(N - P) \times P$ data matrix has the specific form:

$$\mathbf{X} = \begin{bmatrix} x[P-1] & x[P-2] & \cdots & x[0] \\ x[P] & x[P-1] & \cdots & x[1] \\ \vdots & \vdots & \ddots & \vdots \\ x[N-2] & x[N-3] & \cdots & x[N-P-1] \end{bmatrix} \quad (\text{T:9.113})$$

A variation of this method called the **prewindowed covariance method** chooses $n_I = 0$ and $n_F = N - 1$, and results in a data matrix that consists of the first N rows of Equation T:9.112. Moreover, the **postwindowed covariance method** chooses $n_I = P$ and $n_F = N + P - 1$. In the **autocorrelation method**, the data is said to be both *prewindowed* and *postwindowed*.

With the covariance method, or the prewindowed covariance method, the resulting correlation matrix is positive semidefinite, but it is *not* Toeplitz. Thus, the Yule-Walker equations are more difficult to solve. Moreover, the resulting AR model *may not be stable*, since the poles corresponding to the estimated parameters may not lie within the unit circle. Nevertheless, unstable cases rarely seem to occur in practice, and the covariance method is often preferred because it makes use of only the measured data. This avoids any bias in the estimation of the AR filter coefficients.

Example 12.2 ([Therrien:1991, Example 9.6, Page 539]). It is desired to estimate the parameters of a second-order AR model for the sequence $\{x[n]\}_0^4 = \{1, -2, 3, -4, 5\}$ by using both the autocorrelation and covariance methods.

SOLUTION. Applying both methods as requested:

Autocorrelation Method The data matrix can be obtained from Equation T:9.112, and is given by:

$$\mathbf{X} = \begin{bmatrix} 0 & 0 \\ x[0] & 0 \\ x[1] & x[0] \\ x[2] & x[1] \\ x[3] & x[2] \\ x[4] & x[3] \\ 0 & x[4] \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ -2 & 1 \\ 3 & -2 \\ -4 & 3 \\ 5 & -4 \\ 0 & 5 \end{bmatrix} \quad (12.75)$$

Hence, it can be shown that:

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} 0 & 1 & -2 & 3 & -4 & 5 & 0 \\ 0 & 0 & 1 & -2 & 3 & -4 & 5 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ -2 & 1 \\ 3 & -2 \\ -4 & 3 \\ 5 & -4 \\ 0 & 5 \end{bmatrix} \quad (12.76)$$

$$= \begin{bmatrix} 55 & -40 \\ -40 & 55 \end{bmatrix} \quad (12.77)$$

Note that the matrix is Toeplitz. The least squares Yule-Walker equations can then be found by solving:

$$\mathbf{a}_{LSE} = -(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{x} \quad (12.78)$$

$$= - \begin{bmatrix} 55 & -40 \\ -40 & 55 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 1 & -2 & 3 & -4 & 5 & 0 \\ 0 & 0 & 1 & -2 & 3 & -4 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \\ 3 \\ -4 \\ 5 \\ 0 \\ 0 \end{bmatrix} \quad (12.79)$$

Solving these equations gives:

$$\mathbf{a}_{LSE} = \begin{bmatrix} \frac{232}{285} \\ \frac{34}{285} \end{bmatrix} \approx \begin{bmatrix} 0.8140 \\ 0.1193 \end{bmatrix} \quad (12.80)$$

The LS error is then given by:

$$J(\mathbf{a}_{LSE}) = \mathbf{x}^T (\mathbf{x} + \mathbf{X}\mathbf{a}_{LSE}) = 25.54 \quad (12.81)$$

Hence, the **prediction error variance** is estimated as:

$$\sigma_e^2 = \frac{J(\mathbf{a}_{LSE})}{N} = \frac{25.54}{7} = 3.64 \quad (12.82)$$

Covariance Method Next, apply the covariance method to the same problem. Since the AR filter stays entirely within the data, the error is evaluated from $n = 2$ to $n = 4$. The data matrix is therefore:

$$\mathbf{X} = \begin{bmatrix} x[1] & x[0] \\ x[2] & x[1] \\ x[3] & x[2] \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 3 & -2 \\ -4 & 3 \end{bmatrix} \quad (12.83)$$

Notice that, in this data matrix, not all the the data has been used, since $x[4]$ does not appear. Hence, the correlation matrix is given by:

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} -2 & 3 & -4 \\ 1 & -2 & 3 \end{bmatrix} \begin{bmatrix} -2 & 1 \\ 3 & -2 \\ -4 & 3 \end{bmatrix} = \begin{bmatrix} 29 & -20 \\ -20 & 14 \end{bmatrix} \quad (12.84)$$

This matrix is not Toeplitz. The LSE estimate is therefore:

$$\mathbf{a}_{LSE} = -(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{x} \quad (12.85)$$

$$= - \begin{bmatrix} 29 & -20 \\ -20 & 14 \end{bmatrix}^{-1} \begin{bmatrix} -2 & 3 & -4 \\ 1 & -2 & 3 \end{bmatrix} \begin{bmatrix} 3 \\ -4 \\ 5 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad (12.86)$$

Moreover, the LS error is then given by:

$$J(\mathbf{a}_{LSE}) = \mathbf{x}^T (\mathbf{x} + \mathbf{X}\mathbf{a}_{LSE}) = 0 \quad (12.87)$$

Hence, the **prediction error variance** is estimated as:

$$\sigma_e^2 = \frac{J(\mathbf{a}_{LSE})}{N} = \frac{0}{3} = 0 \quad (12.88)$$

Evidently, this filter predicts the data perfectly. Indeed, if the prediction error, $e[n]$, is computed over the chosen range $n = 2$ to $n = 4$, it is found to be zero at every point. The price to be paid for this perfect prediction, however, is an unstable AR model. The transfer function for this AR model can be written as:

$$H(z) = \frac{1}{1 + 2z^{-1} + z^{-2}} = \frac{1}{(1 + z^{-1})^2} \quad (12.89)$$

□

which has a double pole at $z = -1$. Therefore, a bounded-input into this filter can potentially produce an unbounded-output. Further, any errors in computation of the model coefficients can easily put a pole outside of the unit circle.

Part IV
Advanced Topics

13

Application: Passive Target Localisation

This handout discusses a general problem of passive target localisation. Using the techniques described throughout this tutorial, it should now be possible to appreciate many of the techniques used in this problem.

13.1 Introduction

- This research tutorial is intended to cover a wide range of aspects which link acoustic source localisation (ASL) and blind source separation (BSS). 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.

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://mod-udrc.org/events/2016-summer-school>
- Thanks to Xionghu Zhong and Ashley Hughes for borrowing some of their diagrams from their dissertations.

13.1.1 Structure of the Tutorial

- Recommended Texts

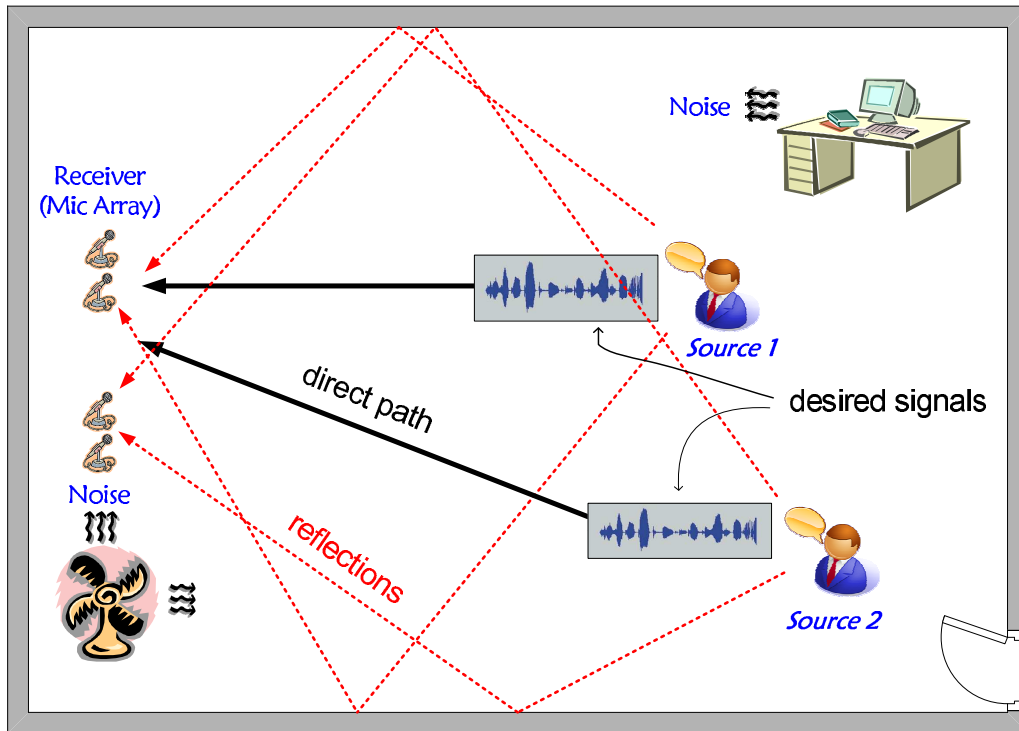


Figure 13.1: Source localisation and blind source separation (BSS).

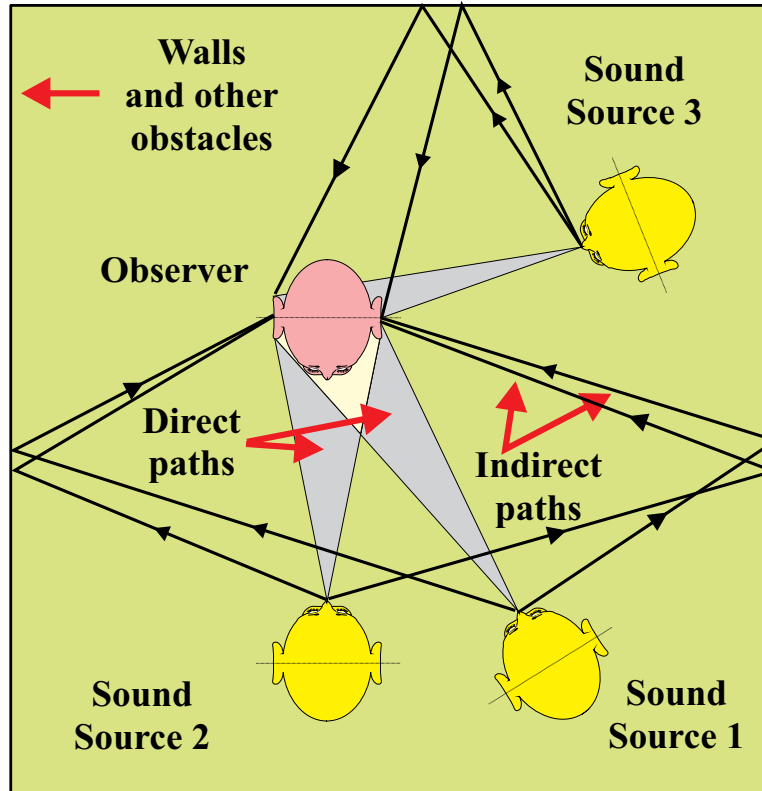


Figure 13.2: Humans turn their head in the direction of interest in order to reduce interference from other directions; *joint detection, localisation, and enhancement*.



Figure 13.3: Recommended book chapters and the references therein.

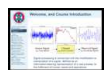
- Conceptual link between ASL and BSS.
- Geometry of source localisation.
- Spherical and hyperboloidal localisation.
- Estimating time-difference of arrivals (TDOAs).
- Steered beamformer response function.
- Multiple target localisation using BSS.
- Conclusions.

