The Mathematics of Medical Imaging

Charles L. Epstein

November 6, 2001

Charles L. Epstein
Department of Mathematics
University of Pennsylvania
Philadelphia, PA 19104
cle@math.upenn.edu

This book is dedicated to my wife, Jane and our children, Leo and Sylvia. They make it all worthwhile.

Preface

Over the past several decades advanced mathematics has quietly insinuated itself into many facets of our day to day life. Mathematics is at the heart of technologies from cellular telephones and satellite positioning systems to online banking and metal detectors. Arguably no technology has had a more positive and profound effect on our lives than advances in medical imaging and in no technology is the role of mathematics more pronounced. X-ray tomography, ultrasound, positron emission tomography and magnetic resonance imaging have fundamentally altered the practice of medicine. At the core of each modality is a mathematical model to interpret the measurements and a numerical algorithm to reconstruct an image. While each modality operates on a different physical principle and probes a different aspect of our anatomy or physiology, there is a large overlap in the mathematics used to model the measurements, design reconstruction algorithms and analyze the effects of noise. In this text we provide a tool kit, with detailed operating instructions, to work on the sorts of mathematical problems which arise in medical imaging. Our treatment steers a course midway between a complete, rigorous, abstract mathematical discussion and a cookbook engineering approach.

The target audience for this book is junior or senior math undergraduates with a firm command of calculus and linear algebra. The book is written in the language of mathematics, which, as I have learned is quite distinct from the language of physics or the language of engineering. Nonetheless, the discussion of every topic begins at an elementary level and the book should, with a little translation, be usable by advanced science and engineering students with a good mathematical background. A large part of the background material is surveyed in two appendices. Our presentation of these topics is non-standard, situating them in the context of measurement and practical computation. This is a book for people who like to know a little more than they have to know.

Our emphasis is squarely on mathematical concepts; only the particulars of X-ray tomography are discussed in any detail. X-ray tomography is employed as a *pedagogical machine*, similar in spirit to the elaborate devices used to illustrate the principles of Newtonian mechanics. The *physical principles* at work in X-ray tomography are simple to describe and require little formal background in physics to understand. This is not the case in any of the other modalities described above or in less developed modalities like infrared imaging and impedence tomography. The *mathematical* problems that arise in X-ray tomography and the tools used to solve them have a great deal in common with those used in the other imaging modalities. This is why our title is "The Mathematics of Medical Imaging" instead of "The Mathematics of X-ray tomography."A student with a thorough understanding of the material in this book should be well prepared, mathematically for further investigations in any subfield of medical imaging. Very good treatments of the

physical principles underlying the other modalities can be found in *Radiological Imaging* by Harrison H. Barrett and William Swindell, [4], *Principles of Computerized Tomographic Imaging* by Avinash C. Kak and Malcolm Slaney, [39], *Foundations of Medical Imaging* by Cho, Jones, Singh, [86], *Image reconstruction from Projections* by Gabor T. Herman, [24] and *Magnetic Resonance Imaging* by E. Mark Haacke, Robert W. Brown, Michael R. Thompson, Ramesh Venkatesan, [80]. Indeed these books were invaluable sources as I learned the subject myself. My treatment of many topics doubtless owes a great deal to these books. Graduate level treatments of the mathematics and algorithms can be found in *The Mathematics of Computerized Tomography* by Frank Natterer, [50] and *Mathematical Methods in Image Reconstruction* by Frank Natterer and Frank Wübbelling, [51].

The book begins with an introduction to the idea of using a mathematical model as a tool to extract the physical state of system from feasible measurements. In medical imaging, the "state of the system" in question is the anatomy and physiology of a *living* human being. To probe it non-destructively requires considerable ingenuity and sophisticated mathematics. After considering a variety of examples, each a toy problem for some aspect of medical imaging we turn to a description of X-ray tomography. This leads us to our first mathematical topic, *integral transforms*. The transform of immediate interest is the Radon transform, though we are quickly led to the Abel transform, Hilbert transform and the Fourier transform. Our study of the Fourier transform is dictated by the applications we have in mind, with a strong emphasis on the connection between the smoothness of a function and the decay of its Fourier transform and vice versa. Many of the basic ideas of functional analysis appear as we consider these examples. The concept of a weak derivative, which is ubiquitous in the engineering literature and essential to a precise understanding of the Radon inversion formula is described in detail. This part of the book culminates in a study of the Radon inversion formula. A major theme in these chapters is the difference between finite and infinite dimensional linear algebra.

The next topics we consider are sampling and filtering theory. These form the basis for applying the mathematics of the Fourier transform to real world problems. In the chapter on sampling theory we discuss the Nyquist theorem, the Shannon-Whittaker interpolation formula, the Poisson summation formula and the consequences of undersampling. In the chapter on filtering theory we recast Fourier analysis as a tool for image and signal processing. We then discuss the mathematics of approximating continuous time, linear shift invariant filters on finitely sampled data, using the finite Fourier transform. The chapter concludes with an overview of image processing and a linear systems analysis of some basic imaging hardware.

In Chapter eight these tools are applied to the problem of image reconstruction in X-ray tomography. Most of the chapter is devoted to the filtered backprojection algorithm, though other methods are briefly considered. After deriving the reconstruction algorithms we analyze the point spread function and modulation transfer function of the full measurement and reconstruction process. We also use this formalism to analyze a variety of imaging artifacts. Chapter nine contains a brief description of "algebraic reconstruction techniques," which are essentially methods for solving large, sparse systems of linear equations

The final topic is noise in the backprojection algorithm. This part of the book begins with an introduction to probability theory. Our presentation uses the ideas of measure theory, though in a metaphoric rather than a technical way. The chapter concludes with a study of specific probability distributions that are important in imaging. The next chapter

introduces the ideas of random processes and their role in signal and image processing. Again the focus is on those processes needed to analyze noise in X-ray imaging. A student with a good grasp of Riemann integration should not have difficulty with the material in these chapters. In the final chapter we study the effects of noise on the image reconstruction process. This chapter culminates with the famous resolution-dosage fourth power relation, which shows that to double the resolution in a CT-image, keeping the SNR constant, the radiation dosage must be increased by a factor of 16!

Each chapter builds from elementary material to more advanced material as is also the case with the longer sections. The elementary material in each chapter (or section) depends only on the elementary material in previous chapters. Sections which cover noticeably more advanced material, with more prerequisites, are marked with an asterisk. Several of these sections assume a familiarity with the elementary parts of the theory of function of a complex variable. All can be omitted without any loss in continuity. Many sections begin with a box containing a list of sections in the appendices, these are recommended background readings. A one semester course in the mathematics of medical imaging can be fashioned from Chapter 1, 2.1-2.4 (omitting the$*$-sections), 2.5 (up to 2.5.1), 3.1, 3.2 (up to 3.2.9), 3.3 (up to 3.3.2), 4.1-4.5 (omitting the $*$-sections), 5.1-5.5 (omitting the$*$-sections), 6.1-6.2.2, 6.3, 7.1 (up to 7.1.7), 7.3, 7.5, 8.1-8.5 (omitting the $*$-sections). Though given the diversity of students interested in this field the subject matter must be tailored, each semester to the suit the needs of those actually sitting in the room. Exercises are collected at the ends of the sections and sub-sections. Most develop ideas presented in the text, only a few are of a standard, computational character.

Acknowledgments

Perhaps the best reward for writing a book of this type is the opportunity it affords for thanking the many people who contributed to it in one way or another. There are a lot of people to thank and I address them in roughly chronological order.

First I would like to thank my parents, Jean and Herbert Epstein, for their encouragement to follow my dreams and the very high standards they set for me from earliest childhood. I would also like to thank my father and Robert M. Goodman for giving me the idea that, through careful thought and observation, the world can be understood and the importance of expressing ideas simply but carefully. My years as an undergraduate at MIT not only provided me with a solid background in mathematics, physics and engineering but also reinforced my belief in the unity of scientific enquiry. I am especially grateful for the time and attention Jerry Lettvin lavished on me. My interest in the intricacies of physical measurement surely grew out of our many conversations. I was fortunate to be a graduate student at the Courant Institute, one of the few places where mathematics and its applications lived together in harmony. In both word and deed, my thesis advisor, Peter Lax placed mathematics and its applications on an absolutely equal footing. It was a privilege to be his student. I am very grateful for the enthusiasm that he and his late wife, Anneli showed for turning my lecture notes into a book.

Coming closer to the present day, I would like to thank Dennis Deturck for his unflagging support, both material (in the form of NSF grant DUE95-52464) and emotional for the development of my course on medical imaging and this book. The first version of these notes were transcribed from my lectures in the spring of 1999 by Hyunsuk Kang. Without her hard work it is very unlikely I would ever have embarked on this project. I am very grateful to my

colleagues in the Radiology department Gabor Herman, Peter Joseph and Felix Wehrli for sharing with me their profound, first hand knowledge of medical imaging. Gabor Herman's computer program, SNARK93 was used to make the simulated reconstructions in this book; Peter Joseph and Felix Wehrli provided many other images. I am most appreciative for the X-ray spectrum provided by Dr. Andrew Karellas (figure 2.7). I would like to thank John D'Angelo and Phil Nelson for their help with typesetting, publishing and the writing process itself and Fred Villars for sharing with me his insights on medicine, imaging and a host of other topics. The confidence my editor, George Lobell, expressed in the importance of this project was an enormous help in the long months it took to finish it.

Finally I would like to thank my wife, Jane and our children, Leo and Sylvia for their love, constant support and daily encouragement through the many, moody months. Without them I would have given up a long time ago.

Charles L. Epstein
Philadelphia, PA
October 17, 2001

Contents

Preface **v**

1 Measurements and modeling **1**
1.1 Mathematical modeling . . . 3
1.1.1 Finitely many degrees of freedom . . . 3
1.1.2 Infinitely many degrees of freedom . . . 7
1.2 A simple model problem for image reconstruction . . . 13
1.2.1 The space of lines in the plane . . . 14
1.2.2 Reconstructing an object from its shadows . . . 15
1.2.3 Approximate reconstructions . . . 18
1.2.4 Can an object be reconstructed from its width? . . . 20
1.3 Linearity . . . 21
1.3.1 Solving linear equations . . . 23
1.3.2 Infinite dimensional linear algebra . . . 28
1.4 Conclusion . . . 31

2 A basic model for tomography **33**
2.1 Tomography . . . 33
2.1.1 Beer's law and X-ray tomography . . . 36
2.2 Analysis of a point source device . . . 41
2.3 Some physical considerations . . . 44
2.4 The definition of the Radon transform . . . 46
2.4.1 Appendix: Proof of Lemma 2.4.1* . . . 51
2.4.2 Continuity of the Radon transform* . . . 52
2.4.3 The backprojection formula . . . 55
2.5 The Radon transform of a radially symmetric function . . . 56
2.5.1 The range of the radial Radon transform* . . . 57
2.5.2 The Abel transform* . . . 59
2.5.3 Fractional derivatives* . . . 61
2.5.4 Volterra equations of the first kind* . . . 62

3 Introduction to the Fourier transform **67**
3.1 The complex exponential function. . . . 67
3.2 Functions of a single variable . . . 68
3.2.1 Absolutely integrable functions . . . 69

3.2.2 Appendix: The Fourier transform of a Gaussian* . . . 72
3.2.3 Regularity and decay . . . 73
3.2.4 Fourier transform on $L^2(\mathbb{R})$ . . . 79
3.2.5 Basic properties of the Fourier Transform on $\mathbb{R}$ . . . 83
3.2.6 Convolution . . . 84
3.2.7 Convolution equations . . . 90
3.2.8 The δ-function . . . 92
3.2.9 Windowing and resolution . . . 94
3.2.10 Functions with L^2-derivatives* . . . 97
3.2.11 Fractional derivatives and L^2-derivatives* . . . 99
3.2.12 Some refined properties of the Fourier transform* . . . 101
3.2.13 The Fourier transform of generalized functions* . . . 105
3.2.14 The Paley-Wiener theorem* . . . 111
3.3 Functions of several variables . . . 112
3.3.1 L^1-case . . . 113
3.3.2 Regularity and decay . . . 116
3.3.3 L^2-theory . . . 119
3.3.4 Basic properties of the Fourier Transform on $\mathbb{R}^n$ . . . 121
3.3.5 Convolution . . . 122
3.3.6 The support of $f * g$. . . . 125
3.3.7 L^2-derivatives* . . . 126
3.3.8 The failure of localization in higher dimensions* . . . 130

4 The Radon transform 131
4.1 The Radon transform . . . 131
4.2 Inversion of the Radon Transform . . . 135
4.2.1 The Central slice theorem . . . 135
4.2.2 The Radon Inversion Formula . . . 138
4.2.3 Backprojection* . . . 141
4.2.4 Filtered Backprojection . . . 143
4.2.5 Inverting the Radon transform, two examples . . . 145
4.2.6 An alternate formula for the Radon inverse* . . . 148
4.3 The Hilbert transform . . . 149
4.3.1 Mapping properties of the Hilbert transform* . . . 154
4.4 Approximate inverses for the Radon transform . . . 154
4.4.1 Addendum* . . . 156
4.5 The range of the Radon transform . . . 157
4.5.1 Data with bounded support . . . 158
4.5.2 More general data* . . . 160
4.6 Continuity of the Radon transform and its inverse . . . 164
4.6.1 Bounded support . . . 164
4.6.2 Estimates for the inverse transform* . . . 166
4.7 The higher dimensional Radon transform* . . . 170
4.8 The Hilbert transform and complex analysis* . . . 173

5 Introduction to Fourier series **177**
5.1 Fourier series in one dimension . . . 177
5.2 Decay of Fourier coefficients . . . 183
5.3 L^2-theory . . . 186
5.3.1 Geometry in $L^2([0,1])$. . . . 186
5.3.2 Bessel's inequality . . . 191
5.3.3 L^2-derivatives* . . . 193
5.4 General periodic functions . . . 196
5.4.1 Convolution and partial sums . . . 197
5.4.2 Dirichlet kernel . . . 199
5.5 The Gibbs Phenomenon . . . 200
5.5.1 The general Gibbs phenomenon . . . 204
5.5.2 Fejer means . . . 206
5.5.3 Resolution . . . 209
5.6 The localization principle* . . . 211
5.7 Higher dimensional Fourier series . . . 213
5.7.1 L^2-theory . . . 216

6 Sampling **219**
6.1 Sampling and Nyquist's theorem . . . 220
6.1.1 Nyquist's theorem . . . 220
6.1.2 Shannon-Whittaker Interpolation . . . 222
6.2 The Poisson Summation Formula . . . 225
6.2.1 The Poisson summation formula . . . 225
6.2.2 Undersampling and aliasing . . . 228
6.2.3 Sub-sampling . . . 234
6.2.4 Sampling periodic functions . . . 235
6.2.5 Quantization errors . . . 237
6.3 Higher dimensional sampling . . . 239

7 Filters **243**
7.1 Basic definitions . . . 243
7.1.1 Examples of filters . . . 244
7.1.2 Linear filters . . . 246
7.1.3 Shift invariant filters . . . 248
7.1.4 Harmonic components . . . 249
7.1.5 The transfer function . . . 252
7.1.6 The δ-function revisited . . . 254
7.1.7 Causal filters . . . 256
7.1.8 Bandpass filters . . . 257
7.1.9 Resolution* . . . 260
7.1.10 Cascaded filters . . . 263
7.1.11 The resolution of a cascade of filters* . . . 265
7.1.12 Filters and RLC-circuits* . . . 267
7.2 Filtering periodic signals . . . 275
7.2.1 Resolution of periodic filters* . . . 278

7.2.2 The comb filter and Poisson summation . . . 279
7.3 The inverse filter . . . 281
7.4 Higher dimensional filters . . . 285
7.5 Implementing shift invariant filters . . . 290
7.5.1 Sampled data . . . 291
7.5.2 The finite Fourier transform . . . 293
7.5.3 Approximation of Fourier coefficients . . . 295
7.5.4 Implementing periodic convolutions on sampled data . . . 297
7.5.5 Implementing filters on finitely sampled data . . . 298
7.5.6 Zero padding reconsidered . . . 301
7.5.7 Higher dimensional filters . . . 302
7.5.8 Appendix: The Fast Fourier Transform . . . 306
7.6 Image processing . . . 308
7.6.1 Basic concepts and operations . . . 309
7.6.2 Discretized images . . . 318
7.7 General linear filters* . . . 323
7.8 Linear filter analysis of imaging hardware* . . . 324
7.8.1 The transfer function of the scanner . . . 325
7.8.2 The resolution of an imaging system . . . 329
7.8.3 Collimators . . . 331

8 Reconstruction in X-ray tomography 337
8.1 Reconstruction formulæ . . . 340
8.2 Scanner geometries . . . 342
8.3 Reconstruction algorithms for a parallel beam machine . . . 347
8.3.1 Direct Fourier inversion . . . 347
8.3.2 Filtered backprojection . . . 348
8.3.3 Ram-Lak filters . . . 350
8.3.4 Shepp-Logan analysis of the Ram-Lak filters . . . 352
8.3.5 Sample spacing in a parallel beam machine . . . 356
8.4 Filtered backprojection in the fan-beam case . . . 358
8.4.1 Fan beam geometry . . . 358
8.4.2 Fan beam filtered backprojection . . . 361
8.4.3 Implementing the fan beam algorithm . . . 363
8.4.4 Data collection for a fan beam scanner . . . 364
8.4.5 Rebinning . . . 366
8.5 The effect of a finite width X-ray beam . . . 366
8.5.1 A non-linear effect . . . 368
8.5.2 The partial volume effect . . . 369
8.5.3 Some mathematical remarks* . . . 371
8.6 The PSF . . . 373
8.6.1 The PSF without sampling . . . 374
8.6.2 The PSF with sampling . . . 381
8.6.3 View sampling . . . 385
8.6.4 Bad rays versus bad views . . . 393
8.6.5 Beam hardening . . . 398

8.7 The gridding method . . . 401
8.8 Concluding remarks . . . 404

9 Algebraic reconstruction techniques 407
9.1 Algebraic reconstruction . . . 407
9.2 Kaczmarz' method . . . 411
9.3 A Bayesian estimate . . . 417
9.4 Variants of the Kaczmarz method . . . 418
9.4.1 Relaxation parameters . . . 418
9.4.2 Other related algorithms . . . 420

10 Probability and Random Variables 423
10.1 Measure theory . . . 424
10.1.1 Allowable events . . . 424
10.1.2 Measures and probability . . . 427
10.1.3 Integration . . . 429
10.1.4 Independent events . . . 436
10.1.5 Conditional probability . . . 437
10.2 Random variables . . . 439
10.2.1 Cumulative distribution function . . . 441
10.2.2 The variance . . . 444
10.2.3 The characteristic function . . . 445
10.2.4 A pair of random variables . . . 446
10.2.5 Several random variables . . . 452
10.3 Some important random variables . . . 454
10.3.1 Bernoulli Random Variables . . . 455
10.3.2 Poisson Random Variables . . . 455
10.3.3 Gaussian Random Variables . . . 456
10.3.4 The Central Limit Theorem . . . 459
10.3.5 Limits of random variables . . . 461
10.3.6 Modeling a source-detector pair . . . 465
10.3.7 Beer's Law . . . 466
10.4 Statistics and measurements . . . 469

11 Random Processes 473
11.1 Random processes in measurements . . . 473
11.2 Basic definitions . . . 474
11.2.1 Statistical properties of random processes . . . 477
11.2.2 Stationary random processes . . . 478
11.2.3 Independent and stationary increments . . . 482
11.3 Examples of random processes . . . 482
11.3.1 Gaussian random process . . . 483
11.3.2 The Poisson counting process . . . 483
11.3.3 Poisson arrival process . . . 486
11.3.4 Fourier coefficients for periodic processes . . . 488
11.3.5 White noise . . . 491

11.4 Random inputs to linear systems . . . 493
11.4.1 The autocorrelation of the output . . . 494
11.4.2 Thermal or Johnson noise . . . 496
11.4.3 Optimal filters . . . 498

12 Resolution and noise **501**
12.1 The continuous case . . . 502
12.2 Sampled data . . . 504
12.3 A computation of the variance . . . 507
12.3.1 The variance of the Radon transform . . . 508
12.3.2 The variance in the reconstructed image . . . 510
12.3.3 Signal-to-noise ratio, dosage and contrast . . . 512

A Background material **515**
A.1 Numbers . . . 515
A.1.1 Integers . . . 515
A.1.2 Rational numbers . . . 517
A.1.3 Real numbers . . . 520
A.1.4 Cauchy sequences . . . 523
A.2 Vector spaces . . . 524
A.2.1 Euclidean n-space . . . 525
A.2.2 General vector spaces . . . 528
A.2.3 Linear Transformations and matrices . . . 531
A.2.4 Norms and Metrics . . . 536
A.2.5 Inner product structure . . . 540
A.2.6 Linear transformations and linear equations . . . 545
A.2.7 Linear algebra with uncertainties . . . 547
A.2.8 The least squares method . . . 549
A.2.9 Complex numbers and the Euclidean plane . . . 550
A.2.10 Complex vector spaces . . . 553
A.3 Functions, theory and practice . . . 554
A.3.1 Infinite series . . . 556
A.3.2 Partial summation . . . 559
A.3.3 Power series . . . 560
A.3.4 Binomial formula . . . 563
A.3.5 The Gamma function . . . 565
A.3.6 Bessel functions . . . 567
A.4 Spaces of functions . . . 569
A.4.1 Examples of function spaces . . . 569
A.4.2 Completeness . . . 573
A.4.3 Linear functionals . . . 575
A.4.4 Measurement, linear functionals and weak convergence . . . 577
A.4.5 The L^2-case . . . 579
A.4.6 Generalized functions on $\mathbb{R}$ . . . 581
A.4.7 Generalized functions on $\mathbb{R}^n$. . . . 587
A.5 Bounded linear operators . . . 589

A.6 Functions in the real world 594
A.6.1 Approximation 594
A.6.2 Sampling and Interpolation 600
A.7 Numerical techniques for differentiation and integration 603
A.7.1 Numerical integration 605
A.7.2 Numerical differentiation 607

B Basic analysis **613**
B.1 Sequences 613
B.2 Rules for Limits 614
B.3 Existence of limits 614
B.4 Series 615
B.5 Limits of Functions and Continuity 618
B.6 Differentiability 620
B.7 Higher Order Derivatives and Taylor's Theorem 621
B.8 Integration 621
B.9 Improper integrals 624

List of Figures

1.1 The world of old fashioned X-rays. 1
1.2 Using to trigonometry to find the height of a mountain. 4
1.3 A more realistic measurement. 5
1.4 Not exactly what we predicted! 6
1.5 Using refraction to determine the height of an interface. 7
1.6 Convex and non-convex regions. 8
1.7 Using particle scattering to determine the boundary of a convex region. . . 9
1.8 Using sound to measure depth. 10
1.9 The shadow of a convex region 13
1.10 Parameterization of oriented lines in the plane. 15
1.11 The measurement of the shadow 16
1.12 Two regions of constant width 2 21

2.1 Parallel slices of an object. 35
2.2 Analysis of an isotropic point source. 38
2.3 The failure of ordinary X-rays to distinguish objects. 40
2.4 A different projection. 40
2.5 A point source device for measuring line integrals of the absorption coefficient. 41
2.6 Collecting data from many views. 43
2.7 A typical X-ray source spectral function, courtesy Dr. Andrew Kavellas. . . 44
2.8 Back-projection does not work! 56

3.1 Furry functions 76
3.2 A furry function at smaller scales 76
3.3 Graphs of φ_ϵ, with $\epsilon = .5, 2, 8$. 87
3.4 Approximate δ-functions 93
3.5 Approximate δ-functions convolved with $\chi_{[-1,1]}$. 94
3.6 Illustration of the FWHM definition of resolution 95
3.7 FWHM vs. side-lobes 96
3.8 Real and imaginary parts of $\exp(i\langle(x,y),(1,1)\rangle)$ 114
3.9 Real and imaginary parts of $\exp(i\langle(x,y),(2,0)\rangle)$ 114
3.10 f is smeared into the ϵ-neighborhood of $\mathrm{supp}(f)$. 126

4.1 Graphs of $\hat{\psi}$ and k_ψ with $W = 40, C = 5$. 167
4.2 Radial graph of $k_\psi * \chi_{D_1}$, with $W = 40, C = 5$. 168

5.1 Periodic extension may turn a continuous function into discontinuous function. 184
5.2 Graph of the Dirichlet kernel, $D_3(x)$ 199
5.3 An example of the Gibbs phenomenon 201
5.4 Detail showing equi-oscillation property in Gibbs phenomenon 204
5.5 Graph of the Fejer kernel, $F_5(x)$ 207
5.6 Graphs comparing the partial sums and Fejer means. 210
5.7 Expanded view showing the loss of resolution in the Fejer means. 210
5.8 Expanded view showing Gibbs phenomenon in the partial sums. 211
5.9 Illustration of the $2d$-Gibbs phenomenon 215

6.1 Window functions in Fourier space and the ordinary sinc-pulse. 224
6.2 Shannon-Whittaker interpolation functions with second order smoothed windows. 224
6.3 Aliasing in MRI 227
6.4 The two faces of aliasing, $d = .05$. 230
6.5 Partial Fourier inverse and Shannon-Whittaker interpolant. 230
6.6 What aliasing looks like for a smoother function, $d = .1$. 231
6.7 What aliasing looks like for a furry function, $d = .1, .05, .025$. 231

7.1 The effects of errors in the amplitude and phase of the Fourier transform on a reconstructed image. 250
7.2 Using magnetic resonance to determine the vibrational modes of a molecule. 251
7.3 Transfer function for a tent filter. 258
7.4 Pointspread functions for lowpass filters 259
7.5 A network 268
7.6 Standard symbols for passive circuit elements 268
7.7 Simple RLC-circuits 269
7.8 The amplitude and phase of the transfer function of an RC-filter. 270
7.9 The amplitude and phase of the transfer function of an RL-filter 272
7.10 A resonant circuit. 273
7.11 An RLC-circuit. 273
7.12 The amplitude of the transfer function. 274
7.13 A second RLC-circuit. 275
7.14 Modified Fourier Transform of the rectangle function 283
7.15 Impulse responses for 2-dimensional low pass filters. 286
7.16 Low pass filters in two dimensions. 287
7.17 Half maximum curves for 2d low pass filters. 289
7.18 Bad interpolation using formula 7.55. 296
7.19 The Fourier transform of an image is not usually an image. 309
7.20 Removing geometric distortion. 310
7.21 The impulse response and transfer function for $\mathcal{A}_{.25}$. 315
7.22 Output of Laplacian edge detection filters. 316
7.23 A CT-head phantom showing the effect of rescaling grey values. 317
7.24 The Moiré effect is directional aliasing. 318
7.25 Arrangement of an imaging device with a source, object and detector. . . . 325
7.26 Computing the solid angle. 326

7.27 The similar triangle calculation. . . . 327
7.28 A pinhole camera. . . . 328
7.29 The image of two dots. . . . 330
7.30 Beam spreading. . . . 331
7.31 The geometry of a collimator 332
7.32 Evaluating the point spread function of a collimator. . . . 333
7.33 The graph of $p(\mathbf{r}_s; z)$, for a fixed z. . . . 334

8.1 A 3-dimensional X-ray beam. . . . 339
8.2 The reconstruction grid. . . . 340
8.3 A parallel beam scanner and sample set. . . . 343
8.5 Parameters for a fan beam machine. . . . 344
8.4 The two different divergent beam geometries. . . . 345
8.6 An example of a sinogram. . . . 346
8.7 The impulse response for a RamLak filter (solid) and a continuous approximation (dotted). . . . 352
8.8 Ram-Lak filters applied to $\mathrm{R}f_1$. . . . 354
8.9 Ram-Lak filters applied to $\mathrm{R}f_1$. . . . 354
8.10 How to choose sample spacings. . . . 357
8.11 Fan beam geometry. . . . 359
8.12 Quantities used in the fan beam, filtered backprojection algorithm. . . . 360
8.13 Collecting data for fan beam scanners. . . . 365
8.14 Absorbing square. . . . 370
8.15 Rectangle with small inclusion 371
8.16 Relative errors with small inclusion 371
8.17 A mathematical phantom. . . . 374
8.18 Examples of PSF and MTF with band limited regularization. . . . 376
8.19 Limits for the PSF and MTF in the filtered backprojection algorithm. . . . 377
8.20 Examples of PSF and MTF with exponential regularization. . . . 378
8.21 Examples of PSF and MTF with Shepp-Logan regularization. . . . 379
8.22 Resolution phantoms are used to gauge the resolution of a CT-machine or reconstruction algorithm. . . . 380
8.23 Reconstructions of a mathematical phantom using filtered backprojection algorithms. . . . 381
8.24 The effect of ray sampling on the PSF. . . . 384
8.25 Filtered backprojection reconstruction of elliptical phantom 385
8.26 Parameters describing the Radon transform of χ_E. . . . 386
8.27 Filtered backprojection reconstruction of square phantom 388
8.28 View sampling artifacts with $\Delta\theta = \frac{2\pi}{8}$ 392
8.29 View sampling artifacts with $\Delta\theta = \frac{2\pi}{32}$ 392
8.30 Examples comparing view aliasing in parallel beam and fan beam scanners. 393
8.31 A reconstruction with a few bad rays. . . . 395
8.32 A systematic bad ray with $\Delta\theta = \frac{2\pi}{8}$. . . . 396
8.33 A reconstruction with one bad view. . . . 398
8.34 Streaks caused by beam hardening. . . . 399
8.35 Beam hardening through water. . . . 400

8.36 Due to beam hardening, dense objects produce dark streaks. 401
8.37 Mathematical analysis has led to enormous improvements in CT-images. . . 405

9.1 Pixel basis 408
9.2 Method of projections 412
9.3 Examples where the projection algorithm does not converge. 413
9.4 One step in the Kaczmarz algorithm. 414
9.5 Reconstructions using ART 415
9.6 Ranges relaxation parameters. 419

10.1 Comparisons of Bernoulli and Gaussian distribution functions with $p = .1$. . 463
10.2 Comparisons of Bernoulli and Gaussian distribution functions with $p = .5$. . 463
10.3 Comparisons of Poisson and Gaussian distribution functions. 465

11.1 An RL-circuit. 497

12.1 Comparison of the image variance using different models for the variance in the measurements. 511

A.1 Multiplication of complex numbers 552
A.2 Some J-Bessel functions. 568
A.3 Polynomial interpolants for $|x - \frac{1}{2}|$ 602
A.4 A random piecewise linear function. 607
A.5 A fairly random function 609

Chapter 1

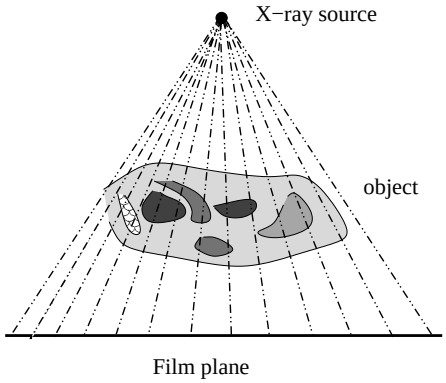
Measurements and modeling

A *quantitative* model of a physical system is expressed in the language of mathematics. A qualitative model often precedes a quantitative model. For many years clinicians used medical X-rays without employing a precise quantitative model. X-rays were thought of as high frequency 'light' with three very useful properties:

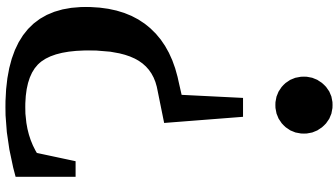
(1). If X-rays are incident on a human body, some fraction of the incident radiation is absorbed, though a sizable fraction is transmitted. The fraction absorbed is proportional to the total 'density' of the material encountered.

(2). A 'beam' of X-ray light travels in a straight line.

(3). X-rays darken photographic film. Taken together, these properties mean that using X-rays one could "see through" a human body to obtain a shadow or projection of the internal anatomy on a sheet of film.



(a) Depth information is lost in a projection.



(b) A old fashioned X-ray.

Figure 1.1: The world of old fashioned X-rays.

The model was adequate given the available technology. In their time, X-rays led to a revolution in the practice of medicine because they opened the door to non-destructive examination of internal anatomy. They are still useful for locating bone fractures, dental caries and foreign objects but their ability to visualize soft tissues and more detailed anatomic structure is very limited. There are several reasons for this. The X-ray image is a two dimensional projection of a three dimensional object which renders it impossible to deduce the spatial ordering in the missing third dimension. Photographic film is not very sensitive to X-rays. To get a usable image, a light emitting phosphor is sandwiched with the film. This increases the sensitivity of the overall 'detector,' but even so, large changes in the intensity of the incident X-rays still produce small differences in the density of film. This means that the contrast between different soft tissues is poor. Because of these limitations a qualitative theory was adequate for the interpretation of X-ray images.

A desire to improve upon this situation led Alan Cormack and Godrey Hounsfield, to independently develop X-ray *tomography* or slice imaging, see [30] and [10]. The first step in their work was to use a quantitative theory for the absorption of X-rays. Such a theory already existed and is little more than a quantitative restatement of (1) and (2). It was not needed for old fashioned X-rays because they are read "by eye," no further processing is done after the film is developed, Both Cormack and Hounsfield realized that mathematics could be used to infer 3-dimensional anatomic structure from a large collection of *different* two dimensional projections. The possibility for making this idea work relied on two technological advances: 1. The availability of scintillation crystals to use as detectors. 2. Powerful, digital computers to process the tens of thousands of measurements needed to form a usable image. A detector using a scintillation crystal is about a hundred times more sensitive than film which makes possible much finer distinctions. As millions of arithmetic operations are needed for each image, fast computers are a necessity for reconstructing an image from the available measurements. It is an interesting historical note that the mathematics underlying X-ray tomography was done in 1917 by Johan Radon, see [59]. It had been largely forgotten and both Hounsfield and Cormack worked out solutions to the problem of reconstructing an image from its projections. Indeed, this problem had arisen and been solved in contexts as diverse as radio astronomy and statistics.

This book is a detailed exploration of the mathematics which underpins the reconstruction of images in X-ray tomography. The list of mathematical topics covered is dictated by their importance and utility in medical imaging. While our emphasis is on understanding these mathematical foundations, we constantly return to the practicalities of X-ray tomography and explore the relationship of the mathematical formulation of a problem and its solution, to the realities of computation and physical measurement. There are many different imaging *modalities* in common use today, X-ray computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), ultrasound, optical imaging, impedence imaging, etc. Because each relies on a different physical principle, each provides different information. In every case the mathematics needed to process and interpret the data has a large overlap with that used in X-ray CT. We concentrate on X-ray CT because of the simplicity and clarity of the physical principles underlying the measurement process. Detailed descriptions of the other modalities can be found in [39] or [4].

1.1 Mathematical modeling

Mathematics is the language in which any quantitative theory or model is eventually expressed. In this introductory chapter we consider a variety of examples of physical systems, measurement processes and the mathematical models used to describe them. These models illustrate different aspects of more complicated models used in medical imaging. The chapter concludes with a consideration of linear models.

Mathematics is used to model physical systems from the formation of the universe to the structure of the atomic nucleus, from the function of the kidney to the opinions of voters. The first step in giving a mathematical description of a "system" is to isolate that system from the universe in which it sits. While it is no doubt true that a butterfly flapping its wings in Siberia in mid-summer will effect the amount of rainfall in the Amazon rain forest a decade hence, it is surely a tiny effect, impossible to accurately quantify. To obtain a practical model such effects are ignored, though they may come back to haunt the model, as measurement error and noise. After delineating a system, we need to find a collection of numerical parameters which describe its state. In this generality these parameters are called *state variables*. In the idealized world of an isolated system the exact measurement of the state parameters would uniquely determine the state of the system. In general the natural state parameters are not directly measurable. The model then describes relations between the state variables which suggest feasible measurements with which one might determine the state of the system.

1.1.1 Finitely many degrees of freedom

See: A.1, B.5, B.6, B.7.

If the state of a system is described by a finite collection of real numbers $\mathbf{x} = (x_1, \ldots, x_n)$ then the system has finite many *degrees of freedom*. Most of the systems encountered in elementary physics and electrical engineering have this property. The mathematical model is then expressed as relations that these variables satisfy, often taking the form of functional relations,

$$\begin{array}{rcl} f_1(x_1, \ldots, x_n) & = & 0 \\ \vdots & & \vdots \\ f_m(x_1, \ldots, x_n) & = & 0 \end{array} \tag{1.1}$$

If there are more state variables than relations, that is $m < n$ then, heuristically $n - m$ measurements are needed to determine the state of the system. Of course this counting parameters arguments assumes that the relations in (1.1) are functionally independent. For linear relations the number of functionally independent equations does not depend on the state. If the relations are non-linear then counting the number of independent conditions can be quite involved and the result generally depends on $\mathbf{x}$. For measurements to be useful they must be expressible as functions of the state variables.

Example 1.1.1. Suppose the system is a ball on a rod. The state of the system is described by (x, y), the coordinates of the ball. If the rod is of length r and one end of it is fixed at

the point $(0, 0)$, then the state variables satisfy the relation

$$x^2 + y^2 = r^2. \tag{1.2}$$

Imagine now that one dimensional creatures, living on the x-axis $\{y = 0\}$ can observe a shadow of the ball, cast by very distant light sources so that the rays of light are perpendicular to the x-axis. The line creatures want to predict whether or not the ball is about to collide with their world. Locating the shadow determines the x-coordinate of the ball, using equation (1.2) gives

$$y = \pm\sqrt{r^2 - x^2}.$$

To determine the sign of the y-coordinate requires additional information not available in the model. On the other hand this information is adequate if one only wants to predict if the ball is about to collide with the x-axis. If the x-axis is illuminated by red light from above and blue light from below, then a ball approaching from below would cast of red shadow while a ball approaching from above would cast a blue shadow. With this additional data, the location of the ball is completely determined.

Example 1.1.2. We would like to find the height of a mountain without climbing it. To that end, the distance l between the point P and the top of the mountain, as well as the angle θ are measured. If l and θ are measured exactly then by a trigonometric identity, the height h of the mountain is given by $l \tan\theta$. Measurements are never exact, let us use the model

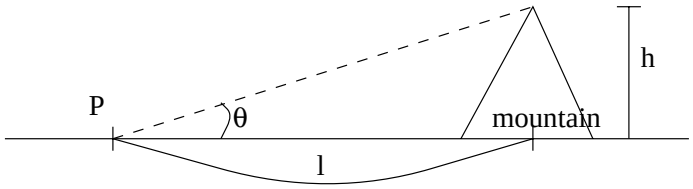


Figure 1.2: Using to trigonometry to find the height of a mountain.

to relate the error in measuring θ to the computed value of h. This requires a basic tool from calculus, the Taylor expansion. If $f(x)$ is a smooth function of the variable x then we can approximate the behavior of f near to x_0 by using the Taylor expansion

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + \cdots + \frac{1}{n!}f^{[n]}(x_0)(x - x_0)^n + R_n(x - x_0).$$

Here R_n is called the n^{th}-remainder, it is a term that goes to zero faster than $|x - x_0|^n$. The complicated function $\tan\theta$ can be replaced by a much simpler function, though at the expense of replacing an exact (but not very useful) formula with an approximate (but more usable) formula.

Suppose that we measure $\theta + \Delta\theta$ where θ is the exact angle. The error in h as a function of $\Delta\theta$ is given approximately by

$$\begin{aligned} \tan(\theta + \Delta\theta) &= \tan(\theta) + \partial_\theta \tan(\theta)\Delta\theta + O(\Delta\theta^2) \\ &= \tan(\theta) + \sec^2(\theta)\Delta\theta + O(\Delta\theta^2). \end{aligned} \tag{1.3}$$

The notation $O(\Delta\theta^2)$ refers to an error term which is bounded by a constant times $\Delta\theta^2$, as $\Delta\theta$ goes to zero. The height predicted from the measurement of the angle is

$$h_m = l\tan(\theta+\Delta\theta) = l(\tan\theta + \frac{\Delta\theta}{\cos^2\theta} + O(\Delta\theta^2)).$$

Disregarding the quadratic error terms $O(\Delta\theta^2)$, the *absolute error* is

$$h_m - h \approx l\frac{\Delta\theta}{\cos^2\theta}.$$

The absolute error is a number with the same units as h; in general it is not a very interesting quantity. If, for example the true measurement is $10,000$m then an error of size 1m would not be too significant whereas if the true measurement is 2m then it would. To avoid this obvious pitfall one normally considers the *relative error*. In this problem the relative error is

$$\frac{h_m - h}{h} = \frac{\Delta\theta}{\cos^2\theta\tan\theta} = \frac{\Delta\theta}{\sin\theta\cos\theta}.$$

Generally the relative error is the absolute error divided by the correct value. It is a dimensionless quantity that gives a quantitative assessment of the accuracy of the measurement. If the angle θ is measured from a point too near to or too far from the mountain, i.e. θ is very close to 0 or $\pi/2$ then small measurement errors result in a substantial loss of accuracy. A useful feature of a precise mathematical model is the possibility of estimating how errors in measurement affect the accuracy of the parameters we wish to determine.

Example 1.1.3. In a real situation we cannot measure the distance l either. Suppose that we measure the angle from two different positions i.e. θ_1 and θ_2 as in the figure below. Then we have $\tan\theta_1 = h/l_1, \tan\theta_2 = h/(l_1+l_2)$, for the same h. Using trigonometry we

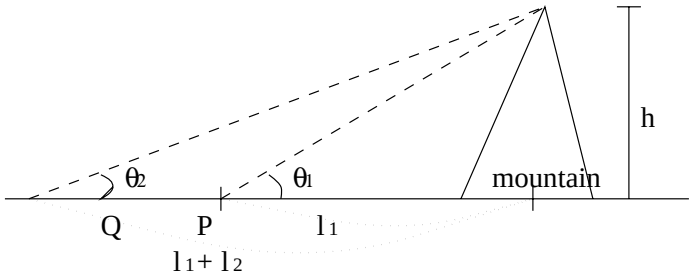


Figure 1.3: A more realistic measurement.

deduce

$$\begin{aligned} l_1 &= \frac{l_2}{\tan\theta_1/\tan\theta_2 - 1}, \\ h &= (l_1+l_2)\tan\theta_2 = \left(\frac{l_2}{\tan\theta_1/\tan\theta_2 - 1} + l_2\right)\tan\theta_2 = l_2\frac{\sin\theta_1\sin\theta_2}{\sin(\theta_1-\theta_2)}. \end{aligned}$$

Assuming that l_2, the distance between P and Q can also be measured, then h can be determined from θ_1 and θ_2. We complete the discussion of this example by listing different ways that this model may fail to capture important features of the physical situation.

- If the shape of a mountain looks like that in figure 1.4and we measure the distance and angle at the point P, we are certainly not finding the real height of the mountain. Some *a priori* information is always incorporated in a mathematical model.

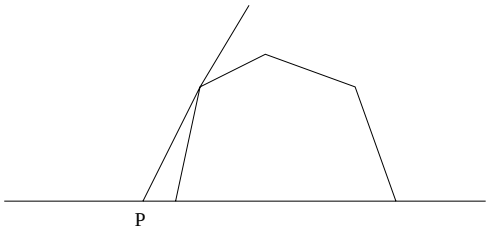


Figure 1.4: Not exactly what we predicted!

- The curvature of the earth is ignored. A more sophisticated geometric model is needed to correct for such errors. This becomes a significant problem as soon as the distances, l, l_1, l_2 are large compared to the distance to the horizon (about 25km for a 2 meter tall person). The approximations used in the model must be adapted to the actual physical conditions of the measurements.

- The geometry of the underlying measurements could be very different from the simple Euclidean geometry used in the model. To measure the angles θ_1, θ_2 one would normally use a transit to sight the peak of the mountain. If the mountain is far away then the light travels on a path from the mountain to the transit which passes through air of varying density. The light is refracted by the air and therefore the ray path is not the straight line assumed in the model. To include this effect would vastly complicate the model. This is an important consideration in the very similar problem of creating a map of the sky from earth based observations of stars.

  Analogous problems arise in medical imaging. If the wavelength of the energy used to probe the human anatomy is very small compared to the size of the structures that are present then it is reasonable to assume that the waves are not refracted by the medium through which they pass, i.e. X-rays can be assumed to travel along straight lines. However for energies with wavelengths comparable to the size of structures present in the human anatomy, this assumption is simply wrong. The waves are then bent and diffracted by the medium and the difficulty of modeling the ray paths is considerable and, in fact largely unsolved! This is an important issue in ultrasound imaging.

Example 1.1.4. Refraction provides another example of a simple physical system. Suppose that we have two fluids in a tank as shown in the figure and would like to determine the height of the interface between them. Suppose first of all that the refractive indices of the fluids are known. Let n_1 be the refractive index of the upper fluid and n_2 the refractive index of the lower one, Snell's law states that

$$\frac{\sin(\theta_1)}{\sin(\theta_2)} = \frac{n_2}{n_1}.$$

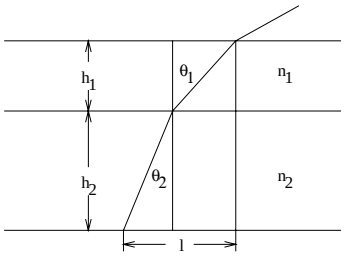


Figure 1.5: Using refraction to determine the height of an interface.

Let h denote the total height of the fluid, then

$$h_1 + h_2 = h.$$

The measurement we make is the total displacement l, of the light ray as it passes through the fluids. It satisfies the relationship

$$h_1 \tan(\theta_1) + h_2 \tan(\theta_2) = l.$$

Using these three relations h_1 and h_2 are easily determined. The assumption that we know n_1 implies, by Snell's law that we can determine θ_1 from a measurement of the angle of the light ray above the fluid. If n_2 is also known, then using these observations we can determine θ_2 as well:

$$\sin(\theta_2) = \frac{n_1}{n_2} \sin(\theta_1).$$

The pair (h_1, h_2) satisfies the 2×2-linear system

$$\begin{pmatrix} 1 & 1 \\ \tan(\theta_1) & \tan(\theta_2) \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = \begin{pmatrix} h \\ l \end{pmatrix}. \tag{1.4}$$

In example 1.3.2 we consider a slightly more realistic situation where the refractive index of the lower fluid in not known. By using more measurements n_2 can also be determined, though the equation determining n_2 is no longer linear.

Exercise 1.1.1. Suppose that in example 1.1.1 light sources are located at $(0, \pm R)$. What is the relationship between the x-coordinate and the shadow?

Exercise 1.1.2. In example 1.1.3 work out how the absolute and relative errors depend on θ_1, θ_2 and l_2.

1.1.2 Infinitely many degrees of freedom

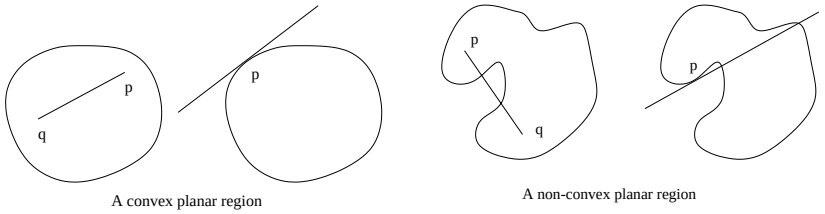
See: A.3, A.6.

In the previous section we examined some simple physical systems described by a finite collection of numbers. Such systems are said to have finitely many degrees of freedom. In

these examples, the problem of determining the state of the system from feasible measurements reduces to solving systems of finitely many equations in finitely many unknowns. In imaging applications the state of a system is usually described by a function or functions of continuous variables. These are systems with infinitely many degrees of freedom. In this section we consider several examples of this type.

Example 1.1.5. Suppose that we would like to determine the shape of a planar object, D that cannot be seen. The object is lying inside a disk and we can fire particles at the object which bounce off. Assume that this scattering process is very simple: each particle strikes the object once and is then scattered along a straight line off to infinity. The outline of the object can be determined by knowing the correspondence between incoming lines, l_{in} and outgoing lines, l_{out}. Each intersection point $l_{\text{in}} \cap l_{\text{out}}$ lies on the boundary of the object. Measuring $\{l^j_{\text{out}}\}$ for finitely many incoming directions $\{l^j_{\text{in}}\}$ determines finitely many points $\{l^j_{\text{in}} \cap l^j_{\text{out}}\}$ on the boundary of D. In order to use this finite collection of points to make any assertions about the rest of the boundary of D, more information is required. If we know that D consists of a single piece or component then these points would lie on a single closed curve, though it might be difficult to decide in what order they should appear on the curve.

Convex obstacles is a class which satisfy these simple assumptions and for which a finite number of points on the boundary might carry a lot of useful information. A region D in the plane is convex if it has the following property: for each pair of points p and q lying in D the line segment $\overline{pq}$ is also contained in D. An equivalent condition is the following: for each point p on the boundary of D there is a line l_p which passes through p but is otherwise disjoint from D. This line is called a *support line* through p. If the boundary is smooth at p then the tangent line to the boundary is the unique support line.



(b)

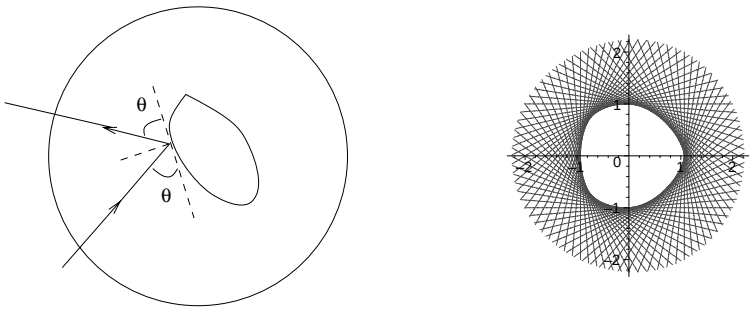
Figure 1.6: Convex and non-convex regions.

If the object is convex and more is known about the scattering process, for example if the angle of incidence is equal to the angle of reflection, then from a finite number of incoming and outgoing pairs, $\{(l^i_{\text{in}}, l^i_{\text{out}}) \quad : i = 1, \ldots, N\}$ we can determine an approximation to D with an estimate for the error. The intersection points, $\{l^i_{\text{in}} \cap l^i_{\text{out}}\}$ lie on the boundary of the convex region, D. If we use these points as the vertices of a polygon, P^{in}_N then the first convexity condition implies that P^{in}_N is completely contained within D. On the other hand, as the angle of incidence equals the angle of reflection we can also determine the tangent lines to the boundary of D at the points of intersection. A line divides the plane into two half planes, since D is convex, it lies entirely in one of the half plane determined by each of

its tangent lines. By intersecting the half planes defined by the tangent lines through the points $\{(l^i_{\text{in}}, l^i_{\text{out}})\}$ we obtain another convex polygon, P^{out}_N which contains D. Thus with these N-measurements we obtain the both an *inner* and *outer* approximation to D :

$$P^{\text{in}}_N \subset D \subset P^{\text{out}}_N.$$

It is clear that the boundary of a convex region has *infinitely many degrees of freedom* as it is conveniently described as the image of a map $s \mapsto (x(s), y(s))$, where s lies in an interval $[0, 1]$. On the other hand the images of such maps can be approximated by polygons. Once the number of sides is fixed, then we are again considering a system with finitely many degrees of freedom. In all practical problems, a system with infinitely many degrees of freedom must eventually be approximated by a system with finitely many degrees of freedom.



(a) The angle of incidence equals the angle of reflection.

(b) The outer approximation as an intersection of half spaces.

Figure 1.7: Using particle scattering to determine the boundary of a convex region.

Remark 1.1.1. For a non-convex body the above method does not work as the correspondence between incoming and outgoing lines can be quite complicated: some incoming lines may undergo multiple reflections before escaping, in fact some lines might become permanently trapped.

Example 1.1.6. Suppose that the *surface* of a sea is mapped by coordinates (x, y) belonging to a region $D \subset \mathbb{R}^2$. The depth of the bottom of the sea is described by a function $h(x, y)$. One way to determine h would be to drop a weighted string until it hits the bottom. There are problems with this method: 1. It is difficult to tell when the weight hits the bottom. 2. Unknown, underwater currents may carry the string so that it does not go straight down. A somewhat less direct approach would be to use sonar to measure the distance to the bottom. The physical principle underlying the measurement is that the speed of sound is determined by the density and temperature of the water which are in turn determined by the depth. Let $c(z)$ denote the *known* speed of sound, as a function of the depth. A speaker underneath the boat emits a loud, short pulse of sound and the time it takes for the sound to return is measured. Here we assume that the sound travels in a straight line

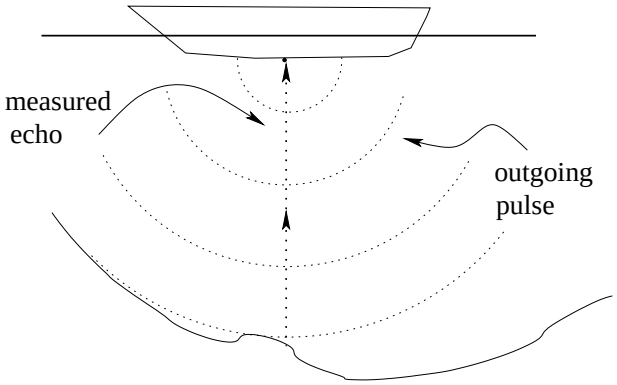


Figure 1.8: Using sound to measure depth.

to the bottom and the microphone only detects the direct reflection, traveling back along the straight line. Using $c(z)$ the transit time can be related to the depth.

A simple model, valid for shallow seas, is that the speed of sound is a constant, c. The measurement, T is the time it takes for the sound pulse to go down and back,

$$2h = cT.$$

This assumes that the boat is stationary from the time the pulse is emitted until the return is received. With such a measurement, $T(x, y)$ for each position $(x, y) \in D$, the depth is determined, everywhere by

$$h(x, y) = \frac{cT(x, y)}{2}.$$

In reality such continuous measurements are not possible. Instead the boat is placed at a finite set of locations $P = \{(x_j, y_j) \, : \, j = 1, \ldots, N\}$ and $T(x_j, y_j)$ is measured. The finite set of values

$$h(x_j, y_j) = \frac{cT(x_j, y_j)}{2}$$

are then determined.

Again, what use is a finite set of values? Without qualitative, *a priori* information about the nature of the function h, this finite data set is indeed useless! On the other hand it is reasonable to assume that h is a *continuous* function of (x, y). With this assumption, values of h for points **not** in P can be *interpolated* from the measured values. The minimum necessary separation between the points in P is determined by a quantitative assessment of how continuous h is expected to be. Suppose it is known that there is a constant M so that

$$|h(x, y) - h(x', y')| \leq M\sqrt{(x - x')^2 + (y - y')^2}.$$

If every point (x, y) is within d of a point (x_j, y_j) in P then we have the estimate

$$|h(x, y) - h(x_j, y_j)| \leq Md.$$

This then gives an estimate for the accuracy of the interpolated values. A small value of M indicates that the depth is varying slowly, while a large value indicates rapid variations. In the former case a larger value of d provides acceptable results, while in the latter case a smaller value of d is needed to get an accurate picture of the bottom.

Example 1.1.7. Now assume that the sea, in the previous example is one dimensional, but that the sound speed is not constant. To use the measurements described above to determine the depth $h(x)$ requires more mathematical apparatus. Let $z(t)$ denote the depth of the sound pulse at a time t after it is emitted. Using calculus we can express the assertion that the 'speed of sound at depth z is $c(z)$' as a differential equation

$$\frac{dz}{dt}(t) = c(z(t)). \tag{1.5}$$

Formally this is equivalent to

$$\frac{dz}{c(z)} = dt.$$

The transit time T is a function of the depth, h integrating this equation gives

$$G(h) \stackrel{d}{=} \int_0^h \frac{dz}{c(z)} = \frac{T(h)}{2}. \tag{1.6}$$

The function G is monotonely increasing and therefore its inverse is well defined. Using G^{-1} we can determine the depth, h from the available measurement, T

$$h = G^{-1}\left(\frac{T}{2}\right).$$

To use this model, the function G^{-1} needs to be explicitly determined. If $c(z)$ is simple enough then an analytic formula for G might be available. Otherwise the integral defining G is computed for a finite collection of depths $\{h_1, \ldots, h_m\}$, with $t_i = G(h_i)$. From this table of values, the inverse function is also known for a finite collection of times

$$h_i = G^{-1}(t_i).$$

If $c(z)$ is a differentiable function, then a linear approximation of the form

$$c(z) \approx c + az$$

is valid for small values of z. Integrating gives

$$G(h) \approx \log(1 + \frac{ah}{c}),$$

solving for $G^{-1}(T)$ we find

$$h(T) \approx \frac{c}{a}(e^{a\frac{T}{2}} - 1).$$

Using Taylor's formula for e^x gives

$$h(T) \approx c\frac{T}{2} + \frac{caT^2}{8} + O(T^3).$$

Here as usual $O(T^3)$ is an error term which goes to zero, as T goes to zero, at the same rate as T^3. This agrees, to leading order with the previous computation.

Example 1.1.8. The one dimensional model in the previous example can be used to solve the two dimensional problem. Suppose that the area we are interested in mapping corresponds to the rectangle $[-1,1]\times[-1,1]$ in the (x,y)-map coordinates. For each y define the function of one variable

$$h_y(x) \stackrel{d}{=} h(x,y).$$

Knowing the collection of functions $\{h_y(x) : y \in [-1,1]\}$ for $x \in [-1,1]$ is evidently exactly the same thing as a knowing $h(x,y)$, for $(x,y) \in [-1,1]\times[-1,1]$. Because the measuring apparatus only observes the sound returning on the straight line from the boat to the bottom of the sea, the analysis in the previous example applies to allow the determination of $h_y(x)$ from measurements of $T_y(x)$,

$$h(x,y) = h_y(x) = G^{-1}\left(\frac{T_y(x)}{2}\right).$$

In this way a two dimensional problem is sliced into simpler one dimensional problems. In real applications, only finitely many measurements are made. A typical strategy if to pick an equally spaced set of y-values,

$$y_k = \frac{k}{N}, \quad k = -N,\ldots,N$$

and determine $h_{y_k}(x_j)$ at finitely many, equally spaced x-values

$$x_j = \frac{j}{N} \quad j = -N,\ldots,N.$$

These examples capture many of the features that we will encounter in X-ray tomography: by using a mathematical model for the measurements, an inaccessible, physical quantity can be determined using feasible measurements. The model is itself an approximation, but is subject to improvements.

Exercise 1.1.3. Describe parameters to describe the set of polygons with n-vertices in the plane. For the case of triangles, find the relations satisfied by your parameters. Find a condition, in terms of your parameters implying that the polygon is convex.

Exercise 1.1.4. Find an example of a planar region such that at least one particle trajectory is trapped forever.

Exercise 1.1.5. Why is G a monotonely increasing function?

Exercise 1.1.6. Suppose that $c(z)$ is piecewise constant, so that

$$c(z)\begin{cases} c_1 & \text{if } 0 \le z \le z_1, \\ c_2 & \text{if } z_1 < z. \end{cases}$$

Find G and G^{-1}.

Exercise 1.1.7. Why is it reasonable to model $c(z)$ as a linear function under the assumption that it is a differentiable function? Suggest a method for determining a.

Exercise 1.1.8. In the examples above it is assumed that all returns not arriving on the straight line path from the bottom of the ocean are ignored. Analyze the problems that result if return signals are accepted from all directions. What impact would this have on using the slicing method to reduce the dimensionality of the problem?

Exercise 1.1.9. Repeat the analysis in example 1.1.7 assuming that the boat is traveling at constant velocity v. Continue assuming that only returns meeting the bottom of the boat at right angles are detected.

1.2 A simple model problem for image reconstruction

The problem of image reconstruction in X-ray tomography is sometimes described as reconstructing an object from its "projections." Of course these are projections under the illumination of X-ray "light." In this section we consider the analogous, but simpler problem, of determining the outline of an object from its shadows. As is also the case in medical applications, we consider a two dimensional problem. Let D be the convex region in the plane. Imagine that a light source is placed very far away from the body. Since the light source is very far away, the rays of light are all traveling in essentially the same direction. We can think of the rays of light as a collection of parallel lines. We want to measure the shadow that D casts for each position of the light source. To describe the measurements imagine that a screen is placed on the "other side" of D perpendicular to the direction of the light rays, see the figure below. The screen is the detector, in a real apparatus sensors would be placed on the screen, allowing us to determine where the shadow begins and ends.

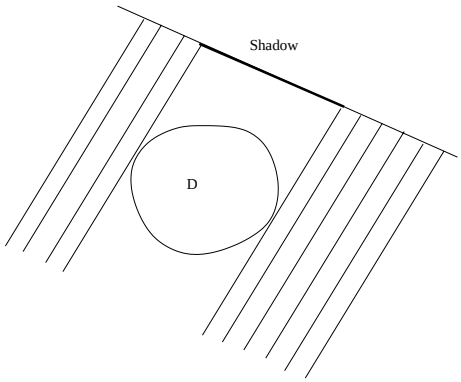


Figure 1.9: The shadow of a convex region

The region, D blocks a certain collection of light rays and allows the rest to pass. Measuring the shadow is therefore determining the "first" and "last" lines in this family of parallel lines to intersect D. To completely describe the object we need to rotate the source and detector through 180°, measuring, at each angle, where the shadow begins and ends.

The first and last lines to intersect a region just meet it along its boundary. These lines are therefore tangent to the boundary of D. The problem of reconstructing a region from its shadows is mathematically the same as the problem of reconstructing a region from a knowledge of the tangent lines to its boundary. As a first step in this direction we need a

good way to organize our measurements. To that end we give a description for the *space of all lines in the plane.*

1.2.1 The space of lines in the plane

A line in the plane is the set of points which satisfy an equation of the form

$$ax + by = c$$

where $a^2 + b^2 \neq 0$. We get the same set of points if we replace this equation by

$$\frac{a}{\sqrt{a^2+b^2}}x + \frac{b}{\sqrt{a^2+b^2}}y = \frac{c}{\sqrt{a^2+b^2}}.$$

The coefficients, $(\frac{a}{\sqrt{a^2+b^2}}, \frac{b}{\sqrt{a^2+b^2}})$ define a point ω on the unit circle and the constant $\frac{c}{\sqrt{a^2+b^2}}$ can be any number. The lines in the plane are parametrized by a pair consisting of a unit vector, ω and a real number t. The line $l_{t,\omega}$ is the set of points satisfying the equation

$$(x, y) \cdot \omega = t.$$

Here $(x, y) \cdot \omega$ is the *dot*-product

$$(x, y) \cdot \omega = x\omega_1 + y\omega_2$$

where $\omega = (\omega_1, \omega_2)$. As the set of points satisfying this equation is unchanged if (t, ω) is replaced by $(-t, -\omega)$ it follows that, as sets, $l_{t,\omega} = l_{-t,-\omega}$.

Very often it is convenient to parametrize the points on the unit circle by a real number, to that end we set

$$\omega(\theta) = (\cos(\theta), \sin(\theta)). \tag{1.7}$$

Since cos and sin are 2π-periodic it clear that $\omega(\theta)$ and $\omega(\theta + 2\pi)$ are the same point on the unit circle. Using this notation the line $l_{t,\theta} = l_{t,\omega(\theta)}$ is the set of solutions to the equation

$$\cos(\theta)x + \sin(\theta)y = t.$$

Both notations are used in the sequel. The vector

$$\hat{\omega}(\theta) = (-\sin(\theta), \cos(\theta)),$$

is perpendicular to $\omega(\theta)$. For any real number s,

$$\omega \cdot (t\omega + s\hat{\omega}) = t$$

and therefore we can describe $l_{t,\omega}$ parametrically as the set of points

$$l_{t,\omega} = \{t\omega + s\hat{\omega} \mid s \in (-\infty, \infty)\}.$$

Both $\hat{\omega}$ and $-\hat{\omega}$ are unit vectors which are perpendicular to ω; $\hat{\omega}$ is singled out by the condition that the 2×2 matrix $(\omega\hat{\omega})$ has determinant $+1$. This shows that the pair (t, ω) determines an *oriented line* The vector $\hat{\omega}$ is the "positive" direction along the line $l_{t,\omega}$.

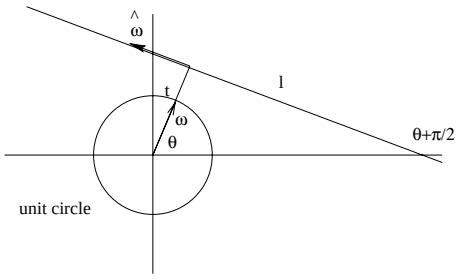


Figure 1.10: Parameterization of oriented lines in the plane.

The vector ω is the direction perpendicular to the line and the number t is called the *affine parameter* of the line, $|t|$ is the distance from the line to the origin of the coordinate system. The pair (t, ω) defines two half planes

$$H^+_{t,\omega} = \{(x, y) \mid (x, y) \cdot \omega > t\} \text{ and } H^-_{t,\omega} = \{(x, y) \mid (x, y) \cdot \omega < t\},$$

the line $l_{t,\omega}$ is the common boundary of these half planes. Facing along the line $l_{t,\omega}$ in the direction specified by $\hat{\omega}$, the half plane $H^-_{t,\omega}$ lies to the left. To summarize, the pairs $(t, \omega) \in \mathbb{R}^1 \times \mathbb{S}^1$ parametrize the *oriented* lines in the plane, which we sometimes call the *space of oriented lines.*

Exercise 1.2.1. Show that

$$|t| = \min\{\sqrt{x^2 + y^2} \ : \ (x, y) \in l_{t,\omega}\}.$$

Exercise 1.2.2. Show that if ω is fixed then the family of lines $\{l_{t,\omega} \ : \ t \in \mathbb{R}\}$ are parallel.

Exercise 1.2.3. Show that every line in the family $\{l_{t,\hat{\omega}} \ : \ t \in \mathbb{R}\}$ is orthogonal to every line in the family $\{l_{t,\omega} \ : \ t \in \mathbb{R}\}$.

Exercise 1.2.4. Each choice of direction ω defines a coordinate system on $\mathbb{R}^2$,

$$(x, y) = t\omega + s\hat{\omega}.$$

Find the inverse, expressing (t, s) as functions of (x, y). Show that the the area element in the plane satisfies

$$dxdy = dtds.$$

1.2.2 Reconstructing an object from its shadows

Now we can quantitatively describe the shadow. Because there are two lines in each family of parallel lines which are tangent to the boundary of D we need a way to select one of them. To do this we choose an orientation for the boundary of D; this operation is familiar from Green's theorem in the plane. The positive direction on the boundary is selected so that, when facing in that direction the region lies to the left; the counterclockwise direction is, by convention the positive direction, see figure 1.11.

Fix a direction $\omega = (\cos(\theta), \sin(\theta))$. In the family of parallel lines $l_{t,\omega}$ there are two values of t, $t_0 < t_1$ such that the lines $l_{t_0,\omega}$ and $l_{t_1,\omega}$ are tangent to the boundary of D,

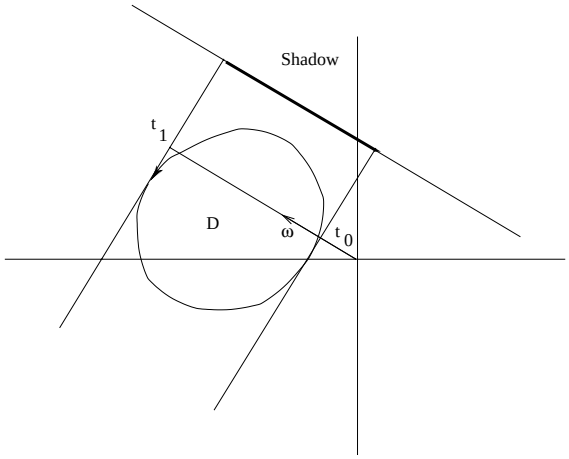


Figure 1.11: The measurement of the shadow

see figure 1.11. Examining the diagram it is clear that the orientation of the boundary at the point of tangency and that of the line agree, for t_1 and are opposite for t_0. For $\omega(\theta) = (\cos(\theta), \sin(\theta))$ define $h_D(\theta) = t_1$. We call $h_D(\theta)$ the *shadow function* for D. The mathematical formulation of reconstruction problem is: Can the boundary of the region D be determined from its shadow function?

The line $l_{h_D(\theta),\omega(\theta)}$ is given parametrically by

$$\{h_D(\theta)(\cos(\theta), \sin(\theta)) + s(-\sin(\theta), \cos(\theta)) \mid s \in (-\infty, \infty)\}.$$

To determine the boundary of D it would suffice to determine the point of tangency of $l_{h_D(\theta),\omega(\theta)}$ with the boundary of D, in other words we would like to find the function $s(\theta)$ so that for each θ,

$$(x(\theta), y(\theta)) = h_D(\theta)(\cos(\theta), \sin(\theta)) + s(\theta)(-\sin(\theta), \cos(\theta)) \tag{1.8}$$

is a point on the boundary of D.

The function $s(\theta)$ is found by recalling that, at the point of tangency, the direction of the tangent line to D is $\hat{\omega}(\theta)$. For a curve in the plane given parametrically by $(x(\theta), y(\theta))$ the direction of the tangent line at a point θ_0 is the same as that of the vector $(x'(\theta_0), y'(\theta_0))$. Differentiating the expression given in (1.8) and using the fact that $\partial_\theta \omega = \hat{\omega}$ we find that

$$(x'(\theta), y'(\theta)) = (h'_D(\theta) - s(\theta))\omega(\theta) + (h_D(\theta) + s'(\theta))\hat{\omega}(\theta). \tag{1.9}$$

Since the tangent line at $(x(\theta), y(\theta))$ is parallel to $\hat{\omega}(\theta)$ it follows from (1.9) that

$$h'_D(\theta) - s(\theta) = 0. \tag{1.10}$$

This gives a parametric representation for the boundary of a convex region in terms of its shadow function: If the shadow function is $h_D(\theta)$ then the boundary of D is given parametrically by

$$(x(\theta), y(\theta)) = h_D(\theta)\omega(\theta) + h'_D(\theta)\hat{\omega}(\theta). \tag{1.11}$$

Note that we have assumed that $h_D(\theta)$ is a differentiable function. This is not always true, for example if the region D is a polygon then the shadow function is not everywhere differentiable.

Let D denote a convex region and h_D its shadow function. We can think of $D \to h_D$ as a mapping from convex regions in the plane to 2π periodic functions. It is reasonable to enquire if every 2π periodic function is the shadow function of a convex region. The answer to this question is **no.** For strictly convex regions with smooth boundaries we are able to characterize the range of this mapping. If h is twice differentiable then the tangent vector to the curve defined by

$$(x(\theta), y(\theta)) = h(\theta)\omega(\theta) + h'(\theta)\hat{\omega}(\theta) \tag{1.12}$$

is given by

$$(x'(\theta), y'(\theta)) = (h''(\theta) + h(\theta))\hat{\omega}(\theta).$$

In our construction of the shadow function we observed that the tangent vector to the curve at $(x(\theta), y(\theta))$ and $\hat{\omega}(\theta)$ point in the same direction. From our formula for the tangent vector we see that this implies that

$$h''(\theta) + h(\theta) > 0 \text{ for all } \theta \in [0, 2\pi]. \tag{1.13}$$

This gives a necessary condition for a twice differentiable function h to be the shadow function for a strictly convex region with a smooth boundary. Mathematically we are determining the of the map that takes a convex body $D \subset \mathbb{R}^2$ to its shadow function h_D, under the assumption that h_D is twice differentiable. This is a convenient mathematical assumption, though in an applied context it is likely to be overly restrictive.

Exercise 1.2.5. Suppose that D_n is a regular n-gon. Determine the shadow function $h_{D_h}(\theta)$.

Exercise 1.2.6. Suppose that $h(\theta)$ is 2π-periodic, twice differentiable function which satisfies (1.13). Show that the curve given by (1.12) is the boundary of a strictly convex region.

Exercise 1.2.7. * Find a characterizations of those functions which are shadow functions of convex regions without assuming that they are twice differentiable or that the region is strictly convex.

Exercise 1.2.8. If $h(\theta)$ is any differentiable function then equation (1.12) defines a curve, by plotting examples, determine what happens if the condition (1.13) is not satisfied.

Exercise 1.2.9. Suppose that h_D is a function satisfying (1.13). Show that the area enclosed by Γ_h is given by the

$$\text{Area}(D_h) = \frac{1}{2}\int_0^{2\pi} [(h(\theta))^2 - (h'(\theta))^2] d\theta.$$

Explain why this implies that a function satisfying (1.13) also satisfies the estimate

$$\int_0^{2\pi} (h'(\theta))^2 d\theta < \int_0^{2\pi} (h(\theta))^2 d\theta.$$

Exercise 1.2.10. Let h be a smooth 2π-periodic function which satisfies (1.13). Prove that the curvature of the boundary of the region with this shadow function, at the point $h(\theta)\omega(\theta) + h'(\theta)\hat{\omega}(\theta)$ is given by

$$\kappa(\theta) = \frac{1}{h(\theta) + h''(\theta)}. \tag{1.14}$$

Exercise 1.2.11. Suppose that h is a function satisfying (1.13). Show that another parametric representation for boundary of the region with this shadow function is

$$\theta \mapsto \left(-\int_0^\theta (h(s) + h''(s)) \sin(s) ds, \int_0^\theta (h(s) + h''(s)) \cos(s) ds \right).$$

Exercise 1.2.12. Which positive functions $\kappa(\theta)$ defined on S^1 are the curvatures of closed convex curves? Prove the following result: A positive function $\kappa(\theta)$ on S^1 is the curvature of a closed, strictly convex curve (parametrized by its tangent direction) if and only if

$$\int_0^\infty \frac{\sin(s) ds}{\kappa(s)} = 0 = \int_0^\infty \frac{\cos(s) ds}{\kappa(s)}.$$

Exercise 1.2.13. Let D be a convex region with shadow function h_D. For a vector $v \in \mathbb{R}^2$ define the translated region

$$D^v = \{(x, y) + v \ : \ (x, y) \in D\}.$$

Find the relation between h_D and h_{D^v}. Explain why this answer is inevitable in light of the formula (1.14), for the curvature.

Exercise 1.2.14. Let D be a convex region with shadow function h_D. For a rotation $A \in SO(2)$ define the rotated region

$$D^A = \{A(x, y) \ : \ (x, y) \in D\}.$$

Find the relation between h_D and h_{D^A}.

Exercise 1.2.15. * If h_1 and h_2 are 2π-periodic functions satisfying (1.13) then they are the shadow functions of convex regions D_1 and D_2. The sum, $h_1 + h_2$ also satisfies (1.13) and so is the shadow function of a convex region, D_3. Describe, geometrically how D_3 is determined by D_1 and D_2.

1.2.3 Approximate reconstructions

See: A.7.2.

In a realistic situation we can only make finitely many measurements. The shadow function is measured at a finite set of angles $\{h_D(\theta_1), \ldots, h_D(\theta_m)\}$. How can this data

be used to construct an approximation to the region D, which cast these shadows? We consider two different strategies both of which rely on the special geometric properties of convex regions. Recall that a convex region always lies in one of the half planes determined by the support line at any point of its boundary. The half plane lying "below" the oriented line $l_{t,\omega}$ is the set defined by

$$H^-_{t,\omega} = \{(x,y) \mid (x,y)\cdot\omega < t\}.$$

Since the boundary of D and $l_{h(\theta),\omega(\theta)}$ have the same orientation at the point of contact, it follows that D lies in each of the half planes

$$H^-_{h(\theta_j),\omega(\theta_j)}, \quad j = 1,\ldots,m.$$

As D lies in each of these half planes it also lies in their intersection. This defines a polygon

$$P_m = \bigcap_{j=1}^{m} H_{h(\theta_j),\omega(\theta_j)}.$$

P_m is a convex polygon which contains D, this then provides one sort of approximation for D from the measurement of a finite set of shadows. This is a stable approximation to D as small changes in the measurements of either the angles θ_j or the corresponding affine parameters $h(\theta_j)$ lead to small changes in the approximating polygon.

The difficulty with using the exact reconstruction formula (1.11) is that h is only known at finitely many values, $\{\theta_j\}$. From this information it is not possible to exactly compute the derivatives, $h'(\theta_j)$. We could use a finite difference approximation for the derivative to determine a finite set of points which approximate points on the boundary of D :

$$(x(\theta_j), y(\theta_j)) = h(\theta_j)\omega(\theta_j) + \frac{h(\theta_j) - h(\theta_{j+1})}{\theta_j - \theta_{j+1}}\hat{\omega}(\theta_j).$$

If the measurements were perfect, the boundary of D smooth and the numbers $\{|\theta_j - \theta_{j+1}|\}$ small then the finite difference approximations to $h'(\theta_j)$ would be accurate and these points would lie close to points on the boundary of D. Joining these points, in the given order gives a polygon, P' which approximates D. If the points could be computed exactly then P' would be contained in D. With approximate values this cannot be asserted with certainty, though under the assumptions above, P' should be largely contained within D.

This gives a different way to reconstruct an approximation to D from a finite set of measurements. This method is not as robust as the first technique because it requires the measured data to be differentiated. In order for the finite difference $\frac{h(\theta_j)-h(\theta_{j+1})}{\theta_j-\theta_{j+1}}$ to be a good approximation to $h'(\theta_j)$ it is generally necessary for $|\theta_j - \theta_{j+1}|$ to be small. Moreover the errors in the measurements of $h(\theta_j)$ and $h(\theta_{j+1})$ must also be small compared to $|\theta_j - \theta_{j+1}|$. This difficulty arises in solution of the reconstruction problem in X-ray CT, the exact reconstruction formula calls for the measured data to be differentiated. In general, measured data is corrupted by random noise, and random noise is usually "non-differentiable."

This means that measurements of a function must be regularized before they can used approximate derivatives. One way to handle this is to improve the quality of individual

measurements. One assumes that the errors in individual measurements have "mean zero:" if the same measurement is repeated many times then the average of the individual measurements should approach the true value. This is the approach taken in magnetic resonance imaging. Another possibility is to make a large number of measurements at closely spaced angles $\{(h_j, j\Delta\theta) \ : \ j = 1, \ldots, N\}$ which are then "averaged" to give less noisy approximations on a coarser grid. There are many ways to do the averaging. One way is to find a differentiable function, H belonging to a family of functions of dimension $M < N$ and minimizes the *square error*

$$e(H) = \sum_{j=1}^{N} (h_j - H(j\Delta\theta))^2.$$

For example H could be taken to be a polynomial of degree $M - 1$, or a continuously differentiable, piecewise cubic function. Using values of H one can find an approximation to the boundary of D which is hopefully less corrupted by noise. Fine structure in the boundary is also blurred by such a procedure. This is closer to the approach used in X-ray CT.

Exercise 1.2.16. Suppose that the angles $\{\theta_j\}$ can be measured exactly but there is an uncertainty of size ϵ in the measurement of the affine parameters, $h(\theta_j)$. Find a polygon $P_{m,\epsilon}$ which gives the best possible approximation to D which certainly contains D.

Exercise 1.2.17. Suppose that we know that $|h''(\theta)| < M$, and the measurement errors are bounded by $\epsilon > 0$. For what angle spacing is the error in using a finite difference approximation for h' due to the uncertainty in the measurements equal to that caused by the non-linearity of h itself.

1.2.4 Can an object be reconstructed from its width?

To measure the location of the shadow requires an expensive detector which can accurately locate a transition from light to dark. It would be much cheaper to build a device, similar to the exposure meter in a camera, to measure the length of the shadow region without determining its precise location. It is therefore an interesting question whether or not the boundary of a region can be reconstructed from measurements of the *widths* of its shadows. Let $w_D(\theta)$ denote the width of the shadow in direction θ, a moments consideration shows that

$$w_D(\theta) = h_D(\theta) + h_D(\theta + \pi).$$

From this formula it follows that w_D does not determine D. Let $e(\theta) \not\equiv 0$ be a function that satisfies

$$e(\theta) + e(\theta + \pi) = 0. \tag{1.15}$$

From the discussion in section 1.2.2 we know that if h_D has two derivatives such that $h_D'' + h_D > 0$ then $h_D(\theta)$ is the shadow function of a strictly convex region. Let e be a smooth function satisfying (1.15) such that

$$h_D'' + h_D + e'' + e > 0$$

as well, then $h_D + e$ is also the shadow function for a different strictly convex region. Observe that $h_D(\theta) + e(\theta)$ is the shadow function for a different region, D' which has the same *width* of shadow for each direction as D. That is

$$w_D(\theta) = (h_D(\theta) + e(\theta)) + (h_D(\theta + \pi) + e(\theta + \pi)) = w_{D'}(\theta).$$

To complete this discussion note that any function with a Fourier representation of the form

$$e(\theta) = \sum_{j=0}^{\infty} [a_j \sin(2j+1)\theta + b_j \cos(2j+1)\theta]$$

satisfies (1.15). This is an infinite dimensional space of functions. This implies that if $w_D(\theta)$ is the "width of the shadow" function for a convex region D then there is an infinite dimensional set of regions with the same "width of the shadow" function. Consequently the simpler measurement is inadequate to reconstruct the boundary of a convex region. The figure below shows the unit disk and another region which has constant "shadow width" equal to 2.

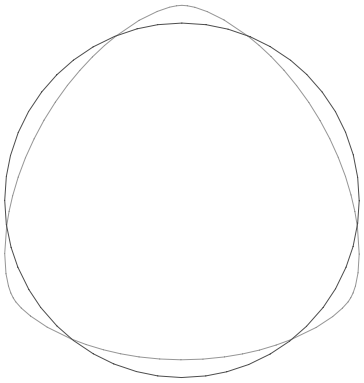


Figure 1.12: Two regions of constant width 2

Exercise 1.2.18. Show that the width function satisfies $w_D'' + w_D > 0$.

Exercise 1.2.19. Is it true that every twice differentiable, π-periodic function, w satisfying $w'' + w > 0$ is the width function of a convex domain?

Exercise 1.2.20. Our motivation for considering whether or not a convex body is determined by the width of its shadows was to replace our expensive detector, which can determine where a shadow begins and ends, with a less expensive detector. The cheaper detector can only measure the width of the covered region. Can you find a way to use a detector which only measures the length of an illuminated region to locate the edge of the shadow? Hint: Only cover half of the detector with photosensitive material.

1.3 Linearity

See: A.2.

As we have seen, using measurements to determine other quantities requires the solution of systems of equations. Sometimes we need to solve differential equations and sometimes algebraic equations. In almost all practical applications one is eventually reduced to solving systems of linear equations. This is true even for physical systems which are described by non-linear equations. Non-linear equations are usually solved iteratively where the iteration step involves the solution of linear equations. There are many reasons why linear equations and linear models are ubiquitous. From the pragmatic point of view, there is a "complete" mathematical theory for systems of linear equations. One has necessary and sufficient conditions for linear equations to have solutions, a description of the space of solutions when they exist and practical algorithms for finding them. This is not true even for the simplest families on non-linear equations. On a more conceptual level, for systems with some sort of intrinsic smoothness, a linear model often suffices to describe small deviations from an equilibrium state.

We give a quick overview of the theory of systems of linear, algebraic equations. This is not intended to serve as a text on linear algebra, merely a review of some important concepts. Detailed discussions of various aspect of this material can be found in [43] or [78].

Consider a system of m equations for n unknowns, $(x_1, \ldots, x_n)$:

$$\begin{array}{ccc} a_{11}x_1 + a_{12}x_2 + & \cdots + a_{1n}x_n & = y_1 \\ a_{21}x_1 + a_{22}x_2 + & \cdots + a_{2n}x_n & = y_2 \\ \vdots & \vdots & \vdots \\ a_{m1}x_1 + a_{m2}x_2 + & \cdots + a_{mn}x_n & = y_m. \end{array} \tag{1.16}$$

There are four questions which require answers:

Existence:
: For a given m-vector $(y_1, \ldots, y_m)$ does there exist an n-vector $(x_1, \ldots, x_n)$ which satisfies the equations in (1.16)?

Uniqueness:
: When a solution exists is it unique? More generally, describe the space of solutions.

Solve in practice:
: Give an algorithm to find approximations for the solutions of (1.16) and criteria to select a solution when there is more than one.

Stability:
: How sensitive is the solution to small variations in the coefficients (a_{ij}) or the right hand side (y_j)?

It is a somewhat unexpected, but very important fact that these issues are in practice, rather independent of one another.

Before proceeding with our analysis we first need to simplify the notation. It is very cumbersome to have to work with complicated expressions like (1.16), instead we use standard matrix and vector notation. Let $\mathbf{a}$ denote the $m \times n$ array of numbers $(a_{ij})_{i=1\ldots m, j=1\ldots n}$ and $\mathbf{x}$ denote the column n-vector. The result of multiplying $\mathbf{x}$ by $\mathbf{a}$ is a column m-vector whose i^{th} entry is

$$(\mathbf{ax})_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n.$$

The system of equations (1.16) is concisely expressed as

$$\mathbf{a}\mathbf{x} = \mathbf{y},$$

Here y is a column m-vector with entries $(y_1, \dots, y_m)$. We briefly recall the properties of matrix multiplication, let $\mathbf{x}^1$ and $\mathbf{x}^2$ be n-vectors, then

$$\mathbf{a}(\mathbf{x}^1 + \mathbf{x}^2) = \mathbf{a}\mathbf{x}^1 + \mathbf{a}\mathbf{x}^2$$

and for any number c

$$\mathbf{a}(c\mathbf{x}^1) = c(\mathbf{a}\mathbf{x}^1).$$

These are the properties that characterize linearity.

1.3.1 Solving linear equations

Suppose that $\mathbf{x}^0$ is a solution of the equation $\mathbf{a}\mathbf{x} = 0$ and $\mathbf{x}^1$ is a solution of the equation $\mathbf{a}\mathbf{x}^1 = \mathbf{y}$ then the rules above show that for any number c we have

$$\mathbf{a}(c\mathbf{x}^0 + \mathbf{x}^1) = c\mathbf{a}\mathbf{x}^0 + \mathbf{a}\mathbf{x}^1 = \mathbf{a}\mathbf{x}^1 = \mathbf{y}.$$

If $\mathbf{y} = 0$ as well then we conclude that the set of solutions to the equation

$$\mathbf{a}\mathbf{x} = 0$$

is a linear space, that is if $\mathbf{x}^0$ and $\mathbf{x}^1$ solve this equation then so does $\mathbf{x}^0 + \mathbf{x}^1$ as well as $c\mathbf{x}^0$, for any number c. This space is called the *null space or kernel*of $\mathbf{a}$. It is denoted by $\ker(\mathbf{a})$ and always contains, at least the zero vector $\mathbf{0} = (0, \dots, 0)$. These observations answer the question above about uniqueness.

Theorem 1.3.1. *Let* $\mathbf{a}$ *be an* $m \times n$ *matrix. Given a vector* $\mathbf{y}$, *if* $\mathbf{x}^1$ *satisfies* $\mathbf{a}\mathbf{x}^1 = \mathbf{y}$ *then every other solution to this equation is of the form* $\mathbf{x}^1 + \mathbf{x}^0$ *where* $\mathbf{x}^0 \in \ker(\mathbf{a})$. *Moreover, every vector of this form solves the equation* $\mathbf{a}\mathbf{x} = \mathbf{y}$.

As a simple corollary it follows that the solution of the equation $\mathbf{a}\mathbf{x} = \mathbf{y}$ is unique only if the null space of $\mathbf{a}$ contains only the 0-vector.

In order to answer the question of existence it is convenient to introduce the notion of a "dot" or inner product. If $\mathbf{x}$ and $\mathbf{y}$ are two n-vectors then define

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{j=1}^{n} x_j y_j = x \cdot y.$$

Suppose that $\mathbf{a}$ is an $m \times n$-matrix, $\mathbf{x}$ is an n-vector and $\mathbf{y}$ is an m-vector then $\mathbf{a}\mathbf{x}$ is an m-vector and

$$\langle \mathbf{a}\mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} y_i a_{ij} x_j.$$

The transpose of the matrix $\mathbf{a}$ is the $n \times m$ matrix $\mathbf{a}^t$ whose ij-entry is a_{ji}. From the previous formula it follows that that

$$\langle \mathbf{a}\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{a}^t \mathbf{y} \rangle.$$

Suppose that $\mathbf{y}$ is a non-zero vector in the null space of $\mathbf{a}^t$ (note that here we are using the transpose!) and the equation $\mathbf{ax} = \mathbf{b}$ has a solution. Using the calculations above we see that

$$\langle \mathbf{b}, \mathbf{y} \rangle = \langle \mathbf{ax}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{a}^t\mathbf{y} \rangle = 0.$$

The last equality follows from the fact that $\mathbf{a}^t\mathbf{y} = 0$. This gives a necessary condition for existence of a solution to the equation $\mathbf{ax} = \mathbf{b}$, the vector $\mathbf{b}$ must satisfy $\langle \mathbf{b}, \mathbf{y} \rangle = 0$ for every solution of the homogeneous equation $\mathbf{a}^t\mathbf{y} = 0$. This also turns out to be sufficient.

Theorem 1.3.2. *Let* $\mathbf{a}$ *be an* $m \times n$*-matrix and* $\mathbf{b}$ *and* m*-vector. The equation* $\mathbf{ax} = \mathbf{b}$ *has a solution if and only if*

$$\langle \mathbf{b}, \mathbf{y} \rangle = 0$$

for every vector $\mathbf{y}$ *satisfying the homogeneous equation*

$$\mathbf{a}^t\mathbf{y} = 0.$$

Putting these two results together we obtain that

Corollary 1.3.1. *Let* $\mathbf{a}$ *be an* $m \times n$*-matrix the equation* $\mathbf{ax} = \mathbf{b}$ *has a* **unique** *solution for any vector* $\mathbf{b}$ *if and only if* $\ker(\mathbf{a}) = \{0\}$ *and* $\ker(\mathbf{a}^t) = \{0\}$.

In a physical situation the vector $\mathbf{x}$ describes the state of a system and the entries of the vector $\mathbf{b}$ are results of measurements made on the system while it is in this state. The matrix $\mathbf{a}$ is a model for the measurement process: it is the assertion that if the physical object is described by the parameters, $\mathbf{x}$ then the results of the experiments performed should be the vector $\mathbf{b} = \mathbf{ax}$. This is a *linear model* because the map from the state of the system to the measurements is a linear map. The problem of determining the state of the system from the measurements is precisely the problem of solving this system of linear equations.

Example 1.3.1. Suppose we have a collection of photons sources, labeled by $1 \leq i \leq n$ and an array of detectors, labeled by$1 \leq j \leq m$. The matrix $\mathbf{P}$ has entries $0 \leq p_{ij} \leq 1$. The ij-entry is the probability that a particle emitted from source i is detected by detector j. Since a given photon can be detected by at most one detector it follows that

$$\sum_{j=1}^{m} p_{ij} \leq 1 \text{ for } i = 1, \ldots n.$$

If $d_j, j = 1, \ldots, m$ is the number of photons detected at detector j and $s_i, i = 1, \ldots, n$ is the number of photons emitted by source i then our model predicts that

$$\mathbf{Ps} = \mathbf{d}.$$

If $m = n$ and $\mathbf{P}$ is an invertible matrix then we can use the measurements $\mathbf{d}$ to obtain a unique vector $\mathbf{s}$. Since the model is probabilistic this should be regarded as an expected value for the distribution of sources. If $m > n$ then we have more measurements than unknowns, so any measurement errors or flaws in the model could make it impossible to find a vector $\mathbf{s}$ so that $\mathbf{Ps} = \mathbf{d}$. This is a frequent situation in image reconstruction problems. One

chooses a way to measure the error, usually a function of the form $e(\mathbf{Ps} - \mathbf{d})$ and seeks a vector s which minimizes the error. Finally we may have more sources than detectors. The measurements are then inadequate, in principle to determine their distribution. This is also a common circumstance in image reconstruction problems and is resolved by making some a priori assumptions about the allowable distribution of sources to obtain a determined (or even overdetermined) problem.

As illustrated by this example and explained in the theorem there are essentially 3 types of linear models for systems with finitely many degrees of freedom.

Determined:

The simplest case arises when the number of *independent* measurements and parameters describing the state of the system are the same. This implies that $n = m$. In this case the measurements uniquely determine the state of the system. Mathematically we say that the matrix, $\mathbf{a}$ is invertible. In the situation that $n = m$ this is equivalent to the statement that the homogeneous equation, $\mathbf{ax} = 0$ has only the trivial solution, $\mathbf{x} = 0$. The inverse matrix is denoted by $\mathbf{a}^{-1}$, it is both a left and a right inverse to $\mathbf{a}$,

$$\mathbf{a}^{-1}\mathbf{a} = \mathrm{Id}_n = \mathbf{a}\mathbf{a}^{-1}.$$

Here Id_n denotes the $n \times n$ identity matrix, that in

$$(\mathrm{Id}_n)_{ij} = \begin{cases} 1 \text{ if } i = j, \\ 0 \text{ if } i \neq j. \end{cases}$$

From the mathematical point of view, the unique solution is obtained by setting

$$\mathbf{x} = \mathbf{a}^{-1}\mathbf{y}.$$

Except in special cases, the inverse matrix $\mathbf{a}^{-1}$ is not computed directly.

Overdetermined:

In this case we have more measurements than parameters, i.e. $m > n$. If the model and measurements are perfect then there should be a unique $\mathbf{x}$ with $\mathbf{ax} = \mathbf{y}$. In general, neither is true and there will not exist any $\mathbf{x}$ exactly satisfying this equation. Having more measurements than parameters can be used to advantage in several different ways. In example 1.3.2 we explain how to use the conditions for solvability given in Theorem 1.3.2 to determine physical parameters. Often times measurements are noisy. A model for the noise in the measurements can be used to select a criterion for a "best approximate solution." The error function is usually defined by picking a norm $\|\cdot\|$ on the space of measurements. We then try to find the vector $\mathbf{x}$ which minimizes the error, $e(\mathbf{x}) = \|\mathbf{ax} - \mathbf{y}\|$. The most commonly used error function is that defined by the *square norm*

$$\|\mathbf{y}\|_2^2 = \sum_{j=1}^{m} y_j^2.$$

There are two reasons why this measure of the error is often employed: 1. It is a natural choice if the noise is normally distributed, 2. The problem of minimizing $\|\mathbf{ax} - \mathbf{y}\|_2$ can be reduced to the problem of solving a system of linear equations.

Underdetermined:

Most of the problems in image reconstruction are underdetermined, that is we do not have enough data to uniquely determine a solution. In mathematical tomography a "perfect reconstruction" requires an infinite number of exact measurements. These are, of course never available. In a linear algebra problem, this is the case where $m < n$. When the measurements $\mathbf{y}$ do not uniquely determine the state $\mathbf{x}$, additional criteria are needed to determine which solution to actually use, for example one might use the solution to $\mathbf{ax} = \mathbf{y}$ which is of smallest norm. Another approach is to assume that $\mathbf{x}$ belongs to a subspace whose dimension is equal to the number of independent measurements. Both of these approaches are used in medical imaging.

Example 1.3.2. In the refraction problem considered in example 1.1.4 we remarked that the refractive index of the lower fluid n_2 could be determined by an additional measurement. Suppose that we shine a beam of light in at a different angle, so that the upper angle is ϕ_1 and the lower angle is ϕ_2. This light beam is displaced by l_2 as it passes through the fluid. We now have 3 equations for the two unknowns:

$$\begin{pmatrix} 1 & 1 \\ \tan(\theta_1) & \tan(\theta_2) \\ \tan(\phi_1) & \tan(\phi_2) \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = \begin{pmatrix} h \\ l_1 \\ l_2 \end{pmatrix}. \tag{1.17}$$

In order for this equation to have a solution the measurements (h, l_1, l_2) must satisfy the condition

$$\begin{pmatrix} 1 \\ \tan(\theta_1) \\ \tan(\phi_1) \end{pmatrix} \times \begin{pmatrix} 1 \\ \tan(\theta_2) \\ \tan(\phi_2) \end{pmatrix} \cdot \begin{pmatrix} h \\ l_1 \\ l_2 \end{pmatrix} = 0.$$

Here $\times$ is the vector cross product. Since

$$\frac{\sin(\theta_1)}{\sin(\theta_2)} = \frac{\sin(\phi_1)}{\sin(\phi_2)} = \frac{n_2}{n_1}$$

and the angles θ_1 and ϕ_1 as well as (h, l_1, l_2) are assumed known, this solvability conditions gives a non-linear equation which allows the determination of $\frac{n_2}{n_1}$ from the measured data.

This brings us to the final questions of giving practical algorithms for finding the solutions of linear equations and their stability. We leave a detailed discussion of algorithms to later chapters for, in practice one needs to select a method that is well adapted to the equations under consideration. We conclude this section with a discussion of the problem of stability. In practical situations many things conspire to limit the accuracy achievable in using measurements to predict the state of a system. There are errors in the model itself as well as noise in the actual measurements. Once the model is made and the measurements are taken one needs to solve systems of equations. In finite time one can only work with numbers having finitely many decimal places, so rounding errors are also an unavoidable problem.

Given all these sources of error one would like to have a *stable* algorithm for solving the system of equations. Suppose that $\mathbf{a}$ is an $n \times n$ invertible matrix that models a

measurement process. If $\mathbf{x}_1$ and $\mathbf{x}_2$ are two states of our system then, because the model is linear the difference in the measurements can easily be computed

$$\mathbf{y}_1 - \mathbf{y}_2 = \mathbf{a}\mathbf{x}_1 - \mathbf{a}\mathbf{x}_2 = \mathbf{a}(\mathbf{x}_1 - \mathbf{x}_2).$$

From this formula we see that nearby states result in nearby measurements. However the reverse is often not true. There may exist states $\mathbf{x}_1$ and $\mathbf{x}_2$ which are not nearby, in the sense that $\|\mathbf{x}_1 - \mathbf{x}_2\|$ is large but $\|\mathbf{a}(\mathbf{x}_1 - \mathbf{x}_2)\|$ is small. Physically, the measurements performed are not sufficiently independent to distinguish certain pairs of states, which are not, in fact very close together. In numerical analysis this is known as an *ill-conditioned* equation. Briefly, a small error in the measurement process can be magnified by applying $\mathbf{a}^{-1}$ to the measurement vector. For an ill-conditioned problem even a good algorithm for solving linear equations can produce meaningless results.

Example 1.3.3. For example, consider the system with $m = n = 2$ and

$$\mathbf{a} = \begin{pmatrix} 1 & 0 \\ 1 & 10^{-5} \end{pmatrix}.$$

Then $\mathbf{x}$ is given by $\mathbf{a}^{-1}\mathbf{y}$ where

$$\mathbf{a} = \begin{pmatrix} 1 & 0 \\ -10^5 & 10^5 \end{pmatrix}.$$

If the actual data is $\mathbf{y} = (1, 1)$ but we make an error in measurement and measure, $\mathbf{y}_m = (1, 1 + \epsilon)$ then the relative error is

$$\frac{|\mathbf{y}_m - \mathbf{y}|}{|\mathbf{y}|} = \epsilon 10^5.$$

Even though the measurements uniquely determine the state of the system, a small error in measurement is vastly amplified.

In image reconstruction the practical problems of solving systems of linear equations are considerable. It is not uncommon to have $10,000$-equations in $10,000$-unknowns. These huge systems arise as finite dimensional approximations to linear equations for functions of continuous variables. We close this section with a short discussion of linear algebra in in infinite dimensional spaces. This is a theme which occupies a large part of this book.

Exercise 1.3.1. Let $\mathbf{a}$ be an $m \times n$ matrix. Show that if $\ker \mathbf{a} = \ker \mathbf{a}^t = \mathbf{0}$ then $n = m$. Is the converse true?

Exercise 1.3.2. Suppose that the state of a system is described by the vector $\mathbf{x}$. The measurements are modeled as inner products $\{\mathbf{a}_j \cdot \mathbf{x} : j = 1, \ldots, m\}$. However the measurements are noisy and each is repeated m_j times leading to measured values $\{y_j^1, \ldots, y_j^{m_j}\}$. Define an error function by

$$e(\mathbf{x}) = \sum_{j=1}^{m} \sum_{i=1}^{m_j} (\mathbf{a}_j \cdot \mathbf{x} - y_j^i)^2.$$

Show that $e(\mathbf{x})$ is minimized by the vector which satisfies the averaged equations

$$\mathbf{a}_j \cdot \mathbf{x} = \frac{1}{m_j} \sum_{i=1}^{m_j} y_j^i.$$

1.3.2 Infinite dimensional linear algebra

The state of a 'system' in medical imaging is described by a function of continuous variables. In this introductory section we consider real valued functions defined on the real line. Let $f(x)$ describe the state of the system. A linear measurement of the state is usually described as an integral

$$\mathcal{M}(f)(x) = \int_{-\infty}^{\infty} m(x,y)f(y)dy.$$

Here $m(x,y)$ is a function on $\mathbb{R} \times \mathbb{R}$ which provides a model for the measurement process. It can be thought of as an infinite 'matrix' with indices x and y. A linear transformation of an infinite dimensional space is called a *linear operator*. A linear transformation which can be expressed as an integral is called an *integral operator*.

Suppose that the function $g(x)$ is the output of the measurement process, to reconstruct f means solving the linear equation

$$\mathcal{M}f = g.$$

This is a concise way to write a system of infinitely many equations in infinitely many unknowns. Theorems 1.3.1 and 1.3.2 contain the complete theory for the existence and uniqueness of solutions to linear equations in finitely many variables. These theorems are entirely algebraic in character. No such theory exists for equations in infinitely many variables. It is usually a very complicated problem to describe both the domain and range of such a transformation. We close this section with a few illustrative examples.

Example 1.3.4. Perhaps the simplest linear operator is the indefinite integral

$$\mathcal{I}(f)(x) = \int_{0}^{x} f(y)dy.$$

If we use the continuous functions on $\mathbb{R}$ as the domain of $\mathcal{I}$ then every function in the range is continuously differentiable. Moreover the null-space of $\mathcal{I}$ is the zero function. Observe that the domain and range of $\mathcal{I}$ are fundamentally different spaces. Because $\mathcal{I}(f)(0) = 0$ not every continuously differentiable function is in the range of $\mathcal{I}$. The derivative is a left inverse to $\mathcal{I}$ as the Fundamental Theorem of Calculus states that if f is continuous then

$$\frac{d}{dx} \circ \mathcal{I}(f)(x) = f(x).$$

On the other hand it is not quite a right inverse because

$$\mathcal{I}(\frac{df}{dx})(x) = f(x) - f(0).$$

The domain of $\mathcal{I}$ can be enlarged to include all *locally integrable* functions. These are functions such that

$$\int_{0}^{x} |f(y)|dy < \infty$$

for every $x \in \mathbb{R}$. Enlarging the domain also enlarges the range. For example the function $|x|$ lies in the enlarged range of $\mathcal{I}$,

$$|x| = \int_0^x \text{sign}(y)dy,$$

where $\text{sign}(y) = 1$ if $y \geq 0$ and -1 if $y < 0$. Even though $|x|$ is not differentiable at $x = 0$ it is still the indefinite integral of a locally integrable function, however the formula

$$\frac{d|x|}{dx} = \text{sign}(x)$$

does not make sense at $x = 0$.

Example 1.3.5. Changing the lower limit of integration to $-\infty$ leads to a very different sort of linear transformation. Initially $\mathcal{I}_\infty$ is defined for continuous functions f, vanishing for sufficiently negative x by

$$\mathcal{I}_\infty(f)(x) = \int_{-\infty}^x f(y)dy.$$

Once again the null-space of $\mathcal{I}_\infty$ consists of the zero function alone. The domain can be enlarged to include locally integrable functions such that

$$\lim_{R\to\infty} \int_{-R}^0 |f(y)|dy < \infty. \tag{1.18}$$

If f is continuous then we can apply the F.T.C. to obtain

$$\frac{d}{dx} \circ \mathcal{I}(f) = f.$$

If a function g belongs to the range of $\mathcal{I}$ then

$$\lim_{x\to-\infty} g(x) = 0. \tag{1.19}$$

There are once differentiable functions satisfying this condition which do not belong to the range of $\mathcal{I}_\infty$. For example,

$$f(x) = \frac{x\cos x - \sin x}{x^2} = \frac{d}{dx}\frac{\sin x}{x}$$

satisfies (1.19) but $\frac{\cos x}{x}$ does not satisfy (1.18). With the domain defined by (1.18) the precise range of $\mathcal{I}_\infty$ is rather difficult to describe.

This example illustrates how a integral operator may have a simple definition on a certain domain, which by a limiting process can be extended to a larger domain. The domain of such an operator is often characterized by a size condition like (1.18).

Example 1.3.6. A real physical measurement is always some sort of an average. If the state of the system is described by a function f of a single variable x then the average of f over an interval of length 2δ is

$$\mathcal{M}_\delta(f)(x) = \frac{1}{2\delta}\int_{x-\delta}^{x+\delta} f(y)dy.$$

A natural domain for $\mathcal{M}_\delta$ is all locally integrable functions. To what extent is f determined by $\mathcal{M}_\delta(f)$? Suppose that f and g are two states, then, because the integral is linear

$$\mathcal{M}_\delta(f) - \mathcal{M}_\delta(g) = \mathcal{M}_\delta(f-g).$$

The extent to which $\mathcal{M}_\delta(f)$ determines f is characterized by the null-space of $\mathcal{M}_\delta$,

$$\mathcal{N}_\delta = \{f \,:\, \mathcal{M}_\delta(f) = 0\}.$$

Proceeding formally, we can differentiate $\mathcal{M}_\delta(f)$ to obtain

$$\frac{d\mathcal{M}_\delta(f)}{dx} = f(x+\delta) - f(x-\delta) \tag{1.20}$$

If $f \in \mathcal{N}_\delta$ then $\mathcal{M}_\delta(f)$ is surely constant and therefore

$$f \in \mathcal{N}_\delta \Rightarrow f(x+\delta) - f(x-\delta) = 0,$$

in other words f is periodic with periodic 2δ. A periodic function has an expansion in terms of sines and cosines, that is

$$f(x) = a_0 + \sum_{j=1}^{\infty}\left[a_j\cos(\frac{\pi jx}{\delta}) + b_j\sin(\frac{\pi jx}{\delta})\right].$$

If $a_0 = 0$ then $\mathcal{M}_\delta(f) = 0$. This shows that the null-space of $\mathcal{M}_\delta$ is infinite dimensional.

In applications one often has additional information about the state of the system, for example one might know that

$$\lim_{|x|\to\infty} f(x) = 0. \tag{1.21}$$

A periodic function that tends to zero at infinity must be identically zero, so among such functions the measurements $\mathcal{M}_\delta(f)$ would appear to determine f completely. To *prove* this statement we need to know somewhat more about f than (1.21). With a more quantitative condition like

$$\|f\|_p = \left[\int_{-\infty}^{\infty}|f(y)|^p dy\right]^{\frac{1}{p}} < \infty, \tag{1.22}$$

for a p between 1 and 2, it is possible to show that $\mathcal{M}_\delta(f) = 0$ implies that $f = 0$. For such functions the measurement $\mathcal{M}_\delta(f)$ uniquely determines f. However, f cannot be *stably* reconstructed from $\mathcal{M}_\delta(f)$. A small error in measurement can lead to a very large error in the reconstructed state.

The integral in (1.22) defines a measure for the size of f called the L^p-norm. It is a generalization of the notion of a norm on a finite dimensional vector space and satisfies the familiar conditions for a norm:

$$\|af\|_p = |a|\|f\|_p \text{ and } \|f+g\|_p \leq \|f\|_p + \|g\|_p.$$

The first step in analyzing linear transformations of infinite dimensional spaces is the introduction of norms on the domain and range. This was not necessary in finite dimensions but is absolutely essential in the infinite dimensional case. In medical image reconstruction there is a small list of linear transformations that are very important, the Fourier transform, Radon transform and Abel transform. A large part of this text is devoted to the analysis of these operators.

Exercise 1.3.3. Prove that the null-space of $\mathcal{I}$ acting on $\mathcal{C}^0(\mathbb{R})$ is the zero function.

1.4 Conclusion

By examining a large collection of examples we have seen how physical systems can be described using mathematical models. The models suggest measurements which one can make to determine the state of the system. It is important to keep in mind that mathematical models are just that, models, often toy models. A good model must satisfy two opposing requirements: the model should accurately depict the system under study while at the same time being simple enough to be usable. Which models are "simple enough to be useful" depends on what you know, one of our goals, in the succeeding chapters is to develop some sophisticated mathematical tools to work with models. The workhorse throughout this book and in most applications of mathematics to problems in measurement and signal processing is the Fourier transform.

The models used in medical imaging usually involve infinitely many degrees of freedom. The state of the system is described by a function of continuous variables. Ultimately of course only a finite number of measurements can be made and only a finite amount of time is available to process them. Our analysis of the reconstruction process in X-ray CT passes through several stages, beginning with a description of the complete, perfect data situation and concluding with an analysis of the effects of noise on the quality of an approximate image, reconstructed from finitely many measurements.

In mathematics, problems of determining the state of a physical system from feasible measurements are gathered under the rubric of *inverse problems*. The division of problems into inverse problems and *direct problems* is often a matter of history. Usually a physical theory which models how the state of the system determines feasible measurements preceded a description of the inverse process: how the state can be determined from measurements. Example 1.1.6 is typical though very simple example. Formula (1.22) describes the solution to the direct problem: the determination of the transit time from a knowledge of the sound speed *and* the depth. The inverse problem asks for a determination of the depth from a knowledge of the sound speed and the transit time. While many of the problems which arise in medical imaging are considered to be inverse problems, we do not give any systematic development of this subject. The curious reader is referred to the very nice article by Joe Keller which contains analyses of many classical inverse problems, see [41].

Chapter 2

A basic model for tomography

We begin our study of medical imaging with a purely mathematical model of the image reconstruction process used in transmission CT. The model begins with a very simplified description of the interaction of X-rays with matter. The physical properties of an object are encoded in a function μ, called the absorption coefficient which quantifies the tendency of an object to absorb X-rays. The mathematical model describes idealized measurements that can be made of certain averages of μ. In mathematical terms, these measurements are described as an integral transform. Assuming that a complete, error free set of measurements can be made, the function μ can be reconstructed. In reality, the data collected is a very limited part of the mathematically "necessary" data and the measurements are subject to a variety of errors. In later chapters we refine the measurement model and reconstruction algorithm to reflect more realistic models of the physical properties of X-rays and the data which is actually collected.

2.1 Tomography

Literally, tomography means *slice imaging*. It is collection of methods for reconstructing a three dimensional object from its two dimensional slices. The objects of interest in medical imaging are described by functions defined on $\mathbb{R}^3$. The function of interest in X-ray tomography is called the *absorption coefficient*; it quantifies the tendency of an object to absorb X-rays. This function varies from point-to-point within the object and is usually taken to vanish outside it. The absorption coefficient is like density, in that it is non-negative. It is useful for medical imaging because different anatomical structures have different absorption coefficients. Bone has a much higher absorption coefficient than soft tissue and different soft tissues have slightly different coefficients. For medical applications it is crucial that normal and cancerous tissues also have slightly different absorption coefficients. However the absorption coefficients of different soft tissues vary over a very small range. Table 2.1 lists typical absorption coefficients for different parts of the body, these are given in *Hounsfield units*. This is a dimensionless quantity defined by comparison with the absorption coefficient of water,

$$H_{\text{tissue}} = \frac{\mu_{\text{tissue}} - \mu_{\text{water}}}{\mu_{\text{water}}} \times 1000.$$

Material	**Absorption coefficient in Hounsfield units**
water	0
air	-1000
bone	1086
blood	53
fat	-61
brain white/gray	-4
breast tissue	9
muscle	41
soft tissue	51

Table 2.1: Absorption coefficients of human tissues for 100keV X-rays, adapted from [31].

The absorption coefficients of air (-1000) and bone (1100) define the range of values present in a typical clinical situation. This means that the *dynamic range* of a clinical CT-measurement is about 2000 Hounsfield units. From the table it is apparent that the variation in the absorption coefficients of soft tissues is about 2% of this range. For X-ray CT to be clinically useful this means that the reconstruction of the absorption coefficient must be accurate to about 10 Hounsfield units or less than a half a percent.

The fundamental idea of the "tomographic method" is that if we know enough two dimensional slices of a function of three variables then we can reconstruct the original function. Let $\mu(\mathbf{x})$ be a function defined on $\mathbb{R}^3$. To define the slices we need to fix a coordinate system $\mathbf{x} = (x_1, x_2, x_3)$. For each fixed value of $x_3 = c$, the x_3-slice of μ is the function of two variables

$$f_c(x_1, x_2) = \mu(x_1, x_2, c).$$

A knowledge of the collection of functions $\{f_c(x_1, x_2) \ : \ c \in [a, b]\}$ is equivalent to a knowledge of $\mu(\mathbf{x})$ for all $\mathbf{x}$ in the set

$$\{(x_1, x_2, x_3) \ : \ -\infty < x_1 < \infty, -\infty < x_2 < \infty, a \leq x_3 \leq b\}.$$

The choice of coordinates is arbitrary, but having a fixed frame of reference is a crucial element of any tomographic method. By convention the slices are defined by fixing the last coordinate. If (x_1', x_2', x_3') is a different coordinate system then the slices would be defined as

$$f_{c'} = \mu(x_1', x_2', c').$$

In general, different coordinate systems lead to different collections of slices. In actual practice the X-ray machine itself fixes the frame of reference.

Example 2.1.1. A simple collection of objects are the subsets of $\mathbb{R}^3$. A subset D defines and is defined by its characteristic function

$$\chi_D(x) = \begin{cases} 1 & \text{if } x \in D, \\ 0 & \text{if } x \notin D. \end{cases}$$

This models an object with constant absorption coefficient. In this case the object is determined by its intersection with the planes

$$H_c = \{(x_1, x_2, c) \ : \ x_1, x_2 \in \mathbb{R}\}.$$

For each c we let $D_c = D \cap H_c$. Figure 2.1 shows sample slices of a 3-dimensional object.

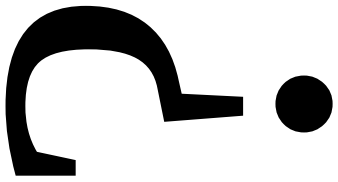


Figure 2.1: Parallel slices of an object.

Example 2.1.2. Suppose that the object is contained in the ball of radius 1 and its absorption coefficient is

$$\mu(\mathbf{x}) = \begin{cases} 1 - \|\mathbf{x}\| & \text{if } \|\mathbf{x}\| \leq 1, \\ 0 & \text{if } \|\mathbf{x}\| > 1. \end{cases}$$

The slices of μ are the functions

$$f_c(x_1, x_2) = \begin{cases} 1 - \sqrt{x_1^2 + x_2^2 + c^2} & \text{if } \sqrt{x_1^2 + x_2^2 + c^2} \leq 1, \\ 0 & \text{if } \sqrt{x_1^2 + x_2^2 + c^2} > 1. \end{cases}$$

Note in particular that if $|c| > 1$ then $f_c \equiv 0$.

For the purposes of medical imaging, air is usually assumed to be transparent to X-rays. This means that the absorption coefficient $\mu(\mathbf{x})$ equals zero outside of the patient. The set where $\mu \neq 0$ can therefore be determined non-destructively. Roughly speaking the *support* of a function is the set of points where the function does not vanish. To get a useful concept we need to add points that are very near to points where the function is non-zero. As the definition is the same in all dimensions we give it for functions defined on $\mathbb{R}^n$.

Definition 2.1.1. Let $f(\mathbf{x})$ be a function defined on $\mathbb{R}^n$. A point $\mathbf{x}$ belong to the *support of* f if there is a sequence of points $< \mathbf{x}_n >$ such that

(1). $f(\mathbf{x}_n) \neq 0$,

(2). $\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}$.

This set is denoted by $\text{supp}(f)$.

Example 2.1.3. The support of the function $f(x) = x$ is the whole real line, even though $f(0) = 0$. The support of the function $f(x, y) = xy$ is the whole plane, even though $f(0, y) = f(x, 0) = 0$. The support of the function $\chi_{(0,1)}(x)$ is $[0, 1]$.

Definition 2.1.2. A function f defined in $\mathbb{R}^n$ is said to have *bounded support* if there is an R so that $f(\mathbf{x}) = 0$ if $\|\mathbf{x}\| > R$. In this case one says that *the support of f is contained in the ball of radius R.*

Most functions in medical imaging have bounded support.

2.1.1 Beer's law and X-ray tomography

We now turn our attention to a simple model of X-ray tomography. This refers to the usage of X-rays to reconstruct a 3-dimensional object by reconstructing its two dimensional slices. X-rays are a very high energy form of electro-magnetic radiation. An X-ray beam can be modeled as a continuous flux of energy. In truth, an X-ray beam is composed of a large number of discrete particles called photons. Each photon has a well defined energy which is often quoted in units of *electron-volts*. In the last chapter of this book we consider the implications of this fact. For the present we use a continuum model and a simplified, though adequate description of the interaction of the X-ray beam with matter. Three physical assumptions are used in the construction of this model:

No refraction or diffraction:
: The X-rays beam travel along straight lines which are not "bent" by the objects they pass through. This is a good approximation to the truth, because X-ray photons have very high energy.

The X-rays used are monochromatic:
: All the photons making up the X-ray beam have the same energy.

Beer's law:
: Each material encountered has a characteristic linear absorption coefficient $\mu(\mathbf{x})$ for X-rays of the given energy. The intensity, $I(\mathbf{x})$ of the X-ray beam satisfies Beer's law

$$\frac{dI}{ds} = -\mu(\mathbf{x})I. \tag{2.1}$$

Here s is the arclength along the straight-line trajectory of an X-ray beam.

Ordinary light is also electro-magnetic radiation, composed of discrete photons. We experience the energy of a photon as the color of the light. The second assumption is that the X-ray beam is "all of one color." This is not a very realistic assumption, but it is very important in the construction of a *linear* model for the measurements. The implications of the failure of this assumption are discussed later in this chapter.

Beer's law requires some explanation. Suppose that an X-ray beam encounters an object. Beer's law describes how the presence of the object affects the intensity of the beam. For the moment suppose that we live in a 1-dimensional world, with s a coordinate in our world. The X-ray beam is described by its intensity $I(s)$, often quoted in units

of electron-volts/sec. Beer's law predicts the change in the flux due to the material lying between s and $s + \Delta s$:

$$I(s + \Delta s) - I(s) \approx -\mu(s)I(s)\Delta s.$$

Think, now of the flux as being composed of a large number, $N(s)$ photons/second, each of the given energy, Beer's law can be rewritten in these terms as

$$N(s + \Delta s) - N(s) \approx -\mu(s)N(s)\Delta s.$$

In this formulation it is clear that $\mu(s)\Delta s$ can be regarded as giving the probability that a photon incident on the material at coordinate s is absorbed. Implicit in Beer's Law is the assumption that X-rays travel along straight lines, it is an essentially 1-dimensional relation. It implicitly asserts that the absorption of X-rays is an isotropic process: it does not depend on the direction of the line along which the X-ray travels.

Because of its one dimensional character, Beer's law is easily applied to 2 and 3-dimensional problems. The 3-dimensional X-ray beam is modeled as a collection of 1-dimensional beams. Beer's law describes how the material attenuates each of these 1-dimensional beams. Suppose that one of the X-ray beams is traveling along the line $l_{t,\omega}$ sitting in the plane. The function

$$i(s) = I(t\omega + s\hat{\omega})$$

gives the intensity of the beam at points along this line and

$$m(s) = \mu(t\omega + s\hat{\omega})$$

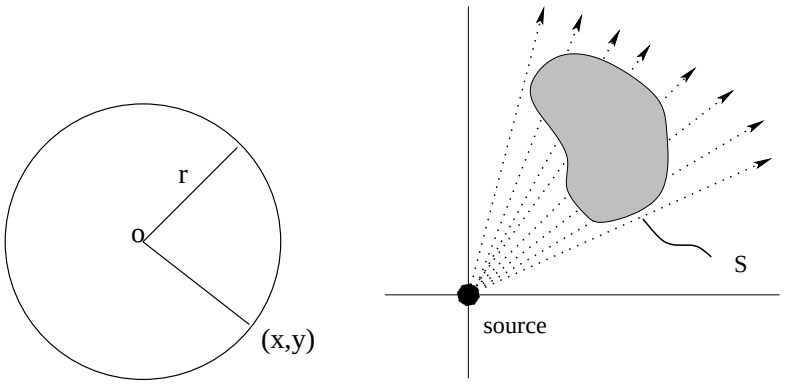
gives the absorption coefficient. Beer's law state that

$$\frac{di}{ds} = -m(s)i(s) \text{ or } \frac{d(\log i)}{ds} = -m(s).$$

Integrating this equation from $s = a$ to $s = b$ gives

$$\log\left[\frac{i(b)}{i(a)}\right] = -\int_a^b m(s)ds.$$

Example 2.1.4. We consider a two dimensional example. Assume that we have a point source of X-rays of intensity I_0 in the plane, see figure 2.2(a). The X-ray source is isotropic which means the outgoing flux is the same in all directions.



(a) A point source of X-rays

(b) The flux through a curve not enclosing the source is zero.

Figure 2.2: Analysis of an isotropic point source.

Because the source is isotropic, the intensity of the beam is only a function of the distance to the source. Let $I(r)$ denote the intensity of the flux at distance r from the source, by the conservation of energy,

$$I_0 = \int\limits_{x^2+y^2=r^2} I(r)ds = 2\pi r I(r). \tag{2.2}$$

The intensity of the beam at distance r from the source is therefore

$$I(r) = \frac{I_0}{2\pi r}. \tag{2.3}$$

Model the X-rays as beams traveling along the rays passing through the source. If I_0 is measured in units of electron-volts/second then $I(r)$ has units electron-volts/(cm$\times$second).

Fixing coordinates so that the source is placed at $(0,0)$, the X-ray flux at a point (x,y) travels along the line from (x,y) to $(0,0)$ and is given by

$$\begin{aligned} I_f(x,y) = I(r)\frac{(x,y)}{\sqrt{x^2+y^2}} = I_0\frac{\hat{r}}{2\pi r}, \\ \text{where } \hat{r} = \frac{(x,y)}{r}, \quad r = \sqrt{x^2+y^2}. \end{aligned} \tag{2.4}$$

If a curve S does not enclose the source then conservation of energy implies that

$$\int\limits_S I_f(x,y)\hat{r}\cdot\hat{n}ds = 0, \tag{2.5}$$

here $\hat{n}$ is the outward normal vector to S. As the curve encloses no sources or sinks, the line integral of the flux is zero: everything which comes into this surface has to go out.

For a point source the intensity of the rays diminish as you move away from the source; this is called "beam spreading." Beer's law can be used to model this effect. Let $\mu_s(x, y)$ denote the attenuation coefficient which accounts for the spreading of the beam. As a guess we let $\mu_s = 1/r$ and see that

$$\frac{dI}{dr} = -\frac{1}{r}I \Rightarrow \frac{d\log I}{dr} = -\frac{1}{r}. \tag{2.6}$$

Integrating equation (2.6) from an $r_0 > 0$ to r gives

$$I(r) = I(r_0)\frac{r_0}{r}.$$

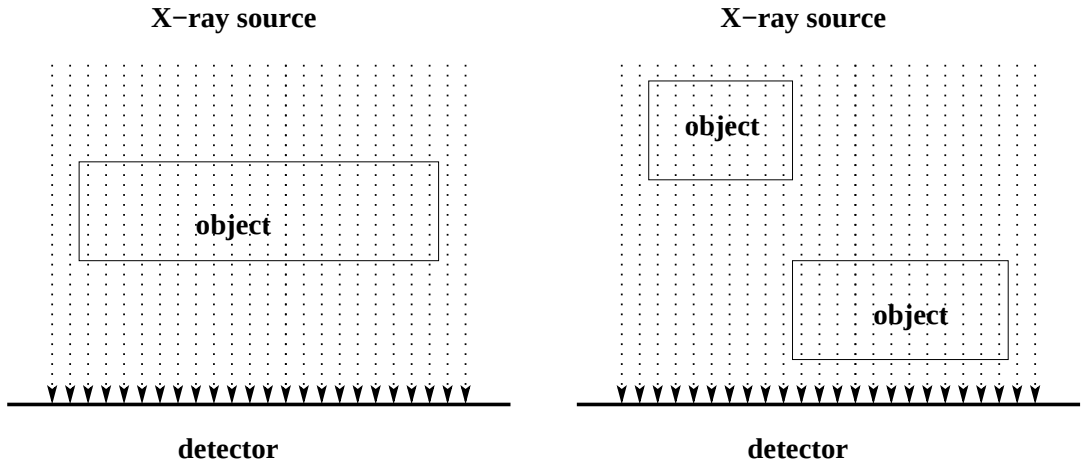
This agrees with (2.1.4). That we cannot integrate down to $r = 0$ is a reflection of the non-physical nature of a point source.

In X-ray tomography, the beam is often assumed to be non-diverging, i.e. the attenuation of the beam due to beam spreading is sufficiently small, compared to the attenuation due to absorption by the object that it can be ignored. In a real measurement, the X-ray source is turned on for an known period of time and the total incident energy I_i is known. The total energy, I_o emerging from the object, along a given line, l is then measured by an X-ray detector. If the X-ray beam is traveling along the line l then Beer's law states that

$$\log \frac{I_o}{I_i} = -\int_l \mu(\mathbf{x})ds. \tag{2.7}$$

Here ds is the arc length parameter along the straight line path l. A perfect measurement of the ratio I_o/I_i would therefore furnish the line integral of the attenuation coefficient along the line l. In this approximation it is precisely these line integrals that can be measured.

An ordinary X-ray image is formed by sending a beam of X-rays through an object, the detector is often a sheet of photographic film. Suppose that the X-rays travel along parallel lines, passing through an object before arriving on a photographic plate as shown in the figure 2.3. By measuring the density of the exposed film we can determine the intensity of the X-ray beam at the surface of the film. More absorbent parts of an object result in fewer X-rays at the surface of the film. If the intensity of the incident beam is known then the density of the film can be used to determine the integrals of the absorption coefficient along this family of parallel lines. The result is a "projection" or shadow of the object. The shadows of the objects in figures 2.3(a) and (b) are the same so it is not possible to distinguish between them using this projection. Placing the X-ray source and detector at a different angle gives different measurements. The measurement in figure 2.4 distinguishes between the objects in figures 2.3(a) and (b).



(a) One object.

(b) Two objects.

Figure 2.3: The failure of ordinary X-rays to distinguish objects.

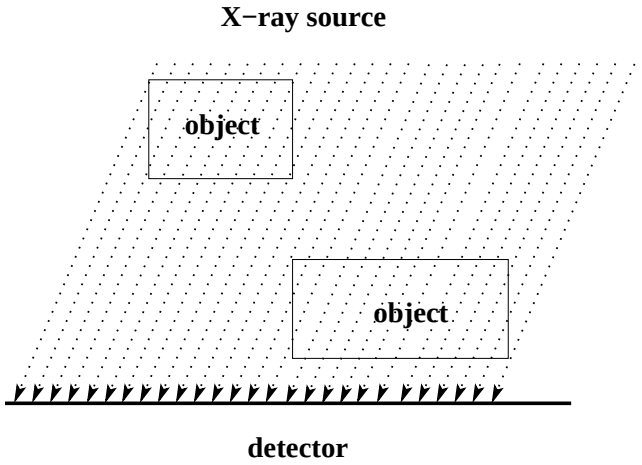


Figure 2.4: A different projection.

The principle is clear: the more directions you make measurements from, the more arrangements of objects you can distinguish. The goal of X-ray tomography is much more ambitious, we would like to use these projections to reconstruct a picture of the slice. This problem is similar to that considered in example 1.1.5. However, it is much more challenging to reconstruct a density function from its averages along lines than to reconstruct the outline of an object from its projections. To accomplish this in principle and in practice requires a great deal more mathematics.

Exercise 2.1.1. Suppose that we have an isotropic point source of X-rays in 3-dimensions of intensity I_0. Find the formula for the intensity of the beams at a distance r from the source. What are the units of $I(r)$?

Exercise 2.1.2. Verify (2.5) by direct computation.

Exercise 2.1.3. Describe an apparatus that would produce a uniform, non-divergent source of X-rays.

2.2 Analysis of a point source device

In this section we use Beer's law to study a simple 2-dimensional apparatus and analyze what it measures. Figure 2.5 shows an apparatus with a point source of X-rays, an absorbing body and a photographic plate. We now derive an expression for the flux at a point P on the photographic plate in terms of the attenuation of the beam caused by absorption as well as beam spreading. The final expression involves the line integral of the attenuation coefficient.

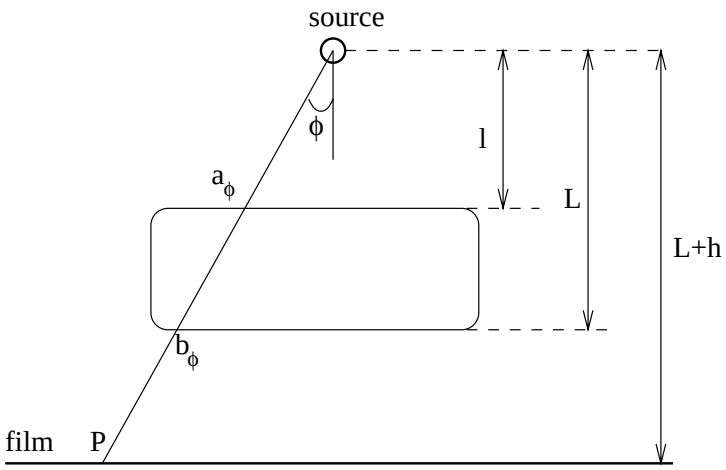


Figure 2.5: A point source device for measuring line integrals of the absorption coefficient.

The effect of beam spreading on the intensity of the flux is analyzed in example 2.1.4. The geometry of our apparatus suggests the use of polar coordinates to label points in the plane. Let r denote the distance from the source and ϕ the angle indicated in the diagram. The attenuation coefficient for the absorbing body in figure 2.5 is then a function of (r, ϕ), denoted by $\mu_a(r, \phi)$. The total attenuation coefficient is obtained by adding $\mu_s(r) = r^{-1}$ to μ_a. For the beam of X-rays traveling along the line through the source, at angle ϕ, the differential equation describing the attenuation of the X-ray beam is

$$\frac{dI}{dr} = -\left(\mu_a(r, \phi) + \frac{1}{r}\right) I. \tag{2.8}$$

The sum, $\mu_a + r^{-1}$ is an *effective* attenuation coefficient as it captures both the attenuation due to absorption and that due to beam spreading. The film is exposed by turning on source for a known period of time. In order to avoid introducing more notation I is also used to denote the total energy resulting from this exposure.

Label the radius of the first point of intersection of the line at angle ϕ with the absorbing body by a_ϕ, and the last by b_ϕ. The other distances, describing the apparatus are labeled

by h, L, and $L+h$ respectively. Integrating equation (2.8) from $r = r_0$ to the film plane $r = r_\phi$ gives

$$\log \frac{I(r_\phi, \phi)}{I(r_0, \phi)} = \log \frac{r_0}{r_\phi} - \int_{a_\phi}^{b_\phi} \mu_a(s, \phi) ds.$$

Using

$$a_\phi = \frac{l}{\cos\phi}, \quad b_\phi = \frac{L}{\cos\phi}, \quad r_\phi = \frac{L+h}{\cos\phi},$$

we get

$$I(r_\phi, \phi) = I_0 \frac{\cos\phi}{2\pi(L+h)} \exp\left[- \int_{a_\phi}^{b_\phi} \mu_a(s, \phi) ds \right].$$

The density of the developed film at a point is proportional to the logarithm of the total energy incident at that point, i.e.

$$\text{density of the film} = \gamma \times \log(\text{total energy intensity}) \tag{2.9}$$

where γ is a constant. We now compute this energy. As the film plane is not perpendicular to the direction of the beam of X-rays, we need to determine the flux across the part of the film subtended by the angle $\Delta\phi$. It is given by

$$\Delta F = \int_{\phi}^{\phi+\Delta\phi} I(r, \phi) \hat{r} \cdot \hat{n} ds, \quad \hat{r} = -(\sin\phi, \cos\phi).$$

Here $\hat{n} = (0, 1)$ is the outward, unit normal vector to the film plane and ds is the arc length element along the film plane. In polar coordinates it is given by

$$ds = \frac{L+h}{\cos^2\phi} d\phi.$$

Since $\Delta\phi$ is small we can approximate the integral by

$$\Delta F \approx \int_{\phi}^{\phi+\Delta\phi} I(r_\phi, \phi) \hat{r} \cdot \hat{n} \frac{L+h}{\cos^2\phi} d\phi \approx I_0 \frac{\cos\phi}{L+h} \exp\left[- \int_{a_\phi}^{b_\phi} \mu_a(s, \phi) ds \right] \frac{L+h}{\cos^2\phi} \Delta\phi. \tag{2.10}$$

The length of film subtended by the angle $\Delta\phi$ is approximately

$$\Delta S = \frac{L+h}{\cos^2\phi} \Delta\phi.$$

The energy density at the point P_ϕ, where the line making angle ϕ with the source, meets the film plane, is ΔF divided by this length. Indeed, letting $\Delta\phi$ tend to zero gives

$$\frac{dF}{ds} = \frac{I_0 \cos\phi}{2\pi(L+h)} \exp\left[- \int_{a_\phi}^{b_\phi} \mu_a(s, \phi) ds \right]$$

According to (2.9) the density of the film at P_ϕ is therefore

$$\gamma \log \frac{dF}{ds} = \gamma \left[\log \frac{I_0 \cos\phi}{2\pi(L+h)} - \int_{a_\phi}^{b_\phi} \mu_a(s,\phi) ds \right].$$

The first term comes from the attenuation due to beam spreading. Subtracting it from the measurement gives the line integral of the attenuation coefficient of the absorbing body along the ray at angle ϕ. Let $\delta(\phi)$ denote the density of the film at P_ϕ, this formula can be rewritten

$$-\int_{a_\phi}^{b_\phi} \mu_a(s,\phi) ds = \gamma^{-1}\delta(\phi) - \log \left[\frac{I_0 \cos\phi}{2\pi(L+h)} \right].$$

On the right hand side are quantities determined by measurement, the left hand side is the line integral of the absorption coefficient. A very important feature of this formula is the fact the measurements are expressed as a *linear function* of the absorption coefficient.

By varying the position of the source we can measure the line integrals of the absorption coefficient along another family of lines. If we move the source and film plane together, around a circle enclosing the absorbent material, making the measurements described above for each position of the source, then we can measure the line integrals of the absorption coefficient for all lines which intercept the object, see figure 2.6. This brings us to an essentially mathematical problem: Can a function be recovered from a knowledge of its line integrals along **all** lines? We shall see that this can in principle be done. That it can also be done, in practice forms the basis for image reconstruction in a transmission CT-machine.