13.2 Recommended Texts

- Huang Y., J. Benesty, and J. Chen, “Time Delay Estimation and Source Localization,” in *Springer Handbook of Speech Processing* by J. Benesty, M. M. Sondhi, and Y. Huang, pp. 1043–1063, , Springer, 2008.
- Chapter 8: DiBiase J. H., H. F. Silverman, and M. S. Brandstein, “Robust Localization in Reverberant Rooms,” in *Microphone Arrays* by M. Brandstein and D. Ward, pp. 157–180, , Springer Berlin Heidelberg, 2001.
- Chapter 10 of Wolfel M. and J. McDonough, *Distant Speech Recognition*, Wiley, 2009.

IDENTIFIERS – Hardback, ISBN13: 978-0-470-51704-8

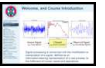
Some recent PhD thesis on the topic include:



New slide

- Zhong X., “*Bayesian framework for multiple acoustic source tracking*,” Ph.D. thesis, University of Edinburgh, 2010.
- Pertila P., “*Acoustic Source Localization in a Room Environment and at Moderate Distances*,” Ph.D. thesis, Tampere University of Technology, 2009.
- Fallon M., “*Acoustic Source Tracking using Sequential Monte Carlo*,” Ph.D. thesis, University of Cambridge, 2008.

13.3 Why Source Localisation?



A number of blind source separation (BSS) techniques rely on knowledge of the desired source position, for example: *New slide*

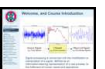
1. Look-direction in beamforming techniques.
2. Camera steering for audio-visual BSS (including Robot Audition).
3. Parametric modelling of the mixing matrix.

Equally, a number of multi-target acoustic source localisation (ASL) techniques rely on BSS. This tutorial will look at the connections and dependencies between ASL and BSS, and discuss how they can be used together. The tutorial will cover some classical well known techniques, as well as some recent advances towards the end.

In particular, the following topics will be considered in detail:

- hyperboloidal (TDOA) based localisation methods;
- TDOA estimation methods;
- steered response power (SRP) based localisation methods;
- computationally efficient SRP methods such as stochastic region contraction (SRC);
- multi-target detection and localisation using BSS algorithms such as degenerate unmixing estimation technique (DUET);

13.4 ASL Methodology



- In general, most ASL techniques rely on the fact that an impinging wavefront reaches one acoustic sensor before it reaches another. *New slide*
- Most ASL algorithms are designed assuming there is no reverberation present, the *free-field assumption*; the performance of each method in the presence of reverberation will be considered after the techniques have been introduced.
- Typically, this acoustic sensor is a microphone; this tutorial will primarily consider *omni-directional pressure sensors*, and therefore many of the techniques discussed will rely on the fact there is a TDOA between the signals at different microphones.

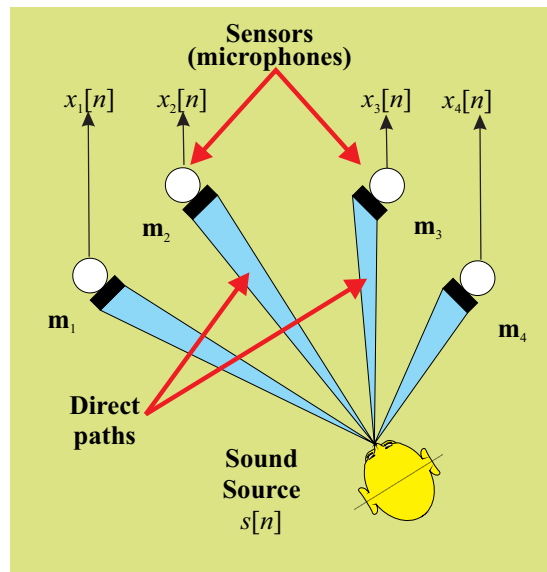


Figure 13.4: Ideal free-field model.



Figure 13.5: An uniform linear array (ULA) of microphones.



Figure 13.6: An acoustic vector sensor.

- Other measurement types include:
 - range difference measurements;
 - interaural level difference;
 - joint TDOA and vision techniques.
- Another sensor modality might include acoustic vector sensors (AVSs) which measure both air pressure and air velocity. Useful for applications such as sniper localisation.

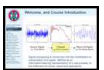
13.4.1 Source Localization Strategies

This section is based on

DiBiase J. H., H. F. Silverman, and M. S. Brandstein, “Robust Localization in Reverberant Rooms,” in *Microphone Arrays* by M. Brandstein and D. Ward, pp. 157–180, Springer Berlin Heidelberg, 2001.

Existing source localisation methods can loosely be divided into three generic strategies:

1. those based on maximising the SRP of a beamformer;
 - location estimate derived directly from a filtered, weighted, and sum version of the signal data received at the sensors.
2. techniques adopting high-resolution spectral estimation concepts (see Stephan Weiss’s talk);
 - any localisation scheme relying upon an application of the signal correlation matrix.
3. approaches employing TDOA information.



New slide

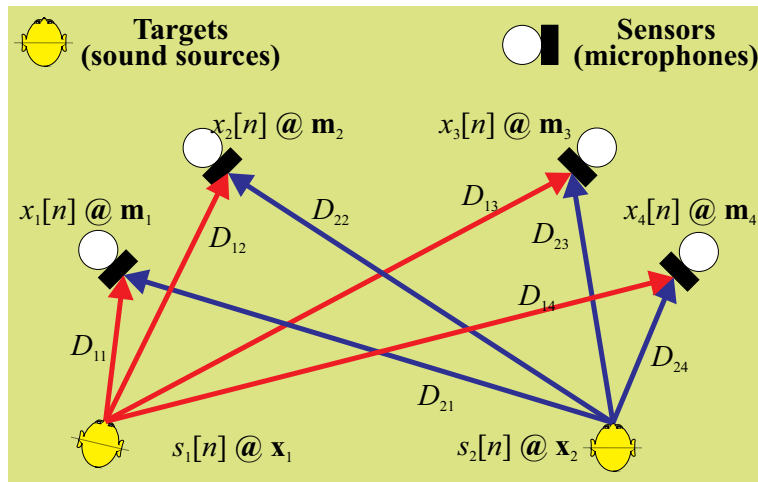


Figure 13.7: Geometry assuming a free-field model.

- source locations calculated from a set of TDOA estimates measured across various combinations of microphones.

Spectral-estimation approaches See Stephan Weiss's talk :-)

TDOA-based estimators Computationally cheap, but suffers in the presence of noise and reverberation.

SBF approaches Computationally intensive, superior performance to TDOA-based methods. However, possible to dramatically reduce computational load.

13.4.2 Geometric Layout

Suppose there is a:

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

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

$$T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k) \triangleq T_{ij}(\mathbf{x}_k) = \frac{|\mathbf{x}_k - \mathbf{m}_i| - |\mathbf{x}_k - \mathbf{m}_j|}{c} \quad (13.1)$$

where c is the speed of sound, which is approximately 344 m/s. More precisely, in air, the speed of sound is given by:

$$c = 331.4 + 0.6\Theta \quad \text{m/s} \quad (13.2)$$

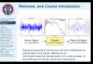
where Θ is the temperature in Centigrade or Celsius. Hence, for instance, at a temperature of 21 Celsius, then $c = 344$ m/s.

The distance from the target at \mathbf{x}_k to the sensor located at \mathbf{m}_i will be defined by D_{ik} , and is called the range. It is given by the expression

$$D_{ik} = |\mathbf{x}_k - \mathbf{m}_i| \quad (13.3)$$

Hence, it follows that

$$T_{ij}(\mathbf{x}_k) = \frac{1}{c} (D_{ik} - D_{jk}) \quad (13.4)$$



13.4.3 Ideal Free-field Model

- In an anechoic free-field acoustic environment, as depicted in Figure 13.4, the signal from source k , denoted by $s_k(t)$, propagates to the i -th sensor at time t according to the expression: New slide

$$x_{ik}(t) = \alpha_{ik} s_k(t - \tau_{ik}) + b_{ik}(t) \quad (13.5)$$

where $b_{ik}(t)$ denotes additive noise. Note that, in the frequency domain, this expression is given by:

$$X_{ik}(\omega) = \alpha_{ik} S_k(\omega) e^{-j\omega \tau_{ik}} + B_{ik}(\omega) \quad (13.6)$$

On the assumption of **geometrical room acoustics**, which assumes high frequencies, a point sound source of single frequency ω , at position \mathbf{x}_k in free space, emits a pressure wave $P_{(\mathbf{x}_k, \mathbf{m}_i), t}(\omega)$ at time t and at position \mathbf{m}_i :

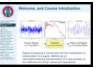
$$P_{(\mathbf{x}_k, \mathbf{m}_i), t}(\omega, t) = P_0 \frac{\exp[j\omega(r/c - t)]}{r} \quad (13.7)$$

where c is the speed of sound, $t \in \mathbb{R}$ is time, and $r = |\mathbf{x}_k - \mathbf{m}_i|$, which can be seen to equate to D_{ik} .

- The additive noise source is assumed to be uncorrelated with the source signal, as well as the noise signals at the other microphones.
- The TDOA between the i -th and j -th microphone is given by:

$$\tau_{ijk} = \tau_{ik} - \tau_{jk} = T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k) \quad (13.8)$$

13.4.4 TDOA and Hyperboloids



It is important to be aware of the geometrical properties that arise from the TDOA relationship given in Equation 13.1: New slide

$$T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k) = \frac{|\mathbf{x}_k - \mathbf{m}_i| - |\mathbf{x}_k - \mathbf{m}_j|}{c} \quad (13.9)$$

- This defines one half of a hyperboloid of two sheets, centered on the midpoint of the microphones, $\mathbf{v}_{ij} = \frac{\mathbf{m}_i + \mathbf{m}_j}{2}$. A generic diagram for the hyperboloid of two sheets is shown in Figure 13.8 and Equation 13.13. Equivalently, as shown in Sidebar 23:

$$(\mathbf{x}_k - \mathbf{v}_{ij})^T \mathbf{V}_{ij} (\mathbf{x}_k - \mathbf{v}_{ij}) = 1 \quad (13.10)$$

where

$$\tau = cT(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k), \quad \mathbf{V}_{ij} = \frac{\mathbf{I}_3 - \frac{4}{\tau^2} \boldsymbol{\mu}_{ij} \boldsymbol{\mu}_{ij}^T}{\tau^2 - |\boldsymbol{\mu}_{ij}|^2} \quad \text{and} \quad \boldsymbol{\mu}_{ij} = \frac{\mathbf{m}_i - \mathbf{m}_j}{2} \quad (13.11)$$

- For source with a large source-range to microphone-separation ratio, the hyperboloid may be well-approximated by a cone with a constant direction angle relative to the axis of symmetry. The corresponding estimated direction angle, ϕ_{ij} for the microphone pair (i, j) is given by

$$\phi_{ij} = \cos^{-1} \left(\frac{cT(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k)}{|\mathbf{m}_i - \mathbf{m}_j|} \right) \quad (13.12)$$

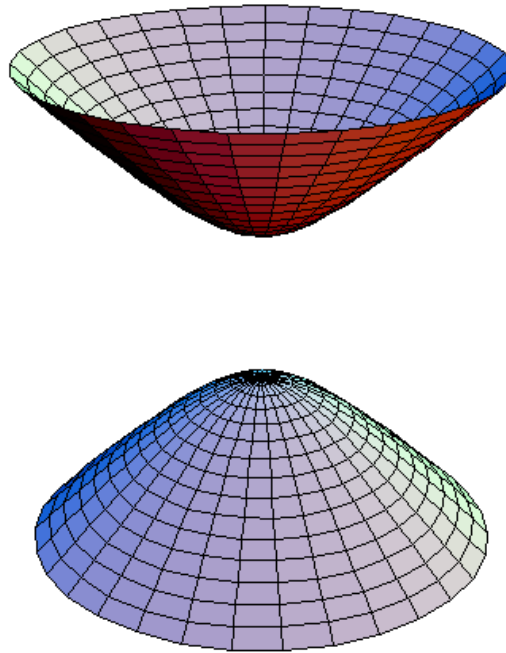


Figure 13.8: Hyperboloid of two sheets

KEYPOINT! (Hyperboloid of two sheets). General expression for a Hyperboloid of two sheets is given by:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = -1 \quad (13.13)$$

□

An example of the resulting hyperboloid for a typical case is shown in Figure 13.9, where the two-dimensional (2-D) equation is simplified in Sidebar 24. This case is for a microphone separation of $d = 0.1$, and a time-delay of $\tau_{ij} = \frac{d}{4c}$.