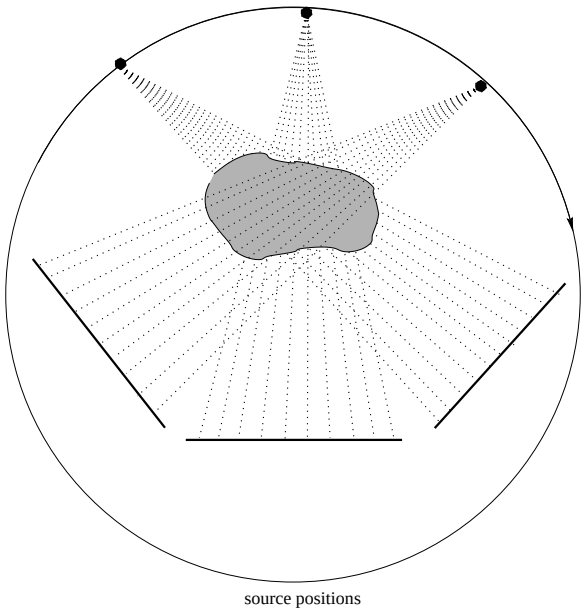


Figure 2.6: Collecting data from many views.

2.3 Some physical considerations

Before proceeding with the mathematical development we briefly revisit the assumptions underlying our model for the absorption of X-rays. This discussion previews topics considered in later chapters and is not essential to the remainder of this chapter. The X-ray source is assumed to be monochromatic. In fact the beam of X-rays is made up of photons having a wide range of energies. The distribution of photons according to energy is described by its *spectral function*, $S(\mathcal{E})$. If $\mathcal{E}_1$ and $\mathcal{E}_2$ are nearby energies then the energy in the beam due to photons with energies lying in the interval $[\mathcal{E}_1, \mathcal{E}_2]$ is about $S(\mathcal{E}_1)(\mathcal{E}_2 - \mathcal{E}_1)$, or more precisely

$$\int_{\mathcal{E}_1}^{\mathcal{E}_2} S(\mathcal{E}) d\mathcal{E}.$$

The graph of a typical spectral function is shown in figure 2.7. The total energy output of the source is given by

$$\Psi_i = \int_0^{\infty} S(\mathcal{E}) d\mathcal{E}.$$

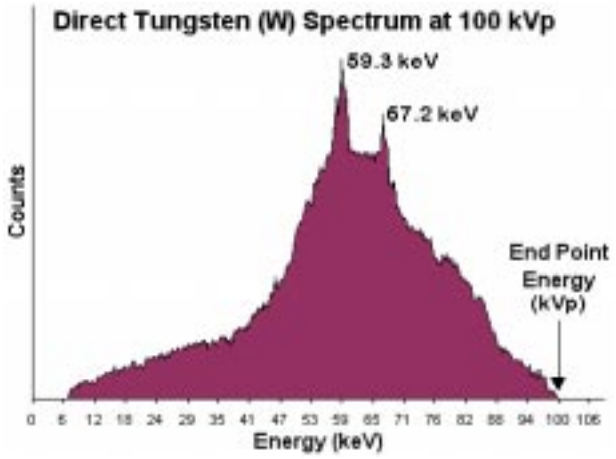


Figure 2.7: A typical X-ray source spectral function, courtesy Dr. Andrew Kavellas.

The attenuation coefficient of a given material is a complicated function of the energy, monotonely decreasing as the energy increases. The absorption coefficient is the sum total of the results of several physical processes that X-rays undergo. A discussion of the physics behind the absorption coefficient can be found in [4]. Let $\mu(\mathbf{x}, \mathcal{E})$ denote the absorption coefficient of the object for photons of energy $\mathcal{E}$. Beer's law, applied to the photons of energy $\mathcal{E}$, traveling along a line l states that the ratio $I_o(\mathcal{E})/I_i(\mathcal{E})$ of emitted flux to incident flux at this energy is

$$\frac{I_o(\mathcal{E})}{I_i(\mathcal{E})} = \exp\left[-\int_l \mu(\mathbf{x}, \mathcal{E}) ds\right].$$

The incident flux at energy $\mathcal{E}$ is $S(\mathcal{E})d\mathcal{E}$ and therefore

$$I_o(\mathcal{E}) = S(\mathcal{E})d\mathcal{E} \exp\left[-\int_l \mu(\mathbf{x}, \mathcal{E})ds\right].$$

Because low energy (or soft) X-rays are absorbed more efficiently than high energy (or hard) X-rays, the distribution of energies in the output beam is skewed towards higher energies. Along a given line the spectral function at energy $\mathcal{E}$ of the output beam is

$$S_{\text{out}}(\mathcal{E}) = S(\mathcal{E}) \exp\left[-\int \mu(x, \mathcal{E})ds\right].$$

In medical imaging, this is called *beam hardening.*

Integrating the output over the energy gives the measured output

$$\Psi_o = \int_0^\infty S(\mathcal{E}) \exp\left[-\int_l \mu(\mathbf{x}, \mathcal{E})ds\right] d\mathcal{E}.$$

As before we would like to reconstruct $\mu(\mathbf{x}, \mathcal{E})$ or perhaps some average of this function over energies. Mathematically this is a *very* difficult problem as the measurement, Ψ_o is a non-linear function of $\mu(\mathbf{x}, \mathcal{E})$. We have avoided this problem by assuming that the X-ray beam used to make the measurements is monochromatic. This provides the much simpler *linear* relationship (2.7) between the measurements and the absorption coefficient. In Chapter 8 we briefly consider the artifacts which result from using polychromatic X-rays and methods used to ameliorate them.

The fact that the X-ray "beam" is not a continuous flux, but is composed of discrete particles produces random errors in the measurements. This type of error is called Poisson noise, quantum noise or photon noise. In Chapter 12 we analyze this effect, showing that the available information in the data is proportional to the square root of the *number* of photons used to make the measurement. The accuracy of the measurements is the ultimate limitation on the number of significant digits in the reconstructed absorption coefficient. Table 2.1 lists the absorption coefficients of different structures encountered in medical CT. The absorption coefficients of air (-1000) and bone (990) define the range of values present in a typical clinical situation. The dynamic range of a clinical CT-measurement is about 2000 Hounsfield units. From the table it is apparent that the variation in the absorption coefficients of soft tissues is about 2% of this range. For X-ray CT to be clinically useful this means that the reconstruction of the absorption coefficient needs to be accurate to less than a half a percent or about 10 Hounsfield units.

An obvious solution to this problem would be to increase the number of photons. Since each X-ray photons carries a very large amount of energy, considerations of patient safety preclude this solution. The number of X-ray photons involved in forming a CT image is approximately $10^7/\,\text{cm}^2$. This should be compared with the 10^{11} to $10^{12}/\,\text{cm}^2$ photons, needed to make a usable photographic image. In ordinary photography, quantum noise is not a serious problem because the number of photons involved is very large. In X-ray tomography, quantum noise places a definite limit on the distinctions which can be made in a CT image.

In Chapter 4 we give a formula for determining a function from its integrals on *all* lines in the plane. Of course it is not possible to make an infinite number of measurements. This means that we need a method for reconstructing an approximation to a function from a finite collection of line integrals. In Chapter 8, beginning with the exact reconstruction formula, we derive algorithms for use with finitely many measurements.

2.4 The definition of the Radon transform

See: A.3, A.4.1, A.5.

The first step in determining a function from its integrals along all straight lines is to organize this information in a usable fashion. We use the parameterization for the set of oriented lines in the plane described in section 1.2.1. Recall that a line in the plane is determined by a pair consisting of a unit vector ω and a real number t. The line $l_{t,\omega}$ is the set

$$\{(x, y) \mid \langle (x, y), \omega \rangle = t\},$$

see figure 1.10. The set of unit vectors, or unit circle in $\mathbb{R}^2$ is denoted by S^1. Let $\hat{\omega}$ be the unit vector orthogonal to ω such that $\det(\omega\hat{\omega}) = 1$. The line $l_{t,\omega}$ has the parametric representation

$$l_{t,\omega} = \{t\omega + s\hat{\omega} \mid s \in (-\infty, \infty)\}.$$

The vector $\hat{\omega}$ specifies an orientation for the line. The pair $(t, \omega) \in \mathbb{R} \times S^1$ determines an *oriented* line in the plane. From either representation it is clear that, as sets,

$$l_{t,\omega} = l_{-t,-\omega}. \tag{2.11}$$

It is sometimes convenient to parametrize the points on the unit circle by a real number, to that end we set

$$\begin{aligned} \omega(\theta) &= (\cos(\theta), \sin(\theta)), \\ \hat{\omega}(\theta) &= (-\sin(\theta), \cos(\theta)) \end{aligned} \tag{2.12}$$

Sometimes we use the notation $l_{t,\theta}$ to denote $l_{t,\omega(\theta)}$, the meaning should be clear from the context. To summarize, the pairs $(t, \omega) \in \mathbb{R} \times \mathbb{S}^1$ parametrize the *oriented* lines in the plane.

Suppose that f is a function defined in the plane, which for simplicity, we assume is continuous with bounded support. The integral of f along the line $l_{t,\omega}$ is denoted by

$$\mathrm{R}f(t, \omega) = \int_{l_{t,\omega}} f ds. \tag{2.13}$$

The collection of integrals of f along the lines in the plane defines a function on $\mathbb{R} \times S^1$, called the *Radon transform* of f.

Definition 2.4.1. The *Radon transform* is the integral transform defined in (2.13) mapping functions defined in $\mathbb{R}^2$ to functions defined on $\mathbb{R} \times S^1$.

The parametric representation of the line gives a more explicit formula

$$\mathrm{R}f(t,\omega) = \int f(s\hat{\omega} + t\omega)ds. \tag{2.14}$$

In terms of Cartesian coordinates in $\mathbb{R}^2$ and (t,θ)-coordinates for the set of oriented lines this can be expressed as

$$\mathrm{R}f(t,\theta) = \int_{-\infty}^{\infty} f(t\cos\theta - s\sin\theta, t\sin\theta + s\cos\theta)ds. \tag{2.15}$$

It is not necessary for f to be either continuous or of bounded support. The Radon transform can be defined, *a priori* for functions whose restriction to each line can be integrated,

$$\int_{-\infty}^{\infty} |f(t\cos\theta - s\sin\theta, t\sin\theta + s\cos\theta)ds| < \infty \text{ for all } (t,\theta) \in \mathbb{R} \times \mathbb{S}^1. \tag{2.16}$$

With these conditions the improper integrals in (2.15) are unambiguously defined. We say that functions which satisfy (2.16) are in the *natural domain* of the Radon transform. This is really two different conditions:

(1). The function is regular enough so that restricting it to any line gives a locally integrable function.

(2). The function goes to zero rapidly enough for the improper integrals to converge.

The function $f(x,y) = 1$ is not in the natural domain of the Radon transform because it does not decay at infinity. The function $f(x,y) = (x^2+y^2)^{-1}$ is not in the natural domain of R because the integrals in (2.16) diverge if $t = 0$. An understanding of the domain of Radon transform is a very important part of its mathematical analysis. Though in applications to medical imaging, functions of interest are usually piecewise continuous and zero outside of some disk and therefore belong to the natural domain of R..

We compute the Radon transforms for a several simple classes functions.

Example 2.4.1. For $E \subset \mathbb{R}^2$ recall that

$$\chi_E(x,y) = \begin{cases} 1 & \text{if } (x,y) \in E, \\ 0 & \text{if } (x,y) \notin E. \end{cases}$$

For functions of this type the Radon transform has a very simple geometric description.

$$\mathrm{R}\chi_E(t,\omega) = \text{ the length of the intersection } l_{t,\omega} \cap E.$$

If E is a closed, bounded subset then χ_E belongs to the natural domain of the R. These functions model objects with constant absorption coefficient

Definition 2.4.2. The disk of radius r centered at (a, b) is denoted

$$D_r(a, b) = \{(x, y) \mid (x - a)^2 + (y - b)^2 < r^2\}.$$

Often $D_r(0, 0)$ is denoted by D_r.

Example 2.4.2. The function $\chi_{D_1}(x, y)$ is a special case of the general class considered in the previous example. The formula for the Radon transform of χ_{D_1} is

$$\mathrm{R}\chi_{D_1}(t, \omega) = \begin{cases} 2\sqrt{1 - t^2} & \text{if } |t| \leq 1, \\ 0 & \text{if } |t| > 1. \end{cases}$$

Note that $|t| > 1$ corresponds to lines $l_{t,\omega}$ which do not intersect the disk of radius one.

Example 2.4.3. A function, $f(x, y)$ is *radial* if it only depends on the distance of (x, y) to $(0, 0)$, In this case there exists a function F, of a single variable so that

$$f(x, y) = F(x^2 + y^2).$$

for the radial functions the Radon transform has a simpler form. From geometric considerations it is clear that $\mathrm{R}f(t, \omega)$ does not depend on ω but only on t, the distance of a line to the origin. Fixing a convenient direction, for example $\omega = (1, 0)$ we obtain

$$\begin{aligned} \mathrm{R}f(t, \omega) &= \int_{-\infty}^{\infty} f(t, s)ds \\ &= \int_{-\infty}^{\infty} F(t^2 + s^2)dy. \end{aligned} \tag{2.17}$$

Using the change of variable, $r^2 = t^2 + y^2, 2rdr = 2ydy$, we obtain

$$\mathrm{R}f(t) = 2\int_t^{\infty} \frac{F(r^2)rdr}{\sqrt{r^2 - t^2}}. \tag{2.18}$$

As $\mathrm{R}f$ does not depend on ω, we have omitted it from the formula. Formula (2.18) expresses the Radon transform of a radial function as a 1-dimensional *integral transform*.

The Radon transform is a *linear* map from functions on $\mathbb{R}^2$ to functions on $\mathbb{R} \times S^1$: If f and g are two functions and a and b are real numbers then

$$\mathrm{R}(af + bg) = a\,\mathrm{R}f + b\,\mathrm{R}g.$$

This observation is of central importance for what follows. It means that the Radon transform is an infinite dimensional analogue of a linear transformation.

Since $\mathrm{R}f(t, \omega)$ is defined as the integral of f over the line $l_{t,\omega}$ and $l_{t,\omega} = l_{-t,-\omega}$ it follows that

$$\mathrm{R}f(t, \omega) = \mathrm{R}f(-t, -\omega). \tag{2.19}$$

Such a function is called an *even function* on the space of oriented lines. Our goal is the recovery of a function, $f(x,y)$ from a knowledge of its Radon transform, $\mathrm{R}f(t,\omega)$. Since R is a linear map one might hope that there is a linear map R^{-1} from functions on $\mathbb{R}\times S^1$ to functions on $\mathbb{R}^2$ satisfying

$$\mathrm{R}^{-1}\circ \mathrm{R}f = f.$$

Ideally the inverse map should also be given by an integral formula. This turns out to be the case, but the derivation and analysis of this formula are rather involved. Because these spaces of functions are infinite dimensional, finding the inverse is not just a problem in linear algebra. The domain of R^{-1} is the range of R and neither the domain of R or R^{-1} is easy to describe explicitly. These issues are studied in Chapter 4. The remainder of this section is devoted to further properties of the Radon transform and its inverse.

Naively one would expect that in order for R^{-1} to exist it would be necessary that $\mathrm{R}f(t,\omega)=0$ for all pairs (t,ω) only if $f\equiv 0$. In light of its definition it is easy to construct examples of functions which are not zero, but have zero Radon transform.

Example 2.4.4. Define the function

$$f(x,y)=\begin{cases}1 & \text{if } (x,y)=(0,0),\\ 0 & \text{if } (x,y)\neq(0,0).\end{cases}$$

Clearly $\mathrm{R}f(t,\omega)=0$ for all (t,ω).

From the point of view of measurement, this is a very trivial example. A somewhat more interesting example arises as follows.

Example 2.4.5. Define a function f by setting $f(x,y)=1$ if $x\in[-1,1]$ and $y=0$ and zero otherwise. Then $\mathrm{R}f(t,\omega)=0$ if $\omega\neq(0,\pm1)$ and $\mathrm{R}f(0,(0,\pm1))=2$. In this case the Radon transform is usually zero, but for certain special lines it is not. Observe that if we replace f by a function $\tilde{f}$ which is 1 on some other subset of $\mathbb{R}\times 0$ of total length 2 then $\mathrm{R}f=\mathrm{R}\tilde{f}$. This gives examples, which are not entirely trivial. where the Radon transform does not contain enough information to distinguish between two functions.

The concept that underlies these examples is that of a set of measure zero.

Definition 2.4.3. A subset $E\subset\mathbb{R}^n$ is said to be of **measure zero** if for any $\epsilon>0$ there is a collection of balls $B(x_i,r_i)$ so that

$$E\subset\bigcup_{i=1}^{\infty}B(x_i,r_i)$$

and

$$\sum_{i=1}^{\infty}r_i^n<\epsilon.$$

A set of measure zero carries no n-dimensional mass. For example, a point is set of measure in the line, a line is a set of measure zero in the plane, a plane is a set of measure zero in $\mathbb{R}^3$, etc.

A basic fact about sets of measure zero is the following: if f is a function defined in $\mathbb{R}^n$ and the set of points where $f \neq 0$ is a set of measure zero then

$$\int_{\mathbb{R}^n} |f(x)|dx = 0.$$

With this concept we can state a basic result about the Radon transform

Proposition 2.4.1. *If f is a function defined in the plane such that*

$$\int_{\mathbb{R}^2} |f(x)|dx = 0$$

then the set of values $(t, \omega) \in \mathbb{R} \times S^1$ for which $\mathrm{R}f(t, \omega) \neq 0$ is itself a set of measure zero.

As example 2.4.5 shows, a function supported on a set of measure zero cannot, in general be reconstructed from its Radon transform. Since the Radon transform is linear, it cannot distinguish functions which differ only on a set of measure zero. This is a feature common to any set of measurements defined by integrals. While it is important to keep in mind, it does not lead to serious difficulties in medical imaging.

The support properties of f are reflected in the support properties of $\mathrm{R}f$.

Proposition 2.4.2. *Suppose that $f(x, y)$ is a function defined in the plane with $f(x, y) = 0$ if $x^2 + y^2 > R^2$ then*

$$\mathrm{R}f(t, \omega) = 0 \text{ if } |t| > R. \tag{2.20}$$

Proof. Any line $l_{t,\omega}$ with $|t| > R$ lies entirely outside of the support of f. From the definition it follows that $\mathrm{R}f(t, \omega) = 0$ if $|t| > R$. □

If f is *known* to vanish outside a certain disk then we do not need to compute its Radon transforms for lines that are disjoint from the disk. It would be tempting to assert that the converse statement is also true, that is "If $\mathrm{R}f(t, \omega) = 0$ for $|t| > R$ then $f(x, y) = 0$ if $x^2 + y^2 > R^2$." As the next set of examples show, this is false. We return to this question in Chapter 4.

Example 2.4.6. For each integer $n > 1$ define a function, in polar coordinates by setting

$$f_n(r, \theta) = r^{-n} \cos(n\theta).$$

These functions all blow up at $r = 0$ faster than r^{-1} and therefore do not belong to the natural domain of R. This is because f_n cannot be integrated along any line which passes through $(0, 0)$. On the other hand, since f_n goes to zero as $r \to \infty$ like r^{-n} and $n > 1$, the integrals defining $\mathrm{R}f_n(t, \omega)$ converge absolutely for any $t \neq 0$. We use the following result:

Lemma 2.4.1. *For each $n > 1$ and $t \neq 0, \omega \in \mathbb{S}^1$ the integrals*

$$\int_{l_{t,\omega}} f_n(t\omega + s\hat{\omega})ds$$

converge absolutely **and** *equal zero.*

The proof of the lemma is at the end of this section. It already indicates the difficulty of inverting the Radon transform. These functions are not in the natural domain of the Radon transform because

$$\int_{-\infty}^{\infty} |f_n(-s\sin\theta, s\cos\theta)|ds = \infty$$

for any value of θ. On the other hand $\mathrm{R}f_n(t,\omega) = 0$ for all $t \neq 0$. So in some sense, $\mathrm{R}f_n$ is supported on the set of measure zero $\{0\} \times S^1$.

For each n we modify f_n to obtain a function F_n in the natural domain of R such that $\mathrm{R}F_n(t,\omega) = 0$ for all (t,ω), with $|t| > 1$. On the hand, the functions F_n do not vanish outside the disk or radius 1. The modified functions are defined by

$$F_n(r,\theta) = \begin{cases} f_n(r,\theta) & \text{for } r > 1, \\ 0 & \text{for } r \leq 1. \end{cases}$$

A line $l_{t,\omega}$ with $|t| > 1$ lies entirely outside the unit disk. On such a line, the lemma applies to show that

$$\mathrm{R}F_n(t,\omega) = \int_{l_{t,\omega}} f_n ds = 0.$$

On the other hand F_n is bounded in a neighborhood of $(0,0)$ and therefore $\mathrm{R}F_n(t,\omega)$ is defined for all $(t,\omega) \in \mathbb{R} \times \mathbb{S}^1$. This shows that the Radon transform of a function may vanish for all t with $|t| > r$ without the function being zero outside disk of radius r.

Exercise 2.4.1. Let $f(x,y) = 1$ if $x^2 + y^2 = 1$ and zero otherwise. Show that $\mathrm{R}f(t,\omega) = 0$ for all $(t,\omega) \in \mathbb{R} \times S^1$.

Exercise 2.4.2. Suppose that $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ are n-distinct points on the unit circle. For $i \neq j$, let l_{ij} denote the line segment joining $\mathbf{x}_i$ to $\mathbf{x}_j$ and r_{ij} denote a real number. Show that if $r_{ij} = r_{ji}$ for all $i \neq j$ then there is function f supported on the line segments $\{l_{ij}\}$ such that

$$\int_{l_{ij}} f ds = r_{ij} \text{ for all } i \neq j.$$

Exercise 2.4.3. Show that a line segment has measure zero as a subset of the plane.

Exercise 2.4.4. Show that the x-axis has measure zero as a subset of the plane.

Exercise 2.4.5. Show that the set $\{(0,\omega) \ : \ \omega \in S^1\}$ has measure zero as a subset of $\mathbb{R} \times S^1$.

2.4.1 Appendix: Proof of Lemma 2.4.1*

The proof of the theorem makes use of the elementary theory of complex variables and is not needed for the subsequent development of the book.

Proof. Let $z = x + iy$ and observe that by Euler's formula it follows that

$$f_n = \operatorname{Re} z^{-n}.$$

This means that for $t \neq 0$

$$\mathrm{R}f_n(t,\omega) = \operatorname{Re} \int\limits_{l_{t,\omega}} z^{-n} ds,$$

where ds is the arc element along the line. The line $(t, \omega(\theta))$ can be represented as

$$z = (t + is)e^{i\theta}, \quad t \in \mathbb{R}.$$

Using this complex parameterization, the Radon transform of f_n can be re-expressed as a complex contour integral:

$$\mathrm{R}f_n(t,\theta) = \int\limits_{-\infty}^{\infty} f_n((t+is)e^{i\theta})ds = \operatorname{Re}\left[-ie^{-i\theta} \int\limits_{\operatorname{Re}(e^{-i\theta}z)=t} z^{-n}dz\right], \tag{2.21}$$

where the arc length element, along the line is written in complex notation as

$$ds = -ie^{-i\theta}dz.$$

As $\int z^{-n} = (1-n)^{-1}z^{1-n}$ the theorem follows from (2.21). □

2.4.2 Continuity of the Radon transform*

See: A.4.2, A.4.4.

The Radon transform is a linear transformation from functions defined in the plane to even functions on the space of lines. In medical applications, the function $\mathrm{R}f(t,\omega)$ is an idealization for what is measured. Generally speaking f is taken to be a bounded, though possibly discontinuous function which vanishes outside of the patient. Suppose that f vanishes outside the disk of radius L and $f \leq M$. The lengths of the intersections of $l_{t,\omega}$ with the support of f are bounded above $2L$, giving a crude estimate $\mathrm{R}f$,

$$|\,\mathrm{R}f(t,\omega)| = \left| \int\limits_{l_{t,\omega}} f ds \right| \leq 2ML. \tag{2.22}$$

The fact that it takes time to make the measurements means that, in the course of acquiring a full set of samples, the patient often moves. For a vector $\tau \in \mathbb{R}^2$, let f_τ denote the function f translated by the vector τ :

$$f_\tau(x,y) = f(x - \tau^1, y - \tau^2).$$

Suppose that we attempt to measure the $\mathrm{R}f(t,\omega)$ but the patient moves, a little, during the measurement process. A better model for what is measured is $\{\mathrm{R}f_{\tau(t,\omega)}(t,\omega)\}$, where

as indicated $\{\tau(t,\omega)\}$ are vectors in $\mathbb{R}^2$ describing the position of the patient as a function of (t,ω).

How sensitive are the measurements to errors? This is a question about the continuity properties of the map $f \mapsto \mathrm{R}f$. It it important to know the answer to this question, but in the final analysis is also not quite the correct question. What we really want to know is how sensitive the *reconstruction method* is to such errors. In other words, we want to understand the continuity properties of R^{-1}. Since we have not yet constructed R^{-1} we consider the somewhat easier question of the continuity of R. That is, how sensitive are the measurements to the data. We need to choose a way to measure the size of the errors in both the data and the measurements. For the present we make the following simple choices: For the measurements we use the maximum norm:

$$\|\mathrm{R}f(t,\omega) - \mathrm{R}g(t,\omega)\|_\infty = \max_{(t,\omega)} |\mathrm{R}f(t,\omega) - \mathrm{R}g(t,\omega)|.$$

As a norm on the data we use

$$\|f\|_{1,\infty} = \max_{(t,\omega)} \int\limits_{l_{t,\omega}} |f ds|.$$

With these definitions it is not difficult to see that

$$\|\mathrm{R}f(t,\omega) - Rg(t,\omega)\|_\infty \leq \|f-g\|_{1,\infty} \tag{2.23}$$

For the problem of patient motion (2.23) implies that

$$\|\mathrm{R}f - \mathrm{R}f_\tau\|_\infty \leq \max_{\tau(t,\omega)} \|f - f_\tau\|_{1,\infty}.$$

If *on average* f does not vary too quickly and the motions which arise are not too large then this estimate shows that the "actual" measurements $\{\mathrm{R}f_{\tau(t,\omega)}(t,\omega)\}$ are close to the model measurements $\{\mathrm{R}f(t,\omega)\}$. Since the functions which arise in imaging are not continuous it is important that only the average variation needs to be controlled and not pointwise variation. This point is illustrated by a one dimensional example.

Example 2.4.7. If $\tau \neq 0$ then

$$\max_x |\chi_{[0,1]}(x) - \chi_{[0,1]}(x-\tau)| = 1$$

on the other hand for $|\tau| < 1$ it is also true that

$$\int\limits_{-\infty}^{\infty} |\chi_{[0,1]}(x) - \chi_{[0,1]}(x-\tau)| dx = 2\tau.$$

If $f(x,y) = \chi_{[0,1]}(x)\chi_{[0,1]}(y)$ then a small change in the position of the patient can lead to a large error in $\mathrm{R}f$, when measured in the maximum-norm. If $\tau = (\epsilon, 0)$ for any $\epsilon < 0$ then

$$\mathrm{R}f(1,(1,0)) - \mathrm{R}f_\tau(1,(1,0)) = 1.$$

Choosing different norms leads to different estimates for the continuity of the map $f \mapsto \mathrm{R}f$. For a function h defined on $\mathbb{R} \times S^1$ define

$$\|h\|_{1,\infty} = \max_{\omega \in S^1} \int_{-\infty}^{\infty} |h(t,\omega)| dt. \tag{2.24}$$

Proposition 2.4.3. *Suppose that f is an absolutely integrable function in the natural domain of the Radon transform then*

$$\|\mathrm{R}f\|_{1,\infty} \leq \int_{\mathbb{R}^2} |f(x,y)| dxdy. \tag{2.25}$$

Proof. The proof of this proposition is simply the change of variables formula and the Fubini theorem. For each $\omega \in S^1$

$$\begin{aligned} \int_{-\infty}^{\infty} |\mathrm{R}f(t,\omega)| dt &\leq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |f(t\omega + s\hat{\omega})| ds dt \\ &= \int_{\mathbb{R}^2} |f(x,y)| dxdy. \end{aligned} \tag{2.26}$$

In the second line we use the fact that $(s,t) \mapsto t\omega + s\hat{\omega}$ is an orthogonal change of variables. Since the last line is independent of ω this proves the proposition. □

Because the Radon transform is linear this estimate implies that

$$\|\mathrm{R}f - \mathrm{R}f_\tau\|_{1,\infty} \leq \int_{\mathbb{R}^2} |f(x,y) - f_\tau(x,y)| dxdy. \tag{2.27}$$

It is not hard to show that the right hand side of (2.27) is small whenever $\|f - f_\tau\|_{1,\infty}$ is small, but not conversely. Because of the averaging which occurs in the reconstruction process, it is often sufficient to keep the measurement errors small in a norm like $\|\cdot\|_{1,\infty}$. A one variable example illustrates this point.

Example 2.4.8. A measuring device is often modeled as computing an average. The length of the interval over which we average is an indication of the resolution in the measurement. Let $f(t)$ be a function we would like to measure and model the measurement as the average of f over the interval of length 2ϵ :

$$M_\epsilon f(t) = \frac{1}{2\epsilon} \int_{t-\epsilon}^{t+\epsilon} f(s) ds.$$

To control $M_\epsilon f$ in a pointwise norm it suffices to control the integral of f,

$$|M_\epsilon f(t)| \leq \frac{1}{2\epsilon} \int_{t-\epsilon}^{t+\epsilon} |f(s)| ds \leq \frac{1}{2\epsilon} \int_{-\infty}^{\infty} |f(s)| ds.$$

Because the measurement is linear we see that if f_1 and f_2 are two functions then

$$|M_\epsilon f_1(t) - M_\epsilon f_2(t)| \leq \frac{1}{2\epsilon} \int_{-\infty}^{\infty} |f_1(s) - f_2(s)| ds.$$

As an example of the application of functional analytic methods to the study of the Radon transform we show that the estimate, (2.25) allows the extension of the Radon transform beyond its natural domain to the space of absolutely integrable functions. This space is denoted by $L^1(\mathbb{R}^2)$. The extension of the Radon transform is obtained by observing that continuous functions with bounded support are dense in the space $L^1(\mathbb{R}^2)$ with respect to the usual L^1-norm:

$$\|f\|_{L^1} = \int_{\mathbb{R}^2} |f(x,y)| dx dy.$$

For $f \in L^1(\mathbb{R}^2)$, choose a sequence $< f_n >$ of continuous function, with bounded support which satisfy

$$\lim_{n\to\infty} \|f - f_n\|_{L^1} = 0.$$

The Radon transform of f is the function on $\mathbb{R} \times S^1$ *defined* as the limit of the sequence of functions $< \mathrm{R}f_n >$ with respect to the norm $\|\cdot\|_1$. The limit $\mathrm{R}f$ also has the property that for all $\omega \in S^1$

$$\int_{-\infty}^{\infty} |\mathrm{R}f(t,\omega)| dt \leq \|f\|_{L^1}.$$

On the other hand, $\mathrm{R}f$ is no longer given by the formula (2.14) as it is not known, *a priori* that these integrals converge. Fubini's theorem implies that these integrals are finite for almost every t.

Exercise 2.4.6. Prove that the sequence of functions $< \mathrm{R}f_n >$ has a limit.

Exercise 2.4.7. Compute the Radon transform of

$$f = \frac{\chi_{[0,1]}(x^2+y^2)}{\sqrt{x^2+y^2}}.$$

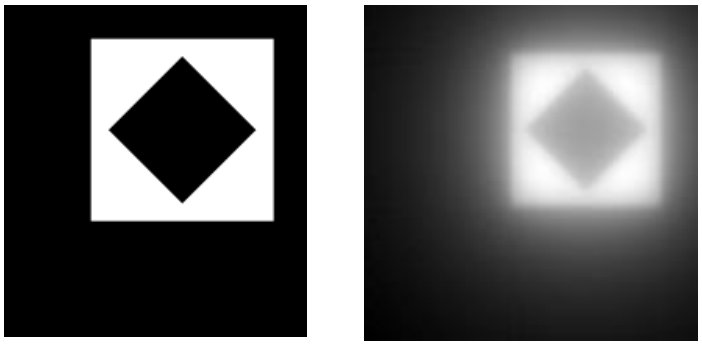
Is f in the natural domain of R? What is $\|\mathrm{R}f\|_1$?

2.4.3 The backprojection formula

Even though the line integrals of a function are very concrete data, it is difficult to use this data directly to reconstruct the function. An obvious thing to try is averaging the values of the $\mathrm{R}f$ over the lines that pass through a point. For a direction ω, the line in the family $\{l_{t,\omega} : t \in \mathbb{R}\}$ passing through a point (x,y) is given by $t = \langle (x,y), \omega \rangle$. Thus we could try setting

$$\tilde{f}(x,y) = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{R}f(\langle (x,y), \omega(\theta) \rangle, \theta) d\theta. \tag{2.28}$$

This is called the *back-projection formula*. While it is a reasonable guess, it does not give the correct answer. Figure 2.8 shows the result of using backprojection to reconstruct a simple black and white image. The object is recognizable but very blurry.



(a) A simple object.

(b) The result of back-projecting its Radon transform.

Figure 2.8: Back-projection does not work!

To find the true inverse of the Radon transform requires an indirect approach passing through the Fourier transform. The Fourier transform, while perhaps more familiar, is a less transparent integral transformation than the Radon transform. On the other hand, the inverse of the Fourier transform is easier to obtain. In the next chapter we consider the Fourier transform in some detail as it is of fundamental importance in the theory of image reconstruction and signal processing. To close this chapter we study the Radon transform acting on radial functions. Inverting the transform in this case is simpler, reducing to a special case of the Abel transform.

2.5 The Radon transform of a radially symmetric function

Recall that a function f defined on $\mathbb{R}^2$ is radial or *radially symmetric* if $f(x,y) = F(x^2+y^2)$. The Radon transform of a radial function does not depend on ω. It is given in (2.18) as an integral transform of a function of one variable. After changing variables we see that this is a special case of an Abel transform. For $0 < \alpha \leq 1$, the *α-Abel transform* of g is defined by

$$\mathrm{A}_\alpha\, g(t) = \frac{1}{\Gamma(\alpha)} \int_t^\infty \frac{g(s)ds}{(s-t)^{1-\alpha}},$$

the coefficient $\Gamma(\alpha)$ is the *Gamma* function, defined in section A.3.5. The theory of the Abel transform is briefly reviewed below.

Changing variables shows that

$$\mathrm{R}f(t) = \sqrt{\pi}(\mathrm{A}_{\frac{1}{2}}\, F)(t^2). \tag{2.29}$$

Using the formula for the inverse of the Abel transform (2.37), which is derived below, and a change of variables, we can solve equation (2.18) to obtain

$$F(r^2) = -\frac{1}{\pi r}\partial_r \left[\int_r^\infty \frac{\mathrm{R}f(t)tdt}{(t^2-r^2)^{1/2}} \right] \tag{2.30}$$

It is a consequence of (2.30) and exercise 2.5.9 that the radial function f vanishes outside the disk of radius L if and only if $\mathrm{R}f(t) = 0$ for $|t| \geq L$.

Example 2.5.1. The characteristic function of an annular region, $\chi_{[a^2,b^2]}(x^2+y^2)$ is a simple model for the sorts of functions encountered in medical imaging. It is piecewise differentiable with jump discontinuities. Using formula 2.18 we easily compute $\mathrm{R}\chi_{[a^2,b^2]}$:

$$\mathrm{R}\chi_{[a^2,b^2]}(t) = \begin{cases} \sqrt{b^2-t^2} - \sqrt{a^2-t^2} & \text{for } |t| \leq a, \\ \sqrt{b^2-t^2} & \text{for } a < |t| \leq b, \\ 0 & \text{for } b < |t|. \end{cases} \tag{2.31}$$

Exercise 2.5.1. Prove formula (2.29).

Exercise 2.5.2. Prove the formulæ in (2.31).

2.5.1 The range of the radial Radon transform*

Due to the simple structure of the inversion formula for radial functions, there are simple sufficient conditions for a function, $\psi(t)$ to be the Radon transform of a bounded continuous, radial function. The next proposition is an example of such a result.

Proposition 2.5.1. *Let $\psi \in \mathcal{C}^2(\mathbb{R})$ satisfy the following conditions*

(1). $\psi(t) = \psi(-t)$.

(2). *There is a constant M so that*

$$|\psi(t)| \leq M \text{ and } |\psi'(t)| \leq M.$$

(3). *Both ψ and ψ' are absolutely integrable.*

Then there is a bounded, continuous function $g(x,y) = G(r^2)$, in the natural domain of the Radon transform, such that

$$\mathrm{R}f = \psi.$$

Proof. For a function satisfying the conditions above we can integrate by parts to show that, for $r > 0$.

$$g(r) = \int_r^\infty \frac{\psi(t)tdt}{\sqrt{t^2-r^2}} = -\int_r^\infty \psi'(t)\sqrt{t^2-r^2}dt. \tag{2.32}$$

As both integrals have the same limit as $r \to 0$ this identity holds for $r \geq 0$. It is not difficult to prove that this function is differentiable, set

$$G(r^2) = -\frac{1}{\pi r}\partial_r g(r) = -\frac{1}{\pi}\int_r^\infty \frac{\psi'(t)dt}{\sqrt{t^2 - r^2}}. \tag{2.33}$$

To show that $G(r)$ is bounded we split the integral into two parts. If $r \geq 1$ then

$$\begin{aligned} |G(r^2)| &\leq \frac{1}{\pi}\left[\int_r^{2r} \frac{M dt}{\sqrt{t^2 - r^2}} + \frac{1}{r}\int_{2r}^\infty |\psi'(t)|dt\right] \\ &\leq C(M + \int_0^\infty |\psi'(t)|dt). \end{aligned} \tag{2.34}$$

If $r < 1$ a different argument is required. Because $\psi(t)$ is twice differentiable and an even function there is a constant M' so that

$$|\psi'(t)| \leq M'|t|.$$

We then have the estimate

$$\begin{aligned} |G(r^2)| &\leq \frac{1}{\pi}\left[\int_r^1 \frac{M'|t|dt}{\sqrt{t^2 - r^2}} + \int_1^\infty |\psi'(t)|dt\right] \\ &\leq C'(M' + \int_0^\infty |\psi'(t)|dt). \end{aligned} \tag{2.35}$$

The continuity of G is left as an exercise. To show that $G(r^2)$ is absolutely integrable we interchange order of the integrations to obtain

$$\begin{aligned} \int_0^\infty |G(r^2)|dr &\leq \frac{1}{\pi}\int_0^\infty\int_0^t \frac{|\psi'(t)|drdt}{\sqrt{t^2 - r^2}} \\ &= \frac{1}{2}\int_0^\infty |\psi'(t)|dt. \end{aligned} \tag{2.36}$$

As $G(r^2)$ is bounded and absolutely integrable it follows that the integrals in (2.18), defining $\mathrm{R}G(t)$, converge absolutely. It is now an elementary calculation to show that $\mathrm{R}G = \psi$. □

Remark 2.5.1. Formula (2.33) gives an alternate form for the inverse of the Radon transform if $\mathrm{R}f$ satisfies the hypotheses of the proposition.

Exercise 2.5.3. Prove that $g(r)$, defined in (2.32) is a differentiable function.

Exercise 2.5.4. Prove that $G(r^2)$ is a continuous function.

Exercise 2.5.5. Prove the fact, used in the proof that

$$\int_0^t \frac{dt}{\sqrt{t^2 - r^2}} = \frac{\pi}{2}.$$

Exercise 2.5.6. Give complete justifications for the statements that G is in the natural domain of the Radon transform and $\mathrm{R}G = \psi$.

2.5.2 The Abel transform*

The Abel transform is a familiar feature of many problems involving measurement. It is also an important example of a non-trivial integral transform which nonetheless admits a fairly explicit analysis. Formally the inverse of the α-Abel transform is

$$\mathrm{A}_\alpha^{-1} = -\partial_x \mathrm{A}_{1-\alpha}\,. \tag{2.37}$$

This formula is a consequence of the identity

$$\int_x^s \frac{dt}{(t-x)^\alpha (s-t)^{1-\alpha}} = \Gamma(\alpha)\Gamma(1-\alpha), \tag{2.38}$$

which holds for $0 < \alpha < 1$. To derive the Abel inversion formula we let

$$g(t) = \mathrm{A}_\alpha f = \frac{1}{\Gamma(\alpha)} \int_t^\infty \frac{f(s)ds}{(s-t)^{1-\alpha}}. \tag{2.39}$$

Changing the order of the integration and using the identity above, gives

$$\begin{aligned} \frac{1}{\Gamma(1-\alpha)} \int_x^\infty \frac{g(t)dt}{(t-x)^\alpha} &= \frac{1}{\Gamma(1-\alpha)\Gamma(\alpha)} \int_x^\infty \int_t^\infty \frac{f(s)dsdt}{(s-t)^{1-\alpha}(t-x)^\alpha} \\ &= \frac{1}{\Gamma(1-\alpha)\Gamma(\alpha)} \int_x^\infty \int_x^s \frac{f(s)dtds}{(s-t)^{1-\alpha}(t-x)^\alpha} \\ &= \int_x^\infty f(s)ds. \end{aligned} \tag{2.40}$$

Taking the derivative we obtain

$$f(x) = -\partial_x \left[\frac{1}{\Gamma(1-\alpha)} \int_x^\infty \frac{g(t)dt}{(t-x)^\alpha} \right].$$

In other words,

$$I = -\partial_x \mathrm{A}_{1-\alpha}\, \mathrm{A}_\alpha\,. \tag{2.41}$$

The operator $-\partial_x \mathrm{A}_{1-\alpha}$ is a *left inverse* to A_α.

Our derivation of the inversion formula is a formal computation, assuming that the various manipulations make sense for the given function f. The main point is the interchange of the order of integrations in the second line of (2.40). If f is absolutely integrable then this step is easily justified. For an absolutely integrable function, f it therefore follows that

$$\mathrm{A}_{1-\alpha} \circ \mathrm{A}_\alpha f = \int_x^\infty f(s)ds.$$

If f is also continuous then this indefinite integral is differentiable and therefore

$$f = \partial_x \, \mathrm{A}_{1-\alpha} \circ \mathrm{A}_\alpha \, f.$$

Data of interest in medical imaging are usually not continuous. Instead, the data is usually piecewise continuous and so can be represented as a sum

$$f(x) = f_c(x) + \sum_{j=1}^{N} \alpha_j \chi_{[a_j,b_j]}(x),$$

where $f_c(x)$ belongs to $\mathcal{C}^0(\mathbb{R})$, and the other term collects the jumps in f. As noted

$$\mathrm{A}_{1-\alpha} \circ \mathrm{A}_\alpha \, \chi_{[a,b]}(x) = \int_x^\infty \chi_{[a,b]}(s) ds.$$

If $x \neq a$ or b then this function is differentiable with derivative 0 or 1. In order to interpret the formula at the exceptional points we need to extend our notion of differentiability.

Definition 2.5.1. A locally integrable function f has a *weak derivative*, if there is a locally integrable function f_1 such that, for every continuously differentiable function g, the following identity holds

$$\int_{-\infty}^{\infty} f(x) g'(x) dx = - \int_{-\infty}^{\infty} f_1(x) g(x) dx. \tag{2.42}$$

In this case f_1 is called the weak derivative of f.

If f is a differentiable function then formula (2.42), with $f_1 = f'$ is just the usual integration by parts formula. The weak derivative of the indefinite integral of a piecewise continuous function is the function itself. This shows that the inversion formula, properly understood, is also valid for the sort of data which arises in imaging applications.

Exercise 2.5.7. Prove (2.38) by using the change of variables

$$t = \lambda x + (1 - \lambda) s$$

and the classical formula

$$\int_0^1 \frac{d\lambda}{\lambda^\alpha (1-\lambda)^{1-\alpha}} = \Gamma(\alpha)\Gamma(1-\alpha),$$

see [82].

Exercise 2.5.8. Let f be a piecewise continuous, absolutely integrable function and $0 < \alpha \leq 1$. Show that if $\mathrm{A}_\alpha \, f(x) = 0$ for $x > R$ then $f(x) = 0$ for $x > R$ as well.

Exercise 2.5.9. Use exercise 2.5.8 to prove the following uniqueness result for the Radon transform. If f is a piecewise continuous, radial function in the natural domain of the Radon transform and $\mathrm{R}f(t) = 0$ for $|t| > R$ then $f(r) = 0$ if $r > R$.

Exercise 2.5.10. Generalize the argument given above to prove that

$$\mathrm{A}_\alpha \circ \mathrm{A}_\beta = \mathrm{A}_{\alpha+\beta}\,.$$

For what range of α and β does this formula make sense?

Exercise 2.5.11. For $0 < a < b$ compute $g_{a,b} = \mathrm{A}_\alpha(\chi_{[a,b]})$ and verify by explicit calculation that $\chi_{[a,b]}$ is the weak derivative of $-\mathrm{A}_{1-\alpha}(g_{a,b})$.

Exercise 2.5.12. Provide the detailed justification for the derivation of (2.41) for f a continuous, absolutely integrable function.

Exercise 2.5.13. Suppose that $f \in \mathcal{C}^1(\mathbb{R})$ and that f and f' are absolutely integrable. Show that

$$\mathrm{A}_\alpha\,[-\partial_x\,\mathrm{A}_{1-\alpha}]\,f = f.$$

Exercise 2.5.14. Suppose that f is a piecewise continuous, absolutely integrable function. Show that f is the weak derivative of

$$F(x) = -\int_x^\infty f(s)ds.$$

2.5.3 Fractional derivatives*

To evaluate the inverse of the Abel transform for a function with a jump discontinuity required an extension of the classical notion of a derivative. The Abel transform A_α is sometimes called an α^{th}-order anti-derivative or *fractional integral*. For example if $\alpha = 1$ then $\mathrm{A}_1\,f$ is just the usual indefinite integral of f. This interpretation of the Abel transforms motivates the following definition for a function with a *fractional order* of differentiability.

Definition 2.5.2. Let $0 < \beta < 1$, a continuous function f has a β^{th}-derivative if

$$\sup_{x\in\mathbb{R}}\sup_h \frac{|f(x+h) - f(x)|}{|h|^\beta} < \infty$$

The collection of such functions is a vector space denoted by $\mathcal{C}^\beta(\mathbb{R})$. A norm is defined on $\mathcal{C}^\beta(\mathbb{R})$ by letting

$$|f|_\beta = \sup_{x\in\mathbb{R}}\sup_h \frac{|f(x+h) - f(x)|}{|h|^\beta} + \sup_{x\in\mathbb{R}}|f(x)|.$$

With this norm, $\mathcal{C}^\beta(\mathbb{R})$ is a complete normed linear space; it is often called the space of β-Hölder continuous functions.

As with the usual notion of differentiability, β-differentiability is a local property of a function. There is an analogous definition for β-Hölder continuous functions defined on an interval $[a, b]$. The space of such functions is denoted by $C^\beta([a, b])$.

It is reasonable to enquire as to the relationship between 1-Hölder continuity and differentiability. If f is a differentiable function then the mean value theorem states that for each $0 < h$ there is a $0 \leq k < h$ so that

$$\frac{f(x+h) - f(x)}{h} = f'(x+k).$$

This implies that if f is differentiable and f' is bounded on $[a, b]$ then f is 1-Hölder continuous on $[a, b]$. On the other hand, the function $f(x) = |x|$ fails to be differentiable at $x = 0$, nonetheless

$$\left| \frac{f(0+h) - f(0)}{h} \right| \leq 1$$

for all h. The function $|x|$ is 1-Hölder continuous on any interval $[-a, a]$, therefore a function can be 1-Hölder continuous without being differentiable at every point. The following proposition describes the precise relationship between these concepts.

Proposition 2.5.2. *A function is 1-Hölder continuous if and only if it is differentiable almost everywhere and has a bounded derivative.*

We have defined a notion of β-differentiability but have not yet defined a "β^{th}-derivative." If $\alpha + \beta$ is not an integer then $\mathrm{A}_\alpha f \in \mathcal{C}^{\alpha+\beta}(\mathbb{R})$, see [22]. This is a sense in which one can interpret the statement that $\mathcal{A}_\alpha$ is an α^{th}-anti-derivative. If $\alpha + \beta$ is an integer then this statement is no longer true. That is $f \in \mathcal{C}^\beta(\mathbb{R})$ does **not** imply that $\mathrm{A}_\alpha f \in \mathcal{C}^{\alpha+\beta}(\mathbb{R})$.

Suppose that $f \in \mathcal{C}^\beta(\mathbb{R})$ and that $0 < \alpha < \beta$. If we define the α^{th}-derivative of f by the formula

$$D_\alpha f = -\partial_x A_{1-\alpha} f$$

then $D_\alpha f \in \mathcal{C}^{\beta-\alpha}(\mathbb{R})$. This is a reasonable definition, however there are other ways to defined fractional orders of differentiability as well as fractional derivatives.

Exercise 2.5.15. Show that $g_{a,b} \in \mathcal{C}^\alpha(\mathbb{R})$. Explain why this shows that $\mathrm{A}_{1-\alpha}$ does not carry $C^\alpha(\mathbb{R})$ to once differentiable functions.

2.5.4 Volterra equations of the first kind*

See: A.2.6, A.7.2.

The Abel transforms are examples of a class of integral operators called Volterra operators. These operators are infinite dimensional generalizations of upper triangular matrices. A linear transformation K is a Volterra operator of the first kind if it can be represented in the form

$$Kf(x) = \int_0^x k(x, y) f(y) dy.$$

This differs a little from the form of the Abel transform as the integral there extends from x to infinity, rather than 0 to x. The function $k(x,y)$ is called the *kernel function* of the integral operator K. The kernel functions for the Abel transforms are singular where $x = y,$. In this section we restrict ourselves to kernel functions which satisfy an estimate of the form

$$|k(x,y)| \leq M,$$

and analyze Volterra operators acting on functions defined on the interval $[0,1]$.

Volterra operators often appear in applications where one is required to solve an equation of the form

$$g = f + Kf = (\mathrm{Id} + K)f.$$

Such equations turn out to be very easy to solve. Formally we would write

$$f = (\mathrm{Id} + K)^{-1} g.$$

Still proceeding formally, we can express $(\mathrm{Id} + K)^{-1}$ as an infinite series:

$$(\mathrm{Id} + K)^{-1} f = \sum_{j=0}^{\infty} (-1)^j K^j f. \tag{2.43}$$

This is called the *Neumann series* for $(\mathrm{Id} + K)^{-1}$; it is obtained from the Taylor expansion of the function $(1+x)^{-1}$ about $x = 0$:

$$(1+x)^{-1} = \sum_{j=0}^{\infty} (-1)^j x^j.$$

Here $K^j f$ means the j-fold composition of K with itself. The sum on the right hand side of (2.43) is an infinite sum of functions and we need to understand in what sense it converges. That is, we need to choose a norm to measure the sizes of the terms in this sum. A useful property of Volterra operators is that this series converges for almost any reasonable choice of norm.

The basic estimates are summarized in the proposition.

Proposition 2.5.3. *Let $1 \leq p \leq \infty$, suppose that $|k(x,y)| \leq M$ and $f \in L^p([0,1])$ then for $x \in [0,1]$ and $j \geq 1$*

$$|K^j f(x)| \leq \frac{M^j x^{j-1}}{(j-1)!} \|f\|_{L^p}. \tag{2.44}$$

Proof. If $f \in L^p([0,1])$ for a $p \geq 1$ then $f \in L^1([0,1])$ and Hölder's inequality implies that

$$\|f\|_{L^1} \leq \|f\|_{L^p}.$$

In light of this, it suffices to prove (2.44) with $p = 1$. The proof is by induction on j. First consider $j = 1$:

$$\begin{aligned} |Kf(x)| &= \left| \int_0^x k(x,y) f(y) dy \right| \\ &\leq \int_0^x M |f(y)| dy \\ &\leq M \|f\|_{L^1}. \end{aligned} \tag{2.45}$$

This verifies (2.44) for $j = 1$; assume it has been proved for j, then

$$\begin{aligned} |K^{j+1}f(x)| &= \left| \int_0^x k(x,y) K^j f(y) dy \right| \\ &\leq \int_0^x M \frac{M^j y^{j-1}}{(j-1)!} \|f\|_{L^1} dy \\ &\leq \frac{M^{j+1} x^j}{j!} \|f\|_{L^1}. \end{aligned} \tag{2.46}$$

This completes the induction step and thereby the proof of the proposition. □

The proposition implies that $(\mathrm{Id} + K)^{-1} f - f$ converges pointwise uniformly, even if f is only in $L^p([0,1])$. Indeed we have the pointwise estimate

$$|(\mathrm{Id} + K)^{-1} f(x) - f(x)| \leq M \|f\|_{L^1} \sum_{j=0}^{\infty} \frac{M^j x^j}{j!} = M \|f\|_{L^1} e^{Mx}. \tag{2.47}$$

Proposition 2.5.4. *If $f \in L^p([0,1])$ then the equation $f = (\mathrm{Id} + K)g$ has a unique solution of the form $g = f + f_0$ where f_0 is a continuous function on $[0,1]$ which satisfies the estimate*

$$|f_0(x)| \leq M \|f\|_{L^p} e^{Mx}.$$

In applications sometimes one encounters equations of the form

$$f = Kg \tag{2.48}$$

where K is a Volterra operator of the first kind. A similar equation arose in the previous section. Provided that $k(x,y)$ is differentiable and $k(x,x)$ does not vanish, this sort of equation can be reduced to the type of equation considered in the last proposition. If (2.48) is solvable, then f must, in some sense, be differentiable. We formally differentiate equation (2.48) to obtain

$$f'(x) = k(x,x) g(x) + \int_0^x k_x(x,y) g(y) dy.$$

If K' denotes the Volterra operator with kernel function $k_x(x,y)/k(x,x)$ then this equation can be rewritten

$$\frac{f'(x)}{k(x,x)} = (\mathrm{Id} + K') g.$$

Applying our earlier result we obtain the solution of the original equation in the form

$$g = (\mathrm{Id} + K')^{-1} \left(\frac{f'(x)}{k(x,x)} \right) = \left(\frac{f'(x)}{k(x,x)} \right) + \sum_{j=1}^{\infty} (-1)^j (K')^j \left(\frac{f'(x)}{k(x,x)} \right). \tag{2.49}$$

In applications K describes a measurement process and f represents measurements. In this context it can be quite difficult to accurately approximate f'. As a result, it is often stated that a problem which involves solving an equation of the form (2.48) is *ill-posed*. Small errors in measurement can lead to substantial errors in the solution of this type of equation. While it is reasonable to expect that we can control measurement errors in the sup-norm, it is usually not possible to control errors in the derivatives of measurements, even in an L^p-norm. The inverse problem is ill-posed because K^{-1} is not continuous as a map from $\mathcal{C}^0([0,1])$ to itself.

Remark 2.5.2. The material in this section is a small sample from the very highly developed field of integral equations. A good introductory treatment can be found in [83] or [61].

Exercise 2.5.16. Suppose that instead of assuming that $k(x,y)$ is uniformly bounded we assume that

$$\int_0^x |k(x,y)|^q dy \leq M$$

for a $1 < q < \infty$ and all $x \in [0,1]$. Show that estimates analogous to (2.44) hold for $f \in L^p([0,1])$ where $p = q(q-1)^{-1}$.

Exercise 2.5.17. Using the previous exercise, show that the equation $g = (\mathrm{Id} + K)f$ is solvable for $g \in L^p([0,1])$.

Exercise 2.5.18. Volterra operators of the first kind are infinite dimensional generalizations of strictly upper triangular matrices. These are matrices a_{ij} such that $a_{ij} = 0$ if $i \leq j$. Suppose that A is an $n \times n$ strictly upper triangular matrix. Show that $A^n = 0$. Prove that $I + A$ is always invertible and give a formula for its inverse.

Chapter 3

Introduction to the Fourier transform

In this chapter we introduce the Fourier transform and review some of its basic properties. The Fourier transform is the "swiss army knife" of mathematical analysis, it is a powerful general purpose tool with many useful special features. In particular the theory of the Fourier transform is largely independent of the dimension: the theory of the Fourier transform for functions of one variable is formally the same as the theory for functions of 2, 3 or n variables. This is in marked contrast to the Radon, or X-ray transforms. For simplicity we begin with a discussion of the basic concepts for functions of a single variable.

3.1 The complex exponential function.

See: A.2.9, A.3.3 .

The building block for the Fourier transform is the complex exponential function, e^{ix}. The basic facts about the exponential function can be found in section A.3.3. Recall that the polar coordinates (r, θ) correspond to the point with rectangular coordinates $(r\cos\theta, r\sin\theta)$. As a complex number this is

$$r(\cos\theta + i\sin\theta) = re^{i\theta}.$$

Multiplication of complex numbers is very easy using the polar representation if $z = re^{i\theta}$ and $w = \rho e^{i\phi}$ then

$$zw = re^{i\theta}\rho e^{i\phi} = r\rho e^{i(\theta+\phi)}.$$

The positive number $r = e^s$ for a real number $s = \log r$. In addition to the usual polar representation we can also express a complex number in the form

$$z = e^{s+i\theta}.$$

The logarithm is then extended to non-zero complex numbers by setting

$$\log z = s + i\theta = \log|z| + i\tan^{-1}\left(\frac{\operatorname{Im} z}{\operatorname{Re} z}\right).$$

As $\exp(2\pi i) = 1$, the imaginary part of the $\log z$ is only determined up to multiplies of 2π.

Out of the basic complex exponential we build a family of functions, $\{e^{ix\xi} : \xi \in \mathbb{R}\}$. Sometimes we think of x as the variable and ξ as a parameter and sometimes their roles are interchanged. Thinking of ξ as a parameter we see that $e^{ix\xi}$ is a $\frac{2\pi}{\xi}$-periodic function, that is

$$\exp(i(x + \frac{2\pi}{\xi})\xi) = \exp(ix\xi).$$

In physical applications $e^{ix\xi}$ describes an oscillatory state with *frequency* $\frac{\xi}{2\pi}$ and *wave length* $\frac{2\pi}{\xi}$. In quantum mechanics $e^{ix\xi}$ is a state with *momentum* ξ and *energy* $\frac{|\xi|^2}{2}$. The goal of Fourier analysis is to represent "arbitrary" functions as linear combinations of these oscillatory states. Using (A.63) we easily derive the fact that

$$\partial_x e^{ix\xi} = i\xi e^{ix\xi}. \tag{3.1}$$

Loosely speaking this formula says that $e^{ix\xi}$ is an *eigenvector* with *eigenvalue* $i\xi$ for the "linear transformation" ∂_x.

Exercise 3.1.1. If a is a real number then it is a consequence of the Fundamental Theorem of Calculus that

$$\int_0^x e^{ay}dy = \frac{e^{ax} - 1}{a} \tag{3.2}$$

Use the power series for the exponential to prove that this formula remains correct, even if a is a complex number.

Exercise 3.1.2. If $\operatorname{Re} a < 0$ then the improper integral is absolutely convergent:

$$\int_0^\infty e^{ax}dx = \frac{-1}{a}.$$

Using the triangle inequality (not the explicit formula) show that

$$\left|\int_0^\infty e^{ax}dx\right| \leq \frac{1}{|\operatorname{Re} a|}.$$

Exercise 3.1.3. Which complex numbers have purely imaginary logarithms?

3.2 The Fourier transform for functions of a single variable

We now turn our attention to the Fourier transform for functions of a single real variable. As the complex exponential itself assumes complex values, it is natural to consider complex valued functions from the outset. The theory for functions of several variables is quite similar and is treated later in the chapter.

3.2.1 Absolutely integrable functions

See: A.2.10, A.4.1.

Let $f(x)$ be a function defined on the real line $\mathbb{R}$. We say that f is *absolutely integrable* if

$$\int_{-\infty}^{\infty} |f(x)|dx < \infty.$$

The set of such functions is a complete, normed linear space denoted by $L^1(\mathbb{R})$.

Definition 3.2.1. The Fourier transform of an absolutely integrable function f, defined on $\mathbb{R}$ is the function $\hat{f}$ defined on $\mathbb{R}$ by the integral

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)e^{-ix\xi}dx. \tag{3.3}$$

The utility of the Fourier transform stems from the fact that f can be "reconstructed" from $\hat{f}$. A result that suffices for most of our applications is the following:

Theorem 3.2.1 (Fourier inversion formula). *Suppose that $f(x)$ is an absolutely integrable function such that $\hat{f}(\xi)$ is also absolutely integrable, then*

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(\xi)e^{ix\xi}d\xi. \tag{3.4}$$

Remark 3.2.1. Formula (3.4) is called the *Fourier inversion formula*. It is *the* prototype of all reconstruction formulæ used in medical imaging. The integral in (3.4) defines an integral transform known as the *inverse Fourier transform*.

Proof. We give a proof of the inversion formula under the additional assumption that $f(x)$ is continuous. We need to show that

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(\xi)e^{i\xi x}d\xi.$$

Because $\hat{f}(\xi)$ is in $L^1(\mathbb{R})$ it is not difficult to show that

$$\begin{aligned}\frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(\xi)e^{i\xi x}d\xi &= \lim_{\epsilon\to 0^+}\frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(\xi)e^{-\epsilon\xi^2}e^{i\xi x}d\xi \\ &= \lim_{\epsilon\to 0^+}\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(y)e^{-\epsilon\xi^2}e^{i\xi(x-y)}dyd\xi.\end{aligned} \tag{3.5}$$

Interchange the integrations in the last formula and use example 3.2.5 to get

$$\begin{aligned}\lim_{\epsilon\to 0^+}\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(y)e^{-\epsilon\xi^2}e^{i\xi(x-y)}dyd\xi &= \lim_{\epsilon\to 0^+}\frac{1}{2\sqrt{\epsilon\pi}}\int_{-\infty}^{\infty} f(y)e^{-\frac{(x-y)^2}{4\epsilon}}dy \\ &= \lim_{\epsilon\to 0^+}\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} f(x-2\sqrt{\epsilon}t)e^{-|t|^2}dt.\end{aligned} \tag{3.6}$$

The $\int \exp(-|t|^2)dt = \sqrt{\pi}$; as f is continuous and integrable it follows that the limit in the last line is $f(x)$. □

Remark 3.2.2. The Fourier transform and its inverse are frequently thought of as mappings. It is then customary to use the notation:

$$\begin{aligned}\mathcal{F}(f) &= \int_{-\infty}^{\infty} f(x)e^{-ix\xi}dx, \\ \mathcal{F}^{-1}(f) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathcal{F}(f)(\xi)e^{ix\xi}d\xi.\end{aligned} \tag{3.7}$$

Observe that the operation performed to recover f from $\hat{f}$ is almost the same as the operation performed to obtain $\hat{f}$ from f. Indeed if $f_r(x) \stackrel{d}{=} f(-x)$ then

$$\mathcal{F}^{-1}(f) = \frac{1}{2\pi}\mathcal{F}(f_r). \tag{3.8}$$

This symmetry accounts for many of the Fourier transform's remarkable properties. Note that the hypothesis $\int |f(x)|dx < \infty$ does not imply that $\int |\hat{f}(\xi)|d\xi < \infty$, see example 3.2.1.

Example 3.2.1. Define the function

$$r_1(x) = \begin{cases} 1 \text{ for } -1 < x < 1, \\ 0 \text{ for } 1 < |x|. \end{cases} \tag{3.9}$$

The Fourier transform of r_1 is

$$\widehat{r_1}(\xi) = \int_{-1}^{1} e^{-\xi x}dx = \frac{1}{-i\xi}e^{-\xi x}\Big|_{-1}^{1} = \frac{2\sin\xi}{\xi},$$

and

$$\int_{-\infty}^{\infty}|\widehat{r_1}(\xi)|d\xi = 2\int_{-\infty}^{\infty}\frac{|\sin\xi|}{|\xi|}$$

diverges. So while $r_1(x)$ is absolutely integrable its Fourier transform $\hat{r}_1(\xi)$ is not. As the Fourier transform of r_1 is such an important function in image processing we define

$$\operatorname{sinc}(x) \stackrel{d}{=} \frac{\sin(x)}{x}.$$

Example 3.2.2. Recall that $\chi_{[a,b)}(x)$ equals 1 for $a \leq x < b$ and zero otherwise. Its Fourier transform is given by

$$\hat{\chi}_{[a,b)}(\xi) = \frac{e^{-ib\xi} - e^{-ia\xi}}{i\xi}. \tag{3.10}$$

Example 3.2.3. A family of functions that arise in MR imaging are those of the form

$$f(x) = \chi_{[0,\infty)}(x)e^{i\alpha x}e^{-\beta x},\ \beta > 0.$$

By simply computing the integral we find that

$$\hat{f}(\xi) = \frac{1}{\beta + i(\xi - \alpha)}$$

Using the fact that $e^{i\alpha x} = \cos(\alpha x) + i\sin(\alpha x)$ it is not difficult to show that

$$\begin{gathered}\mathcal{F}(\cos(\alpha x)e^{-\beta x}\chi_{[0,\infty)}(x)) = \frac{\beta + i\xi}{\beta^2 + \alpha^2 - \xi^2 + 2i\xi\beta} \\ \text{and} \\ \mathcal{F}(\sin(\alpha x)e^{-\beta x}\chi_{[0,\infty)}(x)) = \frac{\alpha}{\beta^2 + \alpha^2 - \xi^2 + 2i\xi\beta}.\end{gathered} \tag{3.11}$$

Example 3.2.4. The "Gaussian" is a function of considerable importance in image processing. For later reference we record its Fourier transform:

$$\begin{aligned}\mathcal{F}(e^{-\frac{x^2}{2}})(\xi) &= \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}e^{-i\xi x}dx \\ &= \sqrt{2\pi}e^{-\frac{\xi^2}{2}},\end{aligned} \tag{3.12}$$

or more generally

$$\mathcal{F}(e^{-ax^2})(\xi) = \sqrt{\frac{\pi}{a}}e^{-\frac{\xi^2}{4a}}. \tag{3.13}$$

Note that $e^{-\frac{x^2}{2}}$ is an eigenvector of the Fourier transform, with eigenvalue $\sqrt{2\pi}$.

Exercise 3.2.1. Show that if $f(x)$ is a continuous, absolutely integrable function then

$$\lim_{\epsilon \to 0^+} \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} f(x - 2\sqrt{\epsilon}t)e^{-|t|^2}dt = f(x).$$

Exercise 3.2.2. Suppose that f is absolutely integrable, show that $\hat{f}(\xi)$ is a bounded, continuous function.

Exercise 3.2.3. Prove the identity (3.8).

Exercise 3.2.4. Prove formula (3.10). Show that for any numbers $a < b$ there is a constant M so that

$$|\hat{\chi}_{[a,b)}(\xi)| \leq \frac{M}{1+|\xi|}.$$

Exercise 3.2.5. Prove the formulæ in (3.11) and show that

$$\mathcal{F}(e^{-\beta|x|}e^{i\alpha x}) = \frac{2\beta}{\beta^2+(\xi-\alpha)^2}.$$

3.2.2 Appendix: The Fourier transform of a Gaussian*

For completeness we include a derivation of the Fourier transform of the Gaussian e^{-x^2}. It uses the Cauchy integral formula for analytic functions of a complex variable. The Fourier transform is given by

$$\begin{aligned}\mathcal{F}(e^{-x^2})(\xi) &= \int_{-\infty}^{\infty} e^{-(x^2+ix\xi)}dx \\ &= e^{-\frac{\xi^2}{4}} \int_{-\infty}^{\infty} e^{-(x+i\xi/2)^2}dx.\end{aligned} \tag{3.14}$$

The second integral is complex contour integral of the analytic function e^{-z^2} along the contour $\operatorname{Im} z = \xi/2$. Because e^{-z^2} decays rapidly to zero as $|\operatorname{Re} z|$ tends to infinity, Cauchy's theorem implies that the contour can be shifted to the real axis without changing the value of the integral, that is

$$\int_{-\infty}^{\infty} e^{-(x+i\xi/2)^2}dx = \int_{-\infty}^{\infty} e^{-x^2}dx.$$

To compute the last integral observe that

$$\begin{aligned}\left[\int_{-\infty}^{\infty} e^{-x^2}dx\right]^2 &= \int_{-\infty}^{\infty} e^{-x^2}dx \int_{-\infty}^{\infty} e^{-y^2}dy \\ &= \int_0^{2\pi}\int_0^{\infty} e^{-r^2}rdrd\theta = \pi.\end{aligned} \tag{3.15}$$

Polar coordinates are used in the second line. Combining these formulæ gives

$$\mathcal{F}(e^{-x^2}) = \sqrt{\pi}e^{-\frac{\xi^2}{4}}.$$

3.2.3 Regularity and decay

See: A.4.1, A.6.1.

It is a general principle that the *regularity* properties of f are reflected in the *decay* properties of its Fourier transform $\hat{f}$ and similarly the regularity of the Fourier transform is a reflection of the decay properties of f. Without any regularity, beyond absolute integrability we have the fundamental result:

Lemma 3.2.1 (The Riemann Lebesgue Lemma). *If $f(x)$ is an absolutely integrable function then its Fourier transform $\hat{f}(\xi)$ is a continuous function which goes to zero at infinity, that is, for $\eta \in \mathbb{R}$,*

$$\lim_{\xi \to \eta} \hat{f}(\xi) = \hat{f}(\eta) \text{ and } \lim_{\xi \to \pm\infty} \hat{f}(\xi) = 0. \tag{3.16}$$

Proof. This result is a consequence of the basic approximation theorem for L^1-functions, Theorem A.6.2. According to this theorem, given $\epsilon > 0$ there is a step function $F(x)$ so that

$$\int_{-\infty}^{\infty} |f(x) - F(x)| < \epsilon.$$

From the estimate for difference of the Fourier transforms

$$\begin{aligned} |\hat{F}(\xi) - \hat{f}(\xi)| &= \left| \int_{-\infty}^{\infty} (F(x) - f(x)) e^{-ix\xi} dx \right| \\ &\leq \int_{-\infty}^{\infty} |F(x) - f(x)| dx \\ &\leq \epsilon, \end{aligned} \tag{3.17}$$

it is clear that it suffices to show that $\lim_{|\xi| \to \infty} \hat{F}(\xi) = 0$. This is elementary as F has a representation as a finite sum

$$F(x) = \sum_{j=1}^{N} c_j \chi_{[a_j, b_j)}(x).$$

The Fourier transform of F is therefore

$$\hat{F}(\xi) = \sum_{j=1}^{N} c_j \hat{\chi}_{[a_j, b_j)}(\xi).$$

As this is a finite sum, the conclusion follows from formula 3.10. The continuity of $\hat{f}(\xi)$ is left as an exercise. □

To go beyond (3.16) we need to introduce quantitative measures of regularity and decay. A simple way to measure regularity is through differentiation, the more derivatives a function has, the more regular it is.

Definition 3.2.2. A function $f \in \mathcal{C}^j(\mathbb{R})$ if it has j-continuous derivatives.

Since the Fourier transform involves integration over the whole real line it is often important to assume that these derivatives are also integrable. To quantify rates of decay we compare a function $f(x)$ to a simpler function such as a power of $|x|$.

Definition 3.2.3. A function $f(x)$ decays like $|x|^{-\alpha}$ if there are constants C and R so that

$$|f(x)| \leq \frac{C}{|x|^{\alpha}} \text{ for } |x| > R.$$

A very important formula for use in Fourier analysis is the integration by parts formula: Let f and g be differentiable functions on the interval $[a, b]$ then

$$\int_a^b f'(x)g(x)dx = f(x)g(x)\Big|_{x=a}^{x=b} - \int_a^b f(x)g'(x)dx. \tag{3.18}$$

We need an extension of this formula with $a = -\infty$ and $b = \infty$. For our purposes it suffices to assume that fg, $f'g$ and fg' are absolutely integrable, the integration by parts formula then becomes

$$\int_{-\infty}^{\infty} f'(x)g(x)dx = -\int_{-\infty}^{\infty} f(x)g'(x)dx. \tag{3.19}$$

Suppose that $f(x)$ is an absolutely integrable function with an absolutely integrable first derivative, that is

$$\int_{-\infty}^{\infty} [|f(x)| + |f'(x)|]dx < \infty.$$

Provided $\xi \neq 0$ we can use (3.19) to obtain a formula for $\hat{f}(\xi)$

$$\begin{aligned} \hat{f}(\xi) &= \int_{-\infty}^{\infty} f(x)e^{-ix\xi}dx \\ &= \int_{-\infty}^{\infty} f'(x)\frac{e^{-ix\xi}}{i\xi}dx. \end{aligned} \tag{3.20}$$

That is

$$\hat{f}(\xi) = \frac{\widehat{f'}(\xi)}{i\xi}.$$

Because f' is absolutely integrable the Riemann Lebesgue lemma implies that $\widehat{f'}(\xi)$ tends to zero as $|\xi|$ tends to ∞. Combining our formula for $\hat{f}(\xi)$ with this observation we see that $\hat{f}(\xi)$ goes to zero **more** rapidly than $|\xi|^{-1}$. This should be contrasted with the computation of the Fourier transform of $r_1(x)$. The function $\hat{r}_1(\xi)$ tends to zero as $|\xi|$ tends to infinity

exactly like $|\xi|^{-1}$. This is a reflection of the fact that $r_1(x)$ is **not** everywhere differentiable, having jump discontinuities at ± 1.

If f has j integrable derivatives then, by repeatedly integrating by parts, we get a formula for $\hat{f}(\xi)$

$$\hat{f}(\xi) = \left[\frac{1}{i\xi}\right]^j \widehat{f^{[j]}}(\xi).$$

Again, because $f^{[j]}(x)$ is absolutely integrable $\widehat{f^{[j]}}(\xi)$ tends to zero as $|\xi| \to \infty$. We state the result of these computations as a proposition.

Proposition 3.2.1. *Let $f(x)$ be an absolutely integrable function with $j > 0$, absolutely integrable derivatives then its Fourier transform goes to zero, as $|\xi|$ tends to infinity faster than $|\xi|^{-j}$.*

This identity can also be viewed as giving a formula for the Fourier transform of $f^{[j]}$ in terms to the Fourier transform for f :

Proposition 3.2.2. *Let $f(x)$ be an absolutely integrable function with $j > 0$ absolutely integrable derivatives then the Fourier transform $f^{[j]}(x)$ is given by*

$$\widehat{f^{[j]}}(\xi) = (i\xi)^j \hat{f}(\xi). \tag{3.21}$$

The rate of decay in $\hat{f}(\xi)$ is reflected in the smoothness of $f(x)$.

Proposition 3.2.3. *Suppose that j is a non-negative integer. If $\hat{f}(\xi)$ decays like $|\xi|^{-(j+1+\epsilon)}$, for an $\epsilon > 0$, then f is continuous and has j continuous derivatives.*

Proof. To prove this statement we use the Fourier inversion formula

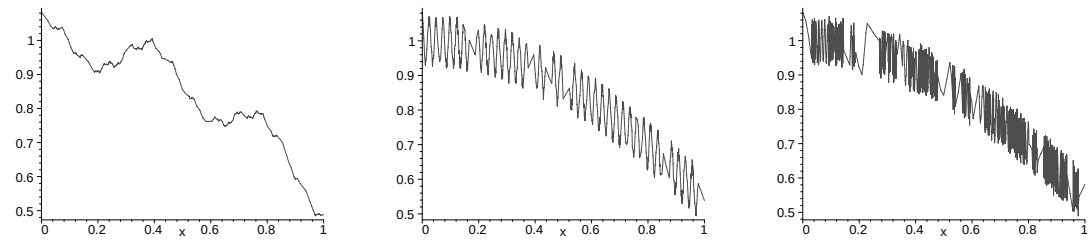
$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) e^{ix\xi} d\xi.$$

The hypothesis of the theorem implies that we can differentiate this formula up to j times. That is

$$f^{[l]}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi)[i\xi]^l e^{ix\xi} d\xi \qquad \text{for } 0 \le l \le j.$$

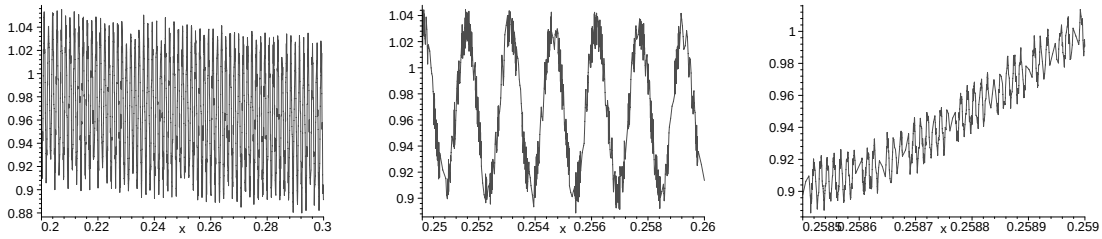
As $[i\xi]^l \hat{f}(\xi)$ is absolutely integrable for $l \le j$ this shows that f has j continuous derivatives. □

Remark 3.2.3. Note that if $f(x)$ has j integrable derivatives then $\hat{f}(\xi)$ decays faster than $|\xi|^{-j}$. The exact rate of decay depends on how continuous $f^{[j]}(x)$ is. We need to assume that $\hat{f}(\xi)$ decays *faster than* $|\xi|^{-(1+j)}$ to deduce that $f(x)$ has j-continuous derivatives. So we appear to "loose" one order of differentiability when inverting the Fourier transform. Both results are actually correct. The function $r_1(x)$ provides an example showing the second result is sharp. It has a jump discontinuity and its Fourier transform decays like $|\xi|^{-1}$. To construct an example to show that the first result is sharp, we now consider the case (not actually covered by the theorem) of $j = 0$. By integrating these examples we obtain functions which show that the Fourier transform of function with an integrable derivative may decay slower than $|\xi|^{-(1+\epsilon)}$, for any fixed positive $\epsilon > 0$.



(a) (b) (c)

Figure 3.1: Furry functions



(a) (b) (c)

Figure 3.2: A furry function at smaller scales

Example 3.2.5. Let $\varphi(x)$ be a smooth, rapidly decaying function with Fourier transform $\hat{\varphi}(\xi)$ which satisfies the following conditions

(1). $0 \leq \hat{\varphi}(\xi) \leq 1$ for all ξ,

(2). $\hat{\varphi}(0) = 1$,

(3). $\hat{\varphi}(\xi) = 0$ if $|\xi| > 1$.

For example we could take

$$\hat{\varphi}(\xi) = \begin{cases} e^{-\frac{1}{1-\xi^2}} & \text{if } |\xi| < 1, \\ 0 & \text{if } |\xi| \geq 1. \end{cases}$$

In fact the details of this function are not important, only the listed properties are needed to construct the examples. For each $k \in \mathbb{N}$ we define the function

$$\hat{f}_k(\xi) = \sum_{n=1}^{\infty} \frac{\hat{\varphi}(\xi - n^k)}{n^2}.$$

For a given ξ at most one term in the sum is non-zero. If $k > 1$ then $\hat{f}_k(\xi)$ is zero "most of the time." On the other hand the best *rate of decay* that is true for all ξ is

$$|\hat{f}_k(\xi)| \leq \frac{C}{|\xi|^{\frac{2}{k}}}.$$

By using a large k we can make this function decay as slowly as we like. Because $\sum n^{-2} < \infty$ we can apply the Fourier inversion formula to obtain

$$f_k(x) = \varphi(x) \sum_{n=1}^{\infty} \frac{e^{ixn^k}}{n^2}.$$

The infinite sum converges absolutely and uniformly in x and therefore $f_k(x)$ is a continuous function. Because $\varphi(x)$ decays rapidly at infinity so does $f_k(x)$. This means that $f_k(x)$ is an absolutely integrable, continuous function whose Fourier transform goes to zero like $|\xi|^{-\frac{2}{k}}$. These examples show that the rate of decay of the Fourier transform of a continuous, absolutely integrable function can be as slow as one likes. The graphs show the real parts of these functions; they are very "furry." The function, $f_k(x)$ is the smooth function $\varphi(x)$, *modulated* by noise, see figure 3.1. The fact that these functions are not differentiable is visible in figure 3.2. These graphs show f_{12} at smaller and smaller scales, observe that f_{12} does not appear smoother at small scales than at large scales.

These examples demonstrate that there are two different phenomena which govern the rate of decay of the Fourier transform of a function. The function $r_1(x)$ is very smooth, except where it has a jump. This kind of very localized failure of smoothness produces a characteristic $|\xi|^{-1}$ rate of decay in the Fourier transform. In the L^1-sense the function $r_1(x)$ is very close to being a continuous function. In fact by using linear interpolation we can find piecewise differentiable functions very close to r_1. These sort of functions frequently arise in medical imaging. The functions $f_k(x)$ are continuous, but very fuzzy. The larger k is, the higher the amplitude of the high frequency components producing the fuzz and the slower the rate of decay for $\hat{f}_k(\xi)$. These functions are not close to differentiable functions in the L^1 or L^2 sense. Such functions are typical of random processes used to model noise.

These results establish the connection between the regularity of f and the decay of its Fourier transform. If on the other hand, we know that f itself decays then this is reflected in increased regularity of its Fourier transform.

Proposition 3.2.4. *Suppose that j is a positive integer and*

$$\int_{-\infty}^{\infty} |f(x)|(1+|x|)^j dx < \infty,$$

then $\hat{f}(\xi)$ has j-continuous derivatives which tend to zero as $|\xi|$ tends to infinity. In fact for $0 \leq k \leq j$

$$\partial_\xi^k \hat{f}(\xi) = \int_{-\infty}^{\infty} (-ix)^k f(x) e^{-ix\xi} dx. \tag{3.22}$$

Of course (3.22) gives a formula for the Fourier transform of $x^k f(x)$ in terms of the Fourier transform of f :

$$\widehat{x^k f}(\xi) = i^k \partial_\xi^k \hat{f}(\xi). \tag{3.23}$$

A special case of this proposition arises if f vanishes outside a bounded interval. In this case $x^k f(x)$ is absolutely integrable for any positive integer k and therefore $\hat{f}(\xi)$ is a function with infinitely many derivatives. The derivatives tend to zero as $|\xi|$ tends to infinity but the *rate* of decay may be the same for all the derivatives, for example

$$\hat{r}_1(\xi) = \frac{2\sin\xi}{\xi}.$$

Differentiating this function repeatedly gives a sum of terms one of which tends to zero exactly like $|\xi|^{-1}$. This further confirms our principle that the rate of decay of the Fourier transform is a reflection of the smoothness of the function.

Example 3.2.6. An important application of the Fourier transform is to study ordinary differential equations with constant coefficients. Suppose that $\{a_0, \ldots, a_n\}$ are complex numbers, we would like to study the solutions of the differential equation

$$Df \stackrel{d}{=} \sum_{j=0}^{n} a_j \partial_x^j f = g.$$

Proceeding formally, take the Fourier transform of both sides, (3.21) gives relation

$$\left[\sum_{j=0}^{n} a_j (i\xi)^j\right] \hat{f}(\xi) = \hat{g}(\xi) \tag{3.24}$$

The polynomial,

$$P_D(\xi) = \sum_{j=0}^{n} a_j (i\xi)^j$$

is called the *characteristic polynomial* for the differential operator D. If a complex number ξ_0 is a root of this equation, i.e. $P_D(\xi_0) = 0$ then the exponential function $v_0 = \exp(i\xi_0 x)$ is a solution of the homogeneous equation $Dv_0 = 0$.

If on the other hand, $P_D(\xi)$ has no real roots and g is absolutely integrable then we can divide in (3.24) to obtain

$$\hat{f}(\xi) = \frac{\hat{g}(\xi)}{P_D(\xi)}.$$

Using the Fourier inversion formula we obtain a particular solution to the equation $Df = g$,

$$f_p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\hat{g}(\xi) e^{i\xi x} d\xi}{P_D(\xi)}. \tag{3.25}$$

The general solution is of the form $f_p(x) + f_0(x)$ where $Df_0 = 0$. If $P_D(\xi)$ has real roots then a more careful analysis is required, see [7].

Exercise 3.2.6. Suppose that $f'g$ and fg' are absolutely integrable. Show that the limits

$$\lim_{t\to\infty} fg(x) \text{ and } \lim_{t\to-\infty} fg(x)$$

both exist.

Exercise 3.2.7. Prove that for any number j the j^{th}-derivative $\partial_\xi^j \hat{r}_1(\xi)$ has a term which decays exactly like $|\xi|^{-1}$.

Exercise 3.2.8. Show that $f_P(x)$, defined in example 3.25 and its first n derivatives tend to zero as $|x|$ tends to infinity.

Exercise 3.2.9. Show that the function, $\varphi(x)$ defined in example 3.2.5 is infinitely differentiable.

3.2.4 Fourier transform on $L^2(\mathbb{R})$

See: A.2.4, A.2.5, A.4.2, A.4.5.

In the foregoing discussion we considered absolutely integrable functions. The Fourier transform is then defined in terms of an absolutely convergent integral. As we observed, this does not imply that the Fourier transform is itself absolutely integrable. In fact, it is very difficult to describe the range of $\mathcal{F}$ if the domain is $L^1(\mathbb{R})$. When using the L^1-norm, there are also discrepancies in the quantitative relationships between the smoothness of a function and the rate of decay of its Fourier transform. A more natural condition, when working with Fourier transform is *square integrability*.

Definition 3.2.4. A function $f(x)$ defined on $\mathbb{R}$ is square integrable if

$$\|f\|_{L^2}^2 = \int_{-\infty}^{\infty} |f(x)|^2 dx < \infty.$$

The set of such functions, with norm defined by $\|\cdot\|_{L^2}$, is denoted $L^2(\mathbb{R})$. With this norm $L^2(\mathbb{R})$ is a complete, normed linear space.

The completeness of L^2 is very important for what follows.

Example 3.2.7. The function $f(x) = (1+|x|)^{-\frac{3}{4}}$ is not absolutely integrable, but it is square integrable. On the other hand the function

$$g(x) = \frac{\chi_{[-1,1]}(x)}{\sqrt{|x|}}$$

is absolutely integrable but not square integrable.

The norm on $L^2(\mathbb{R})$ is defined by an inner product,

$$\langle f, g\rangle_{L^2} = \int_{\mathbb{R}^n} f(x)\overline{g(x)}dx.$$

This inner product satisfies the usual Cauchy-Schwarz inequality.

Proposition 3.2.5. *If $f, g \in L^2(\mathbb{R})$ then*

$$|\langle f, g\rangle_{L^2}| \leq \|f\|_{L^2}\|g\|_{L^2}. \tag{3.26}$$

Proof. The proof of the Cauchy-Schwarz inequality for L^2 is identical to the proof for $\mathbb{C}^n$. To slightly simplify the argument we assume that f and g are real valued. For any $t \in \mathbb{R}$, the squared norm

$$\|f + tg\|^2_{L^2} = \|f\|^2_{L^2} + 2t\langle f, g\rangle_{L^2} + t^2\|g\|^2_{L^2}$$

is a quadratic function. Differentiating shows that this polynomial assumes its minimum value at

$$t_0 = -\frac{\langle f, g\rangle_{L^2}}{\|g\|^2_{L^2}}.$$

As the norm of $f + t_0 g$ is positive it follows that

$$\|f\|^2_{L^2} - \frac{\langle f, g\rangle^2_{L^2}}{\|g\|^2_{L^2}} \geq 0,$$

which proves the proposition. □

An L^2 function is always *locally absolutely integrable*. This means that for any finite interval, $[a, b]$ the integral

$$\int_a^b |f(x)|dx \text{ is finite.}$$

To prove this we use the Cauchy-Schwarz inequality with $g = 1$:

$$\int_a^b |f(x)|dx \leq \sqrt{|b-a|}\sqrt{\int_a^b |f(x)|^2 dx} \leq \sqrt{|b-a|}\|f\|_{L^2}.$$

The reason square integrability is a natural condition is contained in the following theorem.

Theorem 3.2.2 (Parseval formula). *If f is absolutely integrable and also square integrable, then $\hat{f}(\xi)$ is square integrable and*

$$\int_{-\infty}^{\infty} |f(x)|^2 dx = \int_{-\infty}^{\infty} |\hat{f}(\xi)|^2 \frac{d\xi}{2\pi}. \tag{3.27}$$

The proof of the Parseval formula is given in exercise 3.2.13.

In many physical applications the square integral of a function is interpreted as a total energy. Up to the factor of 2π, Parseval's formula says that the total energy in f is the same as that in $\hat{f}$. Often the variable $\xi/2\pi$ is thought of as a frequency, following the quantum mechanical practice, higher frequencies corresponding to higher energies. In this context $|\hat{f}(\xi)|^2$ is interpreted as the energy density of f at frequency $\xi/2\pi$. As we shall see "noise" is essentially a high frequency phenomenon, a noisy signal has a lot of energy at high frequencies.

The Parseval formula shows that the L^2-norm is very intimately connected to the Fourier transform. However there is a price to pay. For a function like $f(x)$ in example 3.2.7 the integral defining $\hat{f}(\xi)$ is not absolutely convergent. Parseval's formula says that $\mathcal{F}$ is a continuous linear transformation, if the L^2-norm is used in both the domain and range. This indicates that it should be possible to extend the Fourier transform to all functions in $L^2(\mathbb{R})$ and this is, indeed the case. Let $f \in L^2(\mathbb{R})$, for each $R > 0$ define

$$\hat{f}_R(\xi) = \int_{-R}^{R} f(x)e^{-ix\xi}dx. \tag{3.28}$$

From Parseval's formula it follows that if $R_1 < R_2$, then

$$\|\hat{f}_{R_1} - \hat{f}_{R_2}\|^2_{L^2} = 2\pi \int_{R_1 \le |x| \le R_2} |f(x)|^2 dx$$

Because f is square integrable the right hand side of this formula goes to zero as R_1, R_2 tend to infinity. This says that, if we measure the distance in the L^2-norm, then the functions $< \hat{f}_R >$ are clustering, closer and closer together as $R \to \infty$. Otherwise put, $< \hat{f}_R >$ is an L^2-Cauchy sequence. Because $L^2(\mathbb{R})$ is a complete, normed vector space, this implies that $\{\hat{f}_R\}$ converges to a limit as $R \to \infty$; this limit *defines* $\hat{f}$. The limit of a sequence in the L^2-norm is called a *limit in the mean*; it is denoted by the symbol LIM.

Definition 3.2.5. If f is a function in $L^2(\mathbb{R})$ then its Fourier transform is defined by

$$\hat{f} = \underset{R\to\infty}{LIM} \hat{f}_R,$$

where $\hat{f}_R$ is defined in (3.28).

Example 3.2.8. The function $f(x) = x^{-1}\chi_{[1,\infty]}(x)$ is square integrable but not absolutely integrable. We use integration by parts to compute $\hat{f}_R(\xi)$:

$$\hat{f}_R(\xi) = \frac{e^{-i\xi}}{i\xi} - \frac{e^{-iR\xi}}{iR\xi} - \int_1^R \frac{e^{-ix\xi}}{i\xi x^2}dx.$$

The integrand is now absolutely integrable and so for $\xi \neq 0$ we can let $R \to \infty$ to obtain the pointwise limit:

$$\hat{f}(\xi) = \frac{e^{-i\xi}}{i\xi} - \int_1^\infty \frac{e^{-ix\xi}}{i\xi x^2}dx. \tag{3.29}$$

Note that each term diverges at $\xi = 0$, but the divergences cancel, giving a function which is actually continuous at $\xi = 0$.

A consequence of Parseval's formula is the identity.

$$\int_{-\infty}^{\infty} f(x)\overline{g(x)}dx = \int_{-\infty}^{\infty} \hat{f}(\xi)\overline{\hat{g}(\xi)}\frac{d\xi}{2\pi}. \tag{3.30}$$

This is proved by applying (3.27) to $f + tg$ and comparing the coefficients of powers t on the right and left hand sides. Up to the factor of 2π, the Fourier transform preserves the inner product. Recall that this is also a property of rotations of Euclidean space. Such a transformation is called unitary. Another consequence of the Parseval formula is a uniqueness statement: a function in L^2 is determined by its Fourier transform.

Corollary 3.2.1. *If $f \in L^2(\mathbb{R})$ and $\hat{f}(\xi) = 0$, then $f \equiv 0$.*

Remark 3.2.4. It would be more accurate to say that the set of x for which $f(x) \neq 0$ has measure 0.

Exercise 3.2.10. Show that $\hat{f}(\xi)$ defined in (3.29) is continuous in a neighborhood of $\xi = 0$. Prove that $\hat{f}_R(\xi)$ converges to $\hat{f}(\xi)$ in the L^2-norm.

Exercise 3.2.11. If $f \in L^2(\mathbb{R})$ then so is $\hat{f}(\xi)$. However $\hat{f}(\xi)$ is usually not absolutely integrable and therefore the Fourier inversion formula cannot be directly applied. Define

$$f_R(x) = \frac{1}{2\pi}\int_{-R}^{R} \hat{f}(\xi)e^{ix\xi}d\xi;$$

prove that $\underset{R\to\infty}{LIM} f_R = f$.

Exercise 3.2.12. Let $f, g \in L^2(\mathbb{R})$ by using the Parseval formula for the functions $f + tg$ where $t \in \mathbb{C}$ show that it implies (3.30).

Exercise 3.2.13. There is a formula which is very similar to (3.30) but is actually much easier to prove. Let f and g be square integrable functions, Show that

$$\int_{-\infty}^{\infty} f(x)\hat{g}(x)dx = \int_{-\infty}^{\infty} \hat{f}(x)g(x)dx.$$

Letting $\hat{g} = (2\pi)^{-1}\widehat{\hat{f}}$ show how to use the Fourier inversion formula to derive the Parseval formula from this identity.

3.2.5 Basic properties of the Fourier Transform on $\mathbb{R}$

The following properties hold for integrable or square integrable functions.

1. Linearity:
   The Fourier transform is a linear operation:
$$\widehat{f+g} = \hat{f} + \hat{g}, \quad \widehat{\alpha f} = \alpha \hat{f},\ \alpha \in \mathbb{C}.$$

2. Scaling:
   The Fourier transform of $f(ax)$, a function dilated by $a \in \mathbb{R}$ is given by
$$\begin{aligned} \int_{-\infty}^{\infty} f(ax) e^{-i\xi x} dx &= \int_{-\infty}^{\infty} f(y) e^{-\frac{i\xi y}{a}} \frac{dy}{a} \\ &= \frac{1}{a} \hat{f}\left(\frac{\xi}{a}\right). \end{aligned} \tag{3.31}$$

3. Translation:
   Let f_t be the function f shifted by t, i.e. $f_t(x) = f(x-t)$. The Fourier transform of $f_t(x)$ is given by
$$\begin{aligned} \widehat{f_t}(\xi) &= \int_{-\infty}^{\infty} f(x-t) e^{-i\xi x} dx \\ &= \int f(y) e^{-i\xi(y+t)} dy \\ &= e^{-i\xi t} \hat{f}(\xi). \end{aligned} \tag{3.32}$$

4. Reality:
   If f is a real valued function then the Fourier transform satisfies $\hat{f}(\xi) = \overline{\hat{f}(-\xi)}$. This shows that the Fourier transform of a real valued function is completely determined by its values for positive (or negative) frequencies.

5. Evenness:
   If $f(x) = f(-x)$ then $\hat{f}(\xi)$ is real valued and if $f(x) = -f(-x)$ then $\hat{f}(\xi)$ takes purely imaginary values. If f is even then its Fourier transform is given by the formula
$$\hat{f}(\xi) = 2 \int_0^{\infty} f(x) \cos(\xi x) dx. \tag{3.33}$$

Exercise 3.2.14. Verify properties (4) and (5).

Exercise 3.2.15. Find a formula like (3.33) for the Fourier transform of an odd function.

3.2.6 Convolution

See: A.6.1.

Another operation intimately connected to the Fourier transform is the convolution product.

Definition 3.2.6. If f is an integrable function and g is bounded and integrable, then the *convolution product* of f and g is defined by the absolutely convergent integral

$$f * g(x) = \int_{-\infty}^{\infty} f(y)g(x-y)dy. \tag{3.34}$$

The convolution, $f * g$ is also an integrable function.

$$\begin{aligned}\int_{-\infty}^{\infty} |f * g(x)|dx &= \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} f(y)g(x-y)dy \right| dx \\ &\leq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |f(y)g(x-y)|dydx \qquad \text{change the order of integration and set} \\ &\qquad\qquad x - y = t \text{ to obtain} \\ &= \int_{-\infty}^{\infty} |f(y)|dy \int_{-\infty}^{\infty} |g(t)|dt.\end{aligned} \tag{3.35}$$

This estimate is called the Hausdorff-Young inequality. Using it the convolution product can be extended, by continuity as a map from $L^1(\mathbb{R}) \times L^1(\mathbb{R})$ to $L^1(\mathbb{R})$, see exercise 3.2.16.

To understand convolution it is useful to think of f as a non-negative weighting function, then $f * g(x)$ is a weighted average of the values of g at points near to x. The value $g(x-y)$ is given weight $f(y)$.

Example 3.2.9. Suppose that $r_1(x)$ is the rectangle function defined in (3.9). If f is a locally integrable function then the integrals in (3.34) make sense. The function $\frac{1}{2}f * r_1$ is the weighted average of f whose value at x is the average of the values of f over the interval $[x-1, x+1]$. For example the function $r_1 * r_1$ is given by

$$\frac{1}{2}r_1 * r_1(x) = \begin{cases} 0 & \text{if } |x| > 2, \\ \frac{|2+x|}{2} & \text{if } -2 \leq x \leq 0, \\ \frac{|2-x|}{2} & \text{if } 0 \leq x \leq 2. \end{cases}$$

A basic property of the convolution is that the Fourier transform of $f * g$ is the product of $\hat{f}(\xi)$ and $\hat{g}(\xi)$.

Proposition 3.2.6. *Suppose that f and g are absolutely integrable then*

$$\widehat{f * g}(\xi) = \hat{f}(\xi)\hat{g}(\xi). \tag{3.36}$$

Proof. Because $f * g(x)$ is absolutely integrable it has a Fourier transform. Since $f(y)g(x-y)$ is an absolutely integrable function of (x, y) the following manipulations are easily justified:

$$\begin{aligned}
\widehat{f * g}(\xi) &= \int_{-\infty}^{\infty} (f * g)(x) e^{-i\xi x} dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y) g(x-y) e^{-i\xi x} dy dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y) g(t) e^{-i\xi(y+t)} dt dy \\
&= \hat{f}(\xi)\hat{g}(\xi).
\end{aligned} \tag{3.37}$$

□

Example 3.2.10. For each $R > 0$ a partial inverse to the Fourier transform is defined by

$$\begin{aligned}
S_R(f)(x) &= \frac{1}{2\pi} \int_{-R}^{R} \hat{f}(\xi) e^{i\xi x} d\xi \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) \chi_{[-\xi,\xi]}(\xi) e^{i\xi x} d\xi
\end{aligned} \tag{3.38}$$

If f is absolutely integrable then a simple change of variables shows that $S_R(f)$ can be expressed as a convolution

$$S_R(f) = \int_{-\infty}^{\infty} f(y) \frac{\sin(R(x-y))}{\pi(x-y)} dy. \tag{3.39}$$

This explicit formula is very useful in studying the behavior of $S_R(f)$ as $R \to \infty$. The subtlety in this problem is caused by the fact that $\frac{\sin(x)}{x}$ is not absolutely integrable.

Example 3.2.11. This allows us to easily compute the Fourier transform of functions like $r_1 * r_1$ or even the j-fold convolution of r_1 with itself

$$\underbrace{r_1 * \cdots * r_1}_{j-\text{times}} \stackrel{d}{=} r_1 *_j r_1.$$

It is given by

$$\widehat{r_1 *_j r_1} = \left[\frac{2\sin\xi}{\xi}\right]^j.$$

Formula (3.37) also suggests that the convolution of an L^1-function and an L^2-function should be defined. For if $f \in L^1(\mathbb{R})$ then $\hat{f}(\xi)$ is a bounded function and so if $g \in L^2(\mathbb{R})$ then $\hat{f}\hat{g} \in L^2(\mathbb{R})$ as well. This is indeed the case, even though the integral defining $f * g$ need not be absolutely convergent, see exercise (3.2.17).

Using the linearity of the integral, simple changes of variable and interchanges in the order of integration it can easily be shown that the convolution product is commutative, associative, and distributive.

Proposition 3.2.7. *If $f \in L^1(\mathbb{R})$ and g, h belong to $L^1(\mathbb{R})$ or $L^2(\mathbb{R})$ then*

$$g * f = f * g, \quad (f * g) * h = f * (g * h), \quad f * (g + h) = f * g + f * h.$$

If either $\hat{f}$ or $\hat{g}$ decreases rapidly then so does $\hat{f}(\xi)\hat{g}(\xi)$. If $\hat{g}$ decreases rapidly then by Proposition 3.2.3 this implies that g is a smooth function with integrable derivatives. Applying this proposition again we see that $f * g$ is also a smooth function with integrable derivatives.

Theorem 3.2.3. *If $\int |f(x)|dx < \infty$ and g has k continuous derivatives for which there is a constant M so that*

$$|\partial_x^j g(x)| < M \text{ for all } j \leq k \text{ and all } x.$$

Then $f * g$ has k continuous derivatives with

$$\partial_x^j (f * g)(x) = f * (\partial_x^j g)(x).$$

Proof. In light of the hypotheses on f and g, for any $j \leq k$ we can interchange the order of differentiation and integration to obtain

$$\partial_x^j (f * g)(x) = \partial_x^j \int_{-\infty}^{\infty} f(y)g(x-y)dy = \int_{-\infty}^{\infty} f(y)\partial_x^j g(x-y)dy = f * (\partial_x^j g)(x).$$

□

As a consequence, convolution can be used to approximate a locally integrable function by a smooth function. To do that we choose an infinitely differentiable function, ϕ satisfying the conditions

- ϕ is positive on $(-1/2, 1/2)$,
- ϕ vanishes outside of $(-1, 1)$.

An example is given by the function

$$\phi(x) = \begin{cases} e^{-\frac{1}{1-x^2}} & \text{if } |x| < 1, \\ 0 & \text{if } |x| \geq 1. \end{cases}$$

Let $\varphi(x)$ be a constant multiple of $\phi(x)$, with the constant chosen so that

$$\int_{-\infty}^{\infty} \varphi(x)dx = 1.$$

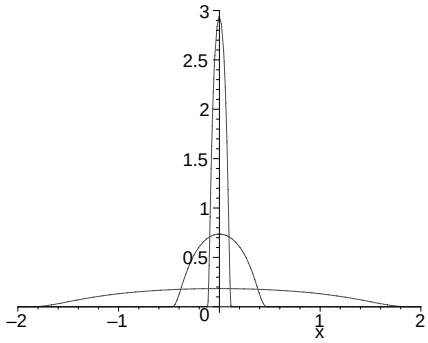


Figure 3.3: Graphs of φ_ϵ, with $\epsilon = .5, 2, 8$.

For $\epsilon > 0$ let

$$\varphi_\epsilon(x) = \epsilon^{-1}\varphi(x/\epsilon),$$

see figure 3.3. Observe that φ_ϵ is supported in $[-\epsilon, \epsilon]$ and

$$\int \varphi_\epsilon(x)dx = \int \frac{1}{\epsilon}\varphi(\frac{x}{\epsilon})dx = \int \varphi(y)dy = 1.$$

The Fourier transform of φ_ϵ is computed using (3.31),

$$\widehat{\varphi_\epsilon}(\xi) = \hat{\varphi}(\epsilon\xi). \tag{3.40}$$

From Proposition 3.2.3 it follows that $\varphi_\epsilon * f$ is an infinitely differentiable function. Note that

$$\hat{\varphi}(0) = \int_{-\infty}^{\infty} \varphi(x)dx = 1.$$

This allows us to understand what happens to $\varphi_\epsilon * f$ as $\epsilon \to 0$:

$$\begin{aligned} \widehat{\varphi_\epsilon * f}(\xi) &= \hat{\varphi}_\epsilon(\xi)\hat{f}(\xi) \\ &= \hat{\varphi}(\epsilon\xi)\hat{f}(\xi) \to \hat{\varphi}(0)\hat{f}(\xi) \\ &= \hat{f}(\xi) \quad \text{as} \quad \epsilon \to 0. \end{aligned} \tag{3.41}$$

In some sense, $\varphi_\epsilon * f$ converges to f as ϵ tends to 0. If we think of f as representing a noisy signal then $\varphi_\epsilon * f$ is a smoothed out version of f. In applications ϵ is a measure of the resolution available in $\varphi_\epsilon * f$. A larger ϵ results in a more blurred, but less noisy signal. A smaller ϵ gives a better approximation, however at the cost of less noise reduction.

At a point, x where f is continuous $\varphi_\epsilon * f(x)$ converges to $f(x)$.

Proposition 3.2.8. *Let f be a locally integrable function and suppose that φ is a non-negative function with bounded support and total integral 1. If f is continuous at x then*

$$\lim_{\epsilon \downarrow 0} \varphi_\epsilon * f(x) = f(x).$$

Proof. As f is continuous at x, given $\eta > 0$ there is a $\delta > 0$ so that

$$|x - y| < \delta \Rightarrow |f(x) - f(y)| < \eta. \tag{3.42}$$

As φ has bounded support there is an ϵ_0 such that if $\epsilon < \epsilon_0$ then the support of φ_ϵ is contained in the interval $(-\delta, \delta)$. Finally since the total integral of φ_ϵ is 1 we have, for an $\epsilon < \epsilon_0$ that

$$\begin{aligned} |\varphi_\epsilon * f(x) - f(x)| &= \left| \int_{-\delta}^{\delta} \varphi_\epsilon(x-y)(f(y) - f(x))dx \right| \\ &\leq \int_{-\delta}^{\delta} \varphi_\epsilon(x-y)|f(y) - f(x)|dx \\ &\leq \int_{-\delta}^{\delta} \varphi_\epsilon(x-y)\eta dx \\ &\leq \eta. \end{aligned} \tag{3.43}$$

In the second line we use that fact that φ_ϵ is non-negative and, in the third line, estimate (3.42). □

There are many variants of this result, the L^2-result is very simple to prove.

Proposition 3.2.9. *If $f \in L^2(\mathbb{R})$ then, as ϵ goes to zero, $\varphi_\epsilon * f$ converges, in the mean to f.*

Proof. Parseval's formula and (3.36) gives

$$\|f - \varphi_\epsilon * f\|_{L^2}^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} |1 - \hat{\varphi}(\epsilon\xi)|^2 |\hat{f}(\xi)|^2 d\xi.$$

Using the Lebesgue dominated convergence theorem it is very easy to show that the right hand side tends to zero as ϵ tends to zero. □

In applications, a function like $\phi(x)$ can be difficult to work with. To simplify computations a finitely differentiable version may be preferred. For example for each $k \in \mathbb{N}$ define

$$\psi_k(x) = \begin{cases} c_k(1 - x^2)^k & \text{if } |x| \leq 1, \\ 0 & \text{if } |x| > 1. \end{cases} \tag{3.44}$$

The constant, c_k is selected so that ψ_k has total integral one. The function ψ_k has $k - 1$ continuous derivatives. If

$$\psi_{k,\epsilon}(x) = \epsilon^{-1}\psi_{k,\epsilon}(\frac{x}{\epsilon})$$

and f is locally integrable, then $< \psi_{k,\epsilon} * f >$ is a sequence of $k - 1$-times differentiable functions, which converge, in an appropriate sense to f.

Using these facts we can complete the proof of the Fourier inversion formula. Thus far Theorem 3.2.1 was proved with the additional assumption that f is continuous.

Proof of the Fourier inversion formula, completed. Suppose that f and $\hat{f}$ are absolutely integrable and φ_ϵ is as above. Note that $\hat{f}(\xi)$ is a continuous function. For each $\epsilon > 0$ the function $\varphi_\epsilon * f$ is absolutely integrable and continuous. Its Fourier transform is $\hat{\varphi}(\epsilon\xi)\hat{f}(\xi)$, which is absolutely integrable. By Proposition 3.2.8 it converges locally uniformly to $\hat{f}(\xi)$. Since these functions are continuous we can use the Fourier inversion formula to conclude that

$$\varphi_\epsilon * f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{\varphi}(\epsilon\xi)\hat{f}(\xi)^{ix\xi} d\xi.$$

This is a locally uniformly convergent family of continuous functions and therefore has a continuous limit. The right hand side converges pointwise to

$$F(x) = \int_{-\infty}^{\infty} \hat{f}(\xi)^{ix\xi} d\xi.$$

On the other hand (3.35) implies that $\|\varphi_\epsilon * f - f\|_{L^1}$ also goes to zero as ϵ tends to 0 and therefore $F(x) = f(x)$. (To be precise we should say that after modification on a set of measure 0, $F(x) = f(x)$.) This completes the proof of the Fourier inversion formula in the special case that both f and $\hat{f}$ are integrable. □

Exercise 3.2.16. For $f \in L^1(\mathbb{R})$ define

$$f_B(x) = \begin{cases} f(x) & \text{if } |f(x)| \leq B, \\ B & \text{if } |f(x)| > B. \end{cases}$$

Show that $\lim_{B\to\infty} \|f - f_B\|_{L^1} = 0$. Use this fact and the Hausdorff-Young inequality, (3.35) to show that the sequence $< f_B * g >$ has a limit in $L^1(\mathbb{R})$. Explain how to extend the convolution product as a bounded map from $L^1(\mathbb{R}) \times L^1(\mathbb{R}) \to L^1(\mathbb{R})$.

Exercise 3.2.17. Using a modification of the argument given in (3.35) and the Cauchy-Schwarz inequality show that if $f \in L^1(\mathbb{R})$ and $g \in L^2(\mathbb{R})$ then $f * g \in L^2(\mathbb{R})$ with

$$\|f * g\|_{L^2} \leq \|f\|_{L^1} \|g\|_{L^2}.$$

In other words $f * g$ is also in $L^2(\mathbb{R})$.

Exercise 3.2.18. Use the previous exercise to show that Proposition 3.2.6 is also true if $f \in L^2(\mathbb{R})$ and $g \in L^1(\mathbb{R})$.

Exercise 3.2.19. Let f be an integrable function with support in the interval $[a, b]$ and g an integrable function with support in $[-\epsilon, \epsilon]$. Show that the support of $f * g(x)$ is contained in $[a - \epsilon, b + \epsilon]$.

Exercise 3.2.20. If we suppose that g has bounded support then Theorem 3.2.3 remains true under the assumption that f is locally integrable. That is for any $-\infty < a < b < \infty$ the integral

$$\int_a^b |f(x)| dx < \infty.$$

Exercise 3.2.21. Use Corollary A.6.1 and Proposition 3.2.8 to prove that if $f \in L^p(\mathbb{R})$ for a $1 \leq p < \infty$ then $\varphi_\epsilon * f$ converges to f in the L^p-norm.

Exercise 3.2.22. For the functions ψ_k, defined in (3.44), find the constants c_k so that

$$\int_{-1}^{1} \psi_k(x)dx = 1.$$

Exercise 3.2.23. Use the Fourier inversion formula to prove that

$$\widehat{fg}(\xi) = \frac{1}{2\pi}\hat{f} * \hat{g}(\xi). \tag{3.45}$$

Exercise 3.2.24. . Give a detailed proof of (3.39). What is the Fourier transform of $\frac{\sin(Rx)}{x}$? Find it without using a direct computation!

3.2.7 Convolution equations

Convolution provides a model for many measurement processes. If f is the state of a system then, for a fixed function ψ, the measurement g is modeled by the convolution

$$f * \psi = g.$$

In order to recover the state of the system from the measurements one must therefore *solve* this equation for f as a function of g. Formally this equation is easy to solve, (3.37) implies that

$$\hat{f}(\xi) = \frac{\hat{g}(\xi)}{\hat{\psi}(\xi)}.$$

There are several problems with this approach. The most obvious problem is that the Fourier transform, $\hat{\psi}$ may vanish for some values of ξ. If the model were perfect then, of course, $\hat{g}(\xi)$ would also have to vanish at the same points. In real applications this leads to serious problems with stability. A second problem is that, if $\psi(x)$ is absolutely integrable, then the Riemann Lebesgue lemma implies that $\hat{\psi}(\xi) \to 0$ as $|\xi| \to \infty$. Unless the measurement $g(x)$ is very smooth and noise free we would be unable to invert this Fourier transform to determine f. In Chapter 7 we discuss how these issues are handled in practice.

Example 3.2.12. The simplest weighting function is perhaps $r_\epsilon(x) = (2\epsilon)^{-1}r_1(\epsilon^{-1}x)$. Its Fourier transform is given by

$$\hat{r}_\epsilon(\xi) = \frac{\sin(\epsilon\xi)}{\epsilon\xi}.$$

This function has zeros at $\xi = \pm(\epsilon^{-1}m\pi)$, where m is any positive integer. If we convolve r_ϵ with itself we get a smoother weighting function. From (3.37) it follows that

$$\widehat{r_\epsilon * r_\epsilon}(\xi) = \left[\frac{\sin(\epsilon\xi)}{\epsilon\xi}\right]^2.$$

This function has the same zeros as $\hat{r}_\epsilon$ however it is always non-negative.

Example 3.2.13. Suppose that ψ is a non-negative function which vanishes outside the interval $[-\epsilon, \epsilon]$ and has total integral 1,

$$\int_{-\infty}^{\infty} \psi(x)dx = 1.$$

If f is a locally integrable function then $f * \psi(x)$ is the weighted average of the values of f over the interval $[x - \epsilon, x + \epsilon]$. Note that $\psi * \psi$ also has total integral 1

$$\begin{aligned} \int_{-\infty}^{\infty} \psi * \psi(x)dx &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(y)\psi(x-y)dydx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(y)\psi(t)dtdy \\ &= 1 \cdot 1 = 1. \end{aligned} \tag{3.46}$$

In the second to last line we reversed the order of the integrations and set $t = x - y$.

Thus $f * (\psi * \psi)$ is again an average of f. Note that $\psi * \psi(x)$ is generally non-zero for $x \in [-2\epsilon, 2\epsilon]$, so convolving with $\psi * \psi$ produces more blurring than convolution with ψ alone. Indeed we know from the associativity of the convolution product that

$$f * (\psi * \psi) = (f * \psi) * \psi,$$

so we are averaging the averages, $f * \psi$. This can be repeated as many times as one likes, the j-fold convolution $\psi *_j \psi$ has total integral 1 and vanishes outside the interval $[-j\epsilon, j\epsilon]$. Of course the Fourier transform of $\psi *_j \psi$ is $[\hat{\psi}(\xi)]^j$ which therefore decays j times as fast as $\hat{\psi}(\xi)$.

We could also use the *scaled* j-fold convolution $\delta^{-1}\psi *_j \psi(\delta^{-1}x)$ to average our data. This function vanishes outside the interval $[-j\delta\epsilon, j\delta\epsilon]$ and has Fourier transform $[\hat{\psi}(\delta\xi)]^j$. If we choose $\delta = j^{-1}$ then convolving with this function will not blur details any more than convolving with ψ itself but better suppresses high frequency noise. By choosing j and δ we can control, to some extent, the trade off between blurring and noise suppression.

Exercise 3.2.25. If a and b are positive numbers then define

$$w_{a,b}(x) = \frac{1}{2}[r_a(x) + r_b(x)].$$

Graph $w_{a,b}(x)$ for several different choices of (a, b). Show that for appropriate choices of a and b the Fourier transform $\hat{w}_{a,b}(\xi)$ does not vanish for any value of ξ.

Exercise 3.2.26. Define a function

$$f(x) = \chi_{[-1,1]}(x)(1 - |x|)^2.$$

Compute the Fourier transform of this function and show that it does not vanish anywhere. Let $f_j = f *_j f$ (the j-fold convolution of f with itself). Show that the Fourier transforms, $\hat{f}_j(\xi)$ are also non-vanishing.

3.2.8 The δ-function

See: A.4.6.

In section 3.2.6 we considered convolution with the family of functions,

$$\varphi_\epsilon(x) = \epsilon^{-1}\varphi(\epsilon^{-1}x),$$

and saw that, as $\epsilon \to 0$ the functions $< \varphi_\epsilon * f >$ tend to f. One might reasonably enquire why we bother with this family, why not just use a single function ψ which satisfies $\psi * f = f$ for all locally integrable functions? The reason is that no such integrable function exists. This is very easy to see by computing the Fourier transform. If $\psi * f = f$ then $\hat{\psi}(x)\hat{f}(\xi) = \hat{f}(\xi)$ for *all* $f \in L^1(\mathbb{R})$ and *all* ξ. This clearly implies that $\hat{\psi}(\xi) = 1$ for all ξ. This shows that ψ cannot be an integrable function, because its Fourier transform violates the Riemann-Lebesgue Lemma.

One can see that no such function exists more intuitively. It is clear that such a function would need to be non-negative and have total integral 1. On the other hand if $\psi * f(x) = f$ for all functions it is also clear that ψ must vanish outside $[-\epsilon, \epsilon]$ for *any* $\epsilon > 0$. In other words $\psi(x) = 0$ for all $x \neq 0$ but $\int \psi = 1$! Again we easily see that no such function exists. Because it is such a useful concept in both engineering and mathematics we introduce an object which is *not* a function but has the desired properties. It is called the δ-function and is denoted by $\delta(x)$. It has the following properties:

(1). If f is a locally integrable function then

$$\delta * f = f,$$

(2). The Fourier transform of $\delta(x)$ is the constant function, 1

$$\mathcal{F}(\delta)(\xi) = 1 \text{ for all } \xi.$$

In the mathematics literature the δ-function is an example of a *distribution* or *generalized function*. The basic properties of generalized functions are introduced in Appendix A.4.6. In the engineering and physics literature it is sometimes called a *unit impulse*. A function like φ_ϵ, for a very small ϵ, is an approximate unit impulse. The Fourier transform of φ_ϵ is $\hat{\varphi}(\epsilon\xi)$. Because φ_ϵ vanishes outside a finite interval its Fourier transform is always a smooth function and $\hat{\varphi}(0) = 1$. Because φ is non-negative, $|\hat{\varphi}(\xi)| < 1$ if $\xi \neq 0$. If φ is an even function then $\hat{\varphi}$ is real valued.

In applications it is usually important that the difference $1 - \hat{\varphi}(\epsilon\xi)$ remain small over a specified interval $[-B, B]$. It is also important that $\hat{\varphi}(\epsilon\xi)$ tend to zero rapidly outside a somewhat larger interval. As φ is non-negative, $\partial_\xi\hat{\varphi}(0) = 0$; this means that the behavior of $\hat{\varphi}(\xi)$ for ξ near to zero is largely governed by

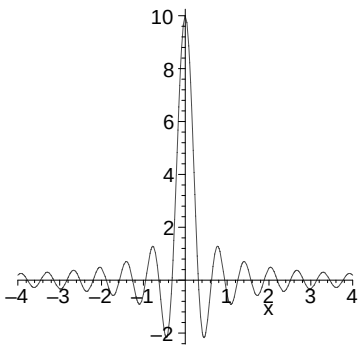
$$\partial_\xi^2\hat{\varphi}(0) = \int_{-\infty}^{\infty} x^2\varphi(x)dx.$$

One would like this number to be small. This is accomplished by putting more of the mass of φ near to $x = 0$. On the other hand the rate at which $\hat{\varphi}(\xi)$ decays as $|\xi| \to \infty$ is determined by the smoothness of $\varphi(x)$. Using the simplest choice, $\varphi(x) = \frac{1}{2}r_1(x)$ leads to a Fourier transform that decays like $|\xi|^{-1}$. Better decay is obtained by using a smoother function. In applications it is usually adequate to have an absolutely integrable Fourier transform as results if φ is continuous and piecewise differentiable.

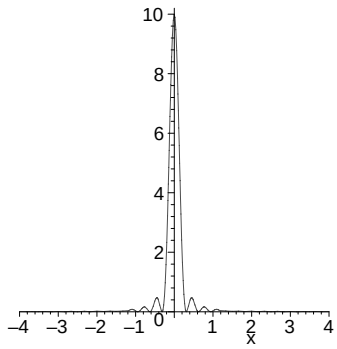
Another approach to constructing approximations to the δ-function is to work on the Fourier transform side. Here one uses a sequence of functions which are approximately 1 in an interval $[-B, B]$ and vanish outside a larger interval. Again the simplest choice is $\chi_{[-B,B]}(\xi)$. The inverse Fourier transform of this function is

$$\psi_B(x) = \frac{\sin(Bx)}{\pi x},$$

see exercise 3.2.24. This is called a *sinc pulse*. At $x = 0$ we have $\psi_B(0) = \pi^{-1}B$. Note that ψ_B assumes both positive and negative values. The fact that the improper integral of ψ_B over the whole real line equals 1 relies on very subtle cancelations between the positive and negative parts of the integral. Because the function ψ_B is not absolutely integrable, it is often not a good choice for approximating the δ-function. For example, using $(2B)^{-1}\chi_{[-B,B]} * \chi_{[-B,B]}(\xi)$ to approximate the Fourier transform of δ gives $(2B)^{-1}\psi_B^2(x)$ as an approximation to $\delta(x)$. This function has much better properties: it does not assume negative values, is more sharply peaked at 0 and tends to zero more rapidly. These functions are graphed in figure 3.4.



(a) A sinc-pulse.

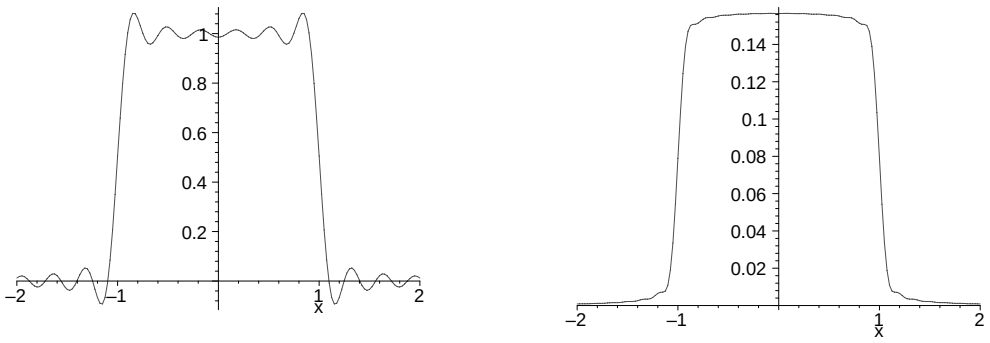


(b) A sinc^2-pulse.

Figure 3.4: Approximate δ-functions

Note that neither of these functions vanishes outside a bounded interval. The graphs of both functions have oscillatory "tails" extending to infinity. In the engineering literature these are called *side lobes*. The presence of side lobes is an inevitable consequence of the fact that the Fourier transforms vanish outside a bounded interval, see section 3.2.14. The results of convolving these functions with $\chi_{[-1,1]}$ are shown in figure 3.5. Notice the large oscillations, near the jump, present in figure 3.5(a). This is an example of the "Gibbs

phenomenon." It results from using a discontinuous cutoff function in the Fourier domain. This effect is analyzed in detail, for the case of Fourier series in section 5.5.



(a) $\mathrm{sinc} * \chi_{[-1,1]}$. (b) $\mathrm{sinc}^2 * \chi_{[-1,1]}$.

Figure 3.5: Approximate δ-functions convolved with $\chi_{[-1,1]}$.

3.2.9 Windowing and resolution

In this section we give a standard definition for the resolution present in a measurement of the form $h * f$. Resolution is a subtle and, in some senses, subjective concept. Suppose that $h(x)$ is a non-negative function with a single hump similar to those shown in figure 3.3. The important features of this function are

$$\begin{aligned} &1.\ \text{It is non-negative,} \\ &2.\ \text{It has a single maximum value, which it attains at 0,} \\ &3.\ \text{It is monotone increasing to the left of the maximum} \\ &\quad \text{and monotone decreasing to the right.} \end{aligned} \tag{3.47}$$

Definition 3.2.7. Let h satisfy these conditions and let M be the maximum value it attains. Let $x_1 < 0 < x_2$ be respectively the smallest and largest numbers so that

$$h(x_1) = h(x_2) = \frac{M}{2}.$$

The difference $x_2 - x_1$ is called the *full width half maximum* of the function h. It is denoted FWHM(h). If f is an input then the resolution available in the measurement, $h * f$ is *defined* to be the FWHM(h).

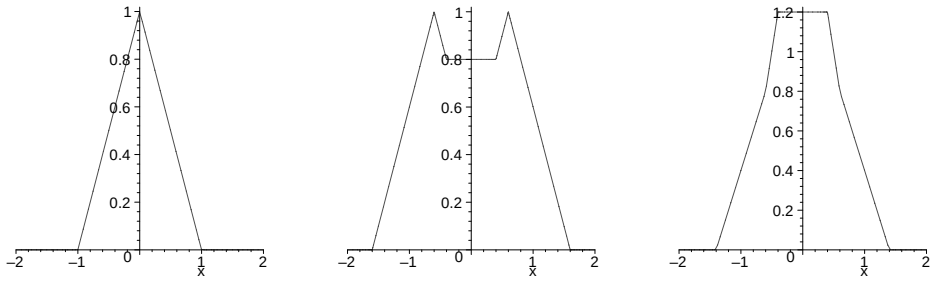
Here is a heuristic explanation for this definition. Suppose that the signal f is pair of unit impulses separated by a distance d,

$$f(x) = \delta(x) + \delta(x - d).$$

Convolving $h(x)$ with f produces two copies of h,

$$h * f(x) = h(x) + h(x - d).$$

If $d >$ FWHM(h) then $h*f$ has two distinct maxima separated by a valley. If $d \leq$ FWHM(h) then the distinct maxima disappear. If the distance between the impulses is greater than the FWHM(h) then we can "resolve" them in the filtered output. In figure 3.6 we use a triangle function for h. The FWHM of this function is 1, the graphs show h and the results of convolving h with a pair of unit impulses separated, respectively by $1.2 > 1$ and $.8 < 1$.



(a) $h(x)$ (b) $h(x+.6)+h(x-.6)$ (c) $h(x+.4)+h(x-.4)$

Figure 3.6: Illustration of the FWHM definition of resolution

This definition is often extended to functions which do not satisfy all the conditions in (3.47) but are qualitatively similar. For example the characteristic function of an interval $\chi_{[-B,B]}(x)$ has a unique maximum value and is monotone to the right and left of the maximum. The FWHM$(\chi_{[-B,B]})$ is therefore $2B$. Another important example is the sinc-function,

$$\text{sinc}(x) = \frac{\sin(x)}{x},$$

the Fourier transform of $2^{-1}\chi_{[-1,1]}$. It has a unique maximum and looks correct near to it. It also has large side-lobes which considerably complicate the behavior of the map $f \mapsto f * \text{sinc}$. The FWHM(sinc$(x)$) is taken to be the full width half maximum of its central peak, it is approximately given by

$$\text{FWHM}(\text{sinc}(x)) = 1.8954942670339809471.$$

As stated before, a measurement is often modeled as a convolution. If we let

$$h_\epsilon(x) = 2\epsilon^{-1}\chi_{[-\epsilon,\epsilon]}(x)$$

then FWHM$(h_\epsilon) = 2\epsilon$. So the resolution in the measurement $h_\epsilon * f$ is 2ϵ. Suppose that $f(x)$ represents an input which is only non-zero in a finite interval, $x \in (-L, L)$ and $\hat{f}(\xi)$ is its Fourier transform. The measurement is then $h * f$. For reasons of computational efficiency, convolutions are usually computed by using the Fourier transform. That is, to calculate

$h * f$ one actually computes $\mathcal{F}^{-1}(\hat{h}\hat{f})$. For practical computations we are only able to use the information in Fourier transform over a finite range of frequencies. So the first step in applying a filter is cutting off $\hat{f}(\xi)$ outside a finite interval. This is called *windowing* the Fourier transform of f.

The simplest way to window $\hat{f}$ is to replace it with $\chi_{[-B,B]}(\xi)\hat{f}(\xi)$. Windowing in the ξ-variable becomes convolution in the x-variable, f is replaced with

$$\mathcal{F}^{-1}(\chi_{[-B,B]}\hat{f})(x) = \frac{B\,\mathrm{sinc}(Bx)}{\pi} * f.$$

As noted above the $\mathrm{FWHM}(\mathrm{sinc}(Bx)) \propto \frac{1}{B}$. So the effect of cutting off the Fourier transform for $|\xi| > B$ is to reduce the resolution of the output to $O(B^{-1})$. This is a very crude description of the consequences of cutting off the Fourier transform of f. The side lobes produce other effects which may be more pronounced than the reduction in resolution. This is because $\int |\,\mathrm{sinc}(x)|dx = \infty$; which is in turn a reflection of the fact $\chi_{[-B,B]}$ is discontinuous.

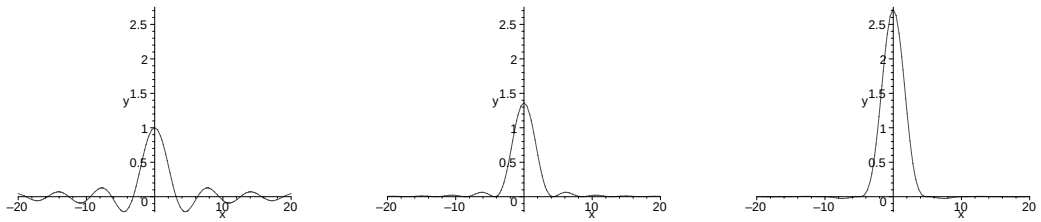
Another possible choice is to use

$$\chi_{2,B}(\xi) = \frac{1}{2B}\chi_{[-B,B]} * \chi_{[-B,B]}(\xi). \tag{3.48}$$

This leads to a sinc^2-pulse, which is non-negative, absolutely integrable and does not suffer from the Gibbs phenomenon. On the other hand it requires a knowledge of the Fourier transform of f over twice the range $[-2B, 2B]$, the

$$\mathrm{FWHM}(\mathcal{F}^{-1}(\chi_{2,B})) \simeq \frac{1.32565430}{B} = \frac{2.65130859}{2B}.$$

Comparing this with the computation of the $\mathrm{FWHM}(\mathrm{sinc}\,Bx)$ we see that, even though the window is twice as long, the resolution has only increased by %22. Figure 3.7 shows the graphs of $[\mathrm{sinc}(B_j x)]^j (B_j)^{1-j}$ for $j = 1, 2, 3$. The values of B_j are selected so that each function has approximately the same FWHM.



(a) j=1 (b) j=2 (c) j=3

Figure 3.7: FWHM vs. side-lobes

Exercise 3.2.27. Suppose that

$$h_j(x) = \left[\frac{\sin(x)}{x}\right]^j.$$

Using the Taylor expansion for sine function show that, as j gets large,

$$\mathrm{FWHM}(h_j) \simeq \sqrt{\frac{6\log 2}{j}}.$$

Exercise 3.2.28. Using the Taylor expansion for the sine, show that as B gets large

$$\mathrm{FWHM}(\mathrm{sinc}(Bx)) \simeq \frac{\sqrt{3}}{B}.$$

3.2.10 Functions with L^2-derivatives*

See: A.4.6.

If f and g are differentiable functions which vanish outside a bounded interval then the integration by parts formula states that

$$\int_{-\infty}^{\infty} f'(x)\overline{g(x)}dx = -\int_{-\infty}^{\infty} f(x)\overline{g'(x)}dx.$$

This formula suggests a way to extend the notion of differentiability to some functions which do not actually have a *classical* derivative. Suppose that f is a locally integrable function and there exists another locally integrable function f_1 such that, for any $\mathcal{C}^1$-function g which vanishes outside a bounded interval, we have the identity

$$\int_{-\infty}^{\infty} f_1(x)\overline{g(x)}dx = -\int_{-\infty}^{\infty} f(x)\overline{g'(x)}dx.$$

From the point of view of measurements, the function f_1 *looks like* the derivative of f. If this condition holds then we say that f has a *weak derivative* and write $f' = f_1$. In this context the function g is called a *test function*. It is clear from the definition that a function which is differentiable in the ordinary sense is weakly differentiable and the two definitions of derivative agree. The notion of weak derivative extends the concept of differentiability to a larger class of functions. In fact this definition can be used to define derivatives of *generalized functions*. This topic is discussed in Appendix A.4.6; the reader is urged to look over this section before proceeding.

It is easy to see from examples that a weak derivative can exist even when f does not have a classical derivative.

Example 3.2.14. The function

$$f(x) = \begin{cases} 0 & \text{if } |x| > 1, \\ |x+1| & \text{if } -1 \le x \le 0, \\ |x-1| & \text{if } 0 \le x \le 1 \end{cases}$$

does not have a classical derivative at $x = -1, 0$ and 1. However the function

$$g(x) = \begin{cases} 0 & \text{if } |x| > 1, \\ 1 & \text{if } -1 \leq x \leq 0, \\ -1 & \text{if } 0 \leq x \leq 1 \end{cases}$$

is the weak derivative of f.

A very useful condition is for the weak derivative to be in L^2.

Definition 3.2.8. Let $f \in L^2(\mathbb{R})$ we say that f has an L^2-*derivative* if there is a function $f_1 \in L^2(\mathbb{R})$ so that for *every* once, differentiable function, φ vanishing outside a finite interval the formula

$$\langle f, \varphi' \rangle = -\langle f_1, \varphi \rangle$$

holds. In this case we say that f_1 is the L^2-derivative of f. We use the usual notations for the L^2-derivative, i.e. f' or $\partial_x f$, etc.

A function $f \in L^2$ which is differentiable in the ordinary sense and whose derivative is $f' \in L^2$ is also differentiable in the L^2-sense. Its L^2-derivative is just f'. An important fact about L^2-derivatives is that they satisfy the fundamental theorem of calculus. Above it was shown that an L^2-function is locally integrable. If f_1 is the L^2-derivative of f then for any $a < b$ we have that

$$f(b) - f(a) = \int_a^b f_1(x)dx.$$

In particular a function with an L^2-derivative is continuous. Using Hölder's inequality we see that

$$|f(b) - f(a)| \leq \sqrt{|b-a|}\|f_1\|_{L^2}.$$

In other words the ratio

$$\frac{|f(b) - f(a)|}{\sqrt{|b-a|}} \leq \|f_1\|_{L^2}.$$

A function for which this ratio is bounded is called a *Hölder-*$\frac{1}{2}$ function. Such a function is said to have a *half a classical derivative*.

If $f \in L^2(\mathbb{R})$ has an L^2-derivative then the Fourier transform of f and f' are related just as they would be if f had a classical derivative

$$\widehat{f'}(\xi) = i\xi\hat{f}(\xi).$$

Moreover the Parseval identity carries over to give

$$\int_{-\infty}^{\infty} |f'(x)|^2 dx = \frac{1}{2\pi}\int_{-\infty}^{\infty} |\xi|^2 |\hat{f}(\xi)|^2 d\xi.$$

On the other hand if $\xi\hat{f}(\xi)$ is square integrable then one can show that f has an L^2-derivative and its Fourier transform is $i\xi\hat{f}(\xi)$. This is what was meant by the statement that the relationship between the smoothness of a function and the decay of the Fourier transform is very tight when these concepts are defined with respect to the L^2-norm.