13.5 Indirect TDOA-based Methods

KEYPOINT! (Executive Summary). This section considers techniques which employ TDOA information directly. The section is broadly split into two sections; localising the source given TDOAs, followed by techniques for estimating TDOAs.

This is typically a two-step procedure in which:

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



New slide

Sidebar 23 Hyperboloids

Consider again Equation 13.1, but change the coordinate system to the center of the microphone pairs, such that:

$$\mathbf{x}_k = \mathbf{x} + \frac{\mathbf{m}_i + \mathbf{m}_j}{2} \quad (13.14)$$

such that:

$$\mathbf{x}_k - \mathbf{m}_i = \mathbf{x} - \underbrace{\frac{\mathbf{m}_i - \mathbf{m}_j}{2}}_{\boldsymbol{\mu}} \quad \text{and} \quad \mathbf{x}_k - \mathbf{m}_j = \mathbf{x} + \underbrace{\frac{\mathbf{m}_i - \mathbf{m}_j}{2}}_{\boldsymbol{\mu}} \quad (13.15)$$

The normalised-TDOA, which $\alpha = c\tau_{ijk}$ is the actual TDOA multiplied by the speed of sound (equivalent to a range) across these two microphones can then be expressed as

$$\alpha = |\mathbf{x} - \boldsymbol{\mu}| - |\mathbf{x} + \boldsymbol{\mu}| \quad (13.16)$$

To show this is a hyperboloid, consider multiplying both sides by $|\mathbf{x} - \boldsymbol{\mu}| + |\mathbf{x} + \boldsymbol{\mu}|$ and dividing by τ such that:

$$|\mathbf{x} - \boldsymbol{\mu}| + |\mathbf{x} + \boldsymbol{\mu}| = \frac{1}{\alpha} (|\mathbf{x} - \boldsymbol{\mu}| + |\mathbf{x} + \boldsymbol{\mu}|) (|\mathbf{x} - \boldsymbol{\mu}| - |\mathbf{x} + \boldsymbol{\mu}|) \quad (13.17)$$

$$= \frac{1}{\alpha} (|\mathbf{x} - \boldsymbol{\mu}|^2 - |\mathbf{x} + \boldsymbol{\mu}|^2) \quad (13.18)$$

$$|\mathbf{x} - \boldsymbol{\mu}| + |\mathbf{x} + \boldsymbol{\mu}| = -\frac{4\boldsymbol{\mu}^T \mathbf{x}}{\alpha} \quad (13.19)$$

Adding Equation 13.16 and Equation 13.19 gives:

$$2|\mathbf{x} - \boldsymbol{\mu}| = \alpha - \frac{4\boldsymbol{\mu}^T \mathbf{x}}{\alpha} \quad (13.20)$$

Squaring both sides again gives:

$$4\mathbf{x}^T \mathbf{x} - 8\boldsymbol{\mu}^T \mathbf{x} + 4\boldsymbol{\mu}^T \boldsymbol{\mu} = \alpha^2 - 8\boldsymbol{\mu}^T \mathbf{x} + \frac{16}{\alpha^2} \mathbf{x}^T \boldsymbol{\mu} \boldsymbol{\mu}^T \mathbf{x} \quad (13.21)$$

$$\mathbf{x}^T \mathbf{x} + \boldsymbol{\mu}^T \boldsymbol{\mu} = \frac{\alpha^2}{4} + \frac{4}{\alpha^2} \mathbf{x}^T \boldsymbol{\mu} \boldsymbol{\mu}^T \mathbf{x} \quad (13.22)$$

$$\mathbf{x}^T \left(\mathbf{I}_3 - \frac{4}{\alpha^2} \boldsymbol{\mu} \boldsymbol{\mu}^T \right) \mathbf{x} = \frac{\alpha^2}{4} - |\boldsymbol{\mu}|^2 \quad (13.23)$$

finally giving:

$$\mathbf{x}^T \mathbf{V} \mathbf{x} = 1 \quad \text{where} \quad \mathbf{V} = \frac{\mathbf{I}_3 - \frac{4}{\alpha^2} \boldsymbol{\mu} \boldsymbol{\mu}^T}{\frac{\alpha^2}{4} - |\boldsymbol{\mu}|^2} \quad (13.24)$$

which is the equation of an arbitrary orientated hyperboloid. The principal directions of the hyperboloid are the eigenvectors of the matrix \mathbf{V} . Since \mathbf{V} is rank-one, it is straightforward to show that the axis of symmetry is $\boldsymbol{\mu} = \frac{\mathbf{m}_i - \mathbf{m}_j}{2}$.

Sidebar 24 Hyperboloids Example

Continuing from the derivation in Sidebar 23, suppose the microphones are at positions $\mathbf{m}_i = [\frac{d}{2} \ 0 \ 0]^T$ and $\mathbf{m}_j = [-\frac{d}{2} \ 0 \ 0]^T$ such that $\boldsymbol{\mu} = [\frac{d}{2} \ 0 \ 0]^T$. Hence, Equation 13.24 becomes:

$$\mathbf{V} = \frac{\mathbf{I}_3 - \frac{4}{\alpha^2} \boldsymbol{\mu} \boldsymbol{\mu}^T}{\frac{\alpha^2}{4} - |\boldsymbol{\mu}|^2} \quad (13.25)$$

$$= \frac{1}{\frac{\alpha^2}{4} - \frac{d^2}{4}} \left\{ \mathbf{I}_3 - \frac{4}{\alpha^2} \begin{bmatrix} \frac{d^2}{4} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \right\} \quad (13.26)$$

$$= \frac{4}{\alpha^2 - d^2} \begin{bmatrix} 1 - \frac{d^2}{\alpha^2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (13.27)$$

This then gives the equation of the hyperboloid as:

$$\mathbf{x}^T \mathbf{V} \mathbf{x} = 1 \quad (13.28)$$

$$\mathbf{x}^T \begin{bmatrix} 1 - \frac{d^2}{\alpha^2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{x} = \frac{\alpha^2 - d^2}{4} \quad (13.29)$$

$$\left(1 - \frac{d^2}{\alpha^2}\right) x^2 + y^2 + z^2 = \frac{\alpha^2 - d^2}{4} \quad (13.30)$$

$$\frac{x^2}{\left(\frac{\alpha}{2}\right)^2} - \frac{y^2 + z^2}{\frac{1}{4}(d^2 - \alpha^2)} = 1 \quad (13.31)$$

Note that the maximum TDOA will occur when the source is on the line through the two microphones, and outside of the microphones. In this case, the maximum observed delay will be $\tau_{ij} = \frac{d}{c}$ or $\alpha = d$. Hence, $d^2 - \alpha^2 \geq 0$.

Writing $r^2 = y^2 + z^2$, which are points in the $x - y$ plane on circles of radius r , this can alternatively be written as:

$$r = \frac{1}{2} \sqrt{d^2 - \alpha^2} \sqrt{\left(\frac{2x}{\alpha}\right)^2 - 1} \quad (13.32)$$

There is no solution for $x < \frac{\alpha}{2}$.

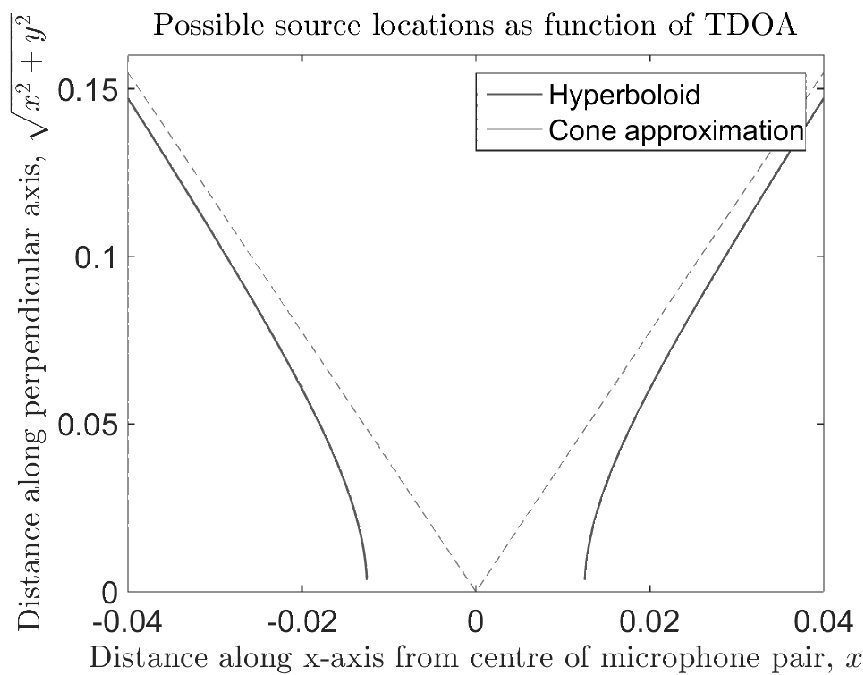
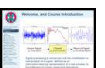


Figure 13.9: Hyperboloid, for a microphone separation of $d = 0.1$, and a time-delay of $\tau_{ij} = \frac{d}{4c}$.

- Accurate and robust TDOA estimation is the key to the effectiveness of this class of ASL methods.
- An alternative way of viewing these solutions is to consider what **spatial positions** of the target could lead to the estimated TDOA.

In the following subsections, two key error functions are considered which can be optimised in a variety of methods.

13.5.1 Spherical Least Squares Error Function



New slide

KEYPOINT! (Underlying Idea). Methods using the least squares error (LSE) function relate the distance or *range* to a target, relative to each microphone, in terms of the range to a coordinate origin and the time-difference of arrival (TDOA) estimates at each microphone.

- Suppose the first microphone is located at the origin of the coordinate system, such that $\mathbf{m}_0 = [0 \ 0 \ 0]^T$.
- The range from target k to sensor i can be expressed as the range from the target to the first sensor plus a correction term:

$$D_{ik} = D_{0k} + D_{ik} - D_{0k} \quad (13.33)$$

$$= R_s + c T_{i0}(\mathbf{x}_k) \quad (13.34)$$

where $R_{sk} = |\mathbf{x}_k|$ is the range to the first microphone which is at the origin. This is shown in Figure 13.10.

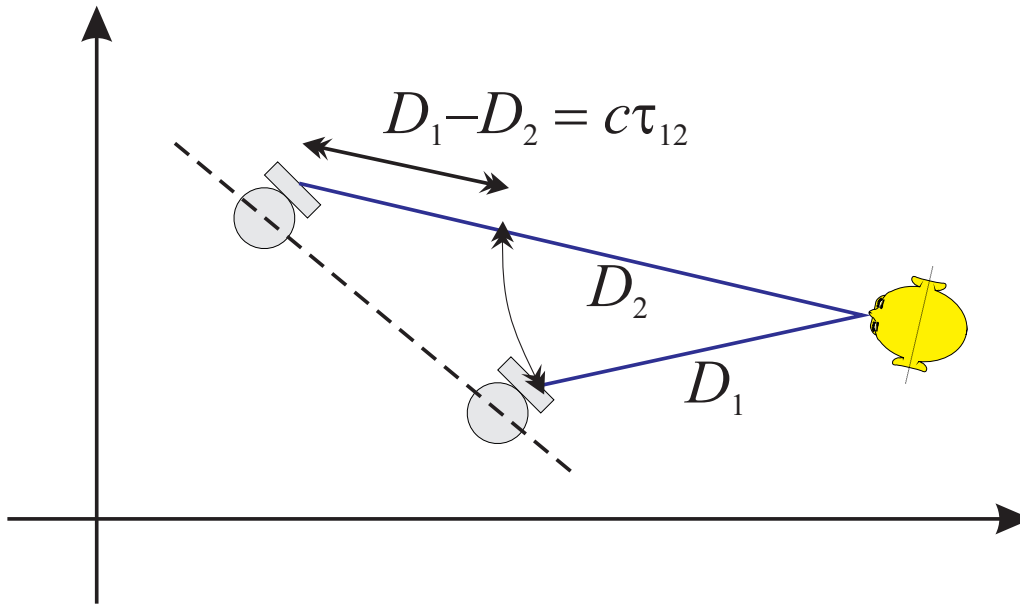


Figure 13.10: Range and TDOA relationship.

- In practice, the observations are the TDOAs and therefore, given R_{sk} , these ranges can be considered the **measurement ranges**.

Of course, knowing R_{sk} is half the solution, but it is just one unknown at this stage. The measurements can be as

$$\hat{D}_{ik} \equiv \hat{R}_s + c\hat{T}_{ij} \quad (13.35)$$

- The source-sensor geometry states that the target lies on a sphere centered on the corresponding sensor. Hence,

$$D_{ik}^2 = |\mathbf{x}_k - \mathbf{m}_i|^2 \quad (13.36)$$

$$= \mathbf{x}_k^T \mathbf{x}_k - 2\mathbf{m}_i^T \mathbf{x}_k + \mathbf{m}_i^T \mathbf{m}_i \quad (13.37)$$

$$= R_s^2 - 2\mathbf{m}_i^T \mathbf{x}_k + R_i^2 \quad (13.38)$$

where $R_i = |\mathbf{m}_i|$ is the distance of the i -th microphone to the origin.

- Define the **spherical error function** for the i th-order-microphone as the difference between the squared measured range and the squared spherical modelled range values. Using Equation 13.34 and Equation 13.38, this spherical error function can be written as:

$$\epsilon_{ik} \triangleq \frac{1}{2} \left(\hat{D}_{ik}^2 - D_{ik}^2 \right) \quad (13.39)$$

$$= \frac{1}{2} \left\{ \left(R_s + c\hat{T}_{i0} \right)^2 - \left(R_s^2 - 2\mathbf{m}_i^T \mathbf{x}_k + R_i^2 \right) \right\} \quad (13.40)$$

$$= \mathbf{m}_i^T \mathbf{x}_k + c R_s \hat{T}_{i0} + \frac{1}{2} \left(c^2 \hat{T}_{i0}^2 - R_i^2 \right) \quad (13.41)$$

- Concatenating the error functions for each microphone gives the expression:

$$\boldsymbol{\epsilon}_{ik} = \mathbf{A} \mathbf{x}_k - \underbrace{(\mathbf{b}_k - R_{sk} \mathbf{d}_k)}_{\mathbf{v}_k} \quad (13.42)$$

$$\equiv \underbrace{[\mathbf{A} \quad \mathbf{d}_k]}_{\mathbf{S}_k} \underbrace{\begin{bmatrix} \mathbf{x}_k \\ R_{sk} \end{bmatrix}}_{\boldsymbol{\theta}_k} - \mathbf{b}_k \quad (13.43)$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{m}_0^T \\ \vdots \\ \mathbf{m}_{N-1}^T \end{bmatrix}, \quad \mathbf{d} = c \begin{bmatrix} \hat{T}_{00} \\ \vdots \\ \hat{T}_{(N-1)0} \end{bmatrix}, \quad \mathbf{b}_k = \frac{1}{2} \begin{bmatrix} c^2 \hat{T}_{00}^2 - R_0^2 \\ \vdots \\ c^2 \hat{T}_{(N-1)0}^2 - R_{N-1}^2 \end{bmatrix} \quad (13.44)$$

- The least-squares estimate (LSE) can then be obtained by forming the sum-of-squared errors term using $J = \boldsymbol{\epsilon}_i^T \boldsymbol{\epsilon}_i$ which simplifies to:

$$J(\mathbf{x}_k) = (\mathbf{A} \mathbf{x}_k - (\mathbf{b}_k - R_{sk} \mathbf{d}_k))^T (\mathbf{A} \mathbf{x}_k - (\mathbf{b}_k - R_{sk} \mathbf{d}_k)) \quad (13.45a)$$

$$J(\mathbf{x}_k, \boldsymbol{\theta}_k) = (\mathbf{S}_k \boldsymbol{\theta}_k - \mathbf{b}_k)^T (\mathbf{S}_k \boldsymbol{\theta}_k - \mathbf{b}_k) \quad (13.45b)$$

- Note that as $R_{sk} = |\mathbf{x}_k|$, these parameters aren't in fact independent. Therefore, the problem to be solved can either be formulated as:

- a nonlinear least-squares problem in \mathbf{x}_k as described by Equation 13.45a;
- a linear minimisation subject to quadratic constraints:

$$\hat{\boldsymbol{\theta}}_k = \arg \min_{\boldsymbol{\theta}_k} (\mathbf{S}_k \boldsymbol{\theta}_k - \mathbf{b}_k)^T (\mathbf{S}_k \boldsymbol{\theta}_k - \mathbf{b}_k) \quad (13.46)$$

subject to the constraint

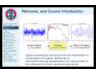
$$\boldsymbol{\theta}_k \Delta \boldsymbol{\theta}_k = 0 \quad \text{where} \quad \Delta = \text{diag}[1, 1, 1, -1] \quad (13.47)$$

The constraint $\boldsymbol{\theta}_k \Delta \boldsymbol{\theta}_k = 0$ is equivalent to

$$x_{sk}^2 + y_{sk}^2 + z_{sk}^2 = R_{sk}^2 \quad (13.48)$$

where (x_{sk}, y_{sk}, z_{sk}) are the Cartesian coordinates of the source position.

13.5.1.1 Two-step Spherical LSE Approaches



New slide

KEYPOINT! (Constrained least-squares). To avoid solving either a nonlinear or a constrained least-squares problem, it is possible to solve the problem in two steps, namely:

1. solving a LLS problem in \mathbf{x}_k assuming the range to the target, R_{sk} , is known;
2. and then solving for R_{sk} given an estimate of \mathbf{x}_k in terms of (i. t. o.) R_{sk} .

This approach is followed in the **spherical intersection (SX)** and **spherical interpolation (SI)** estimators as shown below.

- In both approaches, the range estimate is assumed known, so that the LSE can be expressed as:

$$J(\mathbf{x}_k) = \boldsymbol{\epsilon}_i^T \boldsymbol{\epsilon}_i = (\mathbf{A} \mathbf{x}_k - \mathbf{v}_k)^T (\mathbf{A} \mathbf{x}_k - \mathbf{v}_k) \quad (13.49)$$

Assuming an estimate of R_{sk} , denoted by \hat{R}_{sk} , this can be solved as

$$\hat{\mathbf{x}}_k = \mathbf{A}^\dagger \mathbf{v}_k = \mathbf{A}^\dagger (\mathbf{b}_k - \hat{R}_{sk} \mathbf{d}_k) \quad \text{where} \quad \mathbf{A}^\dagger = [\mathbf{A}^T \mathbf{A}]^{-1} \mathbf{A}^T \quad (13.50)$$

Note that \mathbf{A}^\dagger is the pseudo-inverse of \mathbf{A} .

Again, recall that the only observations are the TDOAs, $\{\hat{T}_{i0}, i \in \{0, N-1\}\}$, and that while R_{sk} is assumed known, clearly it is an unknown parameter. The differences between the following *spherical estimation* techniques essentially reduce to how the unknown range is dealt with. These are covered in the following subsections.

13.5.1.2 Spherical Intersection Estimator

This method uses the physical constraint that the range R_{sk} is the Euclidean distance to the target.

- Writing $\hat{R}_{sk}^2 = \hat{\mathbf{x}}_k^T \hat{\mathbf{x}}_k$, it follows that:

$$\hat{R}_{sk}^2 = (\mathbf{b}_k - \hat{R}_{sk} \mathbf{d}_k)^T \mathbf{A}^{\dagger T} \mathbf{A}^\dagger (\mathbf{b}_k - \hat{R}_{sk} \mathbf{d}_k) \quad (13.51)$$

which can be written as the quadratic:

$$a \hat{R}_{sk}^2 + b \hat{R}_{sk} + c = 0 \quad (13.52)$$

where the individual terms follow through expanding Equation 13.51. These terms are given by:

$$a = 1 - \|\mathbf{A}^\dagger \mathbf{d}_k\|^2, \quad b = 2\mathbf{b}_k \mathbf{A}^{\dagger T} \mathbf{A}^\dagger \mathbf{d}_k, \quad \text{and} \quad c = -\|\mathbf{A}^\dagger \mathbf{b}_k\|^2 \quad (13.53)$$

- The unique, real, positive root of Equation 13.52 is taken as the SX estimator of the source range. Hence, the estimator will fail when:

1. there is no real, positive root, or:
2. if there are two positive real roots.

13.5.1.3 Spherical Interpolation Estimator

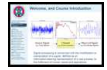
The SI estimator again uses the spherical LSE function, but rather than using the physically intuitive solution of *constraining* the target range relative to the origin to be the actual distance so that $R_{sk} \equiv |\mathbf{x}_k|$, it is estimated in the least-squares sense.

Consider again the **spherical error function**:

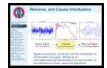
$$\boldsymbol{\epsilon}_{ik} = \mathbf{A} \mathbf{x}_k - (\mathbf{b}_k - R_{sk} \mathbf{d}_k) \quad (13.54)$$

Substituting the LSE from Equation 13.50 into this expression gives:

$$\boldsymbol{\epsilon}_{ik} = \mathbf{A} [\mathbf{A}^T \mathbf{A}]^{-1} \mathbf{A}^T (\mathbf{b}_k - \hat{R}_{sk} \mathbf{d}_k) - (\mathbf{b}_k - R_{sk} \mathbf{d}_k) \quad (13.55)$$



New slide



New slide

Defining the projection matrix as $\mathbf{P}_A = \mathbf{I}_N - \mathbf{A} [\mathbf{A}^T \mathbf{A}]^{-1} \mathbf{A}^T$, then this may be written as:

$$\epsilon_{ik} = R_{sk} \mathbf{P}_A \mathbf{d}_k - \mathbf{P}_A \mathbf{b}_k \quad (13.56)$$

Minimising the LSE using the normal equations gives:

$$R_{sk} = (\mathbf{d}_k^T \mathbf{P}_A \mathbf{P}_A \mathbf{d}_k)^{-1} \mathbf{d}_k^T \mathbf{P}_A \mathbf{P}_A \mathbf{b}_k \quad (13.57)$$

However, the **projection matrix** is symmetric and idempotent, such that $\mathbf{P}_A = \mathbf{P}_A^T$ and $\mathbf{P}_A \mathbf{P}_A = \mathbf{P}_A$. This means that the sum-of-squared errors simplifies to:

$$R_{sk} = (\mathbf{d}_k^T \mathbf{P}_A \mathbf{d}_k)^{-1} \mathbf{d}_k^T \mathbf{P}_A \mathbf{b}_k \quad (13.58)$$

or alternatively, since the quantity in the inverse is a scalar,

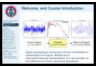
$$R_{sk} = \frac{\mathbf{d}_k^T \mathbf{P}_A \mathbf{b}_k}{\mathbf{d}_k^T \mathbf{P}_A \mathbf{d}_k} \quad (13.59)$$

Substituting back into the LSE for the target position given in Equation 13.50 gives the final estimator:

$$\hat{\mathbf{x}}_k = \mathbf{A}^\dagger \left(\mathbf{I}_N - \mathbf{d}_k \frac{\mathbf{d}_k^T \mathbf{P}_A}{\mathbf{d}_k^T \mathbf{P}_A \mathbf{d}_k} \right) \mathbf{b}_k \quad (13.60)$$

This approach is said to perform better, but is computationally slightly more complex than the SX estimator.