Definition 3.2.9. The higher L^2-derivatives are defined exactly as in the classical case. If $f \in L^2(\mathbb{R})$ has an L^2-derivative, and $f' \in L^2$ also has an L^2-derivative, then we say that f has two L^2-derivatives. This can be repeated to define all higher derivatives. A simple condition for a function $f \in L^2(\mathbb{R})$ to have j L^2-derivatives, is that there are functions $\{f_1, \ldots, f_j\} \subset L^2(\mathbb{R})$ so that for every j-times differentiable function φ, vanishing outside a bounded interval and $1 \leq l \leq j$ we have that

$$\langle f, \varphi^{[l]} \rangle_{L^2} = (-1)^l \langle f_l, \varphi \rangle_{L^2}.$$

The function f_l is then the l^{th} L^2-derivative of f. Standard notations are also used for the higher L^2-derivatives, e.g. $f^{[l]}$, $\partial_x^l f$, etc.

The basic result about L^2-derivatives is.

Theorem 3.2.4. *A function $f \in L^2(\mathbb{R})$ has j L^2-derivatives if and only if $\xi^j \hat{f}(\xi)$ is in $L^2(\mathbb{R})$. In this case*

$$\widehat{f^{[l]}} = (i\xi)^l \hat{f}(\xi), \tag{3.49}$$

moreover

$$\int_{-\infty}^{\infty} |f^{[l]}(x)|^2 dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\xi|^{2l} |\hat{f}(\xi)|^2 d\xi. \tag{3.50}$$

Exercise 3.2.29. In example 3.2.14 prove that g is the weak derivative of f.

Exercise 3.2.30. Suppose that $f \in L^2(\mathbb{R})$ has 2 L^2-derivatives. Show that f has one classical derivative.

Exercise 3.2.31. Suppose that $f \in L^2(\mathbb{R})$ has an $L^2(\mathbb{R})$-derivative f_1. Show that if $f(x)$ vanishes for $|x| > R$ then so does f_1.

Exercise 3.2.32. Prove that if $f \in L^2(\mathbb{R})$ has an L^2-derivative then $\widehat{f'}(\xi) = i\xi \hat{f}(\xi)$.

Exercise 3.2.33. * Suppose that f and $\xi^k \hat{f}(\xi)$ belong to $L^2(\mathbb{R})$. By approximating f by smooth functions of the form $\varphi_\epsilon * f$ show that f has k L^2-derivatives.

3.2.11 Fractional derivatives and L^2-derivatives*

See: A.4.6 .

In the previous section we extended the notion of differentiability to functions which do not have a classical derivative. In the study of the Radon transform it turns out to be useful to have other generalizations of differentiability. In section 2.5.3 we defined the Hölder classes and a notion of fractional derivative. We quickly review the idea and then turn to other ideas of fractional derivatives.

The basic observation is the following: a function f has a derivative if the difference quotients

$$\frac{f(x+h) - f(x)}{h}$$

have a limit as $h \to 0$. In order for this limit to exist it is clearly necessary that the ratios

$$\frac{|f(x+h) - f(x)|}{|h|}$$

be uniformly bounded, for small h. Thus the basic estimate satisfied by a continuously differentiable function is that the ratios

$$\frac{|f(x) - f(y)|}{|x - y|}$$

are locally, uniformly bounded. The function $f(x) = |x|$ shows that these ratios can be bounded without the function being differentiable. However, from the point of view of measurements such a distinction is very hard to make.

Definition 3.2.10. Let $0 \leq \alpha < 1$, we say that a function f, defined in an interval $[a, b]$, has an α^{th}-*classical derivative* if there is a constant M so that

$$\frac{|f(x) - f(y)|}{|x - y|^{\alpha}} \leq M, \tag{3.51}$$

for all $x, y \in [a, b]$. Such a function is also said to be α-*Hölder continuous*.

The same idea can be applied to functions with L^2-derivatives. Recall that an L^2-function has an L^2-derivative if and only if $\xi \hat{f}(\xi) \in L^2(\mathbb{R})$. This is just the estimate

$$\int_{-\infty}^{\infty} |\xi|^2 |\hat{f}(\xi)|^2 d\xi < \infty.$$

By analogy to the classical case we make the following definition.

Definition 3.2.11. A function $f \in L^2(\mathbb{R})$ has an α^{th} L^2-*derivative* if

$$\int_{-\infty}^{\infty} |\xi|^{2\alpha} |\hat{f}(\xi)|^2 d\xi < \infty. \tag{3.52}$$

As before there is no canonical way to define the "α^{th}-L^2-derivative operator." The following definition is sometimes useful. For $\alpha \in (0, 1)$ define the α^{th}-L^2-derivative to be

$$D_{\alpha} f = \underset{R \to \infty}{LIM} \frac{1}{2\pi} \int_{-R}^{R} |\xi|^{\alpha} \hat{f}(\xi) e^{i\xi x} d\xi.$$

It is defined precisely those functions satisfying (3.52) which have an α^{th}-L^2-derivative. Note that this definition with $\alpha = 1$ does *not* give the expected answer.

The relationship between these two notions of fractional differentiability is somewhat complicated. As shown in the previous section: a function with 1 L^2-derivative is Hölder-$\frac{1}{2}$. On the other hand, the function $f(x) = \sqrt{x}$ is Holder-$\frac{1}{2}$. That $(\sqrt{x})^{-1}$ is not square integrable shows that having half a classical derivative does not imply that a function has one L^2-derivative.

Exercise 3.2.34. Suppose that f satisfies the estimate in (3.51) with an $\alpha > 1$. Show that f is constant.

3.2.12 Some refined properties of the Fourier transform*

See: A.3.1, A.3.3.

In this section we consider some properties of the Fourier transform that are somewhat less elementary than those considered so far. The first question we consider concerns the pointwise convergence of the inverse Fourier transform. Let f be a function in either $L^1(\mathbb{R})$ or $L^2(\mathbb{R})$; for each $R > 0$ define

$$f_R(x) = \mathcal{F}^{-1}(\chi_{[-R,R]}\hat{f})(x) = \frac{1}{2\pi}\int_{-R}^{R} \hat{f}(\xi)e^{ix\xi}d\xi.$$

Using Proposition 3.2.6 this can be expressed as a convolution:

$$f_R(x) = \int_{-\infty}^{\infty} f(y)\frac{\sin(R(x-y))}{\pi(x-y)}dy.$$

If $\hat{f}$ is absolutely integrable then Theorem 3.2.1 shows that $f(x)$ is the limit, as $R \to \infty$ of $f_R(x)$. If f is well enough behaved *near to* x then this is always the case, whether or not $\hat{f}$ (or for that matter f) is absolutely integrable. This is Riemann's famous localization principle for the Fourier transform.

Theorem 3.2.5 (Localization principle). *Let $f \in L^1(\mathbb{R}) + L^2(\mathbb{R})$ and suppose that f vanishes in a neighborhood of x_0 then*

$$\lim_{R\to\infty} f_R(x_0) = 0.$$

Proof. The proof of this result is not difficult. The hypothesis that $f \in L^1(\mathbb{R}) + L^2(\mathbb{R})$ means that $f = f_1 + f_2$ where $f_p \in L^p(\mathbb{R})$. The same proof works in either case. Using the convolution representation

$$\begin{aligned} f_R(x_0) &= \int_{-\infty}^{\infty} f(y)\frac{\sin(R(x_0-y))}{\pi(x_0-y)}dy \\ &= \int_{-\infty}^{\infty} [e^{iR(x_0-y)} - e^{-iR(x_0-y)}]\frac{f(y)}{2\pi i(x_0-y)}dy. \end{aligned} \tag{3.53}$$

Because f vanishes in an interval containing x_0, it follows that $f(y)(x_0-y)^{-1}$ is an absolutely integrable function. The conclusion of the theorem is therefore a consequence of the Riemann-Lebesgue lemma. □

Remark 3.2.5. This result has a simple corollary which makes clearer why it is called the "localization principle." Suppose that $g(x)$ is a function in $L^1(\mathbb{R}) + L^2(\mathbb{R})$ such that

$\lim_{R\to\infty} g_R(x_0) = g(x_0)$ and $f(x) = g(x)$ for x in an interval containing x_0. Then, it is also true that $f(x) - g(x) = 0$ in an interval containing x_0 and therefore

$$f(x_0) = \lim_{R\to\infty} f_R(x_0) = \lim_{R\to\infty} g_R(x_0) + \lim_{R\to\infty} (f_R(x_0) - g_R(x_0)).$$

The Fourier inversion process is sensitive to the local behavior of f. It is important to note that this result is special to one dimension. The analogous result is *false* for the Fourier transform in $\mathbb{R}^n$ if $n \geq 2$. This phenomenon is carefully analyzed in [57], see also section 3.3.8.

The next result states that if a function f has bounded support then its Fourier transform cannot.

Proposition 3.2.10. *Suppose* $\operatorname{supp} f \subset (-R, R)$ *if* $\hat{f}$ *also has bounded support then* $f \equiv 0$.

Proof. The radius of convergence of the series $\sum_0^\infty (-ix\xi)^j/j!$ is infinity, and it converges to $e^{-ix\xi}$, uniformly on bounded intervals. Combining this with the fact that f has bounded support, we conclude that we may interchange the integration with the summation to obtain

$$\begin{aligned} \hat{f}(\xi) &= \int_{-\infty}^{\infty} f(x) e^{-i\xi x} dx \\ &= \int_{-R}^{R} \sum_{j=0}^{\infty} f(x) \frac{(-ix\xi)^j}{j!} dx \\ &= \sum_{j=0}^{\infty} \frac{1}{j!} (-i\xi)^j \int_{-R}^{R} f(x) x^j dx. \end{aligned} \tag{3.54}$$

Since

$$\left| \int_{-R}^{R} f(x) x^j dx \right| \leq R^j \int_{-R}^{R} |f(x)| dx,$$

the terms of the series representing $\hat{f}(\xi)$ are bounded by the terms of a series having an infinite radius of convergence; the j^{th} term is bounded by

$$\frac{(R|\xi|)^j}{j!} \int_{-R}^{R} |f(x)| dx.$$

Therefore the series expansion for $\hat{f}(\xi)$ also has an infinite radius of convergence. This argument

can be repeated to obtain the Taylor expansion of $\hat{f}(\xi)$ about an arbitrary ξ_0 :

$$\begin{aligned}\hat{f}(\xi) &= \int_{-R}^{R} e^{-i(\xi-\xi_0)x} f(x) e^{i\xi_0 x} dx \\ &= \int_{-R}^{R} \sum_{j=0}^{\infty} \frac{[-i(\xi-\xi_0)x]^j}{j!} f(x) e^{i\xi_0 x} dx \\ &= \sum_{j=0}^{\infty} \int_{-R}^{R} \frac{[-i(\xi-\xi_0)x]^j}{j!} f(x) e^{i\xi_0 x} dx \\ &= \sum_{j=0}^{\infty} \frac{[-i(\xi-\xi_0)]^j}{j!} \int_{-R}^{R} f(x) x^j e^{i\xi_0 x} dx.\end{aligned} \tag{3.55}$$

If we let $a_j^{\xi_0} = \int f(x) x^j e^{i\xi_0 x} dx$ then

$$\hat{f}(\xi) = \sum_{0}^{\infty} a_j^{\xi_0} \frac{[-i(\xi-\xi_0)]^j}{j!}.$$

As above this expansion is valid for all ξ.

Suppose there exists ξ_0 such that $\partial_\xi^j \hat{f}(\xi_0) = 0$ for all $j = 0, 1, \ldots$ Then $\hat{f}(\xi) \equiv 0$ since all the coefficients, $a_j^{\xi_0} = \partial_\xi^j \hat{f}(\xi_0)$ equal zero. This proves the proposition. □

Remark 3.2.6. The proof actually shows that if f is supported on a finite interval and $\hat{f}$ vanishes on an open interval $(a, b) \subset \mathbb{R}$ then $f \equiv 0$.

This result indicates that one cannot obtain both arbitrarily good resolution and denoising simultaneously. A famous quantitative version of this statement is the Heisenberg uncertainty principle which we now briefly discuss., using physical terms coming from quantum mechanics. Let x be the position of a particle, the *probability* of finding the particle in the interval $[a, b]$ is $\int_a^b |f(x)|^2 dx$. We normalize so that the total probability is 1, this means that the particle is somewhere on the line. By the Parseval formula,

$$\int_{-\infty}^{\infty} |f(x)|^2 dx = \int_{-\infty}^{\infty} |\hat{f}(\xi)|^2 \frac{d\xi}{2\pi} = 1.$$

The expected value of the position of a particle is given by

$$E(x) = \int_{-\infty}^{\infty} x |f(x)|^2 dx.$$

By translating in x we can normalize f to make $E(x)$ zero. In physics, the Fourier transform of f describes the momentum of a particle. The expected value of the momentum is

$$E(\xi) = \int \xi |\hat{f}(\xi)|^2 \frac{d\xi}{2\pi}.$$

By replacing f by $e^{i\xi_0 x} f$ for an appropriate choice of ξ_0 we can also make $E(\xi) = 0$. With these normalizations, the variance of the position and the momentum, $(\Delta x)^2$ and $(\Delta\xi)^2$, are given by

$$\begin{aligned}(\Delta x)^2 &= \int_{-\infty}^{\infty} x^2 |f(x)|^2 dx, \\ (\Delta\xi)^2 &= \int_{-\infty}^{\infty} \xi^2 |\hat{f}(\xi)|^2 \frac{d\xi}{2\pi}.\end{aligned}$$

The Parseval formula implies that

$$(\Delta\xi)^2 = \int_{-\infty}^{\infty} |\partial_x f(x)|^2 dx.$$

The basic result is

Theorem 3.2.6 (The Heisenberg uncertainty principle). *If f and $\partial_x f$ belong to $L^2(\mathbb{R})$ then*

$$\int_{-\infty}^{\infty} |x|^2 |f(x)|^2 dx \int_{-\infty}^{\infty} |\xi|^2 |\hat{f}(\xi)|^2 \frac{d\xi}{2\pi} \geq \frac{1}{4} \left[\int_{-\infty}^{\infty} |f(x)|^2 \right]^2 dx. \tag{3.56}$$

Because the product of the variances has a lower bound, this means that we cannot localize the position and the momentum of a particle, arbitrarily well *at the same time*. The proof of this theorem is a simple integration by parts followed by an application of the Cauchy-Schwarz inequality for square integrable functions, see (A.97).

Proof. If f decays sufficiently rapidly, we can integration by parts to obtain that

$$\begin{aligned}\int_{-\infty}^{\infty} x f f_x dx &= \frac{1}{2}(xf^2)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{1}{2} f^2 dx \\ &= -\frac{1}{2} \int_{-\infty}^{\infty} f^2.\end{aligned} \tag{3.57}$$

The Cauchy-Schwarz inequality implies that

$$\left| \int_{-\infty}^{\infty} x f f_x dx \right| \leq \left[\int_{-\infty}^{\infty} x^2 |f|^2 dx \right]^{\frac{1}{2}} \left[\int_{-\infty}^{\infty} |f_x|^2 dx \right]^{\frac{1}{2}}$$

Using (3.57), the Parseval formula and this estimate we obtain

$$\frac{1}{2} \int_{-\infty}^{\infty} |f|^2 dx \leq \left[\int_{-\infty}^{\infty} x^2 |f|^2 dx \right]^{\frac{1}{2}} \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} \xi^2 |\hat{f}|^2 dx \right]^{\frac{1}{2}}. \tag{3.58}$$

□

With the expected position and momentum normalized to be zero, the variance in the position and momentum are given by

$$\Delta x = \left(\int_{-\infty}^{\infty} x^2 f^2 \right)^{1/2} \text{ and } \Delta \xi = \left(\int_{-\infty}^{\infty} f_x^2 \right)^{1/2}.$$

The estimate (3.58) is equivalent to $\Delta x \cdot \Delta \xi \geq \frac{1}{2}$. If a, b are non-negative numbers then the arithmetic mean-geometric inequality states that

$$ab \leq \frac{a^2 + b^2}{2}.$$

Combining this with the Heisenberg uncertainty principle shows that

$$1 \leq (\Delta x)^2 + (\Delta \xi)^2.$$

That is

$$\int_{-\infty}^{\infty} f^2 dx \leq \int_{-\infty}^{\infty} x^2 f^2 + f_x^2 dx. \tag{3.59}$$

The inequality (3.59) becomes an *equality* if we use the Gaussian function $f(x) = e^{-\frac{x^2}{2}}$. A reason why the Gaussian is often used to smooth measured data is that it provides the optimal resolution (in the L^2-norm) for a given amount of de-noising.

Exercise 3.2.35. Show that both (3.56) and (3.59) are *equalities* if $f = e^{-\frac{x^2}{2}}$. Can you show that the only functions for which this is true are multiples of f?

3.2.13 The Fourier transform of generalized functions*

See: A.4.6.

Initially the Fourier transform is defined, for absolutely integrable functions by an explicit formula, (3.3). It is then extended, in definition (3.2.5), to L^2-functions by using its *continuity* properties. The Parseval formula implies that the Fourier transform is a bounded map from $L^2(\mathbb{R})$ to itself, indeed it is an invertible, isometry. For an L^2-function, the Fourier transform is **not** defined by an integral, nonetheless the Fourier transform on $L^2(\mathbb{R})$ shares all the important properties of the Fourier transform defined earlier for absolutely integrable functions.

It is reasonable to enquire as to the largest class of functions to which the Fourier transform can be extended. In turns out that the answer is *not* a class of functions, but rather the generalized functions (or tempered distributions) defined in section A.4.6. Before proceeding the reader is strongly urged to read this section! The definition of the Fourier transform on generalized functions closely follows the pattern of the definition of the derivative of a generalized function, with the result again a generalized function. To

accomplish this extension we need to revisit the definition of a generalized function. In section A.4.6 we gave the following definition:

Let $\mathcal{C}_c^\infty(\mathbb{R})$ denote infinitely differentiable functions defined on $\mathbb{R}$ which vanish outside of bounded sets. These are called *test functions*.

Definition 3.2.12. A generalized function on $\mathbb{R}$ is a linear function, l defined on the set of test functions such that there is a constant C and an integer k so that, for every $f \in \mathcal{C}_c^\infty(\mathbb{R})$ we have the estimate

$$|l(f)| \leq C \sup_{x\in\mathbb{R}} \left[(1+|x|)^k \sum_{j=0}^{k} |\partial_x^j f(x)| \right] \tag{3.60}$$

These are linear functions on $\mathcal{C}_c^\infty(\mathbb{R})$ which are, in a certain sense continuous. The constants C and k in (3.60) depend on l but do not depend on f. The expression on the right hand side defines a norm on $\mathcal{C}_c^\infty(\mathbb{R})$, for convenience we let

$$\|f\|_k = \sup_{x\in\mathbb{R}} \left[(1+|x|)^k \sum_{j=0}^{k} |\partial_x^j f(x)| \right].$$

The observation that we make is the following: if a generalized function satisfies the estimate

$$|l(f)| \leq C\|f\|_k$$

then it can be extended, by continuity, to any function f which is the limit of a sequence $< f_n > \subset \mathcal{C}_c^\infty(\mathbb{R})$ in the sense that

$$\lim_{n\to\infty} \|f - f_n\|_k = 0.$$

Clearly $f \in \mathcal{C}^k(\mathbb{R})$ and $\|f\|_k < \infty$. This motivates the following definition

Definition 3.2.13. A function $f \in \mathcal{C}^\infty(\mathbb{R})$ belongs to *Schwartz class* if $\|f\|_k < \infty$ for every $k \in \mathbb{N}$. The set of such functions is a vector space denoted by $\mathcal{S}(\mathbb{R})$.

From the definition it is clear that

$$\mathcal{C}_c^\infty(\mathbb{R}) \subset \mathcal{S}(\mathbb{R}). \tag{3.61}$$

Schwartz class does not have a norm with respect to which it is a complete normed linear space, instead each $\|\cdot\|_k$ defines a *semi-norm*. A sequence $< f_n > \subset \mathcal{S}(\mathbb{R})$ converges to $f \in \mathcal{S}(\mathbb{R})$ if and only if

$$\lim \|f - f_n\|_k = 0 \text{ for every } k \in \mathbb{N}.$$

With this notion of convergence, Schwartz class becomes a complete metric space, the distance is defined by

$$d_{\mathcal{S}}(f,g) = \sum_{j=0}^{\infty} 2^{-j} \frac{\|f-g\|_j}{1+\|f-g\|_j}.$$

Remark 3.2.7. Of course each $\|\cdot\|_k$ satisfies all the axioms for a norm. They are called "semi-norms" because each one alone, does not define the topology on $\mathcal{S}(\mathbb{R})$.

Let $\varphi(x) \in \mathcal{C}_c^\infty(\mathbb{R})$ be a non-negative function with the following properties

(1). $\varphi(x) = 1$ if $x \in [-1, 1]$,

(2). $\varphi(x) = 0$ if $|x| > 2$.

Define $\varphi_n(x) = \varphi(n^{-1}x)$, it is not difficult to prove the following proposition.

Proposition 3.2.11. *If $f \in \mathcal{S}(\mathbb{R})$ then $f_n = \varphi_n f \in \mathcal{C}_c^\infty(\mathbb{R}) \subset \mathcal{S}(\mathbb{R})$ converges to f in $\mathcal{S}(\mathbb{R})$. That is*

$$\lim_{n\to\infty} \|f_n - f\|_k = 0 \text{ for every } k. \tag{3.62}$$

The proof is left as an exercise.

From the discussion above it therefore follows that *every* generalized function can be extended to $\mathcal{S}(\mathbb{R})$. Because (3.62) holds for every k, if l is a generalized function and $f \in \mathcal{S}(\mathbb{R})$ then $l(f)$ is defined as

$$l(f) = \lim_{n\to\infty} l(\varphi_n f).$$

To show that this makes sense, it is only necessary to prove that if $< g_n > \subset \mathcal{C}_c^\infty(\mathbb{R})$ which converges to f in Schwartz class then

$$\lim_{n\to\infty} l(g_n - \varphi_n f) = 0. \tag{3.63}$$

This is an immediate consequence of the triangle inequality and the estimate that l satisfies: there is a C and k so that

$$\begin{aligned} |l(g_n - \varphi_n f)| &\leq C\|g_n - \varphi_n f\|_k \\ &\leq C[\|g_n - f\|_k + \|f - \varphi_n f\|_k. \end{aligned} \tag{3.64}$$

Since both terms on the right hand side of the second line tend to zero as $n \to \infty$, equation (3.63) is proved. In fact the generalized functions are exactly the set of continuous linear functions on $\mathcal{S}(\mathbb{R})$. For this reason the set of generalized functions is usually denoted by $\mathcal{S}'(\mathbb{R})$.

Why did we go to all this trouble? How will this help extend the Fourier transform to $\mathcal{S}'(\mathbb{R})$? The integration by parts formula was the "trick" used to extend the notion of derivative to generalized functions. The reason it works is that if $f \in \mathcal{S}(\mathbb{R})$ then $\partial_x f \in \mathcal{S}(\mathbb{R})$ as well. This implies that $l(\partial_x f)$ is defined, and a generalized function whenever l itself is. Schwartz class has a similar property.

Theorem 3.2.7. *The Fourier transform is an isomorphism of $\mathcal{S}(\mathbb{R})$ onto itself, that is if $f \in \mathcal{S}(\mathbb{R})$ then both $\mathcal{F}(f)$ and $\mathcal{F}^{-1}(f)$ also belong to $\mathcal{S}(\mathbb{R})$. Moreover, for each k there is an k' and constant C_k so that*

$$\|\mathcal{F}(f)\|_k \leq C_k \|f\|_{k'} \text{ for all } f \in \mathcal{S}(\mathbb{R}). \tag{3.65}$$

The proof of this theorem is an easy consequence of results in section 3.2.3. We give the proof for $\mathcal{F}$, the proof for $\mathcal{F}^{-1}$ is essentially identical.

Proof. Since $f \in \mathcal{S}(\mathbb{R})$ for any $j, k \in \mathbb{N} \cup \{0\}$ we have the estimates

$$|\partial_x^j f(x)| \leq \frac{\|f\|_k}{(1+|x|)^k}. \tag{3.66}$$

From Propositions 3.2.2 and 3.2.4 it follows that $\hat{f}$ is infinitely differentiable and that, for any k, j,

$$\sup_{\xi \in \mathbb{R}} |\xi|^k |\partial_\xi^j \hat{f}(\xi)| < \infty.$$

To prove this we use the formula

$$\xi^k \partial_\xi^j \hat{f}(\xi) = \int_{-\infty}^{\infty} (i\partial_x)^k \left[(-ix)^j f(x)\right] e^{-ix\xi} dx.$$

Because $f \in \mathcal{S}(\mathbb{R})$ the integrand is absolutely integrable and in fact if $m = \max\{j, k\}$ then

$$|\xi^k \partial_\xi^j \hat{f}(\xi)| \leq C_{k,l} \|f\|_{m+2}, \tag{3.67}$$

here $C_{k,l}$ depends only on k and l. This completes the proof. □

Instead of integration by parts, we now use this theorem and the identity

$$\int_{-\infty}^{\infty} f(x)\hat{g}(x)dx = \int_{-\infty}^{\infty} \hat{f}(x)g(x)dx, \tag{3.68}$$

to extend the Fourier transform to generalized functions. The identity follows by a simple change in the order of integrations which is easily justified if $f, g \in \mathcal{S}(\mathbb{R})$. It is now clear how we should define the Fourier transform of a generalized function.

Definition 3.2.14. If $l \in \mathcal{S}'(\mathbb{R})$ then the Fourier transform of l is the generalized function $\hat{l}$ *defined* by

$$\hat{l}(f) = l(\hat{f}) \text{ for all } f \in \mathcal{S}(\mathbb{R}). \tag{3.69}$$

Theorem 3.2.7 implies that $\hat{f} \in \mathcal{S}(\mathbb{R})$ so that the right hand side in (3.69) define a generalized function.

But why did we need to extend the definition of generalized functions from $\mathcal{C}_c^\infty(\mathbb{R})$ to $\mathcal{S}(\mathbb{R})$? The answer is simple: if $0 \neq f \in \mathcal{C}_c^\infty(\mathbb{R})$ then Proposition 3.2.10 implies that $\hat{f} \notin \mathcal{C}_c^\infty(\mathbb{R})$. This would prevent using (3.69) to define $\hat{l}$ because we would not know that $l(\hat{f})$ made sense! This appears to be a rather abstract definition and it is not at all clear that it can be used to compute the Fourier transform of a generalized function. In fact, it turns out to be very usable.

Example 3.2.15. If φ is an absolutely integrable function then

$$\widehat{l_\varphi} = l_{\hat{\varphi}}.$$

If $f \in \mathcal{S}(\mathbb{R})$ then the identity in (3.68) holds with $g = \hat{\varphi}$, as a simple interchange of integrations shows. Hence, for all $f \in \mathcal{S}(\mathbb{R})$

$$l_\varphi(\hat{f}) = \int_{-\infty}^{\infty} f(x)\hat{\varphi}(x)dx = l_{\hat{\varphi}}(f).$$

This shows that the Fourier transform for generalized functions is indeed an extension of the ordinary transform: if a generalized function l is *represented* by an integrable function in the sense that $l = l_\varphi$ then the definition of the Fourier transform of l is consistent with the earlier definition of the Fourier transform of φ.

Example 3.2.16. If $f \in \mathcal{S}(\mathbb{R})$ then

$$\hat{f}(0) = \int_{-\infty}^{\infty} f(x)dx.$$

This shows that $\hat{\delta} = l_1$ which is *represented* by an ordinary function equal to the constant 1.

Example 3.2.17. On the other hand the Fourier inversion formula implies that

$$\int_{-\infty}^{\infty} \hat{f}(\xi)d\xi = 2\pi f(0)$$

and therefore $\widehat{l_1} = 2\pi\delta$. This is an example of an ordinary function that does not have a Fourier transform, in the usual sense, and whose Fourier transform, as a generalized function is **not** an ordinary function.

Recall that a sequence $< l_n > \subset \mathcal{S}'(\mathbb{R})$ converges to l in $\mathcal{S}'(\mathbb{R})$ provided that

$$l(g) = \lim_{n\to\infty} l_n(g) \text{ for all } g \in \mathcal{S}(\mathbb{R}). \tag{3.70}$$

This is very useful for computing Fourier transforms because the Fourier transform is continuous with respect to the limit in (3.70). It follows from the definition that:

$$\widehat{l_n}(g) = l_n(\hat{g}) \tag{3.71}$$

and therefore

$$\lim_{n\to\infty} \widehat{l_n}(g) = \lim_{n\to\infty} l_n(\hat{g}) = l(\hat{g}) = \hat{l}(g). \tag{3.72}$$

Example 3.2.18. The generalized function $l_{\chi_{[0,\infty)}}$ can be defined as a limit by

$$l_{\chi_{[0,\infty)}}(f) = \lim_{\epsilon\downarrow 0} \int_0^{\infty} e^{-\epsilon x} f(x)dx.$$

The Fourier transform of $l_{e^{-\epsilon x}\chi_{[0,\infty)}}$ is easily computed using example 3.2.15, it is

$$\mathcal{F}(l_{e^{-\epsilon x}\chi_{[0,\infty)}})(f) = \int_{-\infty}^{\infty} \frac{f(x)dx}{ix+\epsilon}.$$

This shows that

$$\mathcal{F}(l_{\chi_{[0,\infty)}})(f) = \lim_{\epsilon \downarrow 0} \int_{-\infty}^{\infty} \frac{f(x)dx}{ix+\epsilon}. \tag{3.73}$$

In fact, it proves that the limit on the right hand side exists!

We close this discussion by verifying that the Fourier transform on generalized functions has many of the properties of the ordinary Fourier transform. Recall that if l is a generalized function and f is an infinitely differentiable function which satisfies estimates

$$|\partial_x^j f(x)| \leq C_j(1+|x|)^k,$$

for a *fixed* k then the product $f \cdot l$ is defined by

$$f \cdot l(g) = l(fg).$$

If $l \in \mathcal{S}'(\mathbb{R})$ then so are all of its derivatives. Using the definition it is not difficult to find formulæ for $\mathcal{F}(l^{[j]})$:

$$\mathcal{F}(l^{[j]})(f) = l^{[j]}(\hat{f}) = (-1)^j l(\partial_x^j \hat{f}) = l(\widehat{(ix)^j f}). \tag{3.74}$$

This shows that

$$\mathcal{F}(l^{[j]}) = (ix)^j \cdot \hat{l}. \tag{3.75}$$

A similar calculation shows that

$$\mathcal{F}((-ix)^j \cdot l) = \hat{l}^{[j]}. \tag{3.76}$$

Exercise 3.2.36. Prove (3.61).

Exercise 3.2.37. Prove that $d_{\mathcal{S}}$ defines a metric. Show that a sequence $< f_n >$ converges in $\mathcal{S}(\mathbb{R})$ to f if and only if

$$\lim_{n\to\infty} d_{\mathcal{S}}(f_n, f) = 0.$$

Exercise 3.2.38. Prove Proposition 3.2.11.

Exercise 3.2.39. Prove (3.68). What is the "minimal" hypothesis on f and g so this formula makes sense, as absolutely convergent integrals.

Exercise 3.2.40. Give a detailed proof of (3.67).

Exercise 3.2.41. Prove, by direct computation that the limit on the right hand side of (3.73) exists for any $f \in \mathcal{S}(\mathbb{R})$.

Exercise 3.2.42. If $l_{1/x}$ is the Cauchy principal value integral

$$l_{1/x}(f) = \text{P. V.} \int_{-\infty}^{\infty} \frac{f(x)dx}{x}$$

then show that $\mathcal{F}(l_{1/x}) = l_{\operatorname{sign} x}$.

Exercise 3.2.43. Prove (3.76).

Exercise 3.2.44. The inverse Fourier transform of a generalized function is defined by

$$[\mathcal{F}^{-1}(l)](g) = l(\mathcal{F}^{-1}(g)).$$

Show that $\mathcal{F}^{-1}(\hat{l}) = l = \widehat{\mathcal{F}^{-1}(l)}$.

3.2.14 The Paley-Wiener theorem*

In imaging applications one usually works with functions of bounded support. The question naturally arises whether it is possible to recognize such a function from its Fourier transform. There are a variety of theorems which relate the support of a function to properties of its Fourier transform. They go collectively by the name of Paley-Wiener theorems.

Theorem 3.2.8 (Paley-Wiener Theorem I). *A square integrable function f satisfies $f(x) = 0$ for $|x| > L$ if and only if its Fourier transform $\hat{f}$ extends to be an analytic function in the whole complex plane which satisfies*

$$\begin{gathered} \int_{-\infty}^{\infty} |\hat{f}(\xi + i\tau)|^2 d\xi \leq Me^{2L|\tau|} \text{ for all } \tau \text{ and} \\ |\hat{f}(\xi + i\tau)| \leq \frac{Me^{L|\tau|}}{\sqrt{|\tau|}} \end{gathered} \tag{3.77}$$

Proof. The proof of the forward implication is elementary. The Fourier transform of f is given by an integral over a finite interval,

$$\hat{f}(\xi) = \int_{-L}^{L} f(x)e^{-ix\xi}d\xi. \tag{3.78}$$

The expression clearly makes sense if ξ is replaced by $\xi + i\tau$, differentiating under the integral shows that $\hat{f}(\xi + i\tau)$ is a analytic function. The first estimate follows from the Parseval formula as $\hat{f}(\xi+i\tau)$ is the Fourier transform of the L^2-function $f(x)e^{-\tau x}$. Using the Cauchy Schwartz inequality we obtain

$$\begin{aligned} |\hat{f}(\xi + i\tau)| &= \left| \int_{-L}^{L} f(x)e^{-ix\xi - x\tau}dx \right| \\ &\leq \frac{e^{L|\tau|}}{\sqrt{|\tau|}} \sqrt{\int_{-L}^{L} |f(x)|^2 dx}; \end{aligned} \tag{3.79}$$

from which the estimate is immediate.

The proof of the converse statement is a little more involved; it uses the Fourier inversion formula and a change of contour. We present the outlines of this argument, the complete justification for the change of contour can be found in [40]. Let $x > L > 0$, the Fourier inversion formula states that

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) e^{ix\xi} d\xi.$$

Since $\hat{f}(z)e^{ixz}$ is an analytic function, satisfying appropriate estimates, we can shift the integration to the line $\xi + i\tau$ for any $\tau > 0$,

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi + i\tau) e^{-x\tau} e^{ix\xi} d\xi.$$

In light of the first estimate in (3.77), we obtain the bound

$$|f(x)| \leq M e^{(L-x)\tau}.$$

Letting τ tend to infinity shows that $f(x) = 0$ for $x > L$. A similar argument using $\tau < 0$ shows that $f(x) = 0$ if $x < -L$. □

For latter applications we state a variant of this result whose proof can be found in [40]

Theorem 3.2.9 (Paley-Wiener II). *A function $f \in L^2(\mathbb{R})$ has an analytic extension $F(x + iy)$ to the upper half plane ($y > 0$) satisfying*

$$\begin{aligned} &\int_{-\infty}^{\infty} |F(x + iy)|^2 dx \leq M, \\ &\lim_{y \downarrow 0} \int_{-\infty}^{\infty} |F(x + iy) - f(x)|^2 = 0 \end{aligned} \tag{3.80}$$

if and only if $\hat{f}(\xi) = 0$ for $\xi < 0$.

3.3 The Fourier transform for functions of several variables.

The Fourier transform can also be defined for functions of several variables. A function defined on $\mathbb{R}^n$ is called a "function of n-variables." This section presents the theory of the Fourier transform for functions of n-variables. In most ways it is quite similar to the one dimensional theory. Some notable differences are discussed in section 3.3.8. As before we begin with the technically simpler case of absolutely integrable functions.

3.3.1 L^1-case

The n-dimensional Euclidean space is the collection of ordered n-tuples of real numbers

$$\mathbb{R}^n = \{(x_1, \ldots, x_n) \,:\, x_j \in \mathbb{R} \text{ for } j = 1, \ldots, n\}.$$

We o use lower case, bold Roman letters $\mathbf{x}, \mathbf{y}$ etc. to denote points in $\mathbb{R}^n$, that is

$$\mathbf{x} = (x_1, \ldots, x_n) \text{ or } \mathbf{y} = (y_1, \ldots, y_n).$$

In this case x_j is called the j^{th}-coordinate of $\mathbf{x}$. The Fourier transform of a function of n-variables is also a function of n-variables. It is customary to use the lower case, bold Greek letters, $\boldsymbol{\xi}$ or $\boldsymbol{\eta}$ as coordinates on the Fourier transform space with

$$\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n) \text{ or } \boldsymbol{\eta} = (\eta_1, \ldots, \eta_n).$$

Definition 3.3.1. If $f(\mathbf{x})$ is an integrable function of n-variables then the Fourier transform, $\hat{f}$ of f is defined by

$$\hat{f}(\boldsymbol{\xi}) = \int_{\mathbb{R}^n} f(\mathbf{x}) e^{-i\boldsymbol{\xi}\cdot\mathbf{x}} d\mathbf{x} \quad \text{for} \quad \boldsymbol{\xi} \in \mathbb{R}^n. \tag{3.81}$$

Note that $\boldsymbol{\xi} \cdot \mathbf{x}$ is the inner product, if $\mathbf{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ then

$$\boldsymbol{\xi} \cdot \mathbf{x} = \mathbf{x} \cdot \boldsymbol{\xi} = \sum_{j=1}^{n} \xi_j x_j.$$

This inner product is sometimes denoted by $\langle \mathbf{x}, \boldsymbol{\xi} \rangle$; the volume form on $\mathbb{R}^n$ is denoted $d\mathbf{x} = dx_1 \ldots dx_n$. Since f is absolutely integrable over $\mathbb{R}^n$ the integral can be computed as an iterated integral

$$\int_{\mathbb{R}^n} f(\mathbf{x}) e^{-i\boldsymbol{\xi}\cdot\mathbf{x}} d\mathbf{x} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) e^{-ix_1\xi_1} dx_1 \cdots e^{-ix_n\xi_n} dx_n; \tag{3.82}$$

changing the order of the one dimensional integrals does not change the result. When thought of as a linear transformation, it is customary to use $\mathcal{F}(f)$ to denote the Fourier transform of f.

It is useful to have a clear geometric picture of the inner product to have a better understanding of the functions $e^{i\langle \boldsymbol{\xi}, \mathbf{x} \rangle}$. To that end we write $\boldsymbol{\xi}$ in polar form as $\boldsymbol{\xi} = r\omega$. Here $r = \|\boldsymbol{\xi}\|$ is the length of $\boldsymbol{\xi}$ and ω its direction. Write $\mathbf{x} = \mathbf{x}' + x_1\omega$ where $\mathbf{x}'$ is orthogonal to ω, (i.e. $\langle \mathbf{x}', \omega \rangle = 0$). As $\langle \mathbf{x}, \omega \rangle = x_1$ the function $\langle \mathbf{x}, \omega \rangle$ depends only x_1. Thus

$$e^{i\langle \mathbf{x}, \boldsymbol{\xi} \rangle} = e^{irx_1}$$

is a function which oscillates in the ω-direction with wave length $\frac{2\pi}{r}$. To illustrate this we give a density plot in the plane of the real and imaginary parts of

$$e^{i\langle \mathbf{x}, \boldsymbol{\xi} \rangle} = \cos\langle \mathbf{x}, \boldsymbol{\xi} \rangle + i \sin\langle \mathbf{x}, \boldsymbol{\xi} \rangle$$

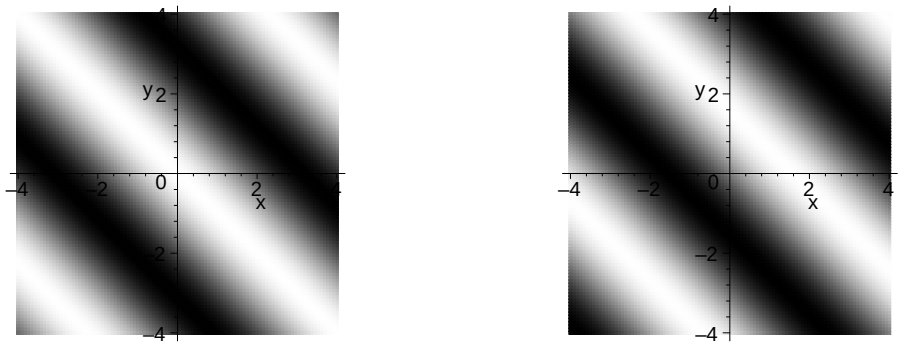
for several choices of $\boldsymbol{\xi}$. In these figures white corresponds to $+1$ and black corresponds to -1. The Fourier transform at $\boldsymbol{\xi} = r\omega$ can be re-expressed as

$$\hat{f}(r\omega) = \int_{-\infty}^{\infty} \int_{L} f(\mathbf{x}' + x_1\omega) e^{-irx_1} d\mathbf{x}' dx_1. \tag{3.83}$$

Here L is the $(n-1)$-dimensional subspace orthogonal to ω :

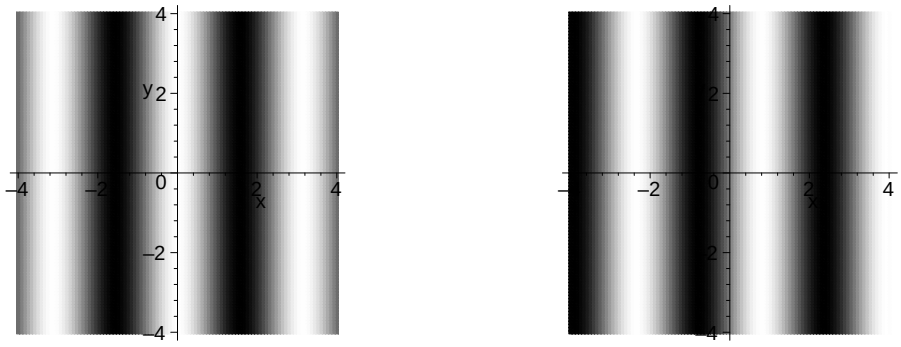
$$L = \{x' \in \mathbb{R}^n \ : \ \langle \mathbf{x}', \omega \rangle = 0\}$$

and $d\mathbf{x}'$ is the $(n-1)$-dimensional Euclidean measure on L.



(a) (b)

Figure 3.8: Real and imaginary parts of $\exp(i\langle (x,y),(1,1)\rangle)$



(a) (b)

Figure 3.9: Real and imaginary parts of $\exp(i\langle (x,y),(2,0)\rangle)$

The Fourier transform is invertible; under appropriate hypotheses there is an explicit formula for the inverse.

Theorem 3.3.1 (Fourier Inversion Formula). *Suppose that f is an absolutely integrable function defined on $\mathbb{R}^n$. If $\int |\hat{f}(\boldsymbol{\xi})| d\boldsymbol{\xi} < \infty$ as well then*

$$f(\mathbf{x}) = \int_{\mathbb{R}^n} \hat{f}(\boldsymbol{\xi}) e^{i\mathbf{x}\cdot\boldsymbol{\xi}} \frac{d\boldsymbol{\xi}}{[2\pi]^n}. \tag{3.84}$$

Here the volume form on Fourier space is $d\boldsymbol{\xi} = d\xi_1 \dots d\xi_n$.

Proof. The proof is formally identical to the proof of the one dimensional result. As before we begin by assuming that f is continuous. The basic fact used is that the Fourier transform of a Gaussian can be computed explicitly:

$$\mathcal{F}(e^{-\epsilon\|x\|^2}) = \left[\frac{\pi}{\epsilon}\right]^{\frac{n}{2}}. \tag{3.85}$$

Because $\hat{f}$ is absolutely integrable

$$\begin{aligned} \int_{\mathbb{R}^n} \hat{f}(\boldsymbol{\xi}) e^{i\mathbf{x}\cdot\boldsymbol{\xi}} \frac{d\boldsymbol{\xi}}{[2\pi]^n} &= \lim_{\epsilon\downarrow 0} \int_{\mathbb{R}^n} \hat{f}(\boldsymbol{\xi}) e^{i\mathbf{x}\cdot\boldsymbol{\xi}} e^{-\epsilon\|\boldsymbol{\xi}\|^2} \frac{d\boldsymbol{\xi}}{[2\pi]^n} \\ &= \lim_{\epsilon\downarrow 0} \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} f(\mathbf{y}) e^{-i\mathbf{y}\cdot\boldsymbol{\xi}} d\mathbf{y} e^{i\mathbf{x}\cdot\boldsymbol{\xi}} e^{-\epsilon\|\boldsymbol{\xi}\|^2} \frac{d\boldsymbol{\xi}}{[2\pi]^n}. \end{aligned} \tag{3.86}$$

The order of the integrations in the last line can be interchanged; using (3.85) gives,

$$\begin{aligned} \int_{\mathbb{R}^n} \hat{f}(\boldsymbol{\xi}) e^{i\mathbf{x}\cdot\boldsymbol{\xi}} \frac{d\boldsymbol{\xi}}{[2\pi]^n} &= \lim_{\epsilon\downarrow 0} \int_{\mathbb{R}^n} f(\mathbf{y}) \left[\frac{\pi}{\epsilon}\right]^{\frac{n}{2}} e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{4\epsilon}} \frac{d\mathbf{y}}{[2\pi]^n} \\ &= \lim_{\epsilon\downarrow 0} \int_{\mathbb{R}^n} f(\mathbf{x} - 2\sqrt{\epsilon}\mathbf{t}) e^{-\|\mathbf{t}\|^2} \frac{d\mathbf{t}}{[2\pi]^{\frac{n}{2}}}. \end{aligned} \tag{3.87}$$

In the last line we use the change of variables $\mathbf{y} = \mathbf{x} - 2\sqrt{\epsilon}\mathbf{t}$. As f is continuous and absolutely integrable this converges to

$$f(\mathbf{x}) \int_{\mathbb{R}^n} e^{-\|\mathbf{t}\|^2} \frac{d\mathbf{t}}{[2\pi]^{\frac{n}{2}}}.$$

As

$$\int_{\mathbb{R}^n} e^{-\|\mathbf{t}\|^2} d\mathbf{t} = [2\pi]^{\frac{n}{2}},$$

this completes the proof of the theorem for continuous functions. As in the one-dimensional case, an approximation argument is used to remove the additional hypothesis. The details are left to the reader. □

Exercise 3.3.1. Prove formula (3.83).

Exercise 3.3.2. If $g_1(x), \dots, g_n(x)$ belong to $L^1(\mathbb{R})$ show that

$$f(x_1, \dots, x_n) = g_1(x_1) \cdots g_n(x_n) \in L^1(\mathbb{R}^n).$$

Show that

$$\hat{f}(\xi_1, \dots, \xi_n) = \hat{g}_1(\xi_1) \cdots \hat{g}_n(\xi_n).$$

Use this to compute the Fourier transform of $e^{-\|\mathbf{x}\|^2}$.

3.3.2 Regularity and decay

Integrable functions on $\mathbb{R}^n$ are described qualitatively in terms of two general properties:

Decay at infinity:
: How fast does $f(\mathbf{x})$ go to zero as $|\mathbf{x}| \to \infty$.

Regularity:
: How regular is f? If f is differentiable then it is also important to know that these derivatives are integrable.

The Riemann-Lebesgue lemma holds in n-variables.

Proposition 3.3.1 (Riemann-Lebesgue Lemma). *Let f be an absolutely integrable function on $\mathbb{R}^n$ then $\hat{f}$ is a continuous function and $\lim_{|\boldsymbol{\xi}|\to\infty} \hat{f}(\boldsymbol{\xi}) = 0$.*

The proof is very similar to the one dimensional case and is left to the reader.

The Fourier Transform is sensitive to decay and regularity. In order to understand how decay at infinity for f is reflected in properties of $\hat{f}$ we first suppose that f vanishes outside the ball of radius R. It can be shown without difficulty that $\hat{f}(\boldsymbol{\xi})$ is a differentiable function, and its derivatives are given by

$$\partial_{\boldsymbol{\xi}_j} \hat{f}(\boldsymbol{\xi}) = \int\limits_{B_R} \partial_{\xi_j}[f(\mathbf{x})e^{-i\boldsymbol{\xi}\cdot\mathbf{x}}]d\mathbf{x} = \int\limits_{B_R} f(\mathbf{x})(-ix_j)e^{-i\boldsymbol{\xi}\cdot\mathbf{x}}d\mathbf{x} = \widehat{(-ix_j f)}(\boldsymbol{\xi}). \tag{3.88}$$

Formulæ in several variables can rapidly become cumbersome and unreadable. Fortunately there is a compact notation which gives formulæ in n-variables that have the simplicity and readability of one variable formulæ . This is called *multi-index* notation.

Definition 3.3.2. A **multi-index** is an ordered n-tuple of non-negative integers usually denoted by a lower case Greek letter. For $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$, a multi-index, set

$$\boldsymbol{\alpha}! = \alpha_1! \cdots \alpha_n! \qquad \text{and} \qquad |\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_n.$$

The function $|\boldsymbol{\alpha}|$ is called the length of $\boldsymbol{\alpha}$. The following conventions are also useful:

$$\mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n} \text{ and } \partial_{\mathbf{x}}^{\boldsymbol{\alpha}} = \partial_{x_1}^{\alpha_1} \partial_{x_2}^{\alpha_2} \cdots \partial_{x_n}^{\alpha_n}.$$

Example 3.3.1. If f is a k-times differentiable function on $\mathbb{R}^n$ then there is a multi-dimensional analogue of Taylor's formula:

$$f(\mathbf{x}) = \sum_{\{\boldsymbol{\alpha}\,:\,|\boldsymbol{\alpha}|\le k\}} \frac{\partial_{\mathbf{x}}^{\boldsymbol{\alpha}} f(0)\mathbf{x}^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} + R_k(\mathbf{x}). \tag{3.89}$$

As in the 1-d case, R_k is the remainder term. It satisfies

$$\lim_{\|\mathbf{x}\|\to 0} \frac{|R_k(\mathbf{x})|}{\|\mathbf{x}\|^k} = 0.$$

Example 3.3.2. The binomial formula also has a higher dimensional analogue

$$(x_1 + \cdots + x_n)^k = k! \sum_{\{\boldsymbol{\alpha} \,:\, |\boldsymbol{\alpha}|=k\}} \frac{\mathbf{x}^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!}.$$

The smoothness of f itself is reflected in the decay properties of its Fourier transform. First, consider the case $|\boldsymbol{\alpha}| = 1$ and suppose $\partial_{x_j} f$ is also integrable, i.e.

$$\int_{\mathbb{R}^n} |\partial_{x_j} f| d\mathbf{x} < \infty.$$

The Fourier transform is bounded by the integral of a derivative of f as follows:

$$|\hat{f}(\boldsymbol{\xi})| \leq \frac{1}{|\xi_j|} \int_{\mathbb{R}^n} |\partial_{x_j} f(\mathbf{x})| d\mathbf{x}.$$

To prove this we integrate by parts in the j$^{\text{th}}$ variable

$$\begin{aligned} \hat{f}(\boldsymbol{\xi}) &= \int_{\mathbb{R}^n} f(\mathbf{x}) e^{-i\boldsymbol{\xi}\cdot\mathbf{x}} d\mathbf{x} \\ &= \frac{1}{-i\xi_j} \int_{\mathbb{R}^{n-1}} \left[f e^{-i\boldsymbol{\xi}\cdot\mathbf{x}} \Big|_{x_j=-\infty}^{x_j=\infty} \right] d\mathbf{x}' + \frac{1}{i\xi_j} \int_{\mathbb{R}^n} \partial_{x_j} f e^{-i\boldsymbol{\xi}\cdot\mathbf{x}} d\mathbf{x} \qquad (3.90) \\ &= \frac{1}{i\xi_j} \int \partial_{x_j} f e^{-i\boldsymbol{\xi}\cdot\mathbf{x}} d\mathbf{x}. \end{aligned}$$

The hypothesis that $\int |f| d\mathbf{x} < \infty$ implies that the boundary term is zero. Equation (3.90) can be rewritten

$$\widehat{\partial_{x_j} f}(\boldsymbol{\xi}) = i\xi_j \hat{f}(\boldsymbol{\xi}).$$

The integration by parts argument can be iterated to obtain formulæ for the Fourier transform of $\partial_{\mathbf{x}}^{\boldsymbol{\alpha}} f$ for any multi-index $\boldsymbol{\alpha}$.

Theorem 3.3.2. *If $\int |\partial_{\mathbf{x}}^{\boldsymbol{\alpha}} f(\mathbf{x})| d\mathbf{x} < \infty$ for all $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}| \leq k$, then*

$$|\hat{f}(\boldsymbol{\xi})| \leq \frac{C}{(1 + \|\boldsymbol{\xi}\|)^k},$$

and

$$\widehat{\partial_{\mathbf{x}}^{\boldsymbol{\alpha}} f}(\boldsymbol{\xi}) = (i\boldsymbol{\xi})^{\boldsymbol{\alpha}} \hat{f}(\boldsymbol{\xi}).$$

The theorem relates the rate of decay of the Fourier transform to the smoothness of f. As in the one dimensional case, this theorem has a partial converse.

Proposition 3.3.2. *Suppose that f is an integrable function on $\mathbb{R}^n$ such that, for an $\epsilon > 0$,*

$$|\hat{f}(\boldsymbol{\xi})| \leq \frac{C}{(1 + \|\boldsymbol{\xi}\|)^{k+n+\epsilon}} < \infty.$$

Then f has k-continuous derivatives which tend to zero at infinity.

Proof. The proof is a consequence of the Fourier inversion formula. The decay hypothesis implies that

$$f(\mathbf{x}) = \frac{1}{[2\pi]^n} \int\limits_{\mathbb{R}^n} \hat{f}(\boldsymbol{\xi}) e^{i\mathbf{x}\cdot\boldsymbol{\xi}} d\boldsymbol{\xi}.$$

The estimate satisfied by $\hat{f}$ implies that this expression can be differentiated k-times and that $\partial_{\mathbf{x}}^{\boldsymbol{\alpha}} f$ is the Fourier transform of $(-i\boldsymbol{\xi})^{\alpha}\hat{f}(-\boldsymbol{\xi})$, an L^1-function. The last statement then follows from the Riemann-Lebesgue lemma. □

It is apparent that the discrepancy between this estimate and that in the theorem grows as the dimension increases.

The decay of f is reflected in the regularity of its Fourier transform. Iterating (3.88) gives

$$\partial_{\boldsymbol{\xi}}^{\boldsymbol{\alpha}} \hat{f}(\boldsymbol{\xi}) = (-i)^{|\boldsymbol{\alpha}|} \int\limits_{\mathbb{R}^n} \mathbf{x}^{\boldsymbol{\alpha}} f(\mathbf{x}) e^{-i\boldsymbol{\xi}\cdot\mathbf{x}} d\mathbf{x} = (-i)^{|\boldsymbol{\alpha}|} \widehat{(\mathbf{x}^{\boldsymbol{\alpha}} f)}(\boldsymbol{\xi}). \tag{3.91}$$

To summarize these computations we have:

Proposition 3.3.3. *If $\int |f(\mathbf{x})|(1+\|\mathbf{x}\|)^k d\mathbf{x} < \infty$ for a positive integer k then its Fourier transform, $\hat{f}(\boldsymbol{\xi})$ has k continuous derivatives and its partial derivatives are given by*

$$\partial_{\boldsymbol{\xi}}^{\boldsymbol{\alpha}} \hat{f}(\boldsymbol{\xi}) = (-i)^{|\boldsymbol{\alpha}|} \widehat{\mathbf{x}^{\boldsymbol{\alpha}} f}(\xi).$$

They satisfy the estimates

$$|\partial_{\boldsymbol{\xi}}^{\boldsymbol{\alpha}} \hat{f}(\boldsymbol{\xi})| \leq \int\limits_{\mathbb{R}^n} \|\mathbf{x}\|^{|\alpha|} |f(\mathbf{x})| d\mathbf{x}$$

and tend to zero as $\|\boldsymbol{\xi}\|$ tends to infinity.

Exercise 3.3.3. Suppose that f is an integrable function which vanishes outside the ball of radius R. Show that $\hat{f}(\boldsymbol{\xi})$ is a differentiable function and justify the interchange of the derivative and the integral in (3.88).

Exercise 3.3.4. Suppose that f is an integrable function which vanishes outside the ball of radius R. Show that $\hat{f}(\boldsymbol{\xi})$ is an infinitely differentiable function.

Exercise 3.3.5. Prove the n-variable binomial formula.

Exercise 3.3.6. Find a function f of n-variables so that

$$|\hat{f}(\boldsymbol{\xi})| \leq \frac{C}{(1+\|\boldsymbol{\xi}\|)^n}$$

but f is **not** continuous.

3.3.3 L^2-theory

See: A.4.6, A.4.5.

As in the 1-dimensional case, the n-dimensional Fourier transform extends to square integrable functions. A function f of n-variables is square integrable provided

$$\int_{\mathbb{R}^n} |f(\mathbf{x})|^2 d\mathbf{x} < \infty.$$

$L^2(\mathbb{R}^n)$ is again a complete, normed linear space with norm defined by the inner product,

$$\langle f, g\rangle_{L^2} = \int_{\mathbb{R}^n} f(\mathbf{x})\overline{g(\mathbf{x})} d\mathbf{x}.$$

The inner product satisfies the Cauchy-Schwarz inequality

$$|\langle f, g\rangle_{L^2}| \leq \|f\|_{L^2}\|g\|_{L^2}. \tag{3.92}$$

The proof is exactly the same as the 1-dimensional case.

If we suppose that f and g are bounded functions vanishing outside a ball of radius R then the order of integrations in the first line can be interchanged to obtain the identity

$$\begin{aligned}\int_{\mathbb{R}^n} f(\mathbf{x})\hat{g}(\mathbf{x}) d\mathbf{x} &= \int_{\mathbb{R}^n} f(\mathbf{x}) \int_{\mathbb{R}^n} e^{-i\mathbf{x}\cdot\mathbf{y}} g(\mathbf{y}) d\mathbf{y} d\mathbf{x} \\ &\int_{\mathbb{R}^n} \hat{f}(\mathbf{y}) g(\mathbf{y}) d\mathbf{y}.\end{aligned} \tag{3.93}$$

Combining (3.93) with the Fourier inversion formula we obtain Parseval's formula.

Theorem 3.3.3 (Parseval formula). *If f is absolutely integrable and $\int_{\mathbb{R}^n} |f(\mathbf{x})|^2 d\mathbf{x} < \infty$, then*

$$\int_{\mathbb{R}^n} |f(\mathbf{x})|^2 dx = \int_{\mathbb{R}^n} |\hat{f}(\boldsymbol{\xi})|^2 \frac{d\boldsymbol{\xi}}{[2\pi]^n}. \tag{3.94}$$

The details of the proof are left to the reader.

As in the single variable case, the integral defining the Fourier transform may not converge absolutely for an L^2-function. A similar, somewhat roundabout definition is required, as before set

$$\hat{f}_R(\boldsymbol{\xi}) = \int_{\|\mathbf{x}\|<R} f(\mathbf{x}) e^{-i\boldsymbol{\xi}\cdot\mathbf{x}} d\mathbf{x}.$$

From the Parseval formula it follows that that $\{\hat{f}_R\}$ is a Cauchy sequence in $L^2(\mathbb{R}^n)$. Because $L^2(\mathbb{R}^n)$ is complete, the Fourier transform of f can be defined as the L^2-limit of this sequence,

$$\hat{f} = \underset{R\to\infty}{LIM} \hat{f}_R.$$

The Parseval formula evidently extends to all functions $f \in L^2(\mathbb{R}^2)$. This shows that the Fourier transform is a continuous mapping of $L^2(\mathbb{R}^n)$ to itself: if $< f_n >$ is a sequence with $\underset{n\to\infty}{LIM} f_n = f$ then

$$\underset{n\to\infty}{LIM} \hat{f}_n = \hat{f}.$$

The L^2-inversion formula is also a consequence of the Parseval formula.

Proposition 3.3.4 (L^2-inversion formula). *Let $f \in L^2(\mathbb{R}^n)$ and define*

$$F_R(\mathbf{x}) = \frac{1}{[2\pi]^n} \int\limits_{\|\boldsymbol{\xi}\|<R} \hat{f}(\boldsymbol{\xi}) e^{i\mathbf{x}\cdot\boldsymbol{\xi}} d\boldsymbol{\xi},$$

then $f = \underset{R\to\infty}{LIM} F_R$.

Proof. We need to show that $\lim_{R\to\infty} \|F_R - f\|_{L^2} = 0$. Because the norm is defined by an inner product we have

$$\|F_R - f\|_{L^2}^2 = \|F_R\|_{L^2}^2 - 2\,\mathrm{Re}\langle F_R, f\rangle_{L^2} + \|f\|_{L^2}.$$

The Parseval formula implies that

$$\|F_R\|_{L^2}^2 = \frac{1}{[2\pi]^n} \int\limits_{\|\boldsymbol{\xi}\|<R} |\hat{f}(\boldsymbol{\xi})|^2 d\boldsymbol{\xi} \text{ and } \|f\|_{L^2}^2 = \frac{1}{[2\pi]^n} \int\limits_{\mathbb{R}^n} |\hat{f}(\boldsymbol{\xi})|^2 d\boldsymbol{\xi}.$$

The proof is completed by using the following lemma.

Lemma 3.3.1. *Let $g \in L^2(\mathbb{R}^n)$ then*

$$\langle F_R, g\rangle = \frac{1}{[2\pi]^n} \int\limits_{\|\boldsymbol{\xi}\|<R} \hat{f}(\boldsymbol{\xi})\overline{\hat{g}(\boldsymbol{\xi})} d\boldsymbol{\xi}. \tag{3.95}$$

The proof of the lemma is a consequence of the Parseval formula, it is left as an exercise for the reader. Using (3.95) gives

$$\|F_R - f\|_{L^2}^2 = \frac{1}{[2\pi]^n} \int\limits_{\|\boldsymbol{\xi}\|\geq R} |\hat{f}(\boldsymbol{\xi})|^2 d\boldsymbol{\xi}.$$

This implies that $\underset{R\to\infty}{LIM} F_R = f$. □

Remark 3.3.1. The extension of the Fourier transform to functions in $L^2(\mathbb{R}^n)$ has many nice properties. In particular the range of the Fourier transform on L^2 is exactly $L^2(\mathbb{R}^n)$. However the formula for the Fourier transform as an integral is *purely symbolic*. The Fourier transform itself is only defined as a LIM; for a given $\boldsymbol{\xi}$ the pointwise limit

$$\lim_{R\to\infty} \int\limits_{\|\mathbf{x}\|<R} f(\mathbf{x})^{-i\mathbf{x}\cdot\boldsymbol{\xi}} d\mathbf{x}$$

may or may not exist.

Exercise 3.3.7. Show that (3.94) implies that

$$\int\limits_{\mathbb{R}^n} f(\mathbf{x})\overline{g(\mathbf{x})} d\mathbf{x} = \int\limits_{\mathbb{R}^n} \hat{f}(\boldsymbol{\xi})\overline{\hat{g}(\boldsymbol{\xi})} \frac{d\boldsymbol{\xi}}{[2\pi]^n}.$$

Exercise 3.3.8. Prove Lemma 3.3.1.

3.3.4 Basic properties of the Fourier Transform on $\mathbb{R}^n$

See: A.3.6.

The Fourier transform of functions of n-variables has the same formal properties as the Fourier transform on functions of 1 variable. Let f, g be integrable or square integrable functions on $\mathbb{R}^n$.

1. Linearity:
   The Fourier transform is a linear operation, if $\alpha \in \mathbb{C}$ then
   $$\widehat{f+g} = \hat{f} + \hat{g}, \quad \widehat{\alpha f} = \alpha \hat{f}.$$

2. Scaling:
   The Fourier transform of $f(ax)$, a function dilated by $a \in \mathbb{R}$ is given by
   $$\begin{aligned} \int_{\mathbb{R}^n} f(ax) e^{-i\boldsymbol{\xi}\cdot\mathbf{x}} d\mathbf{x} &= \int_{\mathbb{R}^n} f(\mathbf{y}) e^{-i\frac{\boldsymbol{\xi}\cdot\mathbf{y}}{a}} \frac{d\mathbf{y}}{a^n} \\ &= \frac{1}{a^n} \hat{f}(\frac{\boldsymbol{\xi}}{a}). \end{aligned} \tag{3.96}$$

3. Translation:
   Let $f_{\mathbf{t}}$ be the function f shifted by the vector $\mathbf{t}$, $f_{\mathbf{t}}(\mathbf{x}) = f(\mathbf{x} - \mathbf{t})$. The Fourier transform of $f_{\mathbf{t}}(\boldsymbol{\xi})$ is given by
   $$\begin{aligned} \widehat{f_{\mathbf{t}}}(\boldsymbol{\xi}) &= \int_{\mathbb{R}^n} f(\mathbf{x} - \mathbf{t}) e^{-i\boldsymbol{\xi}\cdot\mathbf{x}} d\mathbf{x} \\ &= \int_{\mathbb{R}^n} f(\mathbf{y}) e^{-i\boldsymbol{\xi}\cdot(\mathbf{y}+\mathbf{t})} d\mathbf{y} \\ &= e^{-i\boldsymbol{\xi}\cdot\mathbf{t}} \hat{f}(\boldsymbol{\xi}). \end{aligned} \tag{3.97}$$

4. Reality:
   If $f(\mathbf{x})$ is real valued then $\hat{f}(\boldsymbol{\xi}) = \overline{\hat{f}(-\boldsymbol{\xi})}$.

5. Evenness:
   If $f(\mathbf{x}) = f(-\mathbf{x})$ then the $\hat{f}(\boldsymbol{\xi})$ is real valued.

A function $f(\mathbf{x})$ which only depends on $\|\mathbf{x}\|$ is said to be *radially symmetric*. In this case there is function F of a single variable so that

$$f(\mathbf{x}) = F(\|\mathbf{x}\|).$$

The Fourier transform of f is also radially symmetric and is given by the 1-dimensional integral

$$\hat{f}(\boldsymbol{\xi}) = \frac{c_n}{\|\boldsymbol{\xi}\|^{\frac{n-2}{2}}} \int_0^\infty J_{\frac{n-2}{2}}(r\|\boldsymbol{\xi}\|) F(r) r^{\frac{n}{2}} dr. \tag{3.98}$$

Here c_n is a constant and $J_\nu(z)$ is the order ν Bessel function defined by the integral

$$J_\nu(z) = a_\nu z^\nu \int_0^\pi e^{iz\cos(\theta)} \sin^{2\nu}(\theta) d\theta.$$

Here a_ν is a constant.

Example 3.3.3. The Fourier transform of the characteristic function of the unit ball $B_1 \subset \mathbb{R}^n$ is given by the radial integral

$$\widehat{\chi_{B_1}}(\boldsymbol{\xi}) = \frac{c_n}{\|\boldsymbol{\xi}\|^{\frac{n-2}{2}}} \int_0^1 J_{\frac{n-2}{2}}(r\|\boldsymbol{\xi}\|) r^{\frac{n}{2}} dr.$$

Using formula 6.561.5 in [20] gives

$$\widehat{\chi_{B_1}}(\boldsymbol{\xi}) = \frac{c_n}{\|\boldsymbol{\xi}\|^{\frac{n}{2}}} J_{\frac{n}{2}}(\|\boldsymbol{\xi}\|).$$

As $\|\boldsymbol{\xi}\|$ tends to infinity the Bessel function is a oscillatory term times $[\sqrt{\|\boldsymbol{\xi}\|}]^{-1}$. Overall we have the estimate

$$\widehat{\chi_{B_1}}(\boldsymbol{\xi}) \leq \frac{C}{(1+\|\boldsymbol{\xi}\|)^{\frac{n+1}{2}}}.$$

Exercise 3.3.9. Verify properties (4) and (5).