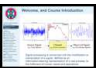
13.5.1.4 Other Approaches



There are several other approaches to minimising the spherical LSE function defined in *New slide* Equation 13.45.

- In particular, the **linear-correction** LSE solves the constrained minimization problem using Lagrange multipliers in a two stage process.
- For further information, see: Huang Y., J. Benesty, and J. Chen, “Time Delay Estimation and Source Localization,” in *Springer Handbook of Speech Processing* by J. Benesty, M. M. Sondhi, and Y. Huang, pp. 1043–1063, , Springer, 2008.

13.5.2 Hyperbolic Least Squares Error Function



KEYPOINT! (Underlying Concept). Suppose that for each pair of microphones i and j , a TDOA corresponding to source k is somehow estimated, and this is denoted by τ_{ijk} . One approach to ASL is to minimise the total error between the measured TDOAs and the TDOAs predicted by the geometry *given* an assumed target position. *New slide*

- If a TDOA is estimated between two microphones i and j , then the error between this and modelled TDOA is given by Equation 13.1:

$$\epsilon_{ij}(\mathbf{x}_k) = \tau_{ijk} - T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k) \quad (13.61)$$

where the error is considered as a function of the source position \mathbf{x}_k .

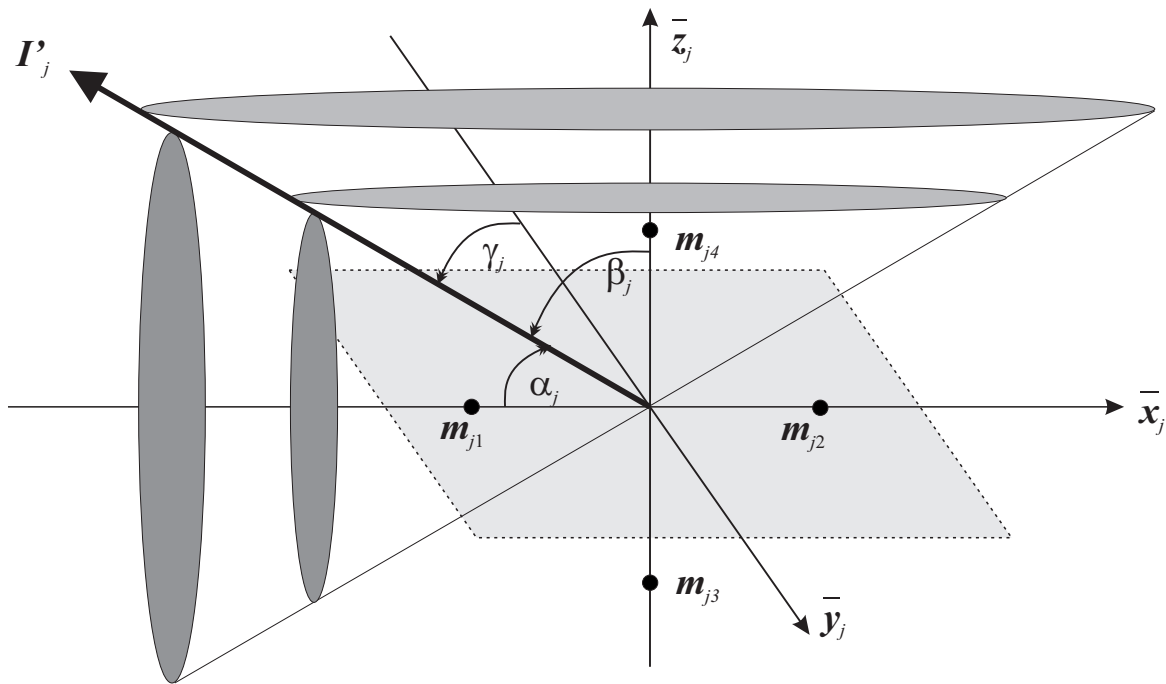


Figure 13.11: Quadruple sensor arrangement and local Cartesian coordinate system.

- The total error as a function of target position

$$J(\mathbf{x}_k) = \sum_{i=1}^N \sum_{j \neq i=1}^N (\tau_{ijk} - T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k))^2 \quad (13.62)$$

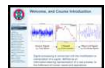
- Unfortunately, since $T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k)$ is a nonlinear function of \mathbf{x}_k , the minimum LSE does not possess a closed-form solution.

13.5.2.1 Linear Intersection Method

KEYPOINT! (Underlying Concept). The linear intersection (LI) algorithm works by utilising a *sensor quadruple* with a common midpoint, which allows a bearing line to be deduced from the intersection of two cones which approximate the hyperboloid. The spatial position that minimises the distance between these bearing lines at the point of nearest intersection is considered the target position.

- Given the bearing lines, it is possible to calculate the points \mathbf{s}_{ij} and \mathbf{s}_{ji} on two bearing lines which give the closest intersection as illustrated in Figure 13.12. This is basic geometry, and for a detailed analysis, see [Brandstein:1997].
- The trick is to note that given these points \mathbf{s}_{ij} and \mathbf{s}_{ji} , the theoretical TDOA, $T(\mathbf{m}_{1i}, \mathbf{m}_{2i}, \mathbf{s}_{ij})$, can be compared with the observed TDOA.

This will then lead to a weighted location estimate, where the weights are related to the likelihood of the target position given the observed TDOA.



New slide

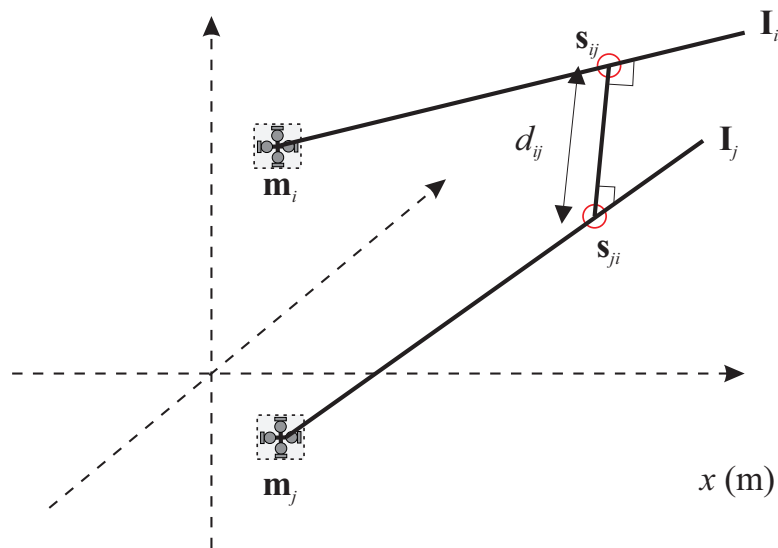


Figure 13.12: Calculating the points of closest intersection.

13.5.3 TDOA estimation methods

Two key methods for TDOA estimation are using the GCC function and the AED algorithm.

GCC algorithm most popular approach assuming an ideal free-field model. It has the advantages that

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

However, GCC-based methods

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

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

- adopts a reverberant rather than free-field model;
- computationally more expensive than GCC;
- can fail when there are common-zeros in the room impulse response (RIR).

Note that both methods assume that the signals received at the microphones arise as the result of a single source, and that if there are multiple sources, the signals will first need to be separated into different contributions of the individual sources.

13.5.3.1 GCC TDOA estimation

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

- The TDOA estimate between two microphones i and j is obtained as the time lag that maximises the cross-correlation between the filtered versions of the microphone outputs:

$$\hat{\tau}_{ij} = \arg \max_{\ell} r_{x_i x_j}[\ell] \quad (13.63)$$

where the signal received at microphone i is given by $x_i[n]$, and where x_i should not be confused with the location of the source k , which is denoted by $\mathbf{x}_k = [x_k, y_k, z_k]^T$.

- The cross-correlation function is given by

$$r_{x_i x_j}[\ell] = \mathcal{F}^{-1} (\Psi_{x_1 x_2} (e^{j\omega T_s})) \quad (13.64)$$

$$= \mathcal{F}^{-1} (\Phi (e^{j\omega T_s}) P_{x_1 x_2} (e^{j\omega T_s})) \quad (13.65)$$

$$= \int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}} \Psi_{x_1 x_2} (e^{j\omega T_s}) e^{j\ell\omega T} d\omega \quad (13.66)$$

$$= \int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}} \Phi (e^{j\omega T_s}) P_{x_1 x_2} (e^{j\omega T_s}) e^{j\ell\omega T} d\omega \quad (13.67)$$

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

$$P_{x_1 x_2} (e^{j\omega T_s}) = \mathbb{E} [X_1 (e^{j\omega T_s}) X_2 (e^{j\omega T_s})] \quad (13.68)$$

The CPSD can be estimated in a variety of means. The choice of the filtering term or frequency domain weighting function, $\Phi (e^{j\omega T_s})$, leads to a variety of different GCC methods for TDOA estimation. In Section 13.5.3.3, some of the popular approaches are listed, but only one is covered in detail, namely the phase transform (PHAT).

13.5.3.2 CPSD for Free-Field Model

For the free-field model in Equation 13.5 and Equation 13.6, it follows that for $i \neq j$ the CPSD in Equation 13.68 is given by:

$$P_{x_i x_j} (\omega) = \mathbb{E} [X_j (\omega) X_i (\omega)] \quad (13.69)$$

$$= \mathbb{E} [(\alpha_{ik} S_k (\omega) e^{-j\omega \tau_{ik}} + B_{ik} (\omega)) (\alpha_{jk} S_k (\omega) e^{-j\omega \tau_{jk}} + B_{jk} (\omega))] \quad (13.70)$$

$$= \alpha_{ik} \alpha_{jk} e^{-j\omega T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k)} \mathbb{E} [|S_k (\omega)|^2] \quad (13.71)$$

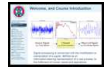
where $\mathbb{E} [B_{ik} (\omega) B_{jk} (\omega)] = 0$ and $\mathbb{E} [B_{ik} (\omega) S_k (\omega)] = 0$ due to the noise being uncorrelated with the source signal and noise signals.

- In particular, note that it follows:

$$\angle P_{x_i x_j} (\omega) = -j\omega T (\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k) \quad (13.72)$$

In other words, all the TDOA information is conveyed in the phase rather than the amplitude of the CPSD. This therefore suggests that the weighting function can be chosen to remove the amplitude information.

These equations can be converted to discrete time as appropriate.



New slide

13.5.3.3 GCC Processors

The most common choices for the GCC weighting term are listed in the table below. In particular, the PHAT is considered in detail.