Exercise 3.3.10. Prove that the Fourier transform of a radial function is also a radial function and formula (3.98).

3.3.5 Convolution

See: A.4.7.

The convolution operation is defined in several variables just as in one dimension.

Definition 3.3.3. Let f be an integrable function and g, a bounded function; the convolution product of f and g is defined by

$$f * g(\mathbf{x}) = \int_{\mathbb{R}^n} f(\mathbf{y}) g(\mathbf{x}-\mathbf{y}) d\mathbf{y}.$$

With these hypotheses the integral converges absolutely.

Once again if g non-negative then $f * g$ can be interpreted as a weighted average of f. If g is also integrable then $f(\mathbf{y})g(\mathbf{x}-\mathbf{y})$ is integrable, as a function of $(\mathbf{y},\mathbf{x}) \in \mathbb{R}^n \times \mathbb{R}^n$:

$$\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} |f(\mathbf{y})g(\mathbf{x}-\mathbf{y})| d\mathbf{y} d\mathbf{x} = \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} |f(\mathbf{y})g(\mathbf{x}-\mathbf{y})| d\mathbf{x} d\mathbf{y} = \|f\|_{L^1}\|g\|_{L^1}.$$

That is

$$\|f * g\|_{L^1} \leq \|f\|_{L^1}\|g\|_{L^1}. \tag{3.99}$$

Any function $g \in L^1(\mathbb{R}^n)$ is the limit, in the L^1-norm of a sequence of functions $< g_n >$, where each g_n is a bounded function of bounded support. The estimate (3.99) shows that

$$\|f * g_n - f * g_m\|_{L^1} = \|f * (g_n - g_m)\|_{L^1} \leq \|f\|_{L^1}\|g_n - g_m\|_{L^1}.$$

Since $L^1(\mathbb{R}^n)$ is complete this implies that the convolution $f * g$ can be defined as the limit of the L^1-Cauchy sequence $< f * g_n >$. Convolution of integrable functions therefore defines a bilinear operation from $L^1(\mathbb{R}^n) \times L^1(\mathbb{R}^n)$ to $L^1(\mathbb{R}^n)$. The convolution product has most of the properties one expects of a multiplication.

Proposition 3.3.5. *The convolution product is commutative, associative and distributive: for $f, g \in L^1(\mathbb{R}^n)$ we have*

$$g * f = f * g, \quad (f * g) * h = f * (g * h), \quad f * (g + h) = f * g + f * h.$$

The proofs are left as exercises for the reader.

As in the one dimensional case there is no locally integrable function ψ so that, for all integrable functions f,

$$\psi * f = f.$$

The n-dimensional δ-function is a generalized function defined by the condition that, for any continuous function f,

$$\int_{\mathbb{R}^n} \delta(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = f(0). \tag{3.100}$$

The argument given in section 3.2.8 shows that no locally integrable function can satisfy this requirement. The Fourier transform of δ is computed, formally as before:

$$\hat{\delta}(\boldsymbol{\xi}) = \int_{\mathbb{R}^n} \delta(\mathbf{x}) e^{-i\mathbf{x}\cdot\boldsymbol{\xi}} d\boldsymbol{\xi} = 1.$$

Applying (3.100) to $f_{\mathbf{x}}(\mathbf{y}) = f(\mathbf{x} - \mathbf{y})$ shows that

$$\delta * f(\mathbf{x}) = \int_{\mathbb{R}^n} \delta(\mathbf{y}) f_{\mathbf{x}}(\mathbf{y}) d\mathbf{y} = f_{\mathbf{x}}(0) = f(\mathbf{x}).$$

The n-dimensional convolution has the same intimate connection with the Fourier transform as in the one dimensional case.

Proposition 3.3.6. *Suppose that f and g are absolutely integrable then*

$$\widehat{f * g}(\boldsymbol{\xi}) = \hat{f}(\boldsymbol{\xi})\hat{g}(\boldsymbol{\xi}).$$

Proof. The proof is a simple change in then order of integrations followed by a change of variable:

$$\begin{aligned}\widehat{f*g}(\boldsymbol{\xi}) = \int\limits_{\mathbb{R}^n} (f*g)(\mathbf{x}) e^{-i\boldsymbol{\xi}\cdot\mathbf{x}} d\mathbf{x} &= \iint\limits_{\mathbb{R}^n\times\mathbb{R}^n} f(\mathbf{y}) g(\mathbf{x}-\mathbf{y}) e^{-i\boldsymbol{\xi}\cdot\mathbf{x}} d\mathbf{y} d\mathbf{x} \\ &= \iint\limits_{\mathbb{R}^n\times\mathbb{R}^n} f(\mathbf{y}) g(\mathbf{t}) e^{-i\boldsymbol{\xi}\cdot(\mathbf{y}+\mathbf{t})} d\mathbf{t} d\mathbf{y} \\ &= \hat{f}(\boldsymbol{\xi})\hat{g}(\boldsymbol{\xi}).\end{aligned} \tag{3.101}$$

□

The convolution satisfies estimates with respect to other L^p-norms.

Proposition 3.3.7. *If $f \in L^1(\mathbb{R}^n)$ and g is a bounded function with bounded support then for any $1 \leq p \leq \infty$ we have the estimate.*

$$\|f*g\|_{L^p} \leq \|f\|_{L^1}\|g\|_{L^p}.$$

Proof. This is a consequence of Hölder's inequality, see (A.91). Let q be the dual exponent so that $p^{-1}+q^{-1}=1$, and note that

$$|f(\mathbf{y})g(\mathbf{x}-\mathbf{y})| = |f(\mathbf{y})|^{\frac{1}{q}}|f(\mathbf{y})|^{\frac{1}{p}}|g(\mathbf{x}-\mathbf{y})|.$$

We use this to estimate the L^p-norm of $f*g$:

$$\begin{aligned}\|f*g\|_{L^p}^p &= \int\limits_{\mathbb{R}^n}\left|\int\limits_{\mathbb{R}^n} f(\mathbf{y})g(\mathbf{x}-\mathbf{y})d\mathbf{y}\right|^p d\mathbf{x} \\ &\leq \int\limits_{\mathbb{R}^n}\left[\int\limits_{\mathbb{R}^n}|f(\mathbf{y})|d\mathbf{y}\right]^{\frac{p}{q}}\left[\int\limits_{\mathbb{R}^n}|f(\mathbf{y})||g(\mathbf{x}-\mathbf{y})|^p d\mathbf{y}\right] d\mathbf{x} \\ &= \|f\|_{L^1}^{\frac{p}{q}}\int\limits_{\mathbb{R}^n}|f(\mathbf{y})|d\mathbf{y}\int\limits_{\mathbb{R}^n}|g(\mathbf{t})|^p d\mathbf{t} \\ &= \|f\|_{L^1}^p\|g\|_{L^p}^p.\end{aligned} \tag{3.102}$$

□

As above with the L^1 case, the convolution can be extended as a bilinear map from $L^1(\mathbb{R}^n)\times L^p(\mathbb{R}^n) \to L^p(\mathbb{R}^n)$.

Convolution can be used to smooth a function.

Proposition 3.3.8. *Suppose that f is locally integrable and that g has bounded support and k continuous derivatives, then $f*g$ also has k continuous derivatives. For any multi-index $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}| \leq k$ we have*

$$\partial_{\mathbf{x}}^{\boldsymbol{\alpha}}(f*g) = f*(\partial_{\mathbf{x}}^{\boldsymbol{\alpha}}g). \tag{3.103}$$

The proof is left to the reader.

Exercise 3.3.11. Suppose that $f \in L^1(\mathbb{R}^n)$ and $g \in L^2(\mathbb{R}^n)$ show that

$$\widehat{f*g} = \hat{f}\hat{g}.$$

Note that neither $\hat{g}$ nor $f*g$ are defined by pointwise convergent integrals.

3.3.6 The support of $f * g$.

Suppose that f and g have bounded support. For applications to medical imaging it is important to understand how the support of $f * g$ is related to the supports of f and g. To that end we define the *algebraic sum* of two subsets of $\mathbb{R}^n$.

Definition 3.3.4. Suppose A and B are subsets of $\mathbb{R}^n$. The algebraic sum of A and B is defined as the set

$$A + B = \{\mathbf{a} + \mathbf{b} \in \mathbb{R}^n \ : \ \mathbf{a} \in A, \text{ and } \mathbf{b} \in B\}.$$

Using this concept we can give a quantitative result describing the way in which convolution "smears" out the support of a function.

Lemma 3.3.2. *The support of $f * g$ is contained in* $\operatorname{supp} f + \operatorname{supp} g$.

Proof. Suppose that $\mathbf{x}$ is not in $\operatorname{supp} f + \operatorname{supp} g$. This means that no matter which $\mathbf{y}$ is selected either $f(\mathbf{y})$ or $g(\mathbf{x} - \mathbf{y})$ is zero. Otherwise $\mathbf{x} = \mathbf{y} + (\mathbf{x} - \mathbf{y})$ would belong to $\operatorname{supp} f + \operatorname{supp} g$. This implies that $f(\mathbf{y})g(\mathbf{x} - \mathbf{y})$ is zero for all $\mathbf{y} \in \mathbb{R}^n$ and therefore

$$f * g(\mathbf{x}) = \int_{\mathbb{R}^n} f(\mathbf{y})g(\mathbf{x} - \mathbf{y})d\mathbf{y} = 0$$

as well. This proves the lemma. □

Suppose that f is a function which represents an image. For example we could imagine that f takes values between 0 and 1 with 0 corresponding to white and 1 to black, with values in between corresponding to shades of grey. Convolution provides a reasonable model for the measurement of an image. Define a smooth function with support in the unit ball by setting

$$\varphi(\mathbf{x}) = \begin{cases} c_n e^{-\frac{1}{1-\|\mathbf{x}\|^2}} & \text{if } \|\mathbf{x}\| < 1, \\ 0 & \text{if } \|\mathbf{x}\| \geq 1. \end{cases}$$

The constant is determined so that

$$\int_{\mathbb{R}^n} \varphi d\mathbf{x} = 1.$$

Scale φ by setting

$$\varphi_\epsilon(\mathbf{x}) = \frac{1}{\epsilon^n}\varphi(\frac{\mathbf{x}}{\epsilon}).$$

The scaled function φ_ϵ has support in a ball of radius ϵ and integral 1. Suppose that f has bounded support then, by Lemma 3.3.2,

$$\operatorname{supp}(f * \varphi_\epsilon) \subset \operatorname{supp} f + \operatorname{supp} \varphi_\epsilon.$$

This is the ϵ-neighborhood of the support of f,

$$\operatorname{supp}(f)_\epsilon = \{\mathbf{x} \in \mathbb{R}^n \ : \ \operatorname{dist}(\mathbf{x}, \operatorname{supp}(f)) \leq \epsilon\}.$$

The figures indicates what happens in the 2-dimensional case.

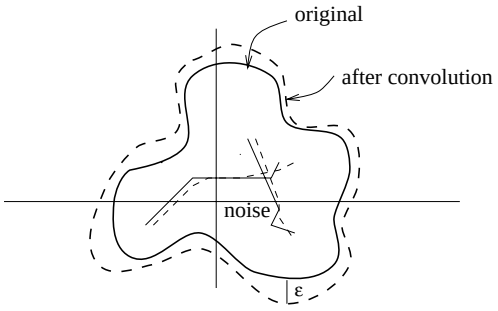


Figure 3.10: f is smeared into the ϵ-neighborhood of supp(f).

If f represents an image then the convolution of f with φ_ϵ smears the image. The value of $\varphi_\epsilon * f(\mathbf{x})$ depends on the values of f in the ball of radius ϵ about $\mathbf{x}$. The parameter ϵ is therefore measure of the resolution of the measuring apparatus. At points where the image is slowly varying the measured image is very close to the actual image. Near points where f is rapidly varying this may not be the case. Noise is usually a high frequency phenomenon with "mean zero;" the smoothing averages out the noise. At the same time, the image is blurred. The size of ϵ determines the degree of blurring.

The Fourier transform of $f * \varphi_\epsilon$ is $\hat{f}(\boldsymbol{\xi})\hat{\varphi}(\epsilon\boldsymbol{\xi})$. Because φ has integral 1 it follows that $\hat{\varphi}(0) = 1$. As the support of φ is a bounded set, its Fourier transform is a smooth function. This shows that $\hat{\varphi}(\epsilon\boldsymbol{\xi}) \approx 1$ if $\|\boldsymbol{\xi}\| << \epsilon^{-1}$. Thus the low frequency part of $f * \varphi_\epsilon$ closely approximates that of f. On the other hand $\hat{\varphi}(\boldsymbol{\xi})$ tend to zero rapidly as $\|\boldsymbol{\xi}\| \to \infty$. Thus the high frequency content of f is strongly suppressed in $f * \varphi_\epsilon$. Unfortunately both noise *and* fine detail are carried by the high frequency components. Noise looks like rapid local variations in the image; convolving f with a smooth function produces a smoother and therefore less noisy image.

Exercise 3.3.12. If f is a locally integrable function and φ has bounded support show that, if f is continuous at $\mathbf{x}$ then

$$\lim_{\epsilon \downarrow 0} \varphi_\epsilon * f(\mathbf{x}) = f(\mathbf{x}).$$

The functions $\{\varphi_\epsilon\}$ are an approximation to the δ-function.

3.3.7 L^2-derivatives*

See: A.3.6, A.4.7.

Just as in the one dimensional case there is a theory of weak derivatives in higher dimensions.

Definition 3.3.5. Let f be a locally integrable function and suppose that there is a locally integrable function f_j such that, for every smooth function φ, vanishing outside a bounded set, we have

$$\int_{\mathbb{R}^n} f(\mathbf{x})\partial_{x_j}\varphi(\mathbf{x})d\mathbf{x} = -\int_{\mathbb{R}^n} f_j(\mathbf{x})\varphi(\mathbf{x})d\mathbf{x}.$$

The function f_j is the weak x_j-partial derivative of f. Higher derivatives are defined recursively as before. For a multi-index $\boldsymbol{\alpha}$ we say that f has an $\boldsymbol{\alpha}^{\text{th}}$ weak derivative if there is a locally integrable function, $f_{\boldsymbol{\alpha}}$ so that for every smooth function, φ of bounded support we have

$$\int_{\mathbb{R}^n} f\partial_{\mathbf{x}}^{\boldsymbol{\alpha}}\varphi d\mathbf{x} = (-1)^{\boldsymbol{\alpha}}\int_{\mathbb{R}^n} f_{\boldsymbol{\alpha}}\varphi d\mathbf{x}.$$

Recall that a smooth function with bounded support is called a *test function.*

The case of principal interest is where f and its weak derivatives are in $L^2(\mathbb{R}^n)$.

Definition 3.3.6. Let $f \in L^2(\mathbb{R}^n)$ and suppose that its weak x_j-partial derivative belongs to $L^2(\mathbb{R}^n)$, we then say that f has an L^2 x_j-partial derivative. For a multi-index $\boldsymbol{\alpha}$ we say that f has an $\boldsymbol{\alpha}^{\text{th}}$ L^2-derivative if its weak $\boldsymbol{\alpha}^{\text{th}}$-derivative belongs to L^2.

Proposition 3.3.9. *A function $f \in L^2(\mathbb{R}^n)$ is k-times L^2-differentiable if and only if*

$$\int_{\mathbb{R}^n} \|\boldsymbol{\xi}\|^{2k}|\hat{f}(\boldsymbol{\xi})|^2 d\boldsymbol{\xi} < \infty. \tag{3.104}$$

In this case, for each $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}| \leq k$ we have the relations

$$\widehat{\partial_{\mathbf{x}}^{\boldsymbol{\alpha}} f} = (i\boldsymbol{\xi})^{\boldsymbol{\alpha}}\hat{f} \text{ and } \int_{\mathbb{R}^n} |\partial_{\mathbf{x}}^{\boldsymbol{\alpha}} f(\mathbf{x})|^2 d\mathbf{x} = \int_{\mathbb{R}^n} |\boldsymbol{\xi}^{\alpha}\hat{f}(\boldsymbol{\xi})|^2 \frac{d\boldsymbol{\xi}}{[2\pi]^n}. \tag{3.105}$$

This proposition contains a statement about the uniqueness of weak derivatives. If f satisfies (3.104) then it is not difficult to show, using the Parseval formula, that $\mathcal{F}^{-1}((i\boldsymbol{\xi})^{\boldsymbol{\alpha}}\hat{f})$ is the $\boldsymbol{\alpha}^{\text{th}}$-weak derivative of f. It is more difficult to prove the converse statement, that if f has weak partial derivatives $< f_{\boldsymbol{\alpha}} >$ of order less than or equal to k which belong to $L^2(\mathbb{R}^n)$ then

$$f_{\boldsymbol{\alpha}} = \mathcal{F}^{-1}\left[(i\boldsymbol{\xi})^{\boldsymbol{\alpha}}\hat{f}\right].$$

As the proof of this statement would take us too far afield we refer the reader to [17].

If on the other hand $f(\mathbf{x})$ decays at a polynomial rate as $\|\mathbf{x}\| \to \infty$, then $\hat{f}$ has L^2-derivatives.

Proposition 3.3.10. *Suppose that $f \in L^2(\mathbb{R}^n)$ and*

$$\int_{\mathbb{R}^n} (1 + \|\mathbf{x}\|^2)^k |f(\mathbf{x})|^2 d\mathbf{x} < \infty,$$

then $\hat{f}$ is k-times L^2-differentiable. If $|\boldsymbol{\alpha}| \leq k$ then

$$\partial_{\boldsymbol{\xi}}^{\boldsymbol{\alpha}}\hat{f} = (-1)^{|\boldsymbol{\alpha}|}\widehat{\mathbf{x}^{\boldsymbol{\alpha}} f} \text{ and } \int_{\mathbb{R}^n} |\mathbf{x}^{\boldsymbol{\alpha}} f(x)|^2 dx = \int_{\mathbb{R}^n} |\partial_{\boldsymbol{\xi}}^{\boldsymbol{\alpha}}\hat{f}(\boldsymbol{\xi})|^2 \frac{d\boldsymbol{\xi}}{[2\pi]^n}. \tag{3.106}$$

There is an important difference between one and several dimensions in the theory of L^2-derivatives. In one dimension we showed that an L^2-differentiable function is continuous (actually Hölder-$\frac{1}{2}$). If the dimension is larger than 1 then this is no longer true. The basic result is the following.

Theorem 3.3.4 (Sobolev embedding theorem). *Suppose that $f \in L^2(\mathbb{R}^n)$ and*

$$\int_{\mathbb{R}^n} (1 + \|\boldsymbol{\xi}\|)^{n+\epsilon} |\hat{f}(\boldsymbol{\xi})|^2 d\xi < \infty \tag{3.107}$$

for some $\epsilon > 0$, then $f(\mathbf{x})$ is a continuous function.

Proof. The estimate, (3.107) implies that $\hat{f}$ is absolutely integrable. We use Hölder's inequality:

$$\int_{\mathbb{R}^n} |\hat{f}(\boldsymbol{\xi})| d\boldsymbol{\xi} \leq \sqrt{\int_{\mathbb{R}^n} |\hat{f}(\boldsymbol{\xi})|^2 (1 + \|\boldsymbol{\xi}\|)^{n+\epsilon} d\boldsymbol{\xi}} \sqrt{\int_{\mathbb{R}^n} (1 + \|\boldsymbol{\xi}\|)^{-(n+\epsilon)} d\boldsymbol{\xi}} < \infty.$$

For $R > 0$ define

$$f_R = \frac{1}{[2\pi]^n} \int_{\|\boldsymbol{\xi}\| < R} \hat{f}(\boldsymbol{\xi}) e^{i\mathbf{x}\cdot\boldsymbol{\xi}} d\boldsymbol{\xi}.$$

The fact that $\hat{f}$ is absolutely integrable implies that $f_R(\mathbf{x})$ converges locally uniformly to a continuous function $F(\mathbf{x})$. On the other hand Proposition 3.3.4 implies that f_R tends to f in $L^2(\mathbb{R}^n)$. This shows that $F = f$ and completes the proof of the theorem. □

The theorem says that, in n-dimensions a function needs to have a little more that $\frac{n}{2}$ L^2-derivatives to be continuous.

Example 3.3.4. To see that these results are sharp and better understand the relationship between the classical notion of smoothness and weak differentiability we consider functions defined on $\mathbb{R}^n$ of the form $f_a(\mathbf{x}) = \psi(\|\mathbf{x}\|)\|\mathbf{x}\|^{-a}$ with $0 \leq a < n$. Here ψ is a fixed infinitely differentiable function with the following properties:

(1). $\psi(r) = 1$ for $r \in [-1, 1]$.

(2). $\psi(r) = 0$ for $|r| > 2$.

The functions, f_a have bounded support and are smooth, but for the singularity at $\|\mathbf{x}\| = 0$. If $0 \leq a < n/2$ then $f_a \in L^2(\mathbb{R}^n)$. The radial derivative of f_a is a sum of two terms

$$\partial_r f_a = \psi'(\|\mathbf{x}\|)\|\mathbf{x}\|^{-a} - a\psi(\|\mathbf{x}\|)\|\mathbf{x}\|^{-(1+a)}.$$

As $\psi'(r) = 0$ if $|r| \leq 1$ the first term is a smooth function on $\mathbb{R}^n$. Inductively it follows that

$$\partial_r^k = f_{k,a}(\|\mathbf{x}\|) + (-1)^k a(1+a)\cdots(k-1+a)\psi(\|\mathbf{x}\|)\|x\|^{-(k+a)}, \tag{3.108}$$

where the function $f_{k,a}$ is smooth and has bounded support.

If $k + a < n$ then the right hand side is the weak radial derivative of f_a. If $k < n/2 - a$ then $\partial_r^k f_a \in L^2(\mathbb{R}^n)$. These examples show that, as the dimension increases, a singular

function can have more and more weak derivatives. There is an interplay between the order of the singularity and the size of the set where the function is singular. In these examples the singular set is a point, which constitutes the smallest possible singular set.

The Fourier transform of f_a can be computed in terms of the $\frac{(n-2)}{2}$-order J-Bessel function:

$$\hat{f}_a(\boldsymbol{\xi}) = \frac{c}{\|\boldsymbol{\xi}\|^{\frac{n-2}{2}}} \int_0^1 \psi(r) r^{-a} J_{\frac{n-2}{2}}(r\|\mathbf{x}\|) r^{\frac{n}{2}} dr.$$

Changing variables and integrating by parts it can be shown that, for large $\boldsymbol{\xi}$,

$$\hat{f}_a(\boldsymbol{\xi}) = \frac{c_{a,n}}{\|\boldsymbol{\xi}\|^{n-a}} + O(\|\boldsymbol{\xi}\|^{-n}).$$

From this asymptotic formula it is clear that

$$\int_{\mathbb{R}^n} |\hat{f}_a(\boldsymbol{\xi})|^2 (1 + \|\boldsymbol{\xi}\|^2)^k d\boldsymbol{\xi} < \infty$$

provided that $k < n/2 - a$. This agrees with our previous computation. It shows that a function with less than $\frac{n}{2}$ L^2-derivatives need not be continuous.

Exercise 3.3.13. Suppose that f is a twice weakly differentiable function. Show that the mixed weak derivatives satisfy

$$\partial_{x_i} \partial_{x_j} f = \partial_{x_j} \partial_{x_i} f.$$

In other words: "mixed weak partial derivatives commute." This shows that weak partial derivatives satisfy one of the most important properties of classical derivatives.

Exercise 3.3.14. Suppose that f satisfies (3.104), show that for any multi-index $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}| < k$ the L^2-function, $\mathcal{F}^{-1}((i\boldsymbol{\xi})^{\boldsymbol{\alpha}} \hat{f})$ is the $\boldsymbol{\alpha}^{\text{th}}$-weak derivative of f.

Exercise 3.3.15. Show that f_a has $n - a$ weak derivatives.

Exercise 3.3.16. Prove that if $a < n$ then

$$\int_0^1 \psi(r) r^{-a} J_{\frac{n-2}{2}}(r\|\mathbf{x}\|) r^{\frac{n}{2}} dr$$

has a limiting value as $\|\boldsymbol{\xi}\|$ tend to infinity. Hint: Split the integral into an integral from 0 to $\epsilon > 0$ and an integral from ϵ to 1 and use integration by parts and the asymptotic expansion for Bessel function in the second part. See [20] for the asymptotic expansions of Bessel functions.

3.3.8 The failure of localization in higher dimensions*

The localization principle is a remarkable feature of the 1-dimensional Fourier transform. Suppose that f is an integrable function defined on $\mathbb{R}$. According to the localization principle the convergence of the partial inverse

$$f_R(x) = \frac{1}{2\pi} \int_{-R}^{R} \hat{f}(\xi) e^{ix\xi} d\xi$$

to $f(x)$ only depends on the behavior of f in an interval about x. This is a uniquely one dimensional phenomenon. In this section we give an example due to Pinsky showing the failure of the localization principle in three dimensions. A complete discussion of this phenomenon can be found in [57].

Pinsky's example is very simple, it concerns $f(\mathbf{x}) = \chi_{B_1}(\mathbf{x})$, the characteristic function of the unit ball. The Fourier transform of f was computed in example 3.3.3, it is

$$\hat{f}(\boldsymbol{\xi}) = \frac{cJ_{\frac{3}{2}}(\|\boldsymbol{\xi}\|)}{\|\boldsymbol{\xi}\|^{\frac{3}{2}}}.$$

In this example c denotes various positive constants. Using formula 8.464.3 in [20] this can be re-expressed in terms of elementary functions by

$$\hat{f}(\boldsymbol{\xi}) = \frac{c[\|\boldsymbol{\xi}\| \cos(\|\boldsymbol{\xi}\|) - \sin(\|\boldsymbol{\xi}\|)]}{\|\boldsymbol{\xi}\|^3}.$$

Using polar coordinates, we compute the partial inverse:

$$\begin{aligned} f_R(0) &= \frac{c}{[2\pi]^3} \int_0^R \left[\cos(r) - \frac{\sin(r)}{r} \right] dr \\ &= c \left[\sin(R) - \int_0^R \frac{\sin(r)}{r} dr \right]. \end{aligned} \tag{3.109}$$

The last integral has a limit as $R \to \infty$ however $\sin(R)$ does not! Thus $f_R(0)$ remains bounded as R tends to infinity but does not converge.

Remark 3.3.2. The reader interested in complete proofs for the results in this section as well as further material is directed to [40], for the one dimensional case or [72], for higher dimensions.

Exercise 3.3.17. Prove the existence of the limit

$$\lim_{R \to \infty} \int_0^R \frac{\sin(r) dr}{r}.$$

Chapter 4

The Radon transform

In Chapter 2 we introduced the Radon transform and discussed its simpler properties. After reviewing its definition we establish several further properties of this transform. The remainder of the chapter is devoted to a study of its inverse.

4.1 The Radon transform

See: A.2.3, A.5.

The Radon transform is a linear operator which maps a function defined in the plane to a function on $\mathbb{R} \times S^1$, the space of oriented lines in $\mathbb{R}^2$. The pair (t, ω) corresponds to the line

$$l_{t,\omega} = \{(x, y) \,:\, \langle \omega, (x, y) \rangle = t\} = \{t\omega + s\hat{\omega} \,:\, s \in \mathbb{R}\}.$$

Here $\hat{\omega}$ is the unit vector perpendicular to ω with the orientation determined by

$$\det(\omega\hat{\omega}) > 0.$$

The variable t is called the *affine parameter*, it is the oriented distance of the line $l_{t,\omega}$ to the origin of the coordinate system. The Radon transform of f at (t, ω) is defined by the integral

$$\mathrm{R}f(t, \omega) = \int_{-\infty}^{\infty} f(t\omega + s\hat{\omega}) ds.$$

For the moment we restrict our attention to piecewise continuous functions with bounded support. Because the geometric lines $l_{t,\omega}$ and $l_{-t,-\omega}$ are the same the Radon transform is an even function,

$$\mathrm{R}f(-t, -\omega) = \mathrm{R}f(t, \omega). \tag{4.1}$$

Representing the point $\omega \in S^1$ as

$$\omega(\theta) = (\cos\theta, \sin\theta)$$

allows an identification of $\mathbb{R}\times S^1$ with $\mathbb{R}\times[0,2\pi)$. With this identification $dtd\theta$ can be used as an area element on the space of lines. The integral of a function $h(t,\theta)=h(t,\omega(\theta))$ is given by

$$\int_0^{2\pi}\int_{-\infty}^{\infty} h(t,\theta)dtd\theta.$$

We often use the notation $dtd\omega$ to denote this measure on $\mathbb{R}\times S^1$.

Definition 4.1.1. The set $L^2(\mathbb{R}\times S^1)$ consists of measurable functions for which

$$\|h\|^2_{L^2(\mathbb{R}\times S^1)}=\int_0^{2\pi}\int_{-\infty}^{\infty}|h(t,\theta)|^2dtd\theta \tag{4.2}$$

is finite.

A function h on $\mathbb{R}\times S^1$ is continuous if $h(t,\theta)$ is 2π-periodic in θ and continuous as a function on $\mathbb{R}\times[0,2\pi]$. Similarly h is differentiable if it is 2π-periodic and differentiable on $\mathbb{R}\times[0,2\pi]$ and $\partial_\theta h$ is also 2π-periodic. Higher orders of differentiability have similar definitions.

The Radon transform has several properties analogous to those established for the Fourier transform in the previous chapter. Suppose that f and g are two functions with bounded supported. There is a simple formula relating $\mathrm{R}(f*g)$ to $\mathrm{R}f$ and $\mathrm{R}g$.

Proposition 4.1.1. *Let f and g be piecewise continuous functions with bounded support then*

$$\mathrm{R}[f*g](t,\omega)=\int_{-\infty}^{\infty}\mathrm{R}f(s,\omega)\,\mathrm{R}g(t-s,\omega)ds. \tag{4.3}$$

Remark 4.1.1. Colloquially one says that the Radon transform converts convolution in the plane to convolution in the affine parameter.

Proof. The proof is a calculation. Fix a direction ω, coordinates (s,t) for the plane are defined by the assignment

$$(s,t)\mapsto s\hat{\omega}+t\omega.$$

This is an orthogonal change of variables so the area element on $\mathbb{R}^2$ is given by $dsdt$. In these variables the convolution of f and g becomes

$$f*g(s\hat{\omega}+t\omega)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(a\hat{\omega}+b\omega)g((s-a)\hat{\omega}+(t-b)\omega)dadb.$$

The Radon transform of $f * g$ is computing by switching the order of the integrations:

$$\begin{aligned} \mathrm{R}f * g(\tau,\omega) &= \int_{-\infty}^{\infty} f * g(\tau\omega + s\hat{\omega})ds \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(a\hat{\omega}+b\omega)g((s-a)\hat{\omega}+(\tau-b)\omega)dadbds \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(a\hat{\omega}+b\omega)g((s-a)\hat{\omega}+(\tau-b)\omega)dsdadb \\ &= \int_{-\infty}^{\infty} \mathrm{R}f(b,\omega)\,\mathrm{R}g(\tau-b,\omega)db. \end{aligned} \tag{4.4}$$

In the second to last line we interchanged the s-integration with the a and b integrations. □

Remark 4.1.2. The smoothness of a function with bounded support is reflected in the decay properties of its Fourier transform. From Proposition 4.1.1 it follows that the smoothness of a function of bounded support is also reflected in the smoothness of its Radon transform in the affine parameter. To see this suppose that f is a continuous function of bounded support and φ is a radially symmetric function, with bounded support and k-continuous derivatives. The convolution $f * \varphi$ has bounded support and k-continuous derivatives. The Radon transform of φ is only a function of t; the Radon transform of the convolution,

$$\mathrm{R}f * \varphi(t,\omega) = \int_{-\infty}^{\infty} \mathrm{R}f(\tau,\omega)\,\mathrm{R}\varphi(t-\tau)d\tau,$$

has the same smoothness in t as $\mathrm{R}\varphi$. Regularity of f is also reflected in smoothness of $\mathrm{R}f$ in the angular variable, though it is more difficult to see explicitly, see exercise 4.1.6.

Let $\mathbf{v} = (v_1, v_2)$ be a fixed vector and define the translate of f by $\mathbf{v}$ to be the function

$$f_{\mathbf{v}}(x,y) = f(x - v_1, y - v_2).$$

There is a simple relation between the Radon transform of f and that of $f_{\mathbf{v}}$.

Proposition 4.1.2. *Let f be a piecewise continuous function with bounded support then*

$$\mathrm{R}f_{\mathbf{v}}(t,\omega) = \mathrm{R}f(t - \langle\omega,\mathbf{v}\rangle,\omega). \tag{4.5}$$

Using this formula we can relate the Radon transform of f to that of its partial derivatives. Let $\mathbf{e}_1 = (1,0)$ and $\mathbf{e}_2 = (0,1)$. The x and y partial derivatives of f are defined by

$$\partial_x f(x,y) = \lim_{h\to 0}\frac{f_{h\mathbf{e}_1}(x,y) - f(x,y)}{h} \text{ and } \partial_y f(x,y) = \lim_{h\to 0}\frac{f_{h\mathbf{e}_2}(x,y) - f(x,y)}{h}.$$

Lemma 4.1.1. *If f is a function with bounded support and bounded, continuous first partial derivatives then $\mathrm{R}f(t,\omega)$ is differentiable in t and*

$$\mathrm{R}\partial_x f(t,\omega) = -\omega_1 \partial_t \mathrm{R}f(t,\omega), \quad \mathrm{R}\partial_y f(t,\omega) = -\omega_2 \partial_t \mathrm{R}f(t,\omega). \tag{4.6}$$

Proof. We consider only the x-derivative, the y-derivative is identical. From (4.5) and the linearity of the Radon transform we conclude that

$$\mathrm{R}\left[\frac{f_{h\mathbf{e}_1} - f}{h}\right](t,\omega) = \frac{\mathrm{R}f(t-h\omega_1,\omega) - \mathrm{R}f(t,\omega)}{h}.$$

The conclusion follows by allowing h to tend to zero. □

This result extends, by induction to higher partial derivatives.

Proposition 4.1.3. *Suppose that f has bounded support and continuous partial derivatives of order k then $\mathrm{R}f(t,\omega)$ is k-times differentiable in t and, for non-negative integers i, j, with $i + j \leq k$ we have the formula*

$$\mathrm{R}\left[\partial_x^i \partial_y^j f\right](t,\omega) = (-1)^{i+j} \omega_1^i \omega_2^j \partial_t^{i+j} \mathrm{R}f(t,\omega). \tag{4.7}$$

Let $A : \mathbb{R}^2 \to \mathbb{R}^2$ be an rigid rotation of the plane, that is A is a linear map such that

$$\langle A\mathbf{v}, A\mathbf{w}\rangle = \langle \mathbf{v}, \mathbf{w}\rangle \text{ for all } \mathbf{v}, \mathbf{w} \in \mathbb{R}^2.$$

If f is a piecewise continuous function with bounded support then

$$f_A(x,y) = f(A(x,y))$$

is as well. The Radon transform of f_A is related to that of f in a simple way.

Proposition 4.1.4. *Let A be an rigid rotation of $\mathbb{R}^2$ and f a piecewise continuous function with bounded support then*

$$\mathrm{R}f_A(t,\omega) = \mathrm{R}f(t, A\omega). \tag{4.8}$$

Proof. The result follows from the fact that $\langle A\omega, A\hat{\omega}\rangle = \langle \omega, \hat{\omega}\rangle = 0$ and therefore

$$\begin{aligned} \mathrm{R}f_A(t,\omega) &= \int_{-\infty}^{\infty} f(tA\omega + sA\hat{\omega})ds \\ &= \mathrm{R}f(t, A\omega). \end{aligned} \tag{4.9}$$

□

Thus far we have only considered the Radon transform for piecewise continuous functions with bounded supported. This transform extends, without difficulty, to sufficiently regular functions with enough decay at infinity. More precisely, a function belongs to the *natural domain* of the Radon transform if the restriction of f to every line $l_{t,\omega}$ is an absolutely integrable function. If for example, $f(x,y)$ is a piecewise continuous function, satisfying an estimate of the form

$$|f(x,y)| \leq \frac{M}{(1 + \|(x,y)\|)^{1+\epsilon}},$$

for an $\epsilon > 0$, then f belongs to the natural domain of the Radon transform. The results in this section extend to functions in the natural domain of R. The proofs in this case are left to the reader. Using functional analytic methods the domain of the Radon transform can be further extended, allowing functions with both less regularity and slower decay. An example of such an extension was already presented in section 2.4.2. We return to this question in section 4.6.

Exercise 4.1.1. Prove formula (4.5). The argument is similar to that used in the proof of (4.3).

Exercise 4.1.2. Give the details of the argument in the proof of Lemma 4.1.1 showing that $\mathrm{R}f(t,\omega)$ is differentiable in the t-variable.

Exercise 4.1.3. Show how to derive formula (4.7) from (4.6).

Exercise 4.1.4. The Laplace operator Δ is defined by $\Delta f = -(\partial_x^2 f + \partial_y^2 f)$. Find a formula for $\mathrm{R}[\Delta f]$ in terms of $\mathrm{R}f$.

Exercise 4.1.5. Suppose that $A : \mathbb{R}^2 \to \mathbb{R}^2$ is an arbitrary linear transformation how is $\mathrm{R}f_A$ related to $\mathrm{R}f$?

Exercise 4.1.6. Let A_θ denote the rotation through the angle θ. Setting $\omega(\phi) = (\cos\phi, \sin\phi)$, let $\mathrm{R}f(t,\phi) = \mathrm{R}f(t,\omega(\phi))$ so that

$$\mathrm{R}f_{A_\theta}(t,\phi) = \mathrm{R}f(t,\theta+\phi).$$

Using these formulæ show that

$$\mathrm{R}[(y\partial_x - x\partial_y)f](t,\phi) = (\partial_\theta \mathrm{R})f(t,\phi).$$

4.2 Inversion of the Radon Transform

Now we are ready to use the Fourier transform to invert the Radon transform.

4.2.1 The Central slice theorem

The Fourier transform and Radon transform are connected in a very simple way. In medical imaging this relationship is called the *Central slice theorem*.

Theorem 4.2.1 (Central slice theorem). *Let f be an absolutely integrable function in the natural domain of* R. *For any real number r and unit vector ω we have the identity*

$$\int_{-\infty}^{\infty} \mathrm{R}f(\omega,t)e^{-itr}dt = \hat{f}(r\omega). \tag{4.10}$$

Proof. From the definition of the Radon transform, the integral on the left is equal to

$$\int_{-\infty}^{\infty} \mathrm{R}f(\omega,t)e^{-itr}dt = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(t\omega+s\hat{\omega})e^{-itr}dsdt. \tag{4.11}$$

This integral is absolutely convergent and therefore we may make the change of variables, $(x,y) = t\omega + s\hat{\omega}$. Checking that the Jacobian determinant is 1 and noting that

$$t = \langle (x,y),\omega\rangle,$$

the above integral therefore becomes

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(t\omega+s\hat{\omega})e^{-itr}dsdt &= \iint_{\mathbb{R}^2} f(x,y)e^{-i\langle(x,y),\omega\rangle r}dxdy \\ &= \hat{f}(r\omega)\end{aligned} \tag{4.12}$$

This completes the proof of the central slice theorem. □

For a given vector $\boldsymbol{\xi} = (\xi_1,\xi_2)$ the inner product, $\langle (x,y),\boldsymbol{\xi}\rangle$ is constant along any line perpendicular to the *direction* of $\boldsymbol{\xi}$. The central slice theorem interprets the computation of the Fourier transform at $\boldsymbol{\xi}$ as a two step process:

(1). First we integrate the function along lines perpendicular to $\boldsymbol{\xi}$, this gives us a function of the affine parameter alone.

(2). Compute the *1-dimensional* Fourier transform of this function of the affine parameter.

To understand this better we consider an example. Let $\mathbf{e}_1 = (1,0)$ and $\mathbf{e}_2 = (0,1)$ and $(t,\omega) = (x,\mathbf{e}_1)$. Since $\hat{\mathbf{e}}_1 = \mathbf{e}_2$, the Radon transform at $(x,\mathbf{e}_1)$ is given by

$$\begin{aligned}\mathrm{R}f(x,\mathbf{e}_1) &= \int_{-\infty}^{\infty} f(x\mathbf{e}_1 + y\hat{\mathbf{e}}_1)dy \\ &= \int_{-\infty}^{\infty} f(x,y)dy.\end{aligned}$$

The Fourier transform of $\mathrm{R}f(x,\mathbf{e}_1)$ is

$$\int_{-\infty}^{\infty} \mathrm{R}f(x,\mathbf{e}_1)e^{-irx}dx = \iint_{\mathbb{R}^2} f(x,y)e^{-irx}dydx.$$

Since $\langle r\mathbf{e}_1,(x,y)\rangle = rx$ this is the definition of $\hat{f}(r\mathbf{e}_1)$.

To simplify the formulæ which follow, we introduce notation for the 1-dimensional Fourier transform, in the affine parameter, of a function $h(t,\omega)$ defined on $\mathbb{R}\times S^1$:

$$\widetilde{h}(r,\omega) = \int_{-\infty}^{\infty} h(t,\omega)e^{-itr}dt. \tag{4.13}$$

If $h(t,\omega)$ belongs to $L^2(\mathbb{R})$ for a fixed ω then the one dimensional Parseval formula implies that

$$\int_{-\infty}^{\infty} |h(t,\omega)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\widetilde{h}(r,\omega)|^2 dr. \tag{4.14}$$

The Parseval formula for the $2d$-Fourier transform and the central slice theorem give a Parseval formula for the Radon transform.

Theorem 4.2.2 (Parseval Formula for the Radon Transform). *Suppose that f is in the natural domain of the Radon transform and is square integrable then*

$$\iint_{\mathbb{R}^2} |f|^2 dxdy = \frac{1}{[2\pi]^2} \int_0^{\pi} \int_{-\infty}^{\infty} |\widetilde{\mathrm{R}f}(r,\omega)|^2 |r| dr d\omega. \tag{4.15}$$

Proof. We begin by assuming that f is also absolutely integrable. The central slice theorem applies to show that

$$\begin{aligned} \iint_{\mathbb{R}^2} |f|^2 dxdy &= \frac{1}{[2\pi]^2} \int_0^{2\pi} \int_0^{\infty} |\hat{f}(r\omega)|^2 r dr d\omega \\ &= \frac{1}{[2\pi]^2} \int_0^{\pi} \int_{-\infty}^{\infty} |\widetilde{\mathrm{R}f}(r,\omega)|^2 |r| dr d\omega. \end{aligned} \tag{4.16}$$

In the last line we use the fact that the evenness of $\mathrm{R}f$ implies that

$$\widetilde{\mathrm{R}f}(r,\omega) = \widetilde{\mathrm{R}f}(-r,-\omega). \tag{4.17}$$

This proves (4.15) with the additional assumption. To remove this assumption we need to approximate f by absolutely integrable functions. Let φ be a non-negative, infinitely differentiable function with support in the disk of radius 1. Suppose further that

$$\int_{\mathbb{R}^2} \varphi(x,y) dxdy = 1.$$

For each $\epsilon > 0$ we let

$$\varphi_\epsilon(x,y) = \frac{1}{\epsilon^2} \varphi(\epsilon^{-1}x, \epsilon^{-1}y)$$

and define

$$f_\epsilon = \left[\chi_{[0,\epsilon^{-1}]}(r) f\right] * \varphi_\epsilon \tag{4.18}$$

The function f_ϵ is smooth and has bounded support. It therefore satisfies the hypotheses of both Theorem 4.2.2 and the central slice theorem. The argument above therefore applies to f_ϵ. The proof is completed by showing that $\lim_\epsilon \mathrm{R}f_\epsilon = \mathrm{R}f$ and that, as $\epsilon \downarrow 0$, f_ϵ converges in $L^2(\mathbb{R}^2)$ to f. These claims are left as exercises for the reader. □

Remark 4.2.1. * Formula (4.15) has two interesting consequences for the map $f \mapsto \mathrm{R}f$ as a map between L^2-spaces. It shows that R does **not** have an extension as a continuous mapping from $L^2(\mathbb{R}^2)$ to $L^2(\mathbb{R} \times S^1)$ and that R^{-1} also cannot be a continuous map from

$L^2(\mathbb{R}\times S^1)$ to $L^2(\mathbb{R}^2)$. These assertions follow from Corollary A.5.1 and the observation that

$$\|h\|^2_{L^2(\mathbb{R}\times S^1)} = \frac{1}{2\pi}\int_0^{2\pi}\int_{-\infty}^{\infty} |\widetilde{h}(r,\omega)|^2 dr d\omega.$$

Because $|r|$ varies between zero and infinity in (4.15) we see that there cannot exist constants M or M' so that either estimate

$$\|\,\mathrm{R}f\|_{L^2(\mathbb{R}\times S^1)} \le M\|f\|_{L^2(\mathbb{R}^2)} \text{ or } \|\,\mathrm{R}f\|_{L^2(\mathbb{R}\times S^1)} \ge M'\|f\|_{L^2(\mathbb{R}^2)}$$

holds for f in a dense subset of $L^2(\mathbb{R}^2)$.

To express the Parseval formula as an integral over the space of oriented lines we define a "half derivative" operator

$$D_{\frac{1}{2}}\,\mathrm{R}f(t,\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{\mathrm{R}f}(r,\omega)|r|^{\frac{1}{2}} e^{irt} dr.$$

The Parseval formula can then be re-written

$$\iint_{\mathbb{R}^2} |f|^2 dx dy = \frac{1}{2\pi}\int_0^{\pi}\int_{-\infty}^{\infty} |D_{\frac{1}{2}}\,\mathrm{R}f(t,\omega)|^2 dt d\omega. \tag{4.19}$$

This implies that in order for a function on the space of lines to be the Radon transform of a square integrable function it must have a "half-derivative" in the affine parameter. Unlike the Fourier transform, the Radon transform is not defined on all of $L^2(\mathbb{R}^2)$.

Exercise 4.2.1. If $f \in L^2(\mathbb{R}^2)$ and f_ϵ is defined in (4.18) show that $\underset{\epsilon\downarrow 0}{LIM} f_\epsilon = f$.

Exercise 4.2.2. If f is in the natural domain of R and f_ϵ is defined in (4.18) show that for every (t,ω)

$$\lim_{epsilon\downarrow 0} \mathrm{R}f_\epsilon(t,\omega) = \mathrm{R}f(t,\omega).$$

4.2.2 The Radon Inversion Formula

The central slice theorem and the inversion formula for the Fourier transform, (3.84) give an inversion formula for the Radon transform.

Theorem 4.2.3 (Radon inversion formula). *If f is an absolutely integrable function in the natural domain of the Radon transform and $\hat{f}$ is absolutely integrable then*

$$f(x,y) = \frac{1}{[2\pi]^2}\int_0^{\pi}\int_{-\infty}^{\infty} e^{ir\langle(x,y),\omega\rangle}\widetilde{\mathrm{R}f}(r,\omega)|r| dr d\omega \tag{4.20}$$

Proof. Because $\mathrm{R}f$ is an even function, it follows that its Fourier transform satisfies

$$\widetilde{\mathrm{R}f}(t,\omega) = \widetilde{\mathrm{R}f}(-t,-\omega). \tag{4.21}$$

As f and $\hat{f}$ are absolutely integrable it follows from Theorem 3.3.1 that

$$f(x,y) = \frac{1}{[2\pi]^2}\int\limits_{\mathbb{R}^2} \hat{f}(\boldsymbol{\xi}) e^{i\langle (x,y),\boldsymbol{\xi}\rangle} d\boldsymbol{\xi}.$$

Re-expressing the Fourier inversion formula using polar coordinates gives

$$\begin{aligned} f(x,y) &= \frac{1}{[2\pi]^2}\int e^{i\langle (x,y),\boldsymbol{\xi}\rangle}\hat{f}(\boldsymbol{\xi})d\boldsymbol{\xi} \\ &= \frac{1}{[2\pi]^2}\int\limits_0^{2\pi}\int\limits_0^{\infty} e^{ir\langle (x,y),\omega\rangle}\hat{f}(r\omega)rdrd\omega \\ &= \frac{1}{[2\pi]^2}\int\limits_0^{2\pi}\int\limits_0^{\infty} e^{ir\langle (x,y),\omega\rangle}\widetilde{\mathrm{R}f}(r\omega)rdrd\omega \end{aligned}$$

The central slice theorem is used in the last line. Using the relation (4.21) we can re-write this as

$$f(x,y) = \frac{1}{[2\pi]^2}\int\limits_0^{\pi}\int\limits_{-\infty}^{\infty} e^{ir\langle (x,y),\omega\rangle}\widetilde{\mathrm{R}f}(r,\omega)|r|drd\omega. \tag{4.22}$$

□

Remark 4.2.2. As was the case with the Fourier transform, the inversion formula for the Radon transform holds under weaker hypotheses than those stated in Theorem 4.2.3. Under these hypotheses all the integrals involved are absolutely convergent and therefore do not require any further interpretation. In imaging applications the data is usually piecewise continuous, vanishing outside a bounded set. As we know from our study of the Fourier transform, this does not imply that $\hat{f}$ is absolutely integrable and so the Fourier inversion formula requires a careful interpretation in this case. Such data is square integrable and therefore it follows from the results in section 3.3.3 that

$$f = \underset{\rho\to\infty}{LIM}\frac{1}{[2\pi]^2}\int\limits_0^{\pi}\int\limits_{-\rho}^{\rho} e^{ir\langle (x,y),\omega\rangle}\widetilde{\mathrm{R}f}(r,\omega)|r|drd\omega. \tag{4.23}$$

In most cases of interest, at points (x,y) where f is continuous, the integrals

$$\frac{1}{[2\pi]^2}\int\limits_{-\infty}^{\infty}\int\limits_0^{\pi} e^{ir\langle (x,y),\omega\rangle}\widetilde{\mathrm{R}f}(r,\omega)|r|d\omega dr$$

exist as improper Riemann integrals. Additional care is required in manipulating these expressions.

Remark 4.2.3. Formula (4.20) allows the determination of f from its Radon transform. This formula completes a highly idealized, mathematical model for medical image reconstruction.

- We consider a two dimensional slice of a three dimensional object, the physical parameter of interest is the attenuation coefficient $f(x, y)$. According to Beer's law, the intensity $I_{(t,\omega)}$ of X-rays (of a given energy) traveling along a line, $l_{t,\omega}$ is attenuated according the differential equation:

$$\frac{dI_{(t,\omega)}}{ds} = -fI_{(t,\omega)}.$$

  Here s is arclength along the line.

- By comparing the intensity of an incident beam of X-rays to that emitted we "measure" the Radon transform of f

$$\mathrm{R}f(t,\omega) = -\log\left[\frac{I_{o,(t,\omega)}}{I_{i,(t,\omega)}}\right].$$

- Using formula (4.20) the attenuation coefficient $f(x, y)$ is reconstructed from the measurements $\mathrm{R}f(t, \omega)$.

The most obvious flaw in this model is that, in practice $\mathrm{R}f(t, \omega)$ can only be measured for a finite set of pairs (t, ω). Nonetheless formula (4.20) provides a good starting point for the development of more practical algorithms.

If we were to omit the $|r|$ factor then it would follow from the 1-dimensional Fourier inversion formula applied to $\widetilde{\mathrm{R}f}$ that $f(x, y)$ would be given by

$$\begin{aligned} f(x,y) =& \frac{1}{[2\pi]^2}\int_0^\pi\int_{-\infty}^{\infty} e^{ir\langle(x,y),\omega\rangle}\hat{f}(r\omega)drd\omega \\ =& \frac{1}{2\pi}\int_0^\pi \mathrm{R}f(\langle(x,y),\omega\rangle,\omega)d\omega \end{aligned}$$

Note that the unique line in the family $\{l_{t,\omega} \,|\, t \in (-\infty, \infty)\}$ which passes through the point (x, y) is the one with affine parameter $t = \langle(x, y), \omega\rangle$. Thus the value of $f(x, y)$ would be half the average of its Radon transform over all lines passing through the point (x, y). This is the backprojection formula introduced in section 2.4.3. By comparison with the true inversion formula (4.20) it is now clear why the backprojection formula cannot be correct.

4.2.3 Backprojection*

See: A.2.5.

The operation of backprojection has a nice mathematical interpretation. If $(X, \langle\cdot,\cdot\rangle_X)$ and $(Y, \langle\cdot,\cdot\rangle_Y)$ are inner product spaces and $A : X \to Y$ is a linear map then the *adjoint* of A, $A^* : Y \to X$ is defined by the relations

$$\langle A\mathbf{x}, \mathbf{y}\rangle_Y = \langle \mathbf{x}, A^*\mathbf{y}\rangle_Y \text{ for all } \mathbf{x} \in X \text{ and } \mathbf{y} \in Y.$$

If we use the L^2-inner product for functions on $\mathbb{R}^2$ and the inner product for functions on $\mathbb{R} \times S^1$ compatible with the L^2-norm defined in (4.2),

$$\langle h, k\rangle_{\mathbb{R}\times S^1} = \int_0^{2\pi}\int_{-\infty}^{\infty} h(t,\omega)k(t,\omega)dtd\omega$$

then backprojection is $[4\pi]^{-1}$ times the formal adjoint of the Radon transform. It is only a formal adjoint because, as noted above, the Radon transform does not extend to define a continuous map from $L^2(\mathbb{R}^2)$ to $L^2(\mathbb{R} \times S^1)$. The proof is a simple calculation; for the sake of simplicity assume that f is a function of bounded support on $\mathbb{R}^2$ and h is a function of bounded support on $\mathbb{R} \times S^1$:

$$\begin{aligned}\langle \mathrm{R}f, h\rangle_{\mathbb{R}\times S^1} &= \int_0^{2\pi}\int_{-\infty}^{\infty} \mathrm{R}f(t,\omega)h(t,\omega)dtd\omega \\ &= \int_0^{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(t\omega + s\hat{\omega})h(t,\omega)dsdtd\omega\end{aligned} \tag{4.24}$$

Let $(x, y) = t\omega + s\hat{\omega}$ so that

$$t = \langle (x, y), \omega\rangle,$$

interchanging the ω and the xy-integrals we obtain

$$\begin{aligned}\langle \mathrm{R}f, h\rangle_{\mathbb{R}\times S^1} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_0^{2\pi} f(x,y)h(\langle (x,y),\omega\rangle,\omega)d\omega dxdy \\ &= \langle f, \mathrm{R}^*h\rangle_{\mathbb{R}^2}.\end{aligned} \tag{4.25}$$

This verifies the assertion that backprojection is $[4\pi]^{-1}$ times the formal adjoint of the Radon transform. The fact that $\mathrm{R}^* \neq \mathrm{R}^{-1}$ is reflection of the fact that R is not a unitary transformation from $L^2(\mathbb{R}^2)$ to $L^2(\mathbb{R} \times S^1)$.

Using the identification of backprojection with the adjoint, along with the Parseval formula, (3.3.3) we can derive an interesting relationship between $\widehat{\mathrm{R}^*\,\mathrm{R}f}$ and $\hat{f}$.

Proposition 4.2.1. *Suppose that f is an absolutely integrable and square integrable function in the natural domain of the Radon transform then*

$$\frac{r}{4\pi}\widehat{\mathrm{R}^*\,\mathrm{R}f}(r\omega) = \hat{f}(r\omega). \tag{4.26}$$

Proof. The proof of this proposition uses the basic principle that, in an inner product space, $(X, \langle\cdot,\cdot\rangle_X)$, an element $\mathbf{x}$ is zero if and only if $\langle \mathbf{x}, \mathbf{y}\rangle = 0$ for all $\mathbf{y}$ belonging to a dense subset of X. Let f and g be two functions satisfying the hypotheses of the proposition. From the definition of the adjoint it follows that

$$\langle \mathrm{R}f, \mathrm{R}g\rangle_{\mathbb{R}\times S^1} = \langle f, \mathrm{R}^*\,\mathrm{R}g\rangle_{\mathbb{R}^2}. \tag{4.27}$$

Using the Parseval formula we get the relations

$$\begin{aligned}\langle f, \mathrm{R}^*\,\mathrm{R}g\rangle_{\mathbb{R}^2} &= \frac{1}{[2\pi]^2}\langle \hat{f}, \widehat{\mathrm{R}^*\,\mathrm{R}g}\rangle_{\mathbb{R}^2} \\ &= \frac{1}{[2\pi]^2}\int_0^{2\pi}\int_0^{\infty} \hat{f}(r\omega)\overline{\widehat{\mathrm{R}^*\,\mathrm{R}g}}(r\omega)r\,dr\,d\omega,\end{aligned} \tag{4.28}$$

and

$$\begin{aligned}\langle \mathrm{R}f, \mathrm{R}g\rangle_{\mathbb{R}\times S^1} &= \frac{1}{2\pi}\int_0^{2\pi}\int_{-\infty}^{\infty} \widetilde{\mathrm{R}f}(r,\omega)\overline{\widetilde{\mathrm{R}g}}(r,\omega)\,dr\,d\omega \\ &= \frac{1}{\pi}\int_0^{2\pi}\int_0^{\infty} \hat{f}(r\omega)\overline{\hat{g}(r\omega)}\,dr\,d\omega.\end{aligned} \tag{4.29}$$

In the last line we use the central slice theorem and the evenness of the Radon transform. Since these formulæ hold for all f and g with bounded support, a dense subset of L^2, it follows that

$$\frac{r}{4\pi}\widehat{\mathrm{R}^*\,\mathrm{R}g}(r\omega) = \hat{g}(r\omega). \tag{4.30}$$

□

This result is used in section 4.2.6 to derive another useful expression for R^{-1}.

Exercise 4.2.3. Let g be a continuous function with bounded support on $\mathbb{R}\times S^1$. Show that there is a constant C so that

$$|\,\mathrm{R}^*g(x,y)| \le \frac{C}{1+|x|+|y|}.$$

Show that if g is a non-negative function which is not identically zero then there is also a constant $C' > 0$ so that

$$|\,\mathrm{R}^*g(x,y)| \ge \frac{C'}{1+|x|+|y|}.$$

Exercise 4.2.4. Explain how we arrived at the limits of integration in the second line of (4.29).

4.2.4 Filtered Backprojection

We now turn our attention to understanding the inversion formula for R. It can be understood as a two step process:

(1). The radial integral is interpreted as a *filter* applied to the Radon transform. The filter acts only in the affine parameter, the output of the filter is denoted by

$$\mathcal{G}\,\mathrm{R}f(t,\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{\mathrm{R}f}(r,\omega)e^{irt}|r|dr. \tag{4.31}$$

(2). The angular integral is then interpreted as the backprojection of the filtered Radon transform. The function f is expressed as

$$f(x,y) = \frac{1}{2\pi}\int_0^{\pi} (\mathcal{G}\,\mathrm{R})f(\langle(x,y),\omega\rangle,\omega)d\omega = \frac{1}{4\pi}\,\mathrm{R}^*\mathcal{G}\,\mathrm{R}f. \tag{4.32}$$

For this reason the Radon inversion formula is often called the *filtered backprojection formula*.

Backprojection is both conceptually and computationally simple, whereas the filtering step requires a more careful analysis. The filter step itself is comprised of two operations. Recall that the Fourier transform of the derivative of a function $g(t)$ is equal to the Fourier transform of g multiplied by $i\xi$: $\widehat{\partial_t g}(\xi) = (i\xi)\hat{g}(\xi)$. If, in the inversion formula (4.20) we had r instead of $|r|$ then $f(x,y)$ would equal

$$\frac{1}{2\pi i}\int_0^{\pi} \partial_t\,\mathrm{R}f(\langle(x,y),\omega\rangle,\omega)d\omega;$$

This is, backprojection of the t-derivative of $\mathrm{R}f$; notice that if f is real valued then this function is purely imaginary! Because differentiation is a *local operation* this is a relatively easy formula to understand. The subtlety in (4.20) therefore stems from the fact that $|r|$ appears and not r itself.

We define another operation on functions of a single variable which is called the *Hilbert transform*, it is defined in terms of the Fourier transform by the formula

$$\mathcal{H}g(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{g}(r)\,\mathrm{sign}(r)e^{itr}dr. \tag{4.33}$$

Recall that

$$\mathrm{sign}(r) = \begin{cases} 1 \text{ if } & r > 0, \\ -1 \text{ if } & r \leq 0. \end{cases}$$

In other words the Hilbert transform of g is the function whose Fourier transform is $\mathrm{sign}(r)\hat{g}(r)$. For any given r_0, the computation of $\mathcal{H}g(r_0)$ requires a knowledge of $g(r)$

for *all* values of r. In other words, the Hilbert transform is not a *local* operation. Conceptually, the Hilbert transform is the most difficult part of the Radon inversion formula. On the other hand, the Hilbert transform has a very simple expression in terms of the Fourier transform and this makes it easy to implement.

We now compute a couple of examples of Hilbert transforms.

Example 4.2.1. Let

$$f(x) = \frac{\sin(x)}{\pi x},$$

its Fourier transform is

$$\hat{f}(\xi) = \chi_{[-1,1]}(\xi) = \begin{cases} 1 \text{ if } & |\xi| \leq 1, \\ 0 \text{ if } & |\xi| > 1. \end{cases}$$

The Hilbert transform of f is expressed as a Fourier integral by

$$\begin{aligned} \mathcal{H}\left(\frac{\sin(x)}{\pi x}\right) =& \frac{1}{2\pi}\left[\int_0^1 e^{ix\xi}dx - \int_{-1}^0 e^{ix\xi}dx\right] \\ =& i\frac{1-\cos(x)}{\pi x}. \end{aligned} \tag{4.34}$$

Example 4.2.2. The next example is of interest in medical imaging. It is difficult to do this example by a direct calculation. A method to do this calculation, using functions of a complex variable is explained in the final section of this chapter. Let

$$f(x) = \begin{cases} \sqrt{1-x^2} & \text{for } |x| < 1, \\ 0 & \text{for } |x| \geq 1. \end{cases}$$

The Hilbert transform of f is given by

$$\mathcal{H}(f) = \begin{cases} ix & \text{for } |x| < 1, \\ i(x + \sqrt{x^2-1}) & \text{for } x < -1, \\ i(x - \sqrt{x^2-1}) & \text{for } x > 1. \end{cases} \tag{4.35}$$

Notice the very different character of $\mathcal{H}f(x)$ for $|x| < 1$ and $|x| > 1$. For $|x| < 1$, $\mathcal{H}f(x)$ is a smooth function with a bounded derivative. Approaching ± 1 from the set $|x| > 1$, the derivative of $\mathcal{H}f(x)$ blows up.

From the differentiation formula for the Fourier transform we conclude that

$$\widetilde{\partial_t \mathrm{R}f}(r) = ir\widetilde{\mathrm{R}f}(r).$$

The Hilbert transform of $\partial_t \mathrm{R}f$ is given by

$$\begin{aligned} \mathcal{H}(\partial_t \mathrm{R}f)(t,\omega) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{\partial_t \mathrm{R}f}(r,\omega)\,\mathrm{sign}(r)e^{itr}dr \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} i|r|\widetilde{\mathrm{R}f}(r,\omega)e^{irt}dr. \end{aligned}$$

Since $\text{sign}(r)r = |r|$ we can identify $\mathcal{G}\,\mathrm{R}f$ as

$$\mathcal{G}\,\mathrm{R}f(t,\omega) = \frac{1}{i}\mathcal{H}(\partial_t\,\mathrm{R}f)(t,\omega). \tag{4.36}$$

Putting this into (4.32) we obtain

$$f(x,y) = \frac{1}{2\pi i}\int_0^\pi \mathcal{H}(\partial_t\,\mathrm{R}f)(\langle(x,y),\omega\rangle,\omega)d\omega \tag{4.37}$$

The function f is reconstructed by backprojecting the Hilbert transform of $\frac{1}{i}\partial_t\,\mathrm{R}f$.

Remark 4.2.4. * The Fourier transform of the function

$$F = \frac{1}{2}(f + \mathcal{H}f)$$

vanishes for $\xi < 0$ and therefore F an analytic extension to the upper half plane, see Theorem 3.2.9. This explains why the Hilbert transform is intimately connected to the theory of analytic functions. Using the Fourier representation, it is easy to see that $\hat{F}(\xi) = \chi_{[0,\infty)}(\xi)\hat{f}(\xi)$ and therefore, if $y > 0$ then

$$F(x+iy) = \frac{1}{2\pi}\int_0^\infty \hat{f}(\xi)e^{-y\xi}e^{ix\xi}d\xi$$

is an absolutely convergent integral. The function $F(x)$ is the boundary value of a analytic function. A basic theorem in function theory states that an analytic function cannot vanish on an open interval, see [52]. This shows that if f has bounded support then $\mathcal{H}f$ cannot. For more on the connection between the Hilbert transform and analytic function theory see section 4.8.

This observation has important implications in image reconstruction. Formula (4.37) expresses f as the backprojection of $-i\mathcal{H}\partial_t\,\mathrm{R}f$. If f has bounded support then so does $\partial_t\,\mathrm{R}f$ and therefore $-i\mathcal{H}\,\mathrm{R}\partial_t f$ *does not*. Suppose (x,y) lies outside the support of f, this means that the integrand in (4.37) is, generally speaking, not zero. The integral vanishes due to subtle cancelations between the positive and negative parts of $-i\mathcal{H}(\partial_t\,\mathrm{R}f)(\langle(x,y),\omega\rangle,\omega)$. We return to this question in section 8.6.3.

Exercise 4.2.5. * Use the Schwarz reflection principle to prove the statement that if $F(x+iy)$ is an analytic function in $y > 0$ such that, for $a < x < b$,

$$\lim_{y\downarrow 0} F(x+iy) = 0$$

then $F \equiv 0$.

4.2.5 Inverting the Radon transform, two examples

Before continuing our analysis of R^{-1} we compute the inverse of the Radon transform in two examples.

Example 4.2.3. In the first example f is the characteristic function on the unit disk. It is defined by

$$f(x,y) = \begin{cases} 1 & \|(x,y)\| \leq 1 \\ 0 & \|(x,y)\| > 1. \end{cases}$$

Using the rotational symmetry, we can check that

$$\mathrm{R}f(t,\omega) = \begin{cases} 2\sqrt{1-t^2} & \|(x,y)\| \leq 1 \\ 0 & \|(x,y)\| > 1. \end{cases} \tag{4.38}$$

Note that $\mathrm{R}f$ satisfies

$$\limsup_{h,t} \left| \frac{\mathrm{R}f(t+h) - \mathrm{R}f(t)}{\sqrt{|h|}} \right| < \infty.$$

In other words $\mathrm{R}f(t,\omega)$ is a Hölder-$\frac{1}{2}$ function.

To apply the filtered backprojection formula we need to compute either $\partial_t \mathcal{H}\,\mathrm{R}f$ or $\mathcal{H}\partial_t\,\mathrm{R}f$. It is instructive to do both. In section 4.8 it is shown that

$$\frac{1}{i}\mathcal{H}\,\mathrm{R}f(t,\omega) = \begin{cases} 2t & \text{for } |t| < 1, \\ 2(t+\sqrt{t^2-1}) & \text{for } t < -1, \\ 2(t-\sqrt{t^2-1}) & \text{for } x > 1. \end{cases} \tag{4.39}$$

Even though this function is not differentiable at $t = \pm 1$, it does have a weak derivative given by

$$\frac{1}{i}\partial_t\mathcal{H}\,\mathrm{R}f(t,\omega) = \begin{cases} 2 - \frac{2|t|}{\sqrt{t^2-1}} & \text{for } |t| \geq 1 \\ 2 & \text{for } |t| < 1. \end{cases} \tag{4.40}$$

This function is absolutely integrable.