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

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

$$\gamma_{x_1x_2}(e^{j\omega T_s}) = \frac{P_{x_1x_2}(e^{j\omega T_s})}{\sqrt{P_{x_1x_1}(e^{j\omega T_s}) P_{x_2x_2}(e^{j\omega T_s})}} \quad (13.73)$$

The PHAT-GCC approach can be written as:

$$r_{x_i x_j}[\ell] = \int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}} \Phi(e^{j\omega T_s}) P_{x_1x_2}(e^{j\omega T_s}) e^{j\ell\omega T} d\omega \quad (13.74)$$

$$= \int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}} \frac{1}{|P_{x_1x_2}(e^{j\omega T_s})|} |P_{x_1x_2}(e^{j\omega T_s})| e^{j\angle P_{x_1x_2}(e^{j\omega T_s})} e^{j\ell\omega T} d\omega \quad (13.75)$$

$$= \int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}} e^{j(\ell\omega T + \angle P_{x_1x_2}(e^{j\omega T_s}))} d\omega \quad (13.76)$$

$$= \delta(\ell T_s + \angle P_{x_1x_2}(e^{j\omega T_s})) \quad (13.77)$$

$$= \delta(\ell T_s - T(\mathbf{m}_i, \mathbf{m}_j, \mathbf{x}_k)) \quad (13.78)$$

- In the absence of reverberation, the GCC-PHAT (GCC-PHAT) algorithm gives an impulse at a lag given by the TDOA divided by the sampling period.

13.5.3.4 Adaptive Eigenvalue Decomposition

KEYPOINT! (Underlying Concept). The AED algorithm adopts the real reverberant rather than free-field model. The AED algorithm actually amounts to a **blind channel identification** problem, which then seeks to identify the channel coefficients corresponding to the direct path elements.

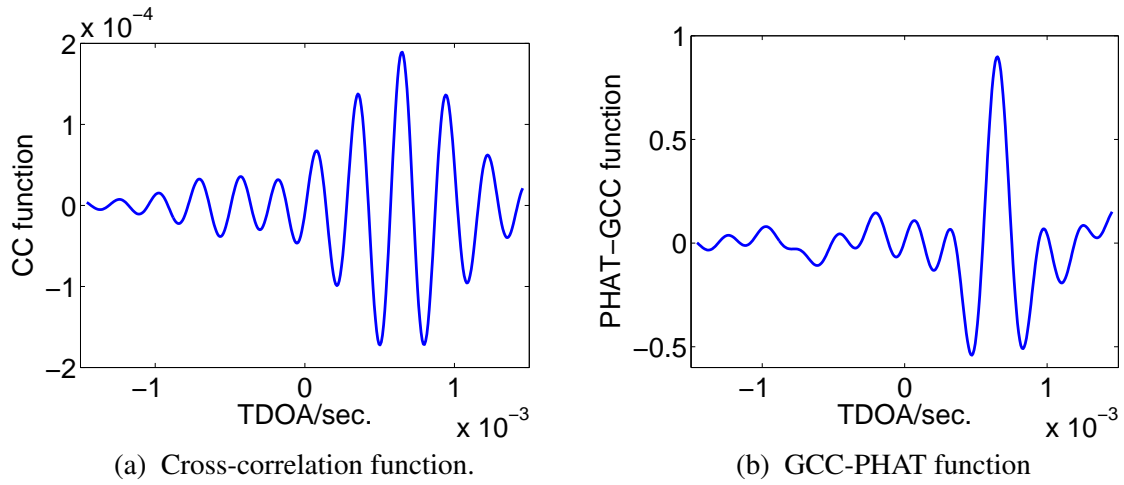


Figure 13.13: Normal cross-correlation and GCC-PHAT functions for a frame of speech.

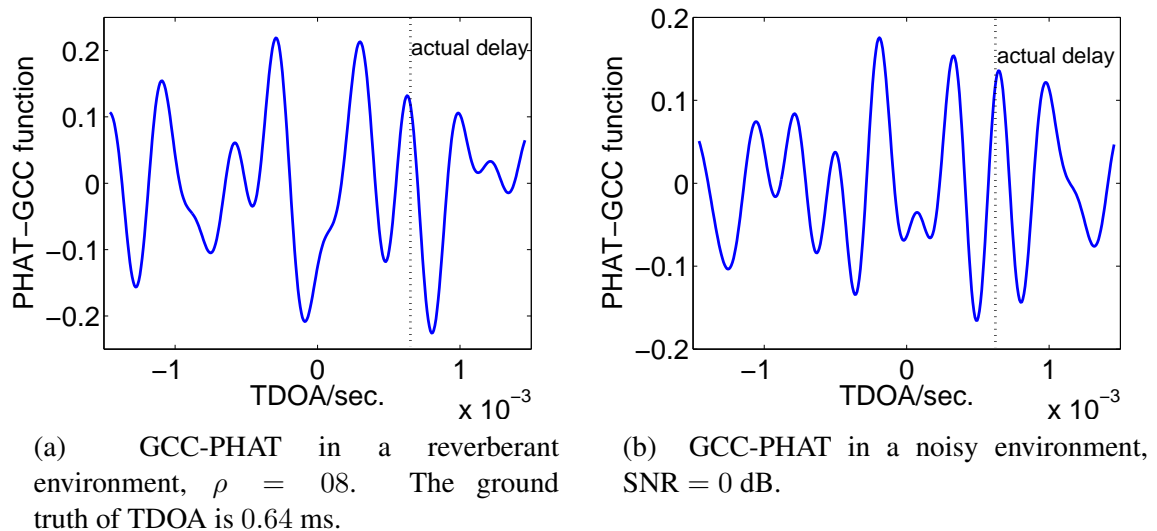


Figure 13.14: The effect of reverberation and noise on the GCC-PHAT can lead to poor TDOA estimates.

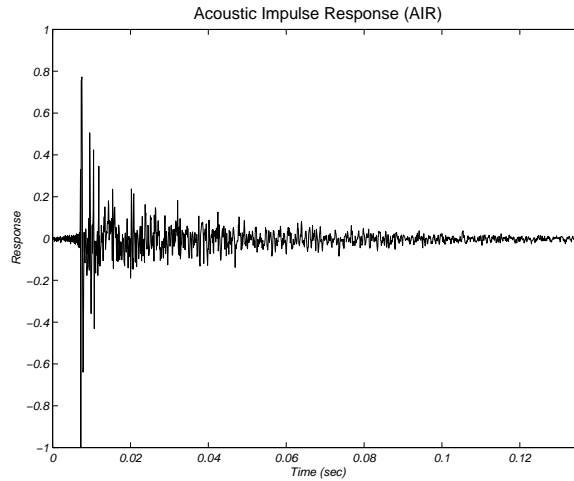


Figure 13.15: A typical room acoustic impulse response.

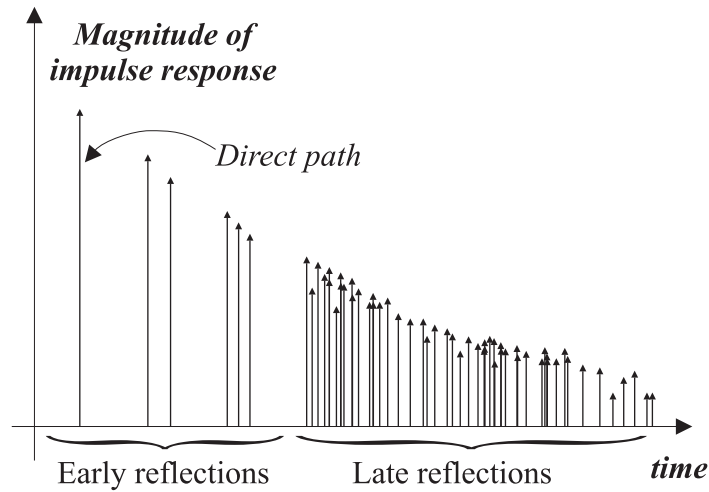


Figure 13.16: Early and late reflections in an AIR.

- Suppose that the acoustic impulse response (AIR) between source k and i is given by $h_{ik}[n]$ such that

$$x_{ik}[n] = \sum_{m=-\infty}^{\infty} h_{ik}[n-m] s_k[m] + b_{ik}[n] \quad (13.79)$$

then the TDOA between microphones i and j is:

$$\tau_{ijk} = \left\{ \arg \max_{\ell} |h_{ik}[\ell]| \right\} - \left\{ \arg \max_{\ell} |h_{jk}[\ell]| \right\} \quad (13.80)$$

This assumes a minimum-phase system, but can easily be made robust to a non-minimum-phase system.

- Reverberation plays a major role in ASL and BSS.
- Consider reverberation as the sum total of all sound reflections arriving at a certain point in a room after room has been excited by impulse.

Trivia: Perceive early reflections to reinforce direct sound, and can help with speech intelligibility. It can be easier to hold a conversation in a closed room than outdoors

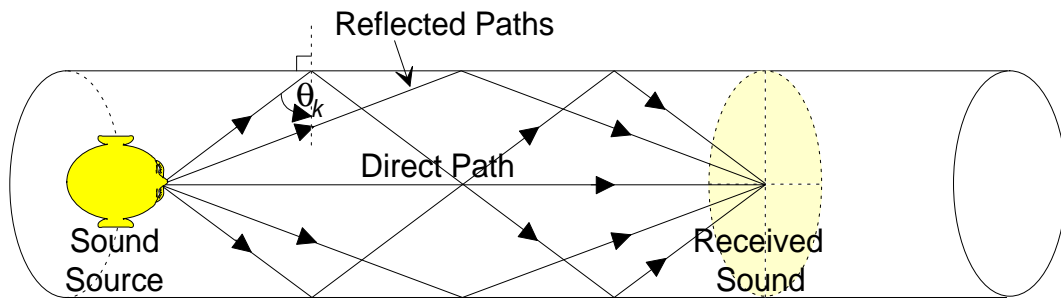


Figure 13.17: In an infinitely long cylindrical tube, the reverberant energy is greater than the energy contained in the sound travelling along a direct path, thus demonstrating the nonminimum-phase properties of room acoustics.

- Room transfer functions are often nonminimum-phase since there is more energy in the reverberant component of the RIR than in the component corresponding to sound travelling along a direct path.
- Therefore AED will need to consider multiple peaks in the estimated AIR.

13.6 Direct Localisation Methods

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

13.6.1 Steered Response Power Function

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

The frequency domain **delay-and-sum beamformer** steered to a spatial position $\hat{\mathbf{x}}_k$ such that $\hat{\tau}_{pk} = |\hat{\mathbf{x}} - \mathbf{m}_p|$, using the notation in Equation 13.8, is given by:

$$S(\hat{\mathbf{x}}) = \int_{\Omega} \left| \sum_{p=1}^N W_p(e^{j\omega T_s}) X_p(e^{j\omega T_s}) e^{j\omega \hat{\tau}_{pk}} \right|^2 d\omega \quad (13.81)$$

Expanding and rearranging the order of integration and summation gives:

$$S(\hat{\mathbf{x}}) = \int_{\Omega} \sum_{p=1}^N \sum_{q=1}^N W_p(e^{j\omega T_s}) W_q^*(e^{j\omega T_s}) X_p(e^{j\omega T_s}) X_q^*(e^{j\omega T_s}) e^{j\omega(\hat{\tau}_{pk} - \hat{\tau}_{qk})} d\omega \quad (13.82)$$

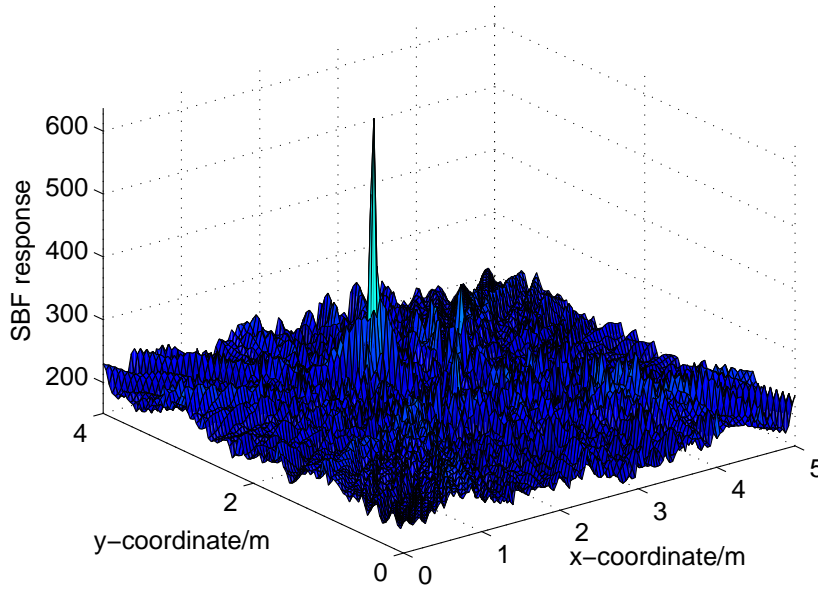


Figure 13.18: SBF response from a frame of speech signal. The integration frequency range is 300 to 3500 Hz (see Equation 13.84). The true source position is at $[2.0, 2.5]m$. The grid density is set to 40 mm.