On the other hand we could first compute the derivative of $\mathrm{R}f$:

$$\partial_t\,\mathrm{R}f(t,\omega) = \begin{cases} \frac{-2t}{\sqrt{1-t^2}} & |(x,y)| < 1 \\ 0 & |(x,y)| > 1. \end{cases}$$

This derivative blows up at $t = 1$ and so again, the derivative needs to be interpreted as a weak derivative. Unfortunately this function does not belong to $L^2(\mathbb{R})$. Thus far we have only defined the Hilbert transform for L^2-functions. It is also possible to define the Hilbert transform of a function in $L^p(\mathbb{R})$ for any $1 < p \leq 2$, see [40]. As

$$\int |\partial_t\,\mathrm{R}f|^p < \infty \text{ for } p < 2$$

the Hilbert transform of $\partial_t\,\mathrm{R}f$ is still defined and can be computed using the complex variable method described in section 4.8. It is given by the formula

$$\frac{1}{i}\mathcal{H}(\partial_t\,\mathrm{R}f)(t) = \begin{cases} 2 - \frac{2|t|}{\sqrt{t^2-1}} & \text{for } |t| \geq 1 \\ 2 & \text{for } |t| < 1. \end{cases} \tag{4.41}$$

This computation is described in greater detail in appendix 4.8.

Now we do the backprojection step. If (x, y) is inside the unit disc then

$$|\langle (x, y), \omega \rangle| \leq 1.$$

At such points, the inverse of the Radon transform is quite easy to compute:

$$\frac{1}{2\pi i} \int_0^\pi \mathcal{H}(\partial_t \, \mathrm{R} f)(\langle (x, y), \omega \rangle, \omega) d\omega = \frac{1}{2\pi} \int_0^\pi 2 d\omega = 1.$$

This is precisely the value of $f(x, y)$ for $\|(x, y)\| \leq 1$. On the other hand, if $\|(x, y)\| > 1$, then the needed calculation is more complicated. Since f is radially symmetric it suffices to consider $f(x, 0)$. If $x > 1$ then there is an angle $0 < \theta_x < \frac{\pi}{2}$ so that $x \cos \theta_x = 1$, the inversion formula can be written

$$f(x, 0) = \frac{1}{2\pi} \left[4 \int_0^{\theta_x} d\theta - 2 \int_{\theta_x}^{\pi - \theta_x} \left(1 - \frac{|r \cos \theta|}{\sqrt{r^2 \cos^2 \theta - 1}} \right) d\theta \right].$$

This is a much more complicated formula. From the point of view of computation it is notable that, for $|x| > 1$, the Radon inversion formula involves an *divergent* integrand. It is of course absolutely integrable, but this nonetheless leads to significant *numerical* difficulties.

The important lesson of this example is the qualitative difference in the backprojection formula between points inside the unit disk and for points outside. This fact has significant consequences in medical imaging, see section 8.6.3.

Example 4.2.4. Our next example is a bit smoother than the characteristic function of the disk. Let $r = \sqrt{x^2 + y^2}$ and define g by

$$g(x, y) = \begin{cases} 1 - r^2 & |r| < 1, \\ 0 & |r| \geq 1. \end{cases}$$

Again using the rotational symmetry, we easily obtain

$$\mathrm{R} g(t, \omega) = \begin{cases} \frac{4}{3}(1 - t^2)^{3/2} & |t| \leq 1, \\ 0 & |t| > 1. \end{cases}$$

This function $\mathrm{R}g$ is classically differentiable, the derivative of $\mathrm{R}g$ is

$$\partial_t \, \mathrm{R} g(t, \omega) = \begin{cases} -4t(1 - t^2)^{1/2} & |t| \leq 1, \\ 0 & |t| > 1. \end{cases}$$

It satisfies

$$\limsup_{h,t} \left| \frac{\partial_t \, \mathrm{R} g(t + h) - \partial_t \, \mathrm{R} g(t)}{|h|^{\frac{1}{2}}} \right| < \infty.$$

This time $\partial_t \mathrm{R}g$ is a Hölder-$\frac{1}{2}$ function. This is a "half" a derivative smoother than g itself. It is a general fact that the Radon transform has better regularity in the affine parameter than the original function by half a derivative. The Hilbert transform of $\partial_t \mathrm{R}g$ is

$$\frac{1}{i}\mathcal{H}(\partial_t \mathrm{R}g)(t) = \begin{cases} 2 - 4t^2 & |t| \leq 1, \\ 4[\ 4|t|(t^2-1)^{1/2} - (2t^2-1)\ ] & |t| > 1. \end{cases}$$

Once again we see that the backprojection formula for points inside the unit disk is, numerically a bit simpler than for points outside. While $\sqrt{t^2-1}$ is continuous, it is not differentiable at $t = \pm 1$. This makes the numerical integration in the backprojection step more difficult for points outside the disk.

Exercise 4.2.6. Prove that (4.40) gives the weak derivative of $\mathcal{H}\,\mathrm{R}f$ defined in (4.39).

Exercise 4.2.7. Used Simpson's rule to numerically integrate $\sqrt{1-t^2}$ from 0 to 1. Determine how the accuracy of the result depends on the mesh size and compare it to the accuracy when instead, $1 - t^2$ is integrated.

4.2.6 An alternate formula for the Radon inverse*

Proposition 4.2.1 leads to an alternate formula for R^{-1}. In this approach, the backprojection is done first and a filter is applied to $\mathrm{R}^*\,\mathrm{R}f$. If f is a piecewise continuous function of bounded support then Proposition 4.2.1 states that

$$\hat{f}(r\omega) = \frac{r}{4\pi}\widehat{\mathrm{R}^*\,\mathrm{R}f}(r\omega).$$

If $\hat{f}$ is absolutely integrable then the Fourier inversion formula therefore implies that

$$\begin{aligned} f(x,y) &= \frac{1}{[2\pi]^2}\iint\limits_{\mathbb{R}^2} \frac{r}{4\pi}\widehat{\mathrm{R}^*\,\mathrm{R}f}(r\omega)e^{i\langle r\omega,(x,y)\rangle} r\,dr\,d\omega \\ &= \frac{1}{[2\pi]^2}\iint\limits_{\mathbb{R}^2} \frac{\|\boldsymbol{\xi}\|}{4\pi}\widehat{\mathrm{R}^*\,\mathrm{R}f}(\boldsymbol{\xi})e^{i\langle \boldsymbol{\xi},(x,y)\rangle} d\boldsymbol{\xi}. \end{aligned} \tag{4.42}$$

The Laplace operator on $\mathbb{R}^2$ is defined as the second order differential operator

$$\Delta f = -(\partial_x^2 f + \partial_y^2 f).$$

As a constant coefficient differential operator it can be expressed in terms of the Fourier transform by

$$\Delta f(x,y) = \frac{1}{[2\pi]^2}\int\limits_{\mathbb{R}^2} \|\boldsymbol{\xi}\|^2 \hat{f}(\boldsymbol{\xi})e^{i\langle \boldsymbol{\xi},(x,y)\rangle} d\boldsymbol{\xi}.$$

This formula motivates a definition for the non-negative powers of the Laplace operator. For $s \geq 0$ and f, a function with $2s$ L^2-derivatives define

$$\Delta^s f(x,y) = \frac{1}{[2\pi]^2}\int\limits_{\mathbb{R}^2} \|\boldsymbol{\xi}\|^{2s} \hat{f}(\boldsymbol{\xi})e^{i\langle \boldsymbol{\xi},(x,y)\rangle} d\boldsymbol{\xi}. \tag{4.43}$$

With this definition we can re-write (4.42) as

$$4\pi f(x,y) = \Delta^{\frac{1}{2}}\,\mathrm{R}^*(\mathrm{R}f)(x,y). \tag{4.44}$$

Remark 4.2.5. Note that $\Delta\,\mathrm{R}^*(\mathrm{R}f)(x,y) = 4\pi\Delta^{\frac{1}{2}}f$. This gives an expression for $\Delta^{\frac{1}{2}}f$ which, given $\mathrm{R}f$ can be computed using entirely elementary operations. The functions f and $\Delta^{\frac{1}{2}}f$ have the same singularities. As edges are discontinuities, this formula gives a straightforward way to find the edges in an image described by a density function f. I thank Gunther Uhlmann for this observation.

Remark 4.2.6. Thus far we have produced a left inverse for the Radon transform. If f is a function in the plane satisfying appropriate regularity and decay hypotheses then, for example,

$$f = (\Delta^{\frac{1}{2}}\,\mathrm{R}^*)\,\mathrm{R}f.$$

We have **not** said that if $h(t,\omega)$ is an even function on $\mathbb{R}\times S^1$ then

$$h = \mathrm{R}(\Delta^{\frac{1}{2}}\,\mathrm{R}^*)h.$$

That is we have not shown that $(\Delta^{\frac{1}{2}}\,\mathrm{R}^*)$ is also a right inverse for R. Under some mild hypotheses on h this is in fact true. The proof of this statement involves characterizing the range of the Radon transform and is discussed further in section 4.5.

Exercise 4.2.8. Using the definition, (4.43) show that

(1). If s is a positive integer then the two definitions of Δ^s agree.

(2). For s and t non-negative numbers, show that

$$\Delta^s\Delta^t = \Delta^{s+t}. \tag{4.45}$$

(3). Conclude from the previous part that

$$\Delta\,\mathrm{R}^*\,\mathrm{R}f = \Delta^{\frac{1}{2}}f.$$

4.3 The Hilbert transform

See: A.4.6, A.5.

To implement the inversion formula for the Radon transform one needs perform the filter operation, in this section we further analyze the Hilbert transform. As above, $\mathcal{F}^{-1}$ denotes the inverse of the Fourier transform, i.e., in one-dimension

$$\mathcal{F}^{-1}(\hat{f}) = \frac{1}{2\pi}\int \hat{f}(\xi)e^{it\xi}d\xi = f(t).$$

The Hilbert transform is defined by

$$\mathcal{H}f = \mathcal{F}^{-1}(\hat{f}(\xi)\,\mathrm{sign}(\xi)) \Rightarrow \widehat{\mathcal{H}f}(\xi) = \mathrm{sign}(\xi)\hat{f}(\xi).$$

The Fourier transform of a convolution is the product of their Fourier transforms, that is

$$\mathcal{F}^{-1}(\hat{f}\hat{g}) = f * g.$$

Hence, if there existed a nice function h such that $\hat{h}(\xi) = \mathrm{sign}(\xi)$, then the Hilbert transform would be just $h*f$. Unfortunately the $\mathrm{sign}(\xi)$ is not the Fourier transform of a nice function because it does not go to zero as $|\xi| \to \infty$. Approximating $\mathrm{sign}(\xi)$ by a function which decays at ∞ gives approximations to the Hilbert transform expressible as convolutions with nice functions.

Modify the signum function by setting

$$\widehat{h_\epsilon}(\xi) := \mathrm{sign}(\xi)e^{-\epsilon|\xi|} \text{ for } \epsilon > 0.$$

The inverse Fourier transform of $\widehat{h_\epsilon}$ is

$$h_\epsilon = \frac{i}{\pi}\frac{t}{t^2+\epsilon^2}.$$

This function behaves like $1/t$ as t goes to infinity which is not fast enough for integrability but at least it goes to zero and has no singularities. Most of the functions encountered in medical imaging have bounded supported and therefore the integrals $h_\epsilon * f$ converge absolutely. For each $\epsilon > 0$ define an approximate Hilbert transform

$$\mathcal{H}_\epsilon f = \mathcal{F}^{-1}(\hat{f}\widehat{h_\epsilon}) = f * h_\epsilon.$$

Letting $\epsilon \downarrow 0$ we see that $h_\epsilon \to i/(t\pi)$. Formally this seems to imply that

$$\mathcal{H}f(t) = \frac{i}{\pi}\int_{-\infty}^{\infty}\frac{f(s)ds}{t-s}.$$

Because $1/|t|$ is not integrable in any neighborhood of 0, this expression is not an absolutely convergent integral. In this instance, the correct interpretation for this formula is as a Cauchy Principal Value:

$$\mathcal{H}f(t) = \frac{i}{\pi}\,\mathrm{P.V.}(f * \frac{1}{t}) = \frac{i}{\pi}\lim_{\epsilon\to 0}\left[\int_{-\infty}^{-\epsilon} + \int_{\epsilon}^{\infty}\frac{f(t-s)}{s}ds\right]. \tag{4.46}$$

This limit is easily seen to be finite, at least if f has bounded support and is once differentiable. Since the function $1/s$ is odd and the interval on which we are integrating is symmetric, we have

$$\left(\int_{-R}^{-\epsilon} + \int_{\epsilon}^{R}\right)\frac{ds}{s} = 0.$$

We can multiply this by $f(t)$ and still get zero:

$$\left(\int_{-R}^{-\epsilon} + \int_{\epsilon}^{R}\right) f(t)\frac{ds}{s} = 0.$$

If we assume that the support of $f(t)$ is contained in $[-\frac{R}{2}, \frac{R}{2}]$ then subtracting this from the above integral we obtain

$$\mathcal{H}f(t) = \frac{i}{\pi}\lim_{\epsilon\to 0}\left(\int_{-R}^{-\epsilon} + \int_{\epsilon}^{R}\right)\frac{f(t-s)-f(t)}{s}ds. \tag{4.47}$$

If f is once differentiable then the integrand in (4.47) remains bounded as $\epsilon \to 0$. If, for some $\alpha > 0$, f satisfies an α-Hölder-condition,

$$\frac{|f(t)-f(s)|}{|t-s|^{\alpha}} \le M, \tag{4.48}$$

then $\mathcal{H}f$ is given by the absolutely convergent integral

$$\mathcal{H}f(t) = \frac{i}{\pi}\int_{-R}^{R}\frac{f(t-s)-f(t)}{s}ds. \tag{4.49}$$

By this process, we have replaced the improper integral, (4.46) with a bounded integral. The cancelation due to the symmetric interval used in the definition of the principal value is critical to obtain a bounded result.

There are other ways to regularize convolution with $1/t$ For example, we could add an imaginary number to the denominator to make it non-vanishing,

$$\lim_{\epsilon\downarrow 0}\frac{i}{\pi}\int_{-R}^{R}\frac{f(t-s)}{s\pm i\epsilon}ds.$$

A computation shows that

$$\frac{i}{\pi}\frac{1}{2}\left(\frac{1}{s+i\epsilon}+\frac{1}{s-i\epsilon}\right) = \frac{i}{\pi}\frac{s}{s^2+\epsilon^2} = h_{\epsilon}(s).$$

This shows that the average of the two regularizations, $(s\pm i\epsilon)^{-1}$ results in the same approximation as before. The difference of these two regularizations is

$$\frac{i}{\pi}\cdot\frac{1}{2}\left(\frac{1}{s+i\epsilon}-\frac{1}{s-i\epsilon}\right) = \frac{1}{\pi}\frac{\epsilon}{s^2+\epsilon^2}$$

which does not tend to zero as ϵ tends to zero. As an example we "test" the characteristic function of the interval $\chi_{[-1,1]}$ by evaluating the limit at $t = 0$,

$$\lim_{\epsilon\downarrow 0}\int_{-\infty}^{\infty}\chi_{[-1,1]}(-s)\frac{\epsilon}{s^2+\epsilon^2}ds = \lim_{\epsilon\downarrow 0}\int_{-1}^{1}\frac{\epsilon}{s^2+\epsilon^2}ds = \lim_{\epsilon\downarrow 0}\int_{-1/\epsilon}^{1/\epsilon}\frac{dt}{t^2+1} = \pi.$$

So we see that in general

$$\lim_{\epsilon \downarrow 0} \frac{i}{\pi} \int_{-\infty}^{\infty} \frac{f(t-s)ds}{s \pm i\epsilon} \neq \mathcal{H}f(t).$$

The lesson is that care must be exercised in choosing a regularization for convolution with $1/t$. Different regularizations may lead to different results.

Things are less delicate using the Fourier representation.

Theorem 4.3.1. *Suppose that $\phi_\epsilon(\xi)$ is a uniformly bounded family of functions which converges pointwise to* $\operatorname{sign}(\xi)$ *as $\epsilon \to 0$. If f and $\hat{f}$ are square integrable then the Hilbert transform of f is given by the limit*

$$\mathcal{H}f(t) = \lim_{\epsilon \downarrow 0} \int \phi_\epsilon(\xi)\hat{f}(\xi)e^{it\xi}\frac{d\xi}{2\pi}.$$

Proof. Since $\hat{f}(\xi)$ is absolutely integrable and $\phi_\epsilon(\xi)$ is uniformly bounded, the conclusion follows from the Lebesgue dominated convergence theorem, see [16]. □

The function $\hat{h}_\epsilon$ satisfies the hypotheses of the theorem. Another important example of a regularization is given by ϕ_L defined by

$$\phi_L(\xi) = \begin{cases} -1 & -L \leq \xi \leq 0, \\ 0 & 0 < \xi \leq L, \\ 1 & \text{otherwise.} \end{cases}$$

Computing the inverse Fourier transform of ϕ_L we obtain a different sequence of kernels which approximately compute the Hilbert transform.

Remark 4.3.1. This discussion shows that there are several different philosophies for approximating the Hilbert transform and therefore the Radon inversion formula. On the one hand we can use the convolution formula for $\mathcal{H}$ and directly approximate $\mathrm{P.V.}(f * \frac{1}{t})$ by $\mathcal{H}_\epsilon$. On the other hand we can use the Fourier integral representation and instead approximate $\operatorname{sign}(\xi)$ as described in Theorem 4.3.1. For sufficiently smooth functions with bounded support we could use (4.49). Mathematically these approaches are equivalent. Computationally they can lead to vastly different results. In a real application one chooses an approximation by considering the competing requirements of resolution, noise reduction and computational efficiency.

Exercise 4.3.1. Suppose that f has bounded support and satisfies an α-Hölder condition for an $0 < \alpha \leq 1$. Show that

$$\lim_{\epsilon \downarrow 0} h_\epsilon * f = \frac{i}{\pi}\,\mathrm{P.V.}(f * \frac{1}{t}).$$

Exercise 4.3.2. Suppose that f and g are continuous functions with bounded support. Show that

$$\mathcal{H}(f * g) = (\mathcal{H}f) * g = f * (\mathcal{H}g).$$

Exercise 4.3.3. Below are linear operators defined in terms of the Fourier transform. Re-express these operators in terms of differentiations and the Hilbert transform. For example, if Af is defined by

$$Af(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \xi \hat{f}(\xi) e^{ix\xi} d\xi$$

then the answer to this question is

$$Af(x) = -i\partial_x f(x).$$

Do not worry about convergence.

(1).

$$A_1 f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\xi|^3 \hat{f}(\xi) e^{ix\xi} d\xi$$

(2).

$$A_2 f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (\xi^4 + |\xi| + 1) \hat{f}(\xi) e^{ix\xi} d\xi$$

(3). In this problem take note of the lower limit of integration.

$$A_3 f(x) = \frac{1}{2\pi} \int_{0}^{\infty} \hat{f}(\xi) e^{ix\xi} d\xi$$

Exercise 4.3.4. This exercise addresses the "spectral theory" of the Hilbert transform.

(1). Which real numbers are eigenvalues of the Hilbert transform? That is, for which real numbers λ does there exist a function f_λ in $L^2(\mathbb{R})$ so that

$$\mathcal{H}f = \lambda f?$$

Hint: Use the Fourier transform.

(2). Can you describe the eigenspaces? That is if λ is an eigenvalue of $\mathcal{H}$ describe the set of all functions in $L^2(\mathbb{R})$ which satisfy

$$\mathcal{H}f = \lambda f.$$

(3). Show that $\mathcal{H} \circ \mathcal{H} f = \mathcal{H}(\mathcal{H}(f)) = f$ for any $f \in L^2(\mathbb{R})$.

4.3.1 Mapping properties of the Hilbert transform*

See: A.4.1.

The Hilbert transform has very good mapping properties with respect to most function spaces. Using the Parseval formula one easily establishes the L^2-result.

Proposition 4.3.1. *If $f \in L^2(\mathbb{R})$ then $\mathcal{H}f \in L^2(\mathbb{R})$ and in fact*

$$\|f\|_{L^2} = \|\mathcal{H}f\|_{L^2}.$$

The Hilbert transform also has good mapping properties on other L^p-spaces as well as Hölder spaces, though the proofs of these results requires more effort.

Proposition 4.3.2. *For each $1 < p < \infty$ the Hilbert transform extends to define a bounded map $\mathcal{H} : L^p(\mathbb{R}) \to L^p(\mathbb{R})$.*

Proposition 4.3.3. *Suppose that f is α-Hölder continuous for an $\alpha \in (0, 1)$ and vanishes outside a bounded interval then $\mathcal{H}f$ is also α-Hölder continuous.*

Notice that the case of $\alpha = 1$ is excluded in this proposition. The result is false in this case. There exist differentiable functions f such that $\mathcal{H}f$ is not even 1-Hölder continuous. Proofs of these propositions can be found in [66].

Exercise 4.3.5. By using formula (4.49), which is valid for a Hölder continuous function vanishing outside a bounded interval, prove Proposition 4.3.3.

Exercise 4.3.6. If $f \in L^2(\mathbb{R})$ then show that

$$\int_{-\infty}^{\infty} |f(x)|^2 dx = \int_{-\infty}^{\infty} |\mathcal{H}f(x)|^2 dx.$$

4.4 Approximate inverses for the Radon transform

To exactly invert the Radon transform we need to compute the Hilbert transform of a derivative. The measured data is a function, $g_m(t, \omega)$ on the space of lines. Measured data is rarely differentiable and the exact Radon inverse entails the computation of $\partial_t g_m$. Indeed the Parseval formula, (4.15) implies that unless g_m has a half an L^2-derivative then it is not the Radon transform of an L^2-function defined in $\mathbb{R}^2$. Thus it important to investigate how to approximate the inverse of the Radon transform in a way that is usable with realistic data. The various approaches to approximating the Hilbert transform lead to different approaches to approximating the Radon inverse. Because the approximate inverses involve some sort of smoothing, they are often called *regularized inverses*.

Recall that a convolution has the following useful properties with respect to derivatives:

$$\partial_x(f * g) = \partial_x f * g = f * \partial_x g.$$

Using formula (4.37) we get an approximate inverse for the Radon transform

$$\begin{aligned} f(x,y) \approx & \frac{1}{2\pi i}\int_0^\pi \mathcal{H}_\epsilon(\partial_t \mathrm{R}f)(\langle(x,y),\omega\rangle,\omega)d\omega \\ = & \frac{1}{2\pi i}\int_0^\pi h_\epsilon * (\partial_t \mathrm{R}f)(\langle(x,y),\omega\rangle,\omega)d\omega \end{aligned} \tag{4.50}$$

Using the the formula for h_ϵ and the fact that $f * \partial_t g = \partial_t f * g$ we get

$$\begin{aligned} f(x,y) \approx & \frac{1}{2\pi i}\int_0^\pi \int_{-\infty}^\infty \mathrm{R}f(s)\partial_t h_\epsilon(t-s)ds \\ = & \frac{1}{2\pi^2}\int_0^\pi \int_{-\infty}^\infty \left[\mathrm{R}f(s,\omega)\frac{\epsilon^2-(t-s)^2}{(\epsilon^2+(t-s)^2)^2}ds\Big|_{s=\langle(x,y),\omega\rangle}\right]d\omega. \end{aligned} \tag{4.51}$$

The expression in (4.51) has an important practical advantage: we have moved the t-derivative from the potentially noisy measurement $\mathrm{R}f(t,\omega)$ over to the smooth, exactly known function h_ϵ. This means that we do not have to approximate the derivatives of $\mathrm{R}f(t,\omega)$.

In most applications convolution operators, such as derivatives and the Hilbert transform are computed using the Fourier representation. Theorem 4.3.1 suggests approximating the filtering step, (4.31) in the exact inversion formula by cutting off the high frequency components. Let $\hat{\psi}(r)$ be a bounded, even function, satisfying the conditions

$$\begin{aligned} &\hat{\psi}(0) = 1, \\ &\hat{\psi}(r) = 0 \text{ for } |r| > W. \end{aligned} \tag{4.52}$$

For l a function on $\mathbb{R}\times S^1$ define

$$\mathcal{G}_\psi(l)(t,\omega) = \frac{1}{2\pi}\int_{-\infty}^\infty \widetilde{l}(r,\omega)e^{irt}\hat{\psi}(r)|r|dr, \tag{4.53}$$

and

$$\mathrm{R}_\psi^{-1}l(x,y) = \frac{1}{2\pi}\int_0^\pi \mathcal{G}_\psi(l)(\langle(x,y),\omega\rangle,\omega)d\omega = \frac{1}{4\pi}\mathrm{R}^*\mathcal{G}_\psi l. \tag{4.54}$$

For notational convenience let

$$f_\psi = \mathrm{R}_\psi^{-1}\circ \mathrm{R}f.$$

How is $\mathrm{R}_\psi^{-1}f$ related to f? The answer to this question is surprisingly simple. The starting point for our analysis is Proposition 4.1.1 which says that if f and g are functions on $\mathbb{R}^2$ then

$$\mathrm{R}f * g(t,\omega) = \int_{-\infty}^\infty \mathrm{R}f(t-\tau,\omega)\,\mathrm{R}g(\tau,\omega)d\tau.$$

Using the convolution theorem for the Fourier transform we see that

$$\widetilde{\mathrm{R}f * g}(r,\omega) = \widetilde{\mathrm{R}f}(r,\omega)\widetilde{\mathrm{R}g}(r,\omega).$$

Suppose now that g is a radial function so that $\mathrm{R}g$ is independent of ω. The filtered backprojection formula for $f * g$ reads

$$f * g(x,y) = \frac{1}{4\pi^2}\int_0^{\pi}\int_{-\infty}^{\infty} \widetilde{\mathrm{R}f}(r,\omega)\widetilde{\mathrm{R}g}(r)e^{irt}|r|dr. \tag{4.55}$$

Comparing (4.55) with the definition of f_ψ we see that, if we can find a radial function k_ψ defined on $\mathbb{R}^2$ so that

$$\mathrm{R}(k_\psi)(t) = \psi(t),$$

then

$$f_\psi(x,y) = k_\psi * f(x,y) \tag{4.56}$$

The existence of such a function is a consequence of the results in section 2.5. Because $\hat{\psi}$ has bounded support, $\psi(t)$ is an infinitely differentiable function, with all derivatives bounded. To apply Proposition 2.5.1 we need to know that $\psi(t)$ and $\psi'(t)$ are absolutely integrable. This translates into a requirement that $\hat{\psi}$ is sufficiently continuous. In that case, the function k_ψ is given by the formula

$$k_\psi(\rho) = -\frac{1}{\pi}\int_\rho^{\infty} \frac{\psi'(t)dt}{\sqrt{t^2-\rho^2}} \tag{4.57}$$

This completes the proof of the following proposition.

Proposition 4.4.1. *Suppose that $\hat{\psi}(r)$ satisfies the conditions in* (4.52) *and $\psi(t)$ is absolutely integrable then*

$$f_\psi(x,y) = k_\psi * f(x,y)$$

where k_ψ is given by (4.57)*.*

Remark 4.4.1. As $\hat{\psi}$ has bounded support its inverse Fourier transform, ψ does not. Replacing f by f_ψ may therefore lead to blurring of the image. Increasing the support of $\hat{\psi}(r)$ leads, in general to a more sharply peaked ψ and therefore a more sharply peaked k_ψ. This discussion is adapted from [69].

4.4.1 Addendum*

See: A.3.6.

The analysis in the previous section is unsatisfactory in one particular: we explicitly exclude the possibility that $\hat{\psi}_W(r) = \chi_{[-W,W]}(r)$. The problem is that $\psi_W(t) = \sin(Wt)/(\pi t)$ is not absolutely integrable and so the general inversion result for radial functions does not

apply. In this special case the integral defining k_ψ is a convergent, improper integral, which can be computed exactly.

We use the formula

$$\int_1^\infty \frac{\sin(xt)dt}{\sqrt{t^2-1}} = \frac{\pi}{2} J_0(x),$$

here J_0 is a Bessel function, see [49]. Putting this into the inversion formula and using the fact that $J_0' = -J_1$ we obtain

$$k_W(x) = \frac{W}{2\pi x} J_1(Wx).$$

The power series for $J_1(x)$ about $x = 0$ is

$$J_1(x) = \frac{x}{2} \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{2^{2k} k!(k+1)!}$$

from which it follows easily that $k_W(x)$ is a smooth function of x^2. The standard asymptotic expansion for $J_1(x)$ as $|x|$ tends to infinity implies that

$$|k_W(x)| \leq \frac{C}{(1+|x|)^{\frac{3}{2}}}$$

and therefore the integrals defining $\mathrm{R}k_W$ converge absolutely. As the Radon transform is linear, we can extend the result of the previous section to allow functions of the form

$$\hat{\psi}(r) = \chi_{[-W,W]}(r) + \hat{\psi}_c(r)$$

where $\psi_c = \mathcal{F}^{-1}\hat{\psi}_c$ satisfies the hypotheses of Proposition 2.5.1. In this case

$$f_\psi = (k_W + k_{\psi_c}) * f. \tag{4.58}$$

Exercise 4.4.1. Justify the computations for the function $\hat{\psi} = \chi_{[-W,W]}$ leading up to formula (4.58).

4.5 The range of the Radon transform

In medical imaging the data under consideration usually has bounded support. The Radon transform of a function with bounded support satisfies an infinite set of *moment* conditions. From the point of view of measurements these can be viewed as consistency conditions. Mathematically this is a part of the problem of characterizing the *range* of the Radon transform on data with bounded support. In this section we begin by considering such data; afterwards we study the range of the Radon transform for a more general class of functions.

4.5.1 Data with bounded support

Suppose that f is a function which vanishes outside the disk of radius R. As observed above this implies that $\mathrm{R}f(t,\omega) = 0$ if $|t| > R$. For a non-negative integer, n consider the integral,

$$M_n(f)(\omega) = \iint_{\mathbb{R}^2} f(x,y)[\langle (x,y),\omega\rangle]^n dxdy. \tag{4.59}$$

If f has bounded support, then these integrals are well defined for any $n \in \mathbb{N}\cup\{0\}$. On the other hand, if f does not vanish outside a disk of finite radius then, for sufficiently large n, these integral may not make sense.

Changing coordinates with $(x,y) = t\omega + s\hat{\omega}$ we can rewrite this integral in terms of $\mathrm{R}f$,

$$\begin{aligned} M_n(f)(\theta) &= \iint_{\mathbb{R}^2} f(t\omega + s\hat{\omega})t^n dsdt \\ &= \int --\infty^{\infty} \mathrm{R}f(t,\omega)t^n dt. \end{aligned} \tag{4.60}$$

The function $M_n(f)(\omega)$ is called the n^{th} *moment* of the Radon transform of f. If $\mathrm{R}f(t,\omega)$ vanishes for $|t| > R$ then this integral is well defined for all n. In example 2.4.6 we showed that there are functions, which do **not** have bounded support, for which the Radon transform is defined and vanishes for large enough values of t. If f does have bounded support then $M_n(f)(\omega)$ depends on ω in a very special way.

It is useful to express ω as a function of the angle θ,

$$\omega(\theta) = (\cos(\theta), \sin(\theta)).$$

Using the binomial theorem we obtain

$$\begin{aligned} \langle (x,y),\omega(\theta)\rangle^n &= (x\cos\theta + y\sin\theta)^n \\ &= \sum_{j=0}^{n} \binom{n}{j} (x\cos\theta)^j (y\sin\theta)^{n-j} \\ &= \sum_{j=0}^{n} \binom{n}{j} \cos^j\theta \sin^{n-j}\theta x^j y^{n-j}. \end{aligned}$$

Putting the sum into formula (4.59) we see that this integral defines a trigonometric polynomial of degree n.

$$\begin{aligned} M_n(f)(\theta) &= \sum_{j=0}^{n} \binom{n}{j} \cos^j\theta \sin^{n-j}\theta \iint_{\mathbb{R}^2} f(x,y)x^j y^{n-j} dxdy \\ &= \sum_{j=0}^{n} a_{nj} \sin^j\theta \cos^{n-j}\theta \end{aligned} \tag{4.61}$$

where

$$a_{nj} = \binom{n}{j} \iint_{\mathbb{R}^2} f(x,y) x^j y^{n-j} dxdy.$$

If f has bounded support then $M_n(f)(\theta)$ is a trigonometric polynomial of degree n. We summarize these computations in a proposition.

Proposition 4.5.1. *Suppose that f is a function with bounded support then*

(1). $\mathrm{R}f(t,\omega)$ *has bounded support.*

(2). *For all non-negative integers, n there exist constants $\{a_{n0}, \ldots, a_{nn}\}$ such that*

$$\int_{-\infty}^{\infty} \mathrm{R}f(t,\omega(\theta)) t^n dt = \sum_{j=0}^{n} a_{nj} \sin^j \theta \cos^{n-j} \theta.$$

The proposition suggests the following question: Suppose that $h(t,\omega)$ is a function on $\mathbb{R} \times S^1$ such that

(1). $h(t,\omega) = h(-t,-\omega)$,

(2). $h(t,\omega) = 0$ if $|t| > R$,

(3). For each non-negative integer n

$$m_n(h)(\theta) = \int_{-\infty}^{\infty} h(t,\omega(\theta)) t^n dt$$

is a trigonometric polynomial of degree n,

(4). $h(t,\omega)$ is a sufficiently smooth function of (t,ω).

Does there exist a function $f(x,y)$ in the domain of the Radon transform vanishing outside of the disk of radius R such that

$$h(t,\omega) = Rf(t,\omega)?$$

In other words: does h belong to the range of the Radon transform, acting on smooth functions with bounded support? According to a theorem of Helgason and Ludwig, the answer to this question turns out to be yes, however the proof of this result requires techniques well beyond the scope of this text. For a detailed discussion of this question the reader is referred to [50]. More material can be found in [23], [44], [48] or [14].

We model the data measured in CT-imaging as the Radon transform of a piecewise continuous function with bounded support. If we could measure $\mathrm{R}f(t,\omega)$ for all (t,ω) then it would probably not be the exact the Radon transform of such a function. This is because all measurements are corrupted by various sources of error and noise. In particular the patient's movements, both internal (breathing, heart beat, blood circulation, etc.) and external, affect the measurements. The measured data would therefore be inconsistent and fail to satisfy the moment conditions prescribed above.

4.5.2 More general data*

In this section we show how to use functional analytic methods to extend the domain of the Radon transform and study its range. The material in this section requires elementary measure theory and is included for completeness. It is not used in the latter parts of the book.

The problem of characterizing the range of the Radon transform is intimately connected with that of determining its domain. As remarked in section 2.4 the Radon transform is defined for any function f such that the restriction of $|f|$ to any line is integrable. A simple sufficient condition is that f is piecewise continuous and satisfies an estimate of the form

$$f(x,y) \leq \frac{C}{(1+|x|+|y|)^{1+\epsilon}}$$

for a constant C and any $\epsilon > 0$. A function satisfying such an estimate has a Radon transform. However, for sufficiently large n, the integral defining the "absolute" n^{th}-moment of f,

$$\int\limits_{\mathbb{R}^2} |f(x,y)||\langle (x,y),\omega\rangle|^n dxdy, \quad n \in \mathbb{N}$$

may not converge. Because of the divergence of these integrals the change of variables argument used in section 4.5.1 to express these moments of f in terms of moments of $\mathrm{R}f$ is not valid. Moreover our formulæ for the inverse transform require interpretation, as the integrals involved do necessarily converge absolutely.

In example 2.4.6 we considered the functions

$$F_n((r\cos(\theta), r\sin(\theta)) = r^{-n}\cos(n\theta)\chi_{[1,\infty)}(r).$$

If $n \geq 2$ then F_n is in the domain of the Radon transform. Moreover $\mathrm{R}F_n(t,\omega) = 0$ if $|t| > 1$ and therefore all the moments of $\mathrm{R}F_n$ are given by convergent integrals. However if $k > n-1$ then the integral defining the k^{th}-moment of F_n does not converge and therefore the k^{th} moment of $\mathrm{R}F_n$ *cannot* be identified with the corresponding moment of F_n. As F_n does not have bounded support, the results of the previous section imply that the higher moments of $\mathrm{R}F_n$ are not trigonometric polynomials of correct degree. For these examples this statement can be verified by direct computation.

Evidently there are functions on $\mathbb{R}\times S^1$, not satisfying the moment conditions, which lie in the range of the Radon transform. It is difficult to give an entirely satisfactory description of this range but a reasonable heuristic is that any measurable, even function on $\mathbb{R}\times S^1$ which is "sufficiently regular in t" and decays "sufficiently rapidly" as $|t|$ goes to infinity, is the Radon transform of a function on $\mathbb{R}^2$. An optimal result to this effect can be found in [50]. We close this section by showing that, under relatively mild hypotheses, a function on $\mathbb{R}\times S^1$ is the "generalized Radon transform" of a function on $\mathbb{R}^2$.

We begin by extending the domain of the Radon transform using functional analytic methods.

Proposition 4.5.2. *Suppose that $f \in L^2(\mathbb{R}^2)$ and its Fourier transform satisfies the following conditions*

(1). $\hat{f}(\boldsymbol{\xi})$ is bounded and continuous away from $|\boldsymbol{\xi}| = 0$,

(2). $\hat{f} \in L^1(\mathbb{R}^2)$.

Then f has a generalized Radon transform, denoted by F with the following properties

(1). *Both F and $D_{\frac{1}{2}}F$ belong to $L^2(\mathbb{R} \times S^1)$.*

(2). *For every $\omega \in S^1$ we have $\widetilde{F}(r, \omega) = \hat{f}(r\omega)$.*

(3). *If, in addition, $f \in L^p(\mathbb{R}^2)$ for a $p < 2$ then, for any continuous function g with bounded support on $\mathbb{R} \times S^1$, we have the identity*

$$\langle F, g\rangle_{\mathbb{R}\times S^1} = \langle f, \mathrm{R}^* g\rangle_{\mathbb{R}^2}. \tag{4.62}$$

Proof. Let $\hat{\varphi}(\boldsymbol{\xi})$ be a smooth, non-negative function, supported in the ball of radius 1, with total integral 1. For $\rho > 0$ let

$$\hat{\varphi}_\rho(\boldsymbol{\xi}) = \frac{1}{\rho^2}\hat{\varphi}(\rho^{-1}\boldsymbol{\xi})$$

and $\varphi_\rho(\mathbf{x}) = \varphi(\rho\mathbf{x})$, its inverse Fourier transform. Because $\hat{f}$ is absolutely integrable it follows, from the Fourier inversion formula, that f is bounded and continuous. Thus for each $\rho > 0$ the function $\varphi_\rho f$ is in the natural domain of the Radon transform. Let

$$F_\rho(t, \omega) = \mathrm{R}(\varphi_\rho f)(t, \omega).$$

To prove the proposition we show that F_ρ converges to a function F in $L^2(\mathbb{R}\times S^1)$ with the conditions enumerated above.

First observe that, as $f \in L^2(\mathbb{R}^2)$, we can apply the Parseval formula and the central slice theorem to conclude that

$$\|(\widetilde{F}_\rho - \hat{f}(r\omega))r\|^2_{L^2(\mathbb{R}\times S^1)} = \|\varphi_\rho f - f\|_{L^2(\mathbb{R}^2)} \tag{4.63}$$

Here we consider $\hat{f}(r\omega)$ as a function on $\mathbb{R} \times S^1$. This shows that

$$\lim_{\rho\to 0} \|(\widetilde{F}_\rho - \hat{f}(r\omega))r\|^2_{L^2(\mathbb{R}\times S^1)} = 0 \tag{4.64}$$

To obtain a limit in $L^2(\mathbb{R} \times S^1)$ we also need to show that

$$\lim_{\rho\to 0} \int_0^{2\pi}\int_{-1}^{1} |\widetilde{F}_\rho(r, \omega) - \hat{f}(r\omega)|^2 dr d\omega = 0. \tag{4.65}$$

Using the central slice theorem and the convolution theorem for the Fourier transform we see that

$$\widetilde{F}_\rho = \hat{\varphi}_\rho * \hat{f}.$$

The convolution takes place in $\mathbb{R}^2$. Because $\hat{\varphi}_\rho$ is non-negative it follows from this formula that

$$|\widetilde{F}_\rho(r, \omega)| \le \|\hat{f}\|_{L^\infty} \int_{\mathbb{R}^2} \hat{\varphi}(\boldsymbol{\xi}) d\boldsymbol{\xi} = \|\hat{f}\|_{L^\infty}.$$

If $\hat{f}$ is continuous at $\boldsymbol{\xi}$ then $\lim \hat{\varphi}_\rho * \hat{f}(\boldsymbol{\xi}) = \hat{f}(\boldsymbol{\xi})$. The bounded convergence theorem and the continuity of $\hat{f}$ away from $\boldsymbol{\xi} = 0$ therefore imply that (4.65) is true. This verifies the existence of

a limit $F(t,\omega)$, in $L^2(\mathbb{R}\times S^1)$ for $< F_\rho \ : \ \rho > 0 >$ which satisfies the first two conditions in the conclusion of the proposition.

To establish the last property we assume that $f \in L^p(\mathbb{R}^2)$ for a $1 < p < 2$. Let g be a bounded, continuous function with bounded support. For each $\rho > 0$ we have the identity

$$\langle F_\rho, g\rangle_{\mathbb{R}\times S^1} = \langle \varphi_\rho f, \mathrm{R}^* g\rangle_{\mathbb{R}^2}. \tag{4.66}$$

In exercise 4.2.3 it is shown that there is a constant C so that

$$|\,\mathrm{R}^* g(x,y)| \leq \frac{C}{1+|x|+|y|}.$$

This implies that $\mathrm{R}^* g \in L^q(\mathbb{R}^2)$ for any $q > 2$. As $f \in L^p(\mathbb{R}^2)$ for a $p < 2$ the right hand side of (4.66) converges to $\langle f, \mathrm{R}^* g\rangle_{\mathbb{R}^2}$. The left hand side converges to $\langle F, g\rangle_{\mathbb{R}\times S^1}$ because F_ρ converges to F in $L^2(\mathbb{R}\times S^1)$ This verifies the last assertion of the proposition. □

We now show that the range of the extended Radon transform is quite large.

Proposition 4.5.3. *Let $h(t,\omega)$ be a continuous, even function defined on $\mathbb{R}\times S^1$ such that*

(1). $h(t,\omega)$ and $r\widetilde{h}(r,\omega)$ are uniformly integrable in t and r respectively.

(2). Letting $\omega(\theta) = (\cos\theta, \sin\theta)$, the function $\widetilde{h}(r,\theta)$ has weak first partial derivatives which satisfy

$$\int_0^{2\pi}\int_{-\infty}^{\infty} |\partial_\theta \widetilde{h}||r|^{-1} dr d\theta < \infty, \quad \int_0^{2\pi}\int_{-\infty}^{\infty} |\partial_r \widetilde{h}||r| dr d\theta < \infty. \tag{4.67}$$

Then there is a bounded continuous function $f \in L^2(\mathbb{R})$ whose generalized Radon transform is h.

Proof. Since h and $r\widetilde{h}$ are uniformly integrable there is a constant M so that for every $\omega \in S^1$ we have

$$\int_{-\infty}^{\infty} |h(t,\omega)| dt < M, \quad \int_{-\infty}^{\infty} |\widetilde{h}(r,\omega)||r| dr < M. \tag{4.68}$$

The first estimate implies that $\widetilde{h}(r,\omega)$ is continuous and bounded. This implies that there is a constant C so that

$$\frac{\widetilde{h}}{C} \leq 1$$

and therefore $|\widetilde{h}|^2 \leq C|\widetilde{h}|$. The second estimate implies that

$$\int_0^{2\pi}\int_0^{\infty} |\widetilde{h}|^2 r dr d\omega < \infty.$$

Define a function f on $\mathbb{R}^2$ by setting

$$f(x,y) = \frac{1}{[2\pi]^2}\int_0^{2\pi}\int_0^{\infty} \widetilde{h}(r,\omega) e^{i\langle (x,y), r\omega\rangle} r dr d\omega. \tag{4.69}$$

From our hypotheses it follows $f \in L^2(\mathbb{R}^2)$ and $\hat{f}(r\omega) = \widetilde{h}(r,\omega)$. Indeed this integral is absolutely convergent and therefore $f(x,y)$ is a bounded, continuous function.

To finish the proof we need to show that $f \in L^p(\mathbb{R}^2)$ for a $p < 2$. This follows from the weak differentiability of $\widetilde{h}$. We identify $\widetilde{h}$ as the polar coordinate representation of a function $H(\xi_1, \xi_2)$ defined on $\mathbb{R}^2$ by

$$H(r\cos\theta, r\sin\theta) = \widetilde{h}(r, \omega(\theta)). \tag{4.70}$$

Our assumptions on $\widetilde{h}$ imply that H is in $L^2(\mathbb{R}^2)$ and has weak derivatives $H_{\xi_1}, H_{\xi_2} \in L^2(\mathbb{R}^2)$ as well. The proof of this statement is left as an exercise. Proposition 3.3.9 implies that there are functions $f_1, f_2 \in L^2(\mathbb{R}^2)$ so that

$$xf(x,y) = f_1(x,y) \text{ and } yf(x,y) = f_2(x,y)$$

and therefore

$$|f(x,y)| = \frac{g(x,y)}{(1+|x|+|y|)}$$

for a third function $g \in L^2(\mathbb{R}^2)$. Fix such a $1 < p < 2$ and apply the Hölder inequality to conclude that

$$\begin{aligned} \int\limits_{\mathbb{R}^2} |f(x,y)|^p dxdy &\leq \left[\int\limits_{\mathbb{R}^2} |g(x,y)|^2 dxdy \right]^{\frac{p}{2}} \left[\int\limits_{\mathbb{R}^2} \frac{dxdy}{(1+|x|+|y|)^{\frac{2}{2-p}}} \right]^{\frac{2-p}{2}} \\ &< \infty. \end{aligned} \tag{4.71}$$

The last estimate follows because $1 < p < 2$. This implies that $f \in L^p(\mathbb{R}^2)$ for any $1 < p < 2$. The function f satisfies the hypotheses of the previous proposition and therefore h is the generalized Radon transform of f. □

Remark 4.5.1. If $\widetilde{h}(r,\omega)$ has three integrable derivatives, when thought of as a function of the *Cartesian coordinates*

$$\xi_1 = r\omega_1, \xi_2 = r\omega_2$$

then f satisfies an estimate of the form

$$|f(x,y)| \leq \frac{C}{(1+|x|+|y|)^2}.$$

In this case f lies in the natural domain of the Radon transform. The facts that

$$\langle \mathrm{R}f, g\rangle_{\mathbb{R}\times S^1} = \frac{1}{2\pi}\langle \widetilde{\mathrm{R}f}, \widetilde{g}\rangle_{\mathbb{R}\times S^1}$$

and $\widetilde{h}(r,\omega) = \hat{f}(r\omega)$ imply that $\mathrm{R}f = h$.

Exercise 4.5.1. Find conditions on $h(t,\omega)$ which imply that $\widetilde{h}(r,\omega)$ has the regularity described in remark 4.5.1.

Exercise 4.5.2. Show that the assumptions on $\widetilde{h}$ imply that $H(\xi)$, defined in (4.70), has L^2 partial derivatives.

4.6 Continuity of the Radon transform and its inverse

In order for the measurement process in X-ray tomography to be stable the map $f \mapsto \mathrm{R}f$ should be continuous in a reasonable sense. Estimates for the continuity of this map quantify the sensitivity of the output, $\mathrm{R}f$ of a CT-scanner to changes in the input. The *less* continuous the map, the *more* sensitive the measurements are to changes in the input. Estimates for the continuity of inverse, $h \mapsto \mathrm{R}^{-1}h$ quantify the effect of errors in the measured data on the quality of the reconstructed image. Because we actually *measure* the Radon transform, estimates for the continuity of R^{-1} are more important for the problem of image reconstruction. To discuss the continuity properties of either transform we need to select norms for functions in the domain and range. Using the L^2-norms on both, the Parseval formula, (4.15) provides a starting point for this discussion.

The Parseval formula says that if $f \in L^2(\mathbb{R}^2)$ then $D_{\frac{1}{2}}\,\mathrm{R}f \in L^2(\mathbb{R}\times S^1)$. This estimate has somewhat limited utility, as $|r|$ vanishes at $r = 0$, we cannot conclude that $\mathrm{R}f$ is actually in $L^2(\mathbb{R}\times S^1)$. In medical applications the data has bounded support and in this case additional estimates are available. The implications of the Parseval formula for the inverse transform are somewhat less desirable. It says that in order to control the L^2-norm of the reconstructed image we need to have control on the half-order derivative of the measured data. Due to noise this is, practically speaking, not possible. After discussing the continuity properties of the forward transform for data with bounded support we consider the continuity properties of the *approximate inverse* described in section 4.4.

4.6.1 Bounded support

Functions with bounded support satisfy better L^2-estimates.

Proposition 4.6.1. *Let $f \in L^2(\mathbb{R}^2)$ and suppose that f vanishes outside the ball of radius L then, for each ω, we have the estimate*

$$\int_{-\infty}^{\infty} |\,\mathrm{R}f(t,\omega)|^2 dt \leq 2L\|f\|_{L^2}^2.$$

Proof. The proof of the proposition is a simple application of Cauchy-Schwarz inequality. Because f vanishes outside the ball of radius L we can express $\mathrm{R}f$ as

$$\mathrm{R}f(t,\omega) = \int_{-L}^{L} f(t\omega + s\hat{\omega})ds.$$

Computing the L^2-norm of $\mathrm{R}f$ in the t-variable we obtain

$$\begin{aligned}\int_{-\infty}^{\infty} |\,\mathrm{R}f(t,\omega)|^2 dt &= \int_{-L}^{L} \left| \int_{-L}^{L} f(t\omega + s\hat{\omega})ds \right|^2 dt \\ &\leq 2L \int_{-L}^{L}\int_{-L}^{L} |f(t\omega + s\hat{\omega})|^2 ds dt.\end{aligned} \tag{4.72}$$

In the second line we used the Cauchy-Schwarz inequality. ∎

The proposition shows that, if f vanishes outside a bounded set, then we control not only the overall L^2-norm of $\mathrm{R}f$ but the L^2-norm in each direction, ω separately. Using the support properties of f more carefully gives a weighted estimate on the L^2-norm of $\mathrm{R}f$.

Proposition 4.6.2. *Let $f \in L^2(\mathbb{R}^2)$ and suppose that f vanishes outside the ball of radius L then, for each ω, we have the estimate*

$$\int_{-\infty}^{\infty} \frac{|\,\mathrm{R}f(t,\omega)|^2 dt}{\sqrt{L^2-t^2}} \leq 2\|f\|_{L^2}^2.$$

Proof. To prove this estimate observe that

$$f(x,y) = \chi_{[0,L^2]}(x^2+y^2) f(x,y).$$

The Cauchy Schwarz inequality therefore implies that, for $|t| \leq L$, we have the estimate

$$\begin{aligned} |\,\mathrm{R}f(t,\omega)|^2 &= \left| \int_{-L}^{L} f(t\omega + s\hat{\omega}) \chi_{[0,L^2]}(s^2+t^2) ds \right|^2 \\ &\leq 2 \int_{-L}^{L} |f(t\omega+s\hat{\omega})|^2 ds \int_0^{\sqrt{L^2-t^2}} ds \\ &= 2\sqrt{L^2-t^2} \int_{-L}^{L} |f(t\omega+s\hat{\omega})|^2 ds. \end{aligned} \tag{4.73}$$

Thus

$$\begin{aligned} \int_{-L}^{L} \frac{|\,\mathrm{R}f(t,\omega)|^2 dt}{\sqrt{L^2-t^2}} &\leq \int_{-L}^{L} \frac{2\sqrt{L^2-t^2}}{\sqrt{L^2-t^2}} \int_{-L}^{L} |f(t\omega+s\hat{\omega})|^2 ds dt \\ &= 2\|f\|_{L^2}^2. \end{aligned} \tag{4.74}$$

□

A function in $f \in L^2(\mathbb{R}^2)$ with support in the disk of radius L can be approximated, in the L^2-norm, by a sequence of smooth functions $< f_n >$. This sequence can also be taken to have support in the disk of radius L. The Radon transforms of these functions satisfy the estimates

$$\int_{-\infty}^{\infty} |\,\mathrm{R}f_n(t,\omega)|^2 dt \leq 2L\|f_n\|_{L^2}^2$$

and

$$\frac{1}{[2\pi]^2} \int_0^{\pi} \int_{-\infty}^{\infty} |\widetilde{\mathrm{R}f_n}(r,\omega)|^2 |r| dr d\omega = \|f_n\|_{L^2(\mathbb{R}^2)}^2.$$

In a manner analogous to that used to extend the Fourier transform to L^2-functions we can now extend the Radon transform to L^2-functions with support in a fixed bounded set.

For bounded functions on $\mathbb{R} \times S^1$ vanishing for $|t| > L$ a norm is defined by

$$\|h\|_{2,L}^2 = \sup_{\omega \in S^1} \int_{-L}^{L} |h(t,\omega)|^2 dt + \frac{1}{[2\pi]^2} \int_0^{\pi} \int_{-\infty}^{\infty} |\widetilde{h}(r,\omega)|^2 |r| dr d\omega.$$

The closure of $\mathcal{C}^0([-L,L] \times S^1)$ in this norm is a Hilbert space which can be identified with a subspace of $L^2([-L,L] \times S^1)$. For f as above, $\mathrm{R}f$ is defined as the limit of $\mathrm{R}f_n$ in this norm. Evidently the estimates above hold for $\mathrm{R}f$. On the other hand the elementary formula for $\mathrm{R}f(t,\omega)$ may not be meaningful as f may not be absolutely integrable over $l_{t,\omega}$.

While it is well beyond the scope of this text, it is nonetheless, true that a function on $\mathbb{R} \times S^1$ with support in the set $|t| \leq L$ and finite $\| \cdot \|_{2,L}$-norm which satisfies the moment conditions is the generalized Radon transform of function in $L^2(\mathbb{R}^2)$ with support in the disk of radius L. A proof can be found in [23] or [50].

Exercise 4.6.1. Suppose that $f \in L^2(\mathbb{R}^2)$ and that f vanishes outside the ball of radius L. Show that $\| \mathrm{R}f(\cdot,\omega_1) - \mathrm{R}f(\cdot,\omega_2)\|_{L^2(\mathbb{R})}$ tends to zero as ω_1 approaches ω_2. In other words the map $\omega \mapsto \mathrm{R}f(\cdot,\omega)$ is a continuous map from the circle into $L^2(\mathbb{R})$. This shows that, if we measure errors in the L^2-norm then the Radon transform is not excessively sensitive to small changes in the measurement environment.

Exercise 4.6.2. For the terms in the approximating sequence, $< \mathrm{R}f_n >$ the moments $\{m_k(\mathrm{R}f_n)\}$ satisfy the conditions in Proposition 4.5.1. Show that for the limiting function, the moments $\{m_k(\mathrm{R}f)\}$ are well defined and also satisfy these conditions.

4.6.2 Estimates for the inverse transform*

The question of more immediate interest is the continuity properties of the *inverse* transform. This is the more important question because we actually measure an approximation, Rf_m to $\mathrm{R}f$. It would appear that to estimate the error in the reconstructed image, we would need to estimate

$$\mathrm{R}^{-1}Rf_m - f = \mathrm{R}^{-1}(Rf_m - \mathrm{R}f). \tag{4.75}$$

There are several problems that immediately arise. The most obvious problem is that Rf_m may not be in the range of the Radon transform. If $Rf_m(t,\omega)$ does not have an L^2-half-derivative in the t-direction, that is,

$$\int_0^{2\pi} \int_{-\infty}^{\infty} |\widetilde{Rf_m}(r,\omega)|^2 |r| dr d\omega = \infty,$$

then according to the Parseval formula, (4.15) Rf_m is **not** the Radon transform of a function in $L^2(\mathbb{R}^2)$. In order to control the L^2-error,

$$\| \mathrm{R}^{-1}(Rf_m - \mathrm{R}f)\|_{L^2(\mathbb{R}^2)}$$

it is necessary that measurements have such a half derivative and the difference

$$\|D_{\frac{1}{2}}(Rf_m - \mathrm{R}f)\|_{L^2(\mathbb{R}\times S^1)}$$

is small. This means that we need to control the high frequency content of Rf_m; in practice this is not possible. If this could be done, it would only give an estimate for the L^2-error. In order for the reconstructed image to "look like" the original it may be important to control the pointwise errors. Even though the L^2-error is small, the pointwise errors can be large on small sets. While the mathematical problem of estimating the Radon inverse is quite interesting and important, it has little bearing on the problem of practical image reconstruction. A very nice treatment of the mathematical question is given in [50]. We now turn our attention to understanding the continuity of the *approximate* inverses defined in section 4.4.

An approximate inverse is denoted by R_ψ^{-1}, where $\psi(t)$ is a regularizing function specified in terms of its Fourier transform by the conditions

$$\begin{aligned} &\hat{\psi}(0) = 1, \\ &\hat{\psi}(r) = 0 \text{ for } |r| > W. \end{aligned} \tag{4.76}$$

It is also assumed that the radial function k_ψ defined in 4.57 is in the domain of the Radon transform and

$$\mathrm{R}k_\psi = \psi.$$

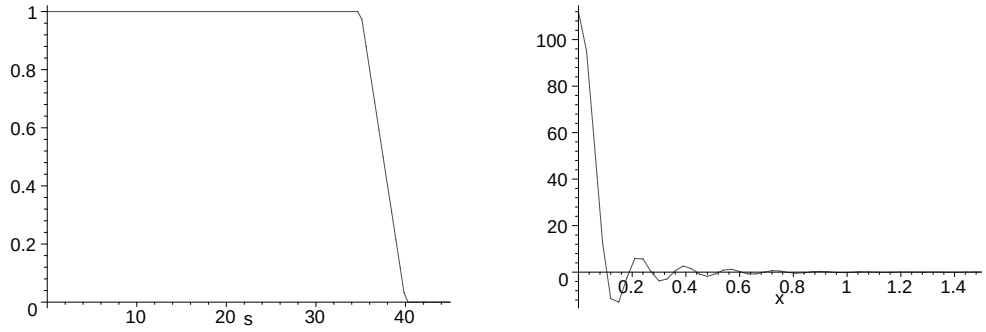
In this case

$$\mathrm{R}_\psi^{-1}\,\mathrm{R}f = k_\psi * f. \tag{4.77}$$

Example 4.6.1. Let $\hat{\psi}$ be the piecewise linear function

$$\hat{\psi}(r) = \begin{cases} 1 & \text{for } |r| < W - C, \\ \frac{W-|r|}{C} & \text{for } W - C \le |r| \le W, \\ 0 & \text{for } |r| > W. \end{cases}$$

Radial graphs of ψ and k_ψ are shown in figure 4.1.



(a) $\hat{\psi}$ (b) k_ψ

Figure 4.1: Graphs of $\hat{\psi}$ and k_ψ with $W = 40, C = 5$.

The reconstructed image is

$$f_\psi = \mathrm{R}_\psi^{-1} R f_m,$$

therefore we need to estimate the difference $f - f_\psi$. As $k_\psi * f = \mathrm{R}_\psi^{-1}\,\mathrm{R} f$ we can rewrite this difference as

$$f - f_\psi = (f - k_\psi * f) + \mathrm{R}_\psi^{-1}(\mathrm{R} f - R f_m). \tag{4.78}$$

The first term on the right hand side is the error caused by using an approximate inverse. It is present even if we have perfect data. Bounds for this term depend in an essential way on the character of the data. If f is assumed to be a continuous function of bounded support then, by taking W very large, the pointwise error,

$$\|f - k_\psi * f\|_\infty = \sup_{(x,y)\in\mathbb{R}^2} |f(x,y) - k_\psi * f(x,y)|$$

can be made as small as desired. It is more realistic to model f as a piecewise continuous function. In this case the difference, $|f(x,y) - k_\psi * f(x,y)|$ can be made small at points where f is continuous. Near points where f has a jump the approximate reconstruction may display an oscillatory artifact. Figure 4.2 shows the reconstruction of $\chi_{D_1}(x,y)$ using the regularizing function graphed in figure 4.1. Robust estimates for the second term are less dependent on the precise nature of f.

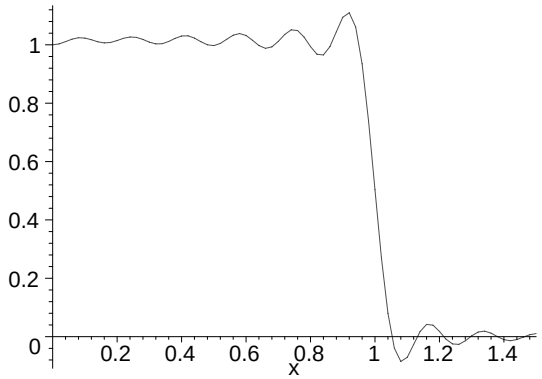


Figure 4.2: Radial graph of $k_\psi * \chi_{D_1}$, with $W = 40, C = 5$.

For $h(t,\omega)$, a function on $\mathbb{R} \times S^1$ with bounded support, the approximate inverse is given by

$$\begin{aligned}(\mathrm{R}_\psi^{-1} h)(x,y) &= \frac{1}{4\pi^2} \int_0^\pi \int_{-\infty}^{\infty} \widetilde{h}(r,\omega) e^{irt} \hat{\psi}(r) |r| dr \\ &= \frac{1}{2\pi} \int_0^\pi g_\psi *_t h(\langle (x,y),\omega\rangle, \omega) d\omega.\end{aligned} \tag{4.79}$$

Here $g_\psi = \mathcal{F}^{-1}(\hat{\psi}(r)|r|)$ and $*_t$ indicates convolution in the t-variable.

A simple estimate for the sup-norm of $\mathrm{R}_\psi^{-1}h$ follows from the sup-norm estimate for a convolution

$$\|l * k\|_{L^\infty} \leq \|l\|_{L^\infty}\|k\|_{L^1}.$$

Applying this estimate gives

$$\|\,\mathrm{R}_\psi^{-1}h\|_{L^\infty} \leq \frac{\|g_\psi\|_{L^\infty}}{2\pi}\int_0^\pi\int_{-\infty}^{\infty}|h(t,\omega)|dtd\omega \tag{4.80}$$

If $\hat{\psi}$ is non-negative then

$$|g_\psi(t)| \leq |g_\psi(0)| = \int_{-\infty}^{\infty}|r|\hat{\psi}(r)dr.$$

Assuming that $0 \leq \hat{\psi}(t) \leq M$ and that it vanishes outside the interval $[-W, W]$ leads to the estimate

$$\|g_\psi\|_{L^\infty} \leq MW^2.$$

Combining this with (4.80) gives

$$\|\,\mathrm{R}_\psi^{-1}h\|_{L^\infty} \leq \frac{MW^2}{2\pi}\|h\|_{L^1(\mathbb{R}\times S^1)}. \tag{4.81}$$

This estimate shows that the sup-norm of the error in the approximate reconstructed image, $\mathrm{R}_\psi^{-1}(\mathrm{R}f - Rf_m)$, can be controlled if the measurement errors can be controlled in the L^1-norm. It also shows that the error increases as W increases.

To summarize, the error in the approximate reconstruction is bounded by

$$|f - f_\psi| \leq |f - k_\psi * f| + \frac{\|g_\psi\|_{L^\infty}}{2\pi}\|\,\mathrm{R}f - Rf_m\|_{L^1(\mathbb{R}\times S^1)}. \tag{4.82}$$

Recall that

$$\mathcal{F}(k_\psi) = \hat{\psi} \text{ and } \mathcal{F}(g_\psi) = |r|\hat{\psi}.$$

The function k_ψ is rapidly decreasing and sharply peaked if $\hat{\psi}$ is smooth and W is taken large. On the other hand g_ψ cannot decay faster than $O(t^{-2})$. This is a consequence of the fact that $|r|\hat{\psi}(r)$ is singular at $r = 0$.

Exercise 4.6.3. Prove that $\|l * k\|_{L^\infty} \leq \|l\|_{L^\infty}\|k\|_{L^1}$.

Exercise 4.6.4. Suppose that $\hat{\psi}(\xi)$ is a smooth function with bounded support such that $\hat{\psi}(0) \neq 0$ and let

$$g_\psi(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\hat{\psi}(\xi)|\xi|e^{it\xi}d\xi.$$

Show that there is a constant $C > 0$ so that the following *lower bound* holds for large enough t :

$$|g_\psi(t)| \geq \frac{C}{1+t^2}. \tag{4.83}$$

Exercise 4.6.5. Use the central slice theorem to give a formula for k_ψ as a Bessel transform of $\hat{\psi}(r)$.

Exercise 4.6.6. Use Hölder's inequality to show that

$$\|l * k\|_{L^\infty} \leq \|l\|_{L^2}\|k\|_{L^2}.$$

Use this estimate to prove that

$$\|\,\mathrm{R}_\psi^{-1}h\|_{L^\infty} \leq \frac{\|g_\psi\|_{L^2(\mathbb{R})}}{\sqrt{4\pi}}\|h\|_{L^2(\mathbb{R}\times S^1)}.$$

Under the assumptions used above to estimate $\|g_\psi\|_{L^\infty}$ show that

$$\|g_\psi\|_{L^2} \leq \sqrt{\frac{2}{3}}MW^{\frac{3}{2}}.$$

4.7 The higher dimensional Radon transform*

See: A.2.1, A.2.5.

For the sake of completeness we briefly present the theory of the Radon transform in higher dimensions. The parameterization of the affine hyperplanes in $\mathbb{R}^n$ is quite similar to that used for lines in $\mathbb{R}^2$. Let ω be a unit vector in $\mathbb{R}^n$, i.e. a point on S^{n-1}, and let $t \in \mathbb{R}$, each affine hyperplane has a representation in the form

$$l_{t,\omega} = \{\mathbf{x} \in \mathbb{R}^n \ : \ \langle \mathbf{x}, \omega\rangle = t\}.$$

As in the two dimensional case $l_{t,\omega} = l_{-t,-\omega}$ and the choice of vector ω defines an orientation on the hyperplane.

In order to define the Radon transform it is useful to choose vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}\}$ so that

$$\langle \omega, \mathbf{e}_j\rangle = 0 \text{ and } \langle \mathbf{e}_i, \mathbf{e}_j\rangle = \delta_{ij} \text{ for } i, j = 1, \ldots, n-1.$$

The n-vectors $\langle \omega, \mathbf{e}_1, \ldots, \mathbf{e}_{n-1}\rangle$ are an orthonormal basis for $\mathbb{R}^n$. Define new orthogonal coordinates, $(t, s_1, \ldots, s_{n-1})$ on $\mathbb{R}^n$ by setting

$$\mathbf{x} = t\omega + \sum_{j=1}^{n-1} s_j\mathbf{e}_j.$$

The n-dimensional Radon transform is defined by

$$\mathrm{R}f(t,\omega) = \int\limits_{l_{t,\omega}} f d\sigma_{n-1} = \int\limits_{\mathbb{R}^{n-1}} f(t\omega + \sum s_j\mathbf{e}_j)ds_1\cdots ds_{n-1}.$$

As before the Radon transform is an even function

$$\mathrm{R}f(t,\omega) = \mathrm{R}f(-t,-\omega).$$

With this definition, the n-dimensional analogue of the Central slice theorem is

Theorem 4.7.1 (Central slice theorem). *If f is an absolutely integrable function on $\mathbb{R}^n$ then*

$$\widetilde{\mathrm{R}f}(r,\omega) = \int_{-\infty}^{\infty} \mathrm{R}f(t,\omega)e^{-irt}dt = \hat{f}(r\omega). \tag{4.84}$$

The central slice theorem and the Fourier inversion formula give the Radon inversion formula.