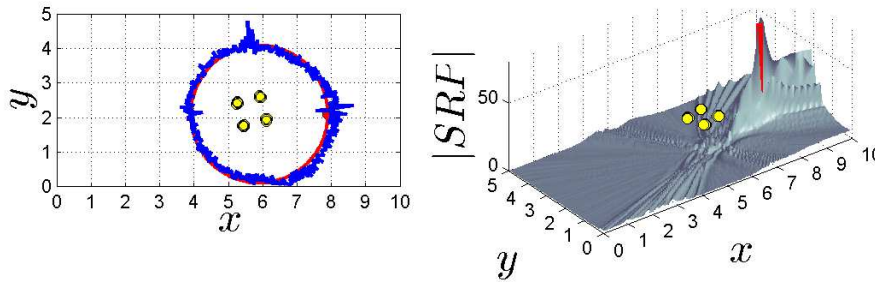


Figure 13.19: An example video showing the SBF changing as the source location moves.

Taking expectations of both sides and setting $\Phi_{pq}(e^{j\omega T_s}) = W_p(e^{j\omega T_s}) W_q^*(e^{j\omega T_s})$ gives

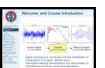
$$\mathbb{E}[S(\hat{\mathbf{x}})] = \sum_{p=1}^N \sum_{q=1}^N \int_{\Omega} \Phi_{pq}(e^{j\omega T_s}) P_{x_p x_q}(e^{j\omega T_s}) e^{j\omega \hat{\tau}_{pqk}} d\omega \quad (13.83)$$

$$= \sum_{p=1}^N \sum_{q=1}^N r_{x_i x_j}[\hat{\tau}_{pqk}] \equiv \sum_{p=1}^N \sum_{q=1}^N r_{x_i x_j} \left[\frac{|\mathbf{x}_k - \mathbf{m}_i| - |\mathbf{x}_k - \mathbf{m}_j|}{c} \right] \quad (13.84)$$

In other words, the SRP is the sum of all possible pairwise GCC functions evaluated at the time delays hypothesised by the target position. This is discussed in Section 13.6.2.

13.6.2 Conceptual Intepretation of SRP

Equation 13.84 gives an elegant conceptual intepretation of the SBF function. Given a candidate spatial position $\hat{\mathbf{x}}_k$, the corresponding TDOA at microphones i and j can be calculated using Equation 13.9:



New slide

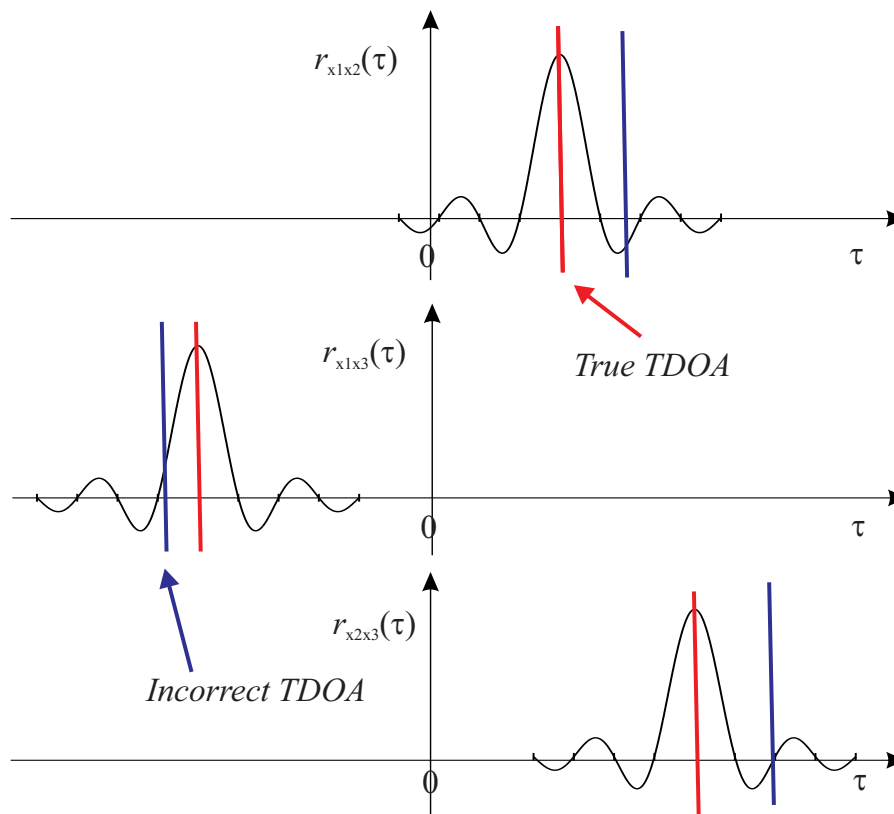


Figure 13.20: GCC-PHAT for different microphone pairs.

$$T(\mathbf{m}_i, \mathbf{m}_j, \hat{\mathbf{x}}_k) = \frac{|\hat{\mathbf{x}}_k - \mathbf{m}_i| - |\hat{\mathbf{x}}_k - \mathbf{m}_j|}{c} \quad (13.85)$$

Since the SBF function in Equation 13.84 is a linear combination of the GCC-PHAT functions, then if $\hat{\mathbf{x}}_k$ is correct, then the GCC-PHAT functions should return a large peak. If $\hat{\mathbf{x}}_k$ is incorrect, then the GCC-PHAT functions return smaller values, and therefore the SBF function in Equation 13.84 is smaller.

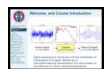
13.7 DUET Algorithm

KEYPOINT! (Summary). The DUET algorithm is an approach to BSS that ties in neatly to ASL. Under certain assumptions and circumstances, it is possible to separate more than two sources using only two microphones.

- DUET is based on the assumption that for a set of signals $x_k[t]$, their time-frequency representations (TFRs) are predominately non-overlapping. This condition is referred to as W-disjoint orthogonality (WDO), and can be stated as follows:

$$S_p(\omega, t) S_q(\omega, t) = 0 \quad \forall p \neq q, \forall t, \omega \quad (13.86)$$

The WDO property is clearly shown in Figure 13.21, where the spectrograms of *clean* speech mixtures are sparse and disjoint. For two speech signals, the product of the corresponding spectrograms is zero at the most area on the time-frequency (TF) domain.



New slide

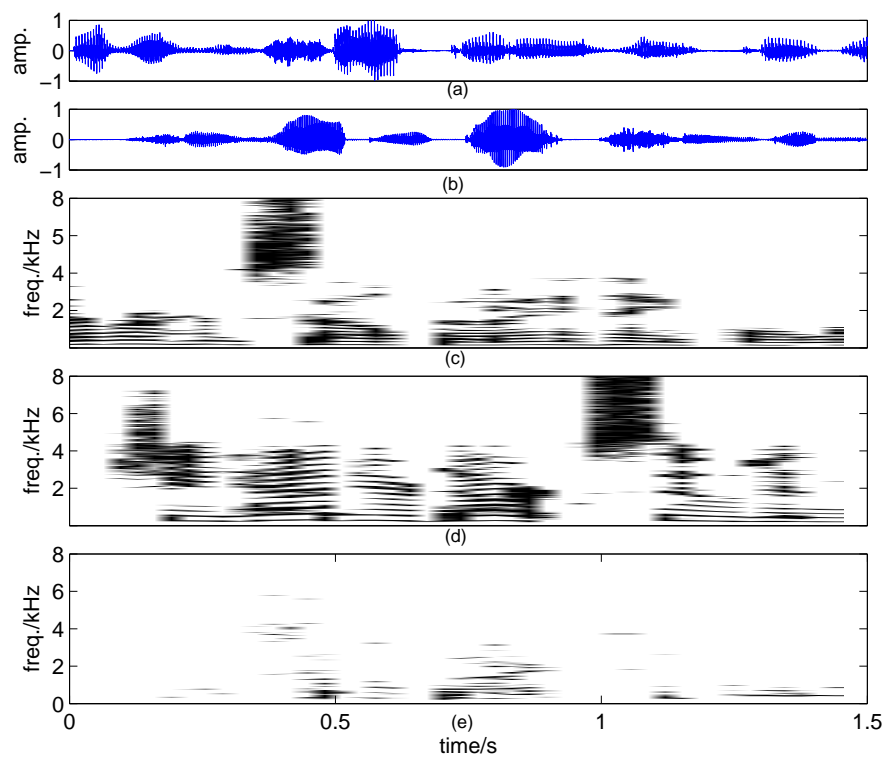


Figure 13.21: W-disjoint orthogonality of two speech signals. Original speech signal (a) $s_1[t]$ and (b) $s_2[t]$; corresponding STFTs (c) $|S_1(\omega, t)|$ and (d) $|S_2(\omega, t)|$; (e) product of the two spectrogram $|S_1(\omega, t) S_2(\omega, t)|$.

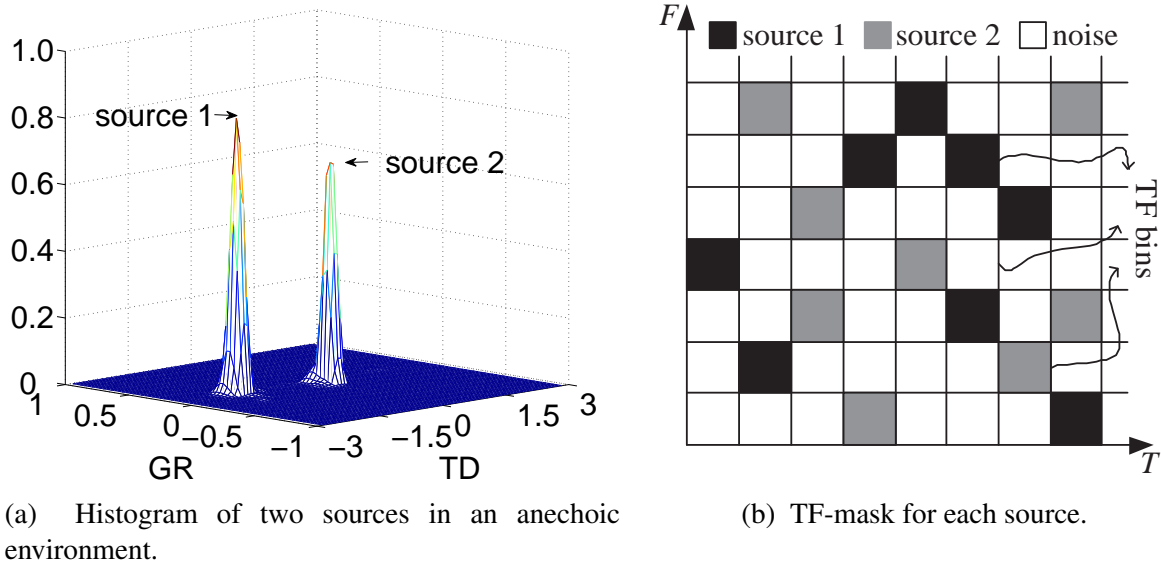


Figure 13.22: Illustration of the underlying idea in DUET.

Consider taking then, a particular TF-bin, (ω, t) , where source p is known to be active. The two received signals at microphones i and j in *that* TF-bin can be written in the TF-domain as:

$$\begin{aligned} X_{ip}(\omega, t) &= \alpha_{ip} e^{-j\omega\tau_{ip}} S_p(\omega, t) + B_i(\omega, t) \\ X_{jp}(\omega, t) &= \alpha_{jp} e^{-j\omega\tau_{jp}} S_p(\omega, t) + B_j(\omega, t) \end{aligned} \quad (13.87)$$

Taking the ratio of these expressions and ignoring the noise terms gives:

$$H_{ikp}(\omega, t) \triangleq \frac{X_{ip}(\omega, t)}{X_{jp}(\omega, t)} = \frac{\alpha_{ip}}{\alpha_{jp}} e^{-j\omega\tau_{ijp}} \quad (13.88)$$

where, again, τ_{ijp} is the TDOA of the signal contribution due to source p between microphones i and j .

KEYPOINT! (Which TF-bins belong to which source?). Of course, which TF-bins belong to which source is unknown, as the source signal and spectrum is unknown. However, if the magnitude and phase terms of the ratio in Equation 13.88 are *histogrammed* over all TF-bins, peaks will occur a distinct magnitude-phase positions, each peak corresponding to a different source.

Hence,

$$\tau_{ijp} = -\frac{1}{\omega} \arg H_{ikp}(\omega, t), \quad \text{and} \quad \frac{\alpha_{ip}}{\alpha_{jp}} = |H_{ikp}(\omega, t)| \quad (13.89)$$

This leads to the essentials of the DUET method which are:

1. Construct the TF representation of both mixtures.
2. Take the ratio of the two mixtures and extract local mixing parameter estimates.
3. Combine the set of local mixing parameter estimates into N pairings corresponding to the true mixing parameter pairings.

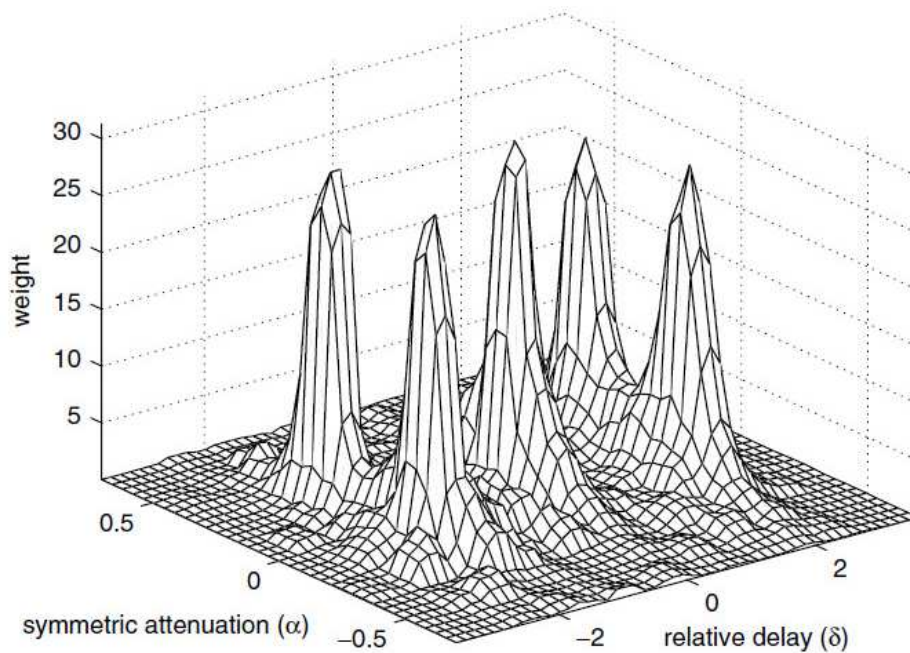


Figure 13.23: DUET for multiple sources.

4. Generate one binary mask for each determined mixing parameter pair corresponding to the TF-bins which yield that particular mixing parameter pair.
5. Demix the sources by multiplying each mask with one of the mixtures.
6. Return each demixed TFR to the time domain.

13.7.1 Effect of Reverberation and Noise

A number of papers have analysed the validity of the WDO property, and anechoic speech often satisfies this. However, while the TFR of speech is very clear in this case, the TFR becomes smeared due to reverberation and noise.

13.7.2 Estimating multiple targets

The underlying idea is shown in Figure 13.25 and Figure 13.26.

13.8 Further Topics

- Reduction in complexity of calculating SRP. This includes SRC and hierarchical searches.
- Multiple-target tracking (see Daniel Clark's Notes)
- Simultaneous (self-)localisation and tracking; estimating sensor and target positions from a moving source.

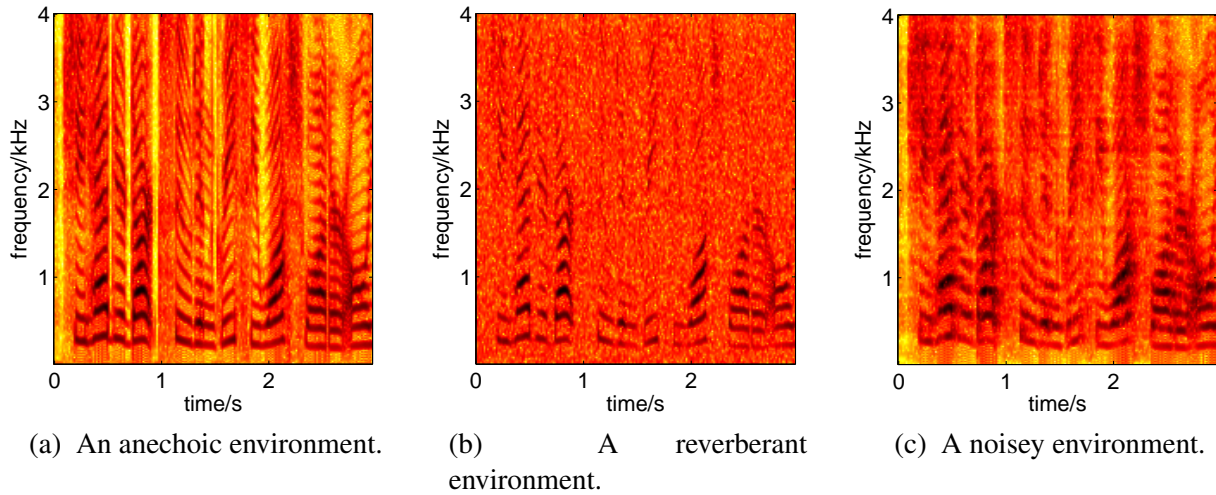


Figure 13.24: The TFR is very clear in the anechoic environment but smeared around by the reverberation and noise.

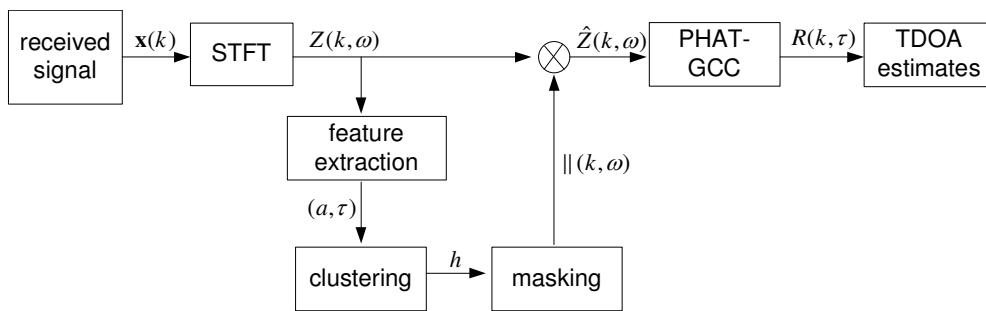


Figure 13.25: Flow diagram of the DUET-GCC approach. Basically, the speech mixtures are separated by using the DUET in the TF domain, and the PHAT-GCC is then employed for the spectrogram of each source to estimate the TDOAs.

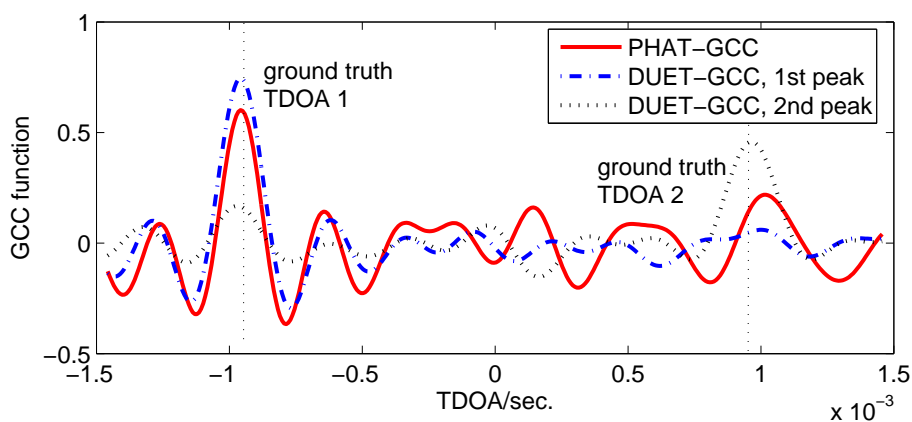


Figure 13.26: GCC function from DUET approach and traditional PHAT weighting. Two sources are located at (1.4, 1.2)m and (1.4, 2.8)m respectively. The GCC function is estimated from the first microphone pair (microphone 1 and microphone 2). The ground truth TDOAs are 0.95 ms.

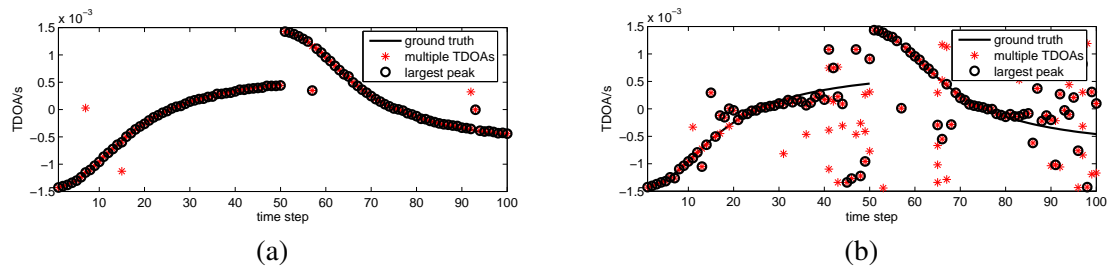


Figure 13.27: Acoustic source tracking and localisation.

- Joint ASL and BSS.
- Explicit signal and channel modelling! (None of the material so forth cares whether the signal is speech or music!)
- Application areas such as gunshot localisation; other sensor modalities; diarisation.