Theorem 4.7.2 (The Radon Inversion Formula). *Suppose that f is a smooth function with bounded support on $\mathbb{R}^n$ then*

$$f(\mathbf{x}) = \frac{1}{2(2\pi)^n} \int_{S^{n-1}} \int_{-\infty}^{\infty} \widetilde{\mathrm{R}f}(r,\omega) r^{n-1} e^{ir\langle \omega, \mathbf{x}\rangle} dr d\omega. \tag{4.85}$$

Remark 4.7.1. This formula holds in much greater generality. Under the hypotheses in the theorem all the integrals converge absolutely and the simplest form of the Fourier inversion formula applies.

This formula takes a very simple form if the dimension is odd, set $n = 2k+1$. In this case the r-integral in (4.85) can be computed explicitly:

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \widetilde{\mathrm{R}f}(r,\omega) r^{n-1} e^{ir\langle \omega, \mathbf{x}\rangle} dr = (-1)^k \partial_t^{2k} \mathrm{R}f(t, \langle \omega, \mathbf{x}\rangle). \tag{4.86}$$

Using this expression in (4.85) we obtain

$$f(\mathbf{x}) = \frac{(-1)^k}{2(2\pi)^{2k}} \int_{S^{n-1}} (\partial_t^{2k} \mathrm{R}f)(\langle \omega, \mathbf{x}\rangle, \omega) d\omega.$$

Thus in odd dimensions the inverse of the Radon transform is essentially a backprojection.

The Laplace operator on $\mathbb{R}^n$ is defined by

$$\Delta_{\mathbb{R}^n} f = \sum_{j=1}^{n} \partial_{x_j}^2 f.$$

It is invariant under rotations so it follows that, for the coordinates $(t, s_1, \ldots, s_{n-1})$ introduced above we also have the formula

$$\Delta_{\mathbb{R}^n} f = \partial_t^2 f + \sum_{j=1}^{n-1} \partial_{s_j}^2 f. \tag{4.87}$$

This formula allows us to establish a connection between $\mathrm{R}(\Delta_{\mathbb{R}^n} f)$ and $\mathrm{R}f$.

Proposition 4.7.1. *Suppose that f is a twice differentiable function of bounded support on $\mathbb{R}^n$ then*

$$\mathrm{R}(\Delta_{\mathbb{R}^n} f) = \partial_t^2 \mathrm{R}f. \tag{4.88}$$

We close our discussion by explaining how the Radon transform can be applied to solve the wave equation. Let τ denote the time variable and c the speed of sound. The "wave equation" for a function $u(\mathbf{x};\tau)$ defined on $\mathbb{R}^n \times \mathbb{R}$ is

$$\partial_\tau^2 u = c^2 \Delta_{\mathbb{R}^n} u.$$

If u satisfies this equation then it follows from the proposition that, for each $\omega \in S^{n-1}$, $\mathrm{R}u(t,\omega;\tau)$ satisfies the equation

$$\partial_\tau^2 Ru = c^2 \partial_t^2 Ru.$$

Here $Ru(t,\omega;\tau)$ is the Radon transform of u in the $\mathbf{x}$-variables with τ the time parameter. In other words, the Radon transform translates the problem of solving the wave equation in n-dimensions into the problem of solving a family of wave equations in 1-dimension.

The one dimensional wave equation is solved by any function of the form

$$v(t;\tau) = g(ct+\tau) + h(ct-\tau).$$

The initial data is usually $v(t;0)$ and $v_\tau(t;0)$; it is related to g and h by

$$\begin{aligned} g(ct) &= \frac{1}{2}\left[v(t;0) + c\int_{-\infty}^{t} v_\tau(s;0)ds \right], \\ h(ct) &= \frac{1}{2}\left[v(t;0) - c\int_{-\infty}^{t} v_\tau(s;0)ds \right]. \end{aligned} \tag{4.89}$$

If $u(\mathbf{x};0) = u_0(\mathbf{x})$ and $u_\tau(\mathbf{x};0) = u_1(\mathbf{x})$ then we see that

$$\mathrm{R}u(t,\omega;\tau) = g(ct+\tau;\omega) + h(ct-\tau;\omega)$$

where

$$\begin{aligned} g(ct;\omega) &= \frac{1}{2}\left[\mathrm{R}u_0(t;\omega) + c\int_{-\infty}^{t} \mathrm{R}u_1(s;\omega)ds \right], \\ h(ct;\omega) &= \frac{1}{2}\left[\mathrm{R}u_0(t;\omega) - c\int_{-\infty}^{t} \mathrm{R}u_1(s;\omega)ds \right]. \end{aligned} \tag{4.90}$$

Using these formulæ along with (4.85) one can obtain an explicit for the solution of the wave equation.

Exercise 4.7.1. Prove the central slice theorem.

Exercise 4.7.2. Let $n = 2k+1$ and suppose that f is a function for which

$$\mathrm{R}f(t,\omega) = 0 \text{ if } |t| < R.$$

Prove that $f(\mathbf{x}) = 0$ if $\|\mathbf{x}\| < R$. Is this true in even dimensions?

Exercise 4.7.3. Prove formula (4.87).

Exercise 4.7.4. Prove Proposition (4.7.1) . Hint: Integrate by parts.

Exercise 4.7.5. Use the simplified version of the Radon inversion formula available for $n = 3$ to derive an explicit formula for the solution of the wave equation in 3 space dimensions in terms of the initial data $u_0(x)$ and $u_1(x)$.

4.8 The Hilbert transform and complex analysis*

In the earlier part of the chapter we used several explicit Hilbert transforms, here we explain how these computations are done. We restrict to the case of square integrable functions. If $f \in L^2(\mathbb{R})$ with Fourier transform f then

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ix\xi} \hat{f}(\xi) d\xi.$$

Define two L^2-functions

$$\begin{aligned} f_+(x) =& \frac{1}{2\pi} \int_0^{\infty} e^{ix\xi} \hat{f}(\xi) d\xi, \\ f_-(x) =& \frac{1}{2\pi} \int_{-\infty}^{0} e^{ix\xi} \hat{f}(\xi) d\xi. \end{aligned} \tag{4.91}$$

Obviously we have that $f = f_+ + f_-$ and $\mathcal{H}f = f_+ - f_-$. This decomposition is useful because the function $f_+(x)$ has an extension as an analytic function in the upper half plane, $H_+ = \{x + iy \ : \ y > 0\}$

$$f_+(x + iy) = \frac{1}{2\pi} \int_0^{\infty} e^{i(x+iy)\xi} \hat{f}(\xi) d\xi.$$

Observe that the Fourier transform of $f_+(x+iy)$ in the x-variable is just $\hat{f}(\xi)\chi[0,\infty)(\xi)e^{-y\xi}$. Since $y\xi > 0$ we see that $f_+(x + iy)$ is in $L^2(\mathbb{R})$ for each $y \geq 0$. A similar analysis shows that f_- has an analytic extension to the lower half plane, $H_- = \{x+iy \ : \ y < 0\}$ such that $f_-(x + iy) \in L^2(\mathbb{R})$ for each $y \leq 0$. Indeed it is not hard to show that this decomposition is unique. The precise statement is the following.

Proposition 4.8.1. *Suppose that $F(x + iy)$ is an analytic function in H_+ such that for $y \geq 0$*

(1).

$$\int_{-\infty}^{\infty} |F(x + iy)|^2 dx < M,$$

(2).

$$\lim_{y\downarrow 0}\int_{-\infty}^{\infty}|F(x+iy)|^2dx=0$$

then $F\equiv 0$.

Proof. By Theorem 3.2.9, a function satisfying the L^2-boundedness condition has the following property

$$\hat{F}(\cdot+iy)=\hat{f}(\xi)e^{-y\xi}$$

where $\hat{f}(\xi)$ is the Fourier transform $F(x)$. Moreover $\hat{f}(\xi)=0$ if $\xi<0$. By the Parseval formula

$$\int_{-\infty}^{\infty}|F(x+iy)|^2dx=\int_{0}^{\infty}|\hat{f}(\xi)|^2e^{-2y\xi}d\xi.$$

The second condition implies that $\hat{f}(\xi)=0$ and therefore $F\equiv 0$. □

If the functions $f_\pm$ can be explicitly determined then $\mathcal{H}f$ can also be computed. If f is a "piece" of an analytic then this determination is often possible. The following example is typical.

Example 4.8.1. Let

$$f(x)=\begin{cases}\sqrt{1-x^2} & \text{for } |x|<1,\\ 0 & \text{for } |x|\geq 1.\end{cases}$$

The analytic function, $\sqrt{1-z^2}$ has a single valued determination in the complex plane minus the subset of $\mathbb{R}$, $\{x\,:\,|x|\geq 1\}$. Denote this function by $F(z)$. Of course $F(x)=f(x)$ for $x\in(-1,1)$ and the restrictions of F to the upper and lower half planes, $F_\pm$ are analytic. Moreover for $|x|>1$ we easily compute that

$$\lim_{\epsilon\downarrow 0}F_+(x+i\epsilon)+F_-(x-i\epsilon)=0.$$

This would solve our problem but for the fact that $F(x+iy)$ is not in L^2 for any $y\neq 0$. To fix this problem we need to add a correction term that reflects the asymptotic behavior of $F(z)$ for large z. Indeed if we set

$$f_\pm(z)=\frac{1}{2}[F_\pm(z)\pm iz]$$

then a simple calculation shows that

$$f_+(x)+f_-(x)=f(x) \text{ for all real } x$$

and that

$$f_\pm(x\pm iy)\simeq\frac{1}{x} \text{ for large } x$$

and therefore $f_\pm(x\pm iy)\in L^2(\mathbb{R})$ for all $y>0$. This allows us to compute the Hilbert transform of f

$$\mathcal{H}f(x)=\begin{cases}ix & \text{for } |x|<1,\\ i(x+\sqrt{x^2-1}) & \text{for } x<-1,\\ i(x-\sqrt{x^2-1}) & \text{for } x>1.\end{cases}\tag{4.92}$$

Exercise 4.8.1. Compute the Hilbert transform of $\chi_{[-1,1]}(x)$. A good place to start is with the formula $\mathcal{H}f = \lim_{\epsilon \downarrow 0} h_\epsilon * f$, see section 4.3.

Chapter 5

Introduction to Fourier series

In engineering applications data is never collected along the whole real line or on the entire plane. Real data can only be collected from a bounded time interval or planar domain. In order to use Fourier analysis to analyze and filter this type of data we can either

- Extend the data by cutting it off to equal zero outside of the set over which the data was collected, or
- Extend the function periodically.

If the data is extended "by zero" then the Fourier transform is available. If the data is extended periodically then it does not vanish at infinity, in any sense, and hence its Fourier transform is not a function. Fourier series provides an alternate tool for the analysis of functions defined on finite intervals in $\mathbb{R}$ or products of intervals in $\mathbb{R}^n$. The goal of Fourier series is to express an "arbitrary" periodic function as a linear combination of complex exponentials. This theory runs parallel to that of the Fourier transform presented in Chapter 3. After running through the basic results in the one dimensional case we give a detailed analysis of the Gibbs phenomenon. The chapter concludes with a rapid presentation of Fourier series in $\mathbb{R}^n$.

5.1 Fourier series in one dimension

See: A.3.1, A.3.2, A.6.1, B.4.

The fundamental tool for analyzing functions defined on finite intervals is the Fourier series. To simplify the exposition, we begin with functions defined on the interval $[0, 1]$. If $g(x)$ is a function defined on an interval $[a, b]$ then letting $f(x) = g(a + (b - a)x)$ gives a function defined on $[0, 1]$ which evidently contains the same information as f.

Definition 5.1.1. Let f be an *absolutely integrable* function defined on $[0, 1]$, that is

$$\|f\|_{L^1} = \int_0^1 |f(x)|dx < \infty.$$

The set of such functions is a complete normed, linear space with norm defined by $\|\cdot\|_{L^1}$. It is denoted by $L^1([0,1])$. Define the *Fourier coefficients* of $f \in L^1([0,1])$ by

$$\hat{f}(n) = \int_0^1 f(x)e^{-2\pi inx}dx \quad \text{for} \quad n \in \mathbb{Z}. \tag{5.1}$$

Example 5.1.1. If $f(x) = \cos(2\pi mx)$ then, using the formula $\cos(y) = 2^{-1}(e^{iy} + e^{-iy})$ we easily compute that

$$\hat{f}(n) = \begin{cases} \frac{1}{2} & \text{if } n = \pm m, \\ 0 & \text{if } n \neq \pm m. \end{cases}$$

Example 5.1.2. Let $0 \leq a < b < 1$, then the Fourier coefficients of $\chi_{[a,b]}(x)$ are

$$\begin{aligned} \widehat{\chi}_{[a,b]}(n) &= \int_a^b e^{-2\pi inx}dx \\ &= \begin{cases} (b-a) & \text{if } n = 0, \\ \frac{e^{-2\pi inb} - e^{-2\pi ina}}{2\pi in}. \end{cases} \end{aligned} \tag{5.2}$$

Example 5.1.3. Let $f(x) = \sin(\pi x)$, again using the expression for the sine in terms of exponentials we compute

$$\hat{f}(n) = \frac{-2}{\pi}\left[\frac{1}{4n^2 - 1}\right]. \tag{5.3}$$

The symmetry properties of Fourier coefficients are summarized in a proposition.

Proposition 5.1.1. *Let f be an integrable function on $[0,1]$.*

(1). If f is real valued then

$$\hat{f}(-n) = \overline{\hat{f}(n)}.$$

(2). f is even if $f(x) = f(1-x)$ for all $x \in [0,1]$. if f is real valued and even then its Fourier coefficients are real.

(3). f is odd if $f(x) = -f(1-x)$ for all $x \in [0,1]$. If f is real valued and odd then its Fourier coefficients are purely imaginary.

The proof of the proposition is left as an exercise.

To study the inverse of the Fourier series we introduce the *partial sum operator*.

Definition 5.1.2. Let f be an absolutely integrable function on $[0,1]$. For each positive integer N, define the N^{th}-partial sum of the Fourier series of f to be

$$S_N(f;x) = \sum_{n=-N}^{N} \hat{f}(n)e^{2\pi inx}. \tag{5.4}$$

For each N the partial sum is linear in f

$$S_N(f+g) = S_N(f) + S_N(g) \text{ and } S_N(af) = aS_N(f) \text{ for } a \in \mathbb{C}.$$

In applications one works with a fixed partial sum, so it is very important to understand in what sense $S_N(f;x)$ is an approximation to $f(x)$. The best one might hope for is that

$$\lim_{N\to\infty} S_N(f;x) = f(x)$$

at every point x. At discontinuities of f such a statement seems very unlikely to be true. In fact, it can even fail at points where f is continuous. The pointwise convergence of Fourier series is a very subtle problem. For the simplest result we make a strong hypothesis about the rate of decay of the Fourier coefficients.

Proposition 5.1.2 (Fourier inversion formula). *If f is a continuous function defined on $[0,1]$ such that the Fourier coefficients of f satisfy*

$$\sum_{n=-\infty}^{\infty} |\hat{f}(n)| < \infty \tag{5.5}$$

then $f(x)$ is represented, at every point by its uniformly convergent Fourier series

$$f(x) = \sum_{n=-\infty}^{\infty} \hat{f}(n) e^{2\pi i n x}, \text{ for all } x \in [0,1]. \tag{5.6}$$

Proof. For $0 < r < 1$, define the absolutely convergent series

$$P_r(x) = \sum_{n=-\infty}^{\infty} r^{|n|} e^{2\pi i n x} = 1 + 2\,\mathrm{Re}\left[\sum_{n=1}^{\infty} r^n e^{2\pi i n x}\right].$$

Using the second expression and the formula for the sum of a geometric series we see that

$$P_r(x) = \frac{1-r^2}{1-2r\cos(2\pi x)+r^2}.$$

For $0 \le r < 1$ this formula implies that $P_r(x) > 0$. For each such r define

$$f_r(x) = \int_0^1 P_r(x-y) f(y)\, dy.$$

From the representation of P_r as an infinite sum we deduce that

$$f_r(x) = \sum_{j=-\infty}^{\infty} \hat{f}(n) r^{|n|} e^{2\pi i n x}.$$

In light of (5.5), the comparison test for infinite sums, Theorem B.4.2 implies that

$$\lim_{r\uparrow 1} f_r(x) = \sum_{j=-\infty}^{\infty} \hat{f}(n) e^{2\pi i n x},$$

with uniform convergence for $x \in \mathbb{R}$. To complete the proof we must show that $\lim_{r\uparrow 1} f_r(x) = f(x)$.

For each r

$$\int_0^1 P_r(x)dx = 1.$$

This fact and the positivity of P_r imply that

$$\begin{aligned} |f_r(x) - f(x)| &= \left| \int_0^1 P_r(x-y)(f(y) - f(x))dy \right| \\ &\leq \int_0^1 P_r(x-y)|f(y) - f(x)|dy. \end{aligned} \tag{5.7}$$

If $x \neq 0$ then

$$\lim_{r\uparrow 1} P_r(x) = 0.$$

In fact if $\epsilon > 0$ is fixed then there is a $0 < \delta$ so that

$$P_r(x) < \epsilon \text{ if } r > 1 - \delta \text{ and } \epsilon < x < 1 - \epsilon. \tag{5.8}$$

To show that the difference $f(x) - f_r(x)$ becomes small we break the integral into two pieces. One piece is small because f is continuous; the other is small because f is bounded and $P_r(x)$ is small, if x is far enough from 0. Since f is continuous, given $\eta > 0$ there is an $\epsilon > 0$ so that

$$|x - y| < \epsilon \text{ implies that } |f(x) - f(y)| < \eta.$$

There is also an M so that $|f(y)| \leq M$ for all y. Let $\epsilon' = \min\{\epsilon, \eta\}$. Using (5.8), with ϵ', there is a $\delta > 0$ so that $r > 1 - \delta$ implies that

$$\begin{aligned} |f_r(x) - f(x)| &\leq \int_0^1 P_r(x-y)|f(y) - f(x)|dy \\ &= \int_{|x-y|<\epsilon'} P_r(x-y)|f(y) - f(x)|dy + \int_{|x-y|>\epsilon'} P_r(x-y)|f(y) - f(x)|dy \\ &\leq \eta \int_{|x-y|<\epsilon'} P_r(x-y)dy + 2M\epsilon' \\ &\leq (1 + 2M)\eta. \end{aligned} \tag{5.9}$$

This estimate shows that

$$\lim_{r\uparrow 1} f_r(x) = f(x)$$

and thereby completes the proof of the Theorem. □

Remark 5.1.1. The argument in the second part of the proof does not require f to be everywhere continuous. It shows that, if f is bounded then $f_r(x)$ converges to $f(x)$ at any point of continuity of f. This does not mean that the Fourier series of f also converges to f at such a point. The hypothesis (5.5), on the Fourier coefficients of f, implies that f is a continuous function. Exercise 5.3.7 outlines a proof of the Fourier inversion formula, without the assumption that f is continuous.

As was the case with the Fourier integral, the Fourier coefficients of an absolutely integrable function f may fail to satisfy (5.5). For example, let

$$f(x) = \begin{cases} 1 & \text{if } x \in [0, \frac{1}{2}], \\ 0 & \text{if } x \in (\frac{1}{2}, 1]. \end{cases}$$

A simple calculation shows that

$$\hat{f}(n) = \begin{cases} 0 & \text{if } n \text{ is even and } n \neq 0, \\ \frac{1}{2} & \text{for } n = 0, \\ \frac{1}{\pi i n} & \text{if } n \text{ is odd.} \end{cases}$$

In this example the function has a jump discontinuity.

If the Fourier coefficients satisfy (5.5) then the partial sums $S_N(f)$ converge uniformly to f. Since each partial sum is a continuous function it follows from an elementary result in analysis that the limit is also a continuous function. Because discontinuous data is common in imaging applications, the difficulties of representing such functions in terms of Fourier series is an important topic. It is considered, in detail, in section 5.5. The only general result on the decay of Fourier coefficients for absolutely integrable functions is the Riemann Lebesgue Lemma.

Theorem 5.1.1 (Riemann Lebesgue Lemma). *If* $\int_0^1 |f(x)|dx < \infty$, *then* $\hat{f}(n) \to 0$ *as* $|n| \to \infty$.

Proof. The proof of this theorem is a bit indirect. Because the class of integrable functions is very large, it is useful to first consider a simpler class of functions. Suppose that $f(x)$ is a step function, that is

$$f(x) = \sum_{j=1}^{N} c_j \chi_{[a_j, b_j)}(x).$$

The Fourier transform of f is therefore

$$\hat{f}(n) = \sum_{j=1}^{N} c_j \widehat{\chi}_{[a_j, b_j)}(n).$$

Since f is a *finite* sum of step functions, formula (5.2) implies that there is a constant M so that

$$|\hat{f}(n)| \leq \frac{M}{(|n| + 1)}.$$

Thus step functions satisfy the Riemann Lebesgue lemma. To complete the proof we use the fact that an integrable function can be approximated by step functions.

Let f be an arbitrary integrable function. Fix an $\epsilon > 0$, according to Theorem A.6.2 there is a step function g so that $\|f - g\|_{L^1} < \epsilon$. We need to compare the Fourier coefficients of f and g :

$$\begin{aligned} |\hat{f}(n) - \hat{g}(n)| &= \left| \int_0^1 (f(x) - g(x)) e^{-2\pi i n x} dx \right| \\ &\leq \int_0^1 |f(x) - g(x)| dx \\ &\leq \epsilon. \end{aligned} \tag{5.10}$$

The triangle inequality shows that

$$|\hat{f}(n)| \leq |\hat{f}(n) - \hat{g}(n)| + |\hat{g}(n)|.$$

Taking the $\limsup_{n\to\infty}$ in this estimate we see that

$$\limsup_{n\to\infty} |\hat{f}(n)| \leq \epsilon.$$

Since ϵ is an arbitrary positive number this shows that

$$\lim_{n\to\infty} |\hat{f}(n)| = 0.$$

□

The proof of the Riemann Lebesgue Lemma uses a very important continuity result for the Fourier coefficients. For later reference we state it as a proposition.

Proposition 5.1.3. *If $f \in L^1([0,1])$ then*

$$|\hat{f}(n)| \leq \|f\|_{L^1} \text{ for all } n \in \mathbb{N}. \tag{5.11}$$

The Riemann Lebesgue does not say that $\hat{f}(n)$ goes to zero at some particular rate, say faster than $n^{-1/3}$. In fact, there is a theorem saying that for any given sequence, $\{a_n\}_{-\infty}^{\infty}$, with

$$\lim_{|n|\to\infty} a_n = 0,$$

there exists an integrable function f, whose Fourier coefficients $\{\hat{f}(n)\}_{-\infty}^{\infty}$, satisfy $|\hat{f}(n)| \geq |a_n|$ for all n. This shows that Fourier coefficients can go to zero arbitrarily slowly, see [40][section I.4]. As with the Fourier integral the smoothness of $f(x)$ and the rate of decay of its Fourier coefficients are intimately related. This is the topic of the next section.

Exercise 5.1.1. Compute the Fourier coefficients of $\sin(2\pi m x)$.

Exercise 5.1.2. Find a more explicit formula for $\widehat{\chi}_{[\frac{1}{4},\frac{3}{4}]}$.

Exercise 5.1.3. Compute the Fourier coefficients of $\cos(\frac{\pi x}{2})$.

Exercise 5.1.4. Prove Proposition 5.1.1.

Exercise 5.1.5. Show that $P_r(x) > 0$ and has total integral 1 for any r.

Exercise 5.1.6. Show that if $\epsilon > 0$ is fixed then there is a $0 < \delta$ so that

$$P_r(x) < \epsilon \text{ if } r > 1 - \delta \text{ and } \epsilon < x < 1 - \epsilon. \tag{5.12}$$

Exercise 5.1.7. Explain why (5.11) is a "continuity" result for the map $f \mapsto < \hat{f}(n) >$.

Exercise 5.1.8. Use summation by parts *twice* to show that

$$f(x) = \sum_{n=2}^{\infty} \frac{\cos(2\pi n x)}{\log n}$$

represents a non-negative, integrable function. In light of this it is a remarkable fact that

$$\sum_{n=2}^{\infty} \frac{\sin(2\pi n x)}{\log n}$$

does **not** represent an absolutely integrable function!

5.2 The decay of Fourier coefficients for differentiable functions

See: A.4.1, B.4.

If the sum in (5.6) converges uniformly, then $f(x)$ is continuous and necessarily satisfies $f(x+1) = f(x)$, for all x since

$$e^{2\pi i n(x+1)} = e^{2\pi i n x} \text{ for all } n.$$

This shows that, *when discussing Fourier series,* we should only consider a function on $[0,1]$ to be continuous if both

$$\lim_{y \to x} f(y) = f(x) \text{ for } x \in (0,1)$$

and

$$\lim_{x \to 0^+} f(x) = f(0) = f(1) = \lim_{x \to 1^-} f(x).$$

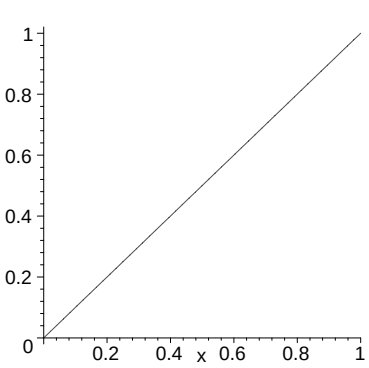
That is, we think of f as a periodic function of period 1. A function, f defined on $[0,1)$ is extended, periodically to the whole real line by letting

$$f(x+n) = f(x) \text{ for all } n \in \mathbb{Z}.$$

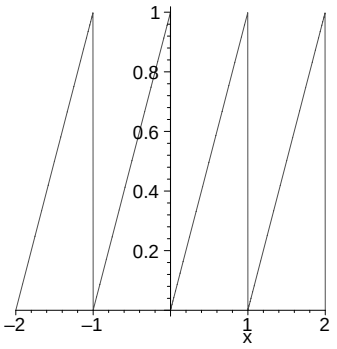
If f is a continuous function on $[0,1]$ in the usual sense then the condition $f(0) = f(1)$ is equivalent to the condition that the 1-periodic extension of f to $\mathbb{R}$ is continuous. These considerations easily extend to the derivatives of f. A k-times differentiable function f defined on $[0,1]$ is k-times differentiable *as a periodic function* provided that

$$f^{[j]}(0) = f^{[j]}(1) \text{ for } j = 0, 1, \ldots, k.$$

Example 5.2.1. The function $f(x) = x$ is a continuous function on $[0, 1]$ however its 1-periodic extension is not, see the graphs.



(a) f.



(b) The 1-periodic extension of f.

Figure 5.1: Periodic extension may turn a continuous function into discontinuous function.

Suppose that f is continuously differentiable on $[0, 1]$; for the moment we do **not** assume that $f(0) = f(1)$. Integrating by parts gives

$$\begin{aligned}\hat{f}(n) = \int_0^1 f(x)e^{-2\pi inx}dx &= \frac{1}{-2\pi in}f(x)e^{-2\pi inx}\Big|_0^1 + \frac{1}{2\pi in}\int_0^1 f'(x)e^{-2\pi inx}dx \\ &= \frac{f(1)-f(0)}{-2\pi in} + \frac{1}{2\pi in}\int_0^1 f'(x)e^{-2\pi inx}dx.\end{aligned} \tag{5.13}$$

In other words

$$\hat{f}(n) = \frac{1}{2\pi in}[f(0) - f(1)] + \frac{\widehat{f'}(n)}{2\pi in}.$$

Since $f(x)$ is continuously differentiable in $[0, 1]$, its first derivative is an integrable function. By the Riemann Lebesgue lemma, the Fourier coefficients of $f'(x)$ go to zero as n goes to infinity. This shows that if $f(0) \neq f(1)$ then $\hat{f}(n)$ decays exactly as $1/n$. The $1/n$ rate of decay is characteristic of a function with a simple jump discontinuity. If $f(0) = f(1)$ then the Fourier coefficients of f are given by

$$\hat{f}(n) = \frac{\widehat{f'}(n)}{2\pi in}$$

which therefore go to zero faster than $1/n$.

If f has $(k-1)$-derivatives, as a periodic function then this integration by parts can be repeated to obtain:

Theorem 5.2.1. *If $f \in C^k([0,1])$ and $f(0) = f(1)$, $f'(0) = f'(1), \ldots, f^{(k-1)}(0) = f^{(k-1)}(1)$, then*

$$\hat{f}(n) = \frac{\widehat{f^{(k)}}(n)}{(2\pi i n)^k} \text{ for } n \neq 0. \tag{5.14}$$

The Riemann-Lebesgue lemma then implies that if f is $k-$times continuously differentiable on $[0,1]$, periodic in the appropriate sense then, $\hat{f}(n)$ decays *faster* than $1/n^k$. This result has a partial converse.

Theorem 5.2.2. *If $f \in L^1([0,1])$ and there is a constant C and an $\epsilon > 0$, so that*

$$|\hat{f}(n)| \leq \frac{C}{(1+|n|)^{k+\epsilon}}$$

then f is in $C^{(k-1)}([0,1])$ with $f(0) = f(1)$, $f'(0) = f'(1), \ldots, f^{(k-1)}(0) = f^{(k-1)}(1)$

Proof. Using the comparison test, Theorem B.4.2, the estimates on $< \hat{f}(n) >$ imply that the series $\sum \hat{f}(n) e^{2\pi i n x}$ and its j^{th} derivatives, for $0 \leq j \leq k-1$, converge uniformly and absolutely. Theorem A.4.1 implies that

$$\sum_{n=-\infty}^{\infty} \hat{f}(n) e^{2\pi i x}$$

represents a $(k-1)$-times, continuously differentiable function and that we can differentiate, term by term to obtain

$$f^{[j]}(x) = \sum_{n=-\infty}^{\infty} (2\pi i n)^j \hat{f}(n) e^{2\pi i n x}.$$

See [67]. □

As before, formula (5.14) can be viewed as a formula for the Fourier coefficients of $f^{[j]}$ in terms of those of f.

Corollary 5.2.1. *If $f(x)$ has k integrable, derivatives on $[0,1]$ with $f^{[j]}(0) = f^{[j]}(1)$ for $0 \leq j \leq k-1$ then, for $j \leq k$*

$$\widehat{f^{[j]}}(n) = [2\pi i n]^j \hat{f}(n). \tag{5.15}$$

For the case of the Fourier series, it is very important that the derivatives are also periodic functions.

Example 5.2.2. Let $f(x) = x(1-x)$ for $x \in [0,1]$, then $f(0) = f(1)$ but $f'(x) = 1 - 2x$, is not continuous as a 1-periodic function. The Fourier coefficients of f are

$$\hat{f}(n) = \begin{cases} \frac{1}{6}, & \text{for } n = 0, \\ \frac{-1}{2\pi^2 n^2}, & \text{for } n \neq 0. \end{cases} \tag{5.16}$$

The Fourier coefficients of f' are

$$\widehat{f'}(n) = \begin{cases} 0, & \text{for } n = 0, \\ \frac{1}{\pi i n}, & \text{for } n \neq 0, \end{cases} \tag{5.17}$$

showing that $\hat{f}(n) = (2\pi i n)\widehat{f'}(n)$. Note also the relationship between the smoothness of f, as a 1-periodic function and the decay of its Fourier coefficients.

Example 5.2.3. If we set

$$f(x) = x - n \text{ for } x \in (n, n+1]$$

then f does not have *any* periodic derivatives. The Fourier coefficients of f are given by

$$\hat{f}(n) = \begin{cases} \frac{1}{2} \text{ for} & n = 0, \\ \frac{i}{2\pi n} \text{ for} & n \neq 0. \end{cases}$$

They display the $\frac{1}{n}$-rate of decay which is characteristic of functions with simple jump discontinuities. On the interval $[0,1]$ $f'(x) = 1$; note however that $\widehat{f'}(n) \neq (2\pi i n)\hat{f}(n)$, for *any* n.

5.3 L^2-theory

See: A.4.5.

Theorem 5.2.2 is not the exact converse to Theorem 5.2.1, though it is closely analogous to the results for the Fourier transform, and reflects the subtlety of pointwise convergence for Fourier series. Simpler statements are obtained by using the L^2-norm. Recall that

$$L^2([0,1]) = \{f \ : \ \int_0^1 |f(x)|^2 dx < \infty\};$$

a norm is defined on $L^2([0,1])$ by setting

$$\|f\|_2 = \left[\int_0^1 |f(x)|^2 dx\right]^{1/2}.$$

With this norm $L^2([0,1])$ is a complete normed linear space. An element of $L^2([0,1])$ is called a square integrable or square summable function.

5.3.1 Geometry in $L^2([0,1])$.

See: A.2.5.

The L^2 norm, $\|\cdot\|_2$ defines a notion of a distance between functions,

$$d_{L^2}(f,g) = \|f - g\|_2 = \sqrt{\int_0^1 |f(x) - g(x)|^2 dx}. \tag{5.18}$$

The norm on $L^2([0,1])$ is defined by the inner product

$$\langle f, g\rangle_{L^2} = \int f(x)\overline{g(x)}dx.$$

Since $z\bar{z} = |z|^2$

$$\langle f, f\rangle_{L^2} = \int |f|^2 dx = \|f\|_2^2.$$

Recall the Cauchy-Schwarz inequality for $L^2([0,1])$.

Theorem 5.3.1 (Cauchy-Schwarz inequality). *If f, g are two functions in $L^2([0,1])$ then*

$$|\langle f, g\rangle_{L^2}| \leq \|f\|_2\|g\|_2. \tag{5.19}$$

Proof. As with earlier cases, the proof uses elementary calculus and the fact that $\langle f, f\rangle_{L^2}$ is always non-negative. Suppose that f and g are non-zero and consider the quadratic function of $t \in \mathbb{R}$ defined by

$$F(t) = \langle f + tg, f + tg\rangle_{L^2} = \|f\|_2^2 + 2t\,\mathrm{Re}\langle f, g\rangle_{L^2} + t^2\|g\|_2^2.$$

Such a function assumes its unique, minimum value where

$$F'(t) = 2\,\mathrm{Re}\langle f, g\rangle_{L^2} + 2t\|g\|_2^2 = 0.$$

That is $t_0 = -\,\mathrm{Re}\langle f, g\rangle_{L^2}\|g\|_2^{-2}$. Since $F(t) \geq 0$ for all t we see that

$$[\mathrm{Re}\langle f, g\rangle_{L^2}]^2 \leq \|f\|_2^2\|g\|_2^2.$$

To complete the proof we replace g by $e^{i\theta}g$ where θ chosen so that

$$\langle f, e^{i\theta}g\rangle_{L^2} = e^{-i\theta}\langle f, g\rangle_{L^2} = |\langle f, g\rangle_{L^2}|.$$

□

The Cauchy-Schwarz inequality, (5.19) implies that

$$-1 \leq \frac{\langle f, g\rangle_{L^2}}{\|f\|_2\|g\|_2} \leq 1.$$

We can therefore define an angle θ between f and g in $L^2([0,1])$ by setting

$$\cos\theta = \frac{\langle f, g\rangle_{L^2}}{\|f\|_2\|g\|_2}.$$

This gives a reasonable notion of orthogonality: Two functions $f, g \in L^2([0,1])$ are orthogonal if the angle between them is $\frac{\pi}{2}$ that is

$$\langle f, g\rangle_{L^2} = 0.$$

The function space L^2 is very special among infinite dimensional spaces because its norm is defined by an inner product. This makes it possible to do many finite dimensional Euclidean geometric constructions in this context as well. For example

$$\langle e^{2\pi inx}, e^{2\pi imx}\rangle_{L^2} = \int_0^1 e^{2\pi i(n-m)x}dx = \begin{cases} 1 & \text{if } n = m, \\ 0 & \text{if } n \neq m. \end{cases} \tag{5.20}$$

In other words the vectors $\{e^{2\pi inx} \mid n \in \mathbb{Z}\}$ are pairwise orthogonal and each has length 1.

A subset $S \subset L^2([0,1])$ is called a subspace if

- Whenever $f, g \in S$ then $f + g \in S$ as well.
- If $f \in S$ and $a \in \mathbb{C}$ then $af \in S$.

If $S \subset L^2([0,1])$ is a subspace then we say that f is orthogonal to S if

$$\langle f, g \rangle_{L^2} = 0 \text{ for every } g \in S.$$

Definition 5.3.1. Let S be a subspace of $L^2([0,1])$. The *orthogonal complement* of S, denoted $S^\perp$ is the subspace of $L^2([0,1])$ consisting of all functions $g \in L^2([0,1])$ orthogonal to S.

Connected to a subspace we have orthogonal projections.

Definition 5.3.2. Let S be a subspace of $L^2([0,1])$. The *orthogonal projection* onto S is a linear map $P_S : L^2([0,1]) \to L^2([0,1])$ with the following properties:

(1). $P_S^2 = P_S$, (a linear map with this property is called a projection),

(2). If $f \in S$ then $P_S(f) = f$,

(3). If $f \in S^\perp$ then $P_S(f) = 0$.

It is a basic result in functional analysis that such an orthogonal projection always exists, see [16].

If f is a square summable function on $[0,1]$ then it is also absolutely integrable, hence its Fourier coefficients are defined. In this case the Fourier coefficients of f go to zero sufficiently fast to make $\sum |\hat{f}(n)|^2$ converge. Once again we have a Parseval formula.

Theorem 5.3.2 (Parseval Formula). *If $f \in L^2([0,1])$ then*

$$\int_0^1 |f(x)|^2 dx = \sum_{n=-\infty}^{\infty} |\hat{f}(n)|^2.$$

Proof. Again the theorem is simple to prove for a special class of L^2-functions, in this instance, the trigonometric polynomials,

$$\mathcal{T} = \{ \sum_{j=-N}^{N} c_j e^{2\pi i j x} \ : \ c_j \in \mathbb{C} \}.$$

If $f \in \mathcal{T}$ then multiplying out the finite sum defining $f\bar{f}$ gives

$$|f(x)|^2 = \sum_{j,k=-N}^{N} c_j \bar{c}_k e^{2\pi i (j-k) x}. \tag{5.21}$$

Integrating both sides of (5.21), using (5.20) gives

$$\int_0^1 |f(x)|^2 dx = \sum_{j=-N}^{N} |c_j|^2.$$

This is the Parseval formula for $f \in \mathcal{T}$.

To complete the proof we need two additional facts. The first is that an arbitrary $f \in L^2([0,1])$ is well approximated by trigonometric polynomials.

Lemma 5.3.1. *If $f \in L^2([0,1])$ and $\epsilon > 0$ is given then there is a $g \in \mathcal{T}$ so that $\|f - g\|_{L^2} < \epsilon$.*

The second is Bessel's inequality. It states that among functions of the form

$$g_N = \sum_{n=-N}^{N} c_n e^{2\pi i n x}$$

the N^{th}-partial sum of the Fourier series of f minimizes the error,

$$\|f - g_N\|_{L^2}.$$

The lemma is proved in section 5.5.2 and Bessel's inequality is proved in section 5.3.2.

The definition of the Fourier coefficients implies that

$$0 \leq \|f - S_N(f)\|_{L^2}^2 = \|f\|_{L^2}^2 - \|S_N(f)\|_{L^2}^2$$

and therefore

$$\|S_N(f)\|_{L^2}^2 \leq \|f\|_{L^2}^2.$$

In particular, using the result for trigonometric polynomials and letting N tend to infinity we deduce that

$$\sum_{n=-\infty}^{\infty} |\hat{f}(n)|^2 \leq \int_0^1 |f(x)|^2 dx. \tag{5.22}$$

On the other hand the triangle inequality gives the estimate

$$\|f\|_{L^2} \leq \|f - S_N(f)\|_{L^2} + \|S_N(f)\|_{L^2}^2.$$

Bessel's inequality, Lemma 5.3.1 and the result for trigonometric polynomials now shows that, for any $\epsilon > 0$

$$\int_0^1 |f(x)|^2 dx \leq \sum_{n=-\infty}^{\infty} |\hat{f}(n)|^2 + \epsilon.$$

Together these inequalities complete the proof of Parseval's formula. □

The Parseval formula gives a criterion for a sequence, $< a_n >$ to be the Fourier coefficients of a square summable function. For example, $a_n = n^{-1/2}$ cannot be the Fourier coefficients of an L^2 function because

$$\sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

The Parseval formula also gives an effective way to compute the values of many infinite sums. A very important example is

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}. \tag{5.23}$$

The proof of this formula is outlined in exercise 5.3.3.

The Parseval formula has a simple geometric interpretation.

Theorem 5.3.3. *The set of exponentials* $\{e^{2\pi inx} | n = -\infty, \ldots, \infty\}$ *is an orthonormal basis for* $L^2([0,1])$.

Remark 5.3.1. The Parseval formula should therefore be regarded as an infinite dimensional version of Pythagoras' theorem.

Parseval's formula also implies that the Fourier series of an L^2-function converges to the function in the L^2-norm.

Proposition 5.3.1. *If* $f \in L^2([0,1])$ *then*

$$\lim_{M,N\to\infty} \|f - \sum_{j=-M}^{N} \hat{f}(j)e^{2\pi ijx}\|_{L^2} = 0. \tag{5.24}$$

Remark 5.3.2. As before it is said that the Fourier series of f converges to f *in the mean*, this is denoted

$$\underset{M,N\to\infty}{LIM} \sum_{j=-M}^{N} \hat{f}(j)e^{2\pi ijx} = f(x).$$

Proof. Given the Parseval formula, the proof is a simple computation using the fact that the L^2-norm is defined by an inner product:

$$\|f - \sum_{j=-M}^{N} \hat{f}(j)e^{2\pi ijx}\|_{L^2}^2 = \|f\|_{L^2}^2 - \sum_{j=-M}^{N} |\hat{f}(j)|^2. \tag{5.25}$$

From the Parseval formula it follows that

$$\|f\|_{L^2}^2 - \sum_{j=M}^{N} |\hat{f}(j)|^2 = \sum_{j=-\infty}^{-(M+1)} |\hat{f}(j)|^2 + \sum_{j=N+1}^{\infty} |\hat{f}(j)|^2. \tag{5.26}$$

As the sum $\sum |\hat{f}(j)|^2$ is finite the right hand side in (5.26) tends to zero as M and N tend to infinity. □

For any $f \in L^2([0,1])$ each partial sum, $S_N(f;x)$ is a very nice function, it is infinitely differentiable and all of its derivatives are periodic. For a general function $f \in L^2([0,1])$ and point $x \in [0,1]$ we therefore do not expect that

$$f(x) = \lim_{N\to\infty} S_N(f;x).$$

In fact the partial sums may not converge pointwise to a limit at all. This means that we need to find a different way to understand the convergence of the Fourier series for L^2-functions.

Example 5.3.1. Define a sequence of coefficients by setting $\hat{f}(n) = n^{-\frac{3}{4}}$ for $n > 0$ and zero otherwise. Because

$$\sum_{n=1}^{\infty} [n^{-\frac{3}{4}}]^2 < \infty$$

these are the Fourier coefficients of an L^2-function. However

$$\sum_{n=1}^{\infty} n^{-\frac{3}{4}} = \infty.$$

The Fourier coefficients $\{\hat{f}(n)\}$ do not, in general go to zero fast enough to make the series

$$\sum_{n=-\infty}^{\infty} \hat{f}(n) e^{2\pi i n x}$$

converge pointwise.

Exercise 5.3.1. Let $S \subset L^2([0,1])$ be a subspace. Show that $S^{\perp}$ is also a subspace.

Exercise 5.3.2. If $\{a_n\}$ is a bi-infinite sequence of complex numbers and

$$\sum_{n=-\infty}^{\infty} |a_n|^2 < \infty.$$

Show that there is a function $f \in L^2[0,1]$ with $\hat{f}(n) = a_n$ for all n.

Exercise 5.3.3. Using the Parseval formula and the function in example 5.2.3 prove (5.23).

Exercise 5.3.4. Prove (5.25).

5.3.2 Bessel's inequality

See: A.6.1.

A fundamental issue, both in proving theorems and applying mathematical techniques to real world problems, is that of approximation. In general a function in $L^2([0,1])$ has an infinite number of non-zero Fourier coefficients. This means that most functions cannot be *exactly* represented by a finite sum of exponentials. As we can only handle a finite amount of data we often need to find the "best" way to approximate a function by a finite sum of exponential functions. What is meant by the "best approximation" is determined by how the error in the approximation is measured.

For each N we define the space of exponential polynomials of degree N to be

$$\mathcal{T}_N = \{ \sum_{n=-N}^{N} a_n e^{2\pi i n x} \mid a_n \in \mathbb{C} \}.$$

Let l denote a norm on a space of functions defined on $[0,1]$, the norm defines a distance by setting $d_l(f,g) = l(f-g)$. Each choice of a norm l gives an approximation problem:

Given a function f with $l(f) < \infty$ find the function $g_f \in \mathcal{T}_N$ such that

$$d_l(f, g_f) = \min\{d_l(f, g) \mid g \in \mathcal{T}_N\}.$$

That is find the point in $\mathcal{T}_N$ whose d_l-distance to f is as small as possible. The minimum value $d_l(f, g_f)$ is called the *error* in the approximation.

The ease with which such a problem is solved depends largely on the choice of l. For most choices of norm this problem is very difficult to solve, indeed is not solvable in practice. One usually has to settle for finding a sequence of approximants $< g_N >$ for which the errors $< d_l(f, g_N) >$ go to zero at essentially the same *rate* as the optimal error. The sole exception is L^2. The following theorem gives the answer if the error is measured in the L^2-norm.

Theorem 5.3.4 (Bessel's inequality). *Given a function $f \in L^2([0,1])$ and constants $\{a_{-N}, \ldots, a_N\}$ the following inequality holds*

$$\|f - \sum_{n=-N}^{N} \hat{f}(n) e^{2\pi i n x}\|_2 \leq \|f - \sum_{n=-N}^{N} a_n e^{2\pi i n x}\|_2$$

with equality if and only if $a_n = \hat{f}(n)$ for all $n \in \{-N, \ldots, N\}$.

Proof. Using the following relation,

$$\|f + g\|_2^2 = \langle f + g, f + g\rangle_{L^2} = \|f\|_2^2 + 2\,\mathrm{Re}\langle f, g\rangle_{L^2} + \|g\|_2^2.$$

we have

$$\|f - \sum_{n=-N}^{N} a_n e^{2\pi i n x}\|_2^2 - \|f - \sum_{n=-N}^{N} \hat{f}(n) e^{2\pi i n x}\|_2^2 = \|\sum_{n=-N}^{N} (a_n - \hat{f}(n)) e^{2\pi i n x}\|_2^2 \geq 0.$$

The equality holds if and only if $a_n = \hat{f}(n)$ for $-N \leq n \leq N$. □

Another way to say this is that for every N the partial sum $S_N(f; x)$ gives the best L^2-approximation to f among functions in $\mathcal{T}_N$. A consequence of the proof of Bessel's inequality is that

$$\langle f - S_N(f), g\rangle_{L^2} = 0 \text{ for any } g \in \mathcal{T}_N.$$

That is the error $f - S_N(f)$ is orthogonal to the subspace $\mathcal{T}_N$. This gives another description of $S_N(f)$ as the L^2-orthogonal projection of f onto the subspace $\mathcal{T}_N$.

Proposition 5.3.2. *The map $f \mapsto S_N(f)$ is the L^2-orthogonal projection onto $\mathcal{T}_N$.*

Exercise 5.3.5. Prove Proposition 5.3.2.

Exercise 5.3.6. For each of the norms

$$\|f\|_p \int_0^1 |f(x)|^p dx, \quad 1 \leq p < \infty$$

find the variational condition characterizing the function $g_N \in \mathcal{T}_N$ which minimizes the error $\|f - g_N\|_p$. Explain why these problems are very difficult to solve if $p \neq 2$.

5.3.3 L^2-derivatives*

See: A.4.3, A.4.6.

In section 3.2.10 we introduced notions of weak and L^2-derivatives for functions defined on $\mathbb{R}$. Here we consider what this should mean for functions defined in a finite interval. We could once again use integration by parts to define a weak derivative, but in the case at hand this is complicated by the presence of boundary terms. There are in fact several different notions of weak differentiability for functions defined on a bounded interval. As we are mostly interested in the relationship between weak differentiability and the behavior of the Fourier coefficients, the appropriate way to handle the boundary terms is to regard f as a 1-periodic function. That is we use the boundary condition $f(0) = f(1)$.

Definition 5.3.3. A function $f \in L^2([0,1])$ is said to have a derivative in $L^2([0,1])$ if there is a function $g \in L^2([0,1])$ such that

$$f(x) = f(0) + \int_0^x g(s)ds \text{ for every } x \in [0,1]$$

and $f(0) = f(1)$.

The definition can be applied recursively, to define the class of functions with k L^2-derivatives.

Definition 5.3.4. A periodic function $f \in L^2([0,1])$ has k L^2-derivatives if there are functions $f_j(x) \in L^2([0,1])$ for $j = 1, \dots k$ such that

- $$f(0) = f(1) \text{ and } f_j(0) = f_j(1) \text{ for } j = 1, \dots, k-1,$$
- $$f(x) = f(0) + \int_0^x f_1(s)ds \text{ for every } x \in [0,1], \tag{5.27}$$
- $$f_{j-1}(x) = f_{j-1}(0) + \int_0^x f_j(s)ds \text{ for every } x \in [0,1] \text{ and } j = 2, \dots, k. \tag{5.28}$$

The function f_j is the j^{th} L^2-derivative of f; we denote the L^2-derivatives of f by $f^{[j]}(x), \partial_x^j f$ etc.

There is a very close connection between having L^2-derivatives and the behavior of the Fourier coefficients.

Theorem 5.3.5. *A function $f \in L^2([0,1])$ has k L^2-derivatives if and only if*

$$\sum_{n=-\infty}^{\infty} (1+|n|)^{2k} |\hat{f}(n)|^2 < \infty. \tag{5.29}$$

In this case, we have

$$\begin{aligned} \int_0^1 |f^{[j]}(x)|^2 dx = \sum_{n=-\infty}^{\infty} |2\pi n|^{2j} |\hat{f}(n)|^2 < \infty \text{ and} \\ \widehat{f^{[j]}}(n) = [2\pi i n]^j \hat{f}(n) \text{ for } j = 1, \ldots, k. \end{aligned} \tag{5.30}$$

Sketch of proof. If (5.29) holds, then Parseval's formula implies that the sequences,

$$< [2\pi i n]^j \hat{f}(n) >, \quad j = 0, \ldots, k$$

are the Fourier coefficients of the functions $f_0, f_1, \ldots, f_k$ in $L^2([0,1])$. Integrating, *formally* it is not difficult to show that these functions are the L^2-derivatives of f. On the other hand, if f has k L^2-derivatives then, using the alternate definition given in exercise 5.3.8, and test functions defined by trigonometric polynomials we deduce that

$$\widehat{f^{[j]}}(n) = [2\pi i n]^j \hat{f}(n) \text{ for } j = 1, \ldots, k. \tag{5.31}$$

The estimate (5.29) is a consequence of these formulæ and Parseval's formula. □

As before there is a relationship between the classical notion of differentiability and having an L^2-derivative. For clarity, let g denote the L^2-derivative of f. It follows from the definition that, for any pair of numbers $0 \leq x < y \leq 1$, we have

$$f(y) - f(x) = \int_x^y g(s) ds.$$

The Cauchy-Schwarz inequality applies to show that

$$\begin{aligned} \left| \int_x^y g(s) ds \right| &\leq \sqrt{\int_x^y 1 \cdot ds} \sqrt{\int_x^y |g(s)|^2 ds} \\ &\leq \sqrt{|x-y|} \|g\|_{L^2}. \end{aligned} \tag{5.32}$$

If we put together this estimate with the previous equation we have that

$$\frac{|f(x) - f(y)|}{\sqrt{|x-y|}} \leq \|g\|_{L^2}.$$

In other words, if a function of one variable has an L^2-derivative then it is Hölder-$\frac{1}{2}$.

In (5.4) we defined the partial sums, $S_N(f)$ of the Fourier series of a function f. If f has an L^2-derivative then it follows from Theorem 5.3.5 and the Cauchy-Schwarz inequality that

$$\sum_{n\neq 0}^{\infty} |\hat{f}(n)| \leq \sqrt{\sum_{n=-\infty}^{\infty} n^2 |\hat{f}(n)|^2} \sqrt{\sum_{n\neq 0} \frac{1}{n^2}} < \infty. \tag{5.33}$$

As we already know that a function with an L^2-derivative is continuous we can apply Proposition 5.1.2 to conclude that the Fourier series of such a function converges pointwise to the function. Indeed it is not difficult to estimate the pointwise error $|f(x) - S_N(f;x)|$,

$$\begin{aligned} |f(x) - S_N(f;x)| &= \left| \sum_{|n|>N} \hat{f}(n) e^{2\pi i n x} \right| \\ &\leq \sum_{|n|>N} |\hat{f}(n)| \\ &\leq \sqrt{\sum_{|n|>N} |n\hat{f}(n)|^2} \sqrt{\sum_{|n|>N} \frac{1}{n^2}} \\ &\leq \|f'\|_{L^2} \sqrt{\frac{2}{N}}. \end{aligned} \tag{5.34}$$

In the first line we used the inversion formula and in the last line, the Parseval formula.

If we measure the distance between two functions in the L^2 sense, then it was shown in Proposition 5.3.1 that the distance between f and the partial sums of its Fourier series does go to zero as $N \to \infty$. This means that, in some average sense $S_N(f)$ converges to f. While not as simple as pointwise convergence, this concept it is well adapted to problems in which measurement is a serious consideration. Given a function f and a point x one cannot exactly measure, the difference

$$|f(x) - S_N(f;x)|.$$

A reasonable mathematical model for what actually can be measured is an average of such differences. For example, we let

$$g_\epsilon(y) = \begin{cases} 0 & \text{if } |x-y| > \epsilon, \\ \frac{1}{2\epsilon} & \text{if } |x-y| \leq \epsilon. \end{cases}$$

Note that $g_\epsilon(y) \geq 0$ for all y and

$$\int_0^1 g_\epsilon(y)dy = 1;$$

the positive number ϵ reflects the resolution of the measuring apparatus. A reasonable model for a *measurement* of the size of the error $|S_N(f;x) - f(x)|$ is given by the average

$$\int_0^1 |f(y) - S_N(f;y)| g_\epsilon(y)dy \leq \|f - S_N(f)\|_2 \|g_\epsilon\|_2 = \frac{1}{\sqrt{2\epsilon}} \|f - S_N(f)\|_2.$$

This estimate for the error is an application of (5.19). For a *fixed resolution*, we see that, as $N \to \infty$, the *measured* difference between f and $S_N(f)$ goes to zero as $N \to \infty$.

Exercise 5.3.7. Show that the hypothesis, in Proposition 5.1.2, that f is continuous is unnecessary by using the observation that $< \hat{f}(n) >$ is a square summable sequence and therefore $f \in L^2([0,1])$. The conclusion needs to be modified to say that f can be modified on a set of measure zero so that (5.6) holds.

Exercise 5.3.8. We can give a different definition using the integration by parts formula. To wit: a function f defined on $[0,1]$ has an L^2-derivative provided that there is a function $f_1 \in L^2([0,1])$ so that

$$\int_0^1 f(x)\varphi'(x)dx = -\int_0^1 f_1(x)\varphi(x)dx$$

for every 1-periodic, once differentiable function φ. Show that this definition is equivalent to the one above.

Exercise 5.3.9. Suppose we use the condition in the previous exercise to define L^2-derivative but without requiring the test functions φ be 1-periodic. Show that we do not get the same class of functions. What boundary condition must a function satisfy to be differentiable in this sense?

Exercise 5.3.10. Provide the details for the derivations the formulæ (5.31).

Exercise 5.3.11. Given any function $g \in L^2([0,1])$, define a *measurement* by setting $l_g(f) = \langle f, g \rangle_{L^2}$. Show that for any $f \in L^2([0,1])$

$$\lim_{N\to\infty} l_g(f - S_N(f)) = 0.$$

5.4 General periodic functions

Up to this point we have only considered functions of period 1. Everything can easily be generalized to functions with arbitrary periods. A function, defined on the real line is periodic of period L, or L-periodic if

$$f(x + L) = f(x)$$

An L-periodic function is determined by its values on any interval of length L. For an integrable function of period L, define the Fourier coefficients by

$$\hat{f}(n) = \int_0^L f(x)e^{-\frac{2\pi inx}{L}}dx.$$

The various results proved above have obvious analogues in this case, for example

1. Inversion Formula:
If f is continuous and $\sum_{n=-\infty}^{\infty} |\hat{f}(n)| < \infty$ then

$$f(x) = \frac{1}{L} \sum_{n=-\infty}^{\infty} \hat{f}(n) e^{\frac{2\pi i n x}{L}}.$$

2. Parseval Formula:
If $f \in L^2([0, L])$ then

$$\int_0^L |f(x)|^2 dx = \frac{1}{L} \sum_{n=-\infty}^{\infty} |\hat{f}(n)|^2.$$

3. Convergence in the mean:
If $f \in L^2([0, L])$ then

$$\lim_{M,N\to\infty} \|f(x) - \sum_{j=-M}^{N} \hat{f}(j) e^{\frac{2\pi i j x}{L}}\|_{L^2}^2 = 0.$$

Exercise 5.4.1. Derive these formulæ from the case $L = 1$ presented above.

5.4.1 Convolution and partial sums

The notion of convolution can be extended to periodic functions. If f and g are defined on $\mathbb{R}$ and are periodic of period L then their convolution is the L-periodic function defined by

$$(f * g)(x) = \int_0^L f(y) g(x - y) dy.$$

Evaluating these integrals requires a knowledge of $g(s)$ for $s \in [-L, L]$; this is where we use the fact that g is L-periodic.

Proposition 5.4.1. *If f and g are L-periodic functions then $f * g$ is also L-periodic. The periodic convolution has the usual properties of a multiplication,*

$$f * g = g * f, \quad f * (g * h) = (f * g) * h, \quad f * (g + h) = f * g + f * h.$$

Periodic convolution and Fourier series are connected in the same way as convolution and the Fourier transform.

Theorem 5.4.1. *The Fourier coefficients of $f * g$ are given by*

$$\widehat{f * g}(n) = \hat{f}(n)\hat{g}(n). \tag{5.35}$$

There is also a notion of convolution for sequences.

Definition 5.4.1. Let $A =< a_n >$ and $B =< b_n >$ be square summable, bi-infinite sequences. The convolution of A with B is the sequence defined by

$$(A \star B)_n = \sum_{j=-\infty}^{\infty} a_j b_{n-j}.$$

Hölder's inequality for l^2 implies that $A \star B$ is a bounded sequence.

This definition is motivated by the result of multiplying trigonometric polynomials, if

$$f = \sum_{j=-N}^{N} a_j e^{2\pi i j x} \text{ and } g = \sum_{j=-N}^{N} b_j e^{2\pi i j x},$$

then

$$f \cdot g = \sum_{l=-2N}^{2N} \left[\sum_{\max\{-N-l,-N\} \le j \le \min\{N,N+l\}} a_j b_{l-j} \right] e^{2\pi i l x}. \tag{5.36}$$

If f and g are square integrable then fg is integrable. Using the notion of convolution of sequences we get obtain a formula for the Fourier coefficients of the pointwise product fg.

Proposition 5.4.2. *If f, g are in $L^2([0, L])$ then the Fourier coefficients of fg are given by*

$$\widehat{fg}(n) = \frac{1}{L} \hat{f} \star \hat{g}(n) = \frac{1}{L} \sum_{j=1}^{\infty} \hat{f}(j)\hat{g}(n-j). \tag{5.37}$$

Proof. For finite sums this result is (5.36). Without worrying about the limits of summation we obtain

$$\begin{aligned} f(x)g(x) &= \frac{1}{L^2} \sum_{j=-\infty}^{\infty} \hat{f}(j) e^{\frac{2\pi i j x}{L}} \sum_{k=-\infty}^{\infty} \hat{g}(k) e^{\frac{2\pi i k x}{L}} \\ &= \frac{1}{L^2} \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \hat{f}(j)\hat{g}(k) e^{\frac{2\pi i (k+j) x}{L}} \\ &= \frac{1}{L} \sum_{l=-\infty}^{\infty} \left[\frac{1}{L} \sum_{j=-\infty}^{\infty} \hat{f}(j)\hat{g}(l-j) \right] e^{\frac{2\pi i l x}{L}}. \end{aligned} \tag{5.38}$$

To get to the last line we set $l = j + k$.

To complete the proof we need to show that if $f, g \in L^2$ then

$$\lim_{N \to \infty} S_N \widehat{(f) S_N}(g)(n) = \widehat{fg}(n). \tag{5.39}$$

Briefly, if $f, g \in L^2$ then the Cauchy-Schwarz inequality implies that $fg \in L^1$. Moreover

$$fg - S_N(f)S_N(g) = (f - S_N(f))g + S_N(f)(g - S_N(g))$$

and therefore the triangle inequality and another application of the Cauchy-Schwarz inequality give

$$\begin{aligned}\|(f - S_N(f))g + S_N(f)(g - S_N(g))\|_{L^1} &\leq \|(f - S_N(f))g\|_{L^1} + \|S_N(f)(g - S_N(g))\|_{L^1} \\ &\leq \|(f - S_N(f))\|_{L^2}\|g\|_{L^2} + \|S_N(f)\|_{L^2}\|(g - S_N(g))\|_{L^2}.\end{aligned} \tag{5.40}$$

Which shows that $S_N(f)S_N(g)$ converges to fg in L^1. The proof is completed by using Proposition 5.1.1 to verify (5.39). □

Exercise 5.4.2. Prove Proposition 5.4.1.

Exercise 5.4.3. Prove Theorem 5.4.1.

Exercise 5.4.4. Show that for square summable sequences A and B the convolution $A \star B$ is a bounded sequence. Hint: Use the Cauchy-Schwarz inequality for l^2.

Exercise 5.4.5. Prove formula (5.36).

Exercise 5.4.6. Give a complete proof of this result by showing that

$$\lim_{N\to\infty} \widehat{S_N(f)S_N}(g)(n) = \widehat{fg}(n).$$

5.4.2 Dirichlet kernel

The partial sums of the Fourier series can be expressed as a convolution. Let

$$\hat{d}_N(n) = \begin{cases} 1 & \text{if } |n| \leq N, \\ 0 & \text{if } |n| > N. \end{cases}$$

The N^{th}-partial sum of the Fourier series of f is just the inverse Fourier transform of the sequence $< \hat{f}\hat{d}_N >$.

Definition 5.4.2. For each N define the *Dirichlet* kernel

$$D_N(x) = \frac{1}{L}\sum_{n=-N}^{N} e^{\frac{2\pi i n x}{L}} = \frac{\sin(\frac{2\pi(N+\frac{1}{2})x}{L})}{L\sin(\frac{\pi x}{L})}. \tag{5.41}$$

The figure shows a graph of D_3.

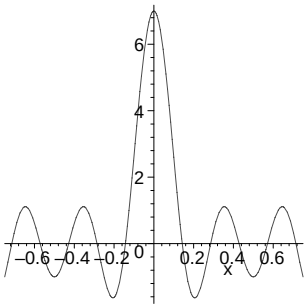


Figure 5.2: Graph of the Dirichlet kernel, $D_3(x)$

It is clear from the definition that $\widehat{D}_N = \hat{d}_N$, Theorem 5.4.1 shows that for $f \in L^1([0, L])$

$$S_N(f; x) = f * D_N(x). \tag{5.42}$$

The zeroth Fourier coefficient of D_N is 1, that is

$$\int_0^L D_N(x)dx = 1.$$

The Dirichlet kernel is oscillatory and assumes both positive and negative values. It is not difficult to show that

$$\lim_{N\to\infty} \int_0^L |D_N(x)|dx = \infty. \tag{5.43}$$

This fact underlies the difficulties in analyzing the pointwise convergence of the partial sums of the Fourier series. Even if f is a continuous function it is **not** always true that

$$\lim_{N\to\infty} S_N(f; x) = f(x).$$

In the next several sections we explore some of these issues in detail. First we consider what happens to the partial sums of the Fourier series near to a jump discontinuity. Then we find a replacement for the partial sums which has better pointwise convergence properties.

Exercise 5.4.7. Prove formula (5.41) by using the formula for the sum of a geometric series.

Exercise 5.4.8. Use the explicit formula for $D_N(x)$ to prove (5.43). Hint: Compare

$$\int_0^L |D_N(x)|dx$$

to the harmonic series.

5.5 The Gibbs Phenomenon

See: A.3.1.

Let f be a function with a jump discontinuity at x_0, that is

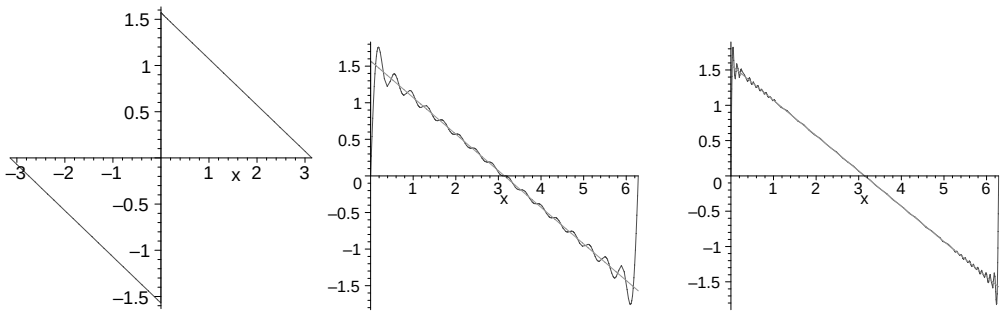
$$\lim_{x\to x_0^+} f(x) \text{ and } \lim_{x\to x_0^-} f(x)$$

both exist but are not equal. Since the partial sums $< S_N(f; x) >$ are continuous functions it is a foregone conclusion that they cannot provide a good pointwise approximation to f near x_0. In fact they do an especially poor job. On the other hand, the way in which they fail does not depend very much on f and can be analyzed completely. We begin by considering an example.

Example 5.5.1. Consider the 2π-periodic function

$$g(x) = \begin{cases} \frac{\pi - x}{2} & 0 \le x \le \pi, \\ -\frac{\pi + x}{2} & -\pi \le x < 0 \end{cases} \tag{5.44}$$

whose graph is shown below along with certain partial sums of its Fourier series.



(a) A periodic function with a jump discontinuity.

(b) Fourier reconstruction with $N = 16$.

(c) Fourier reconstruction with $N = 64$.

Figure 5.3: An example of the Gibbs phenomenon

From the graphs it is apparent that, even as N increases, the pointwise approximation does not improve near the jump. In fact the graphs of the partial sums, $S_N(g)$ "overshoot" the graph of g near to the discontinuity. In the graphs it appears that the amount of overshoot does not decrease as N increases. The partial sums, $S_N(g)$ are also highly oscillatory near to the jump. This collection of bad behaviors is called the *Gibbs phenomenon*. In the engineering literature it is also called *overshoot*.

To analyze the Gibbs phenomenon we first consider, in detail, the partial sums of the function g defined in (5.44). This function has a jump discontinuity of size π at $x = 0$, the Fourier series of g is given by

$$g(x) = \sum_{k=-\infty}^{\infty} \hat{g}(k) e^{ikx} = \sum_{k=1}^{\infty} \frac{\sin kx}{k}.$$

If x is not a multiple of 2π then $S_N(g;x)$ converges to $g(x)$, see section 5.6. At $x = 0$ the series converges to 0 which is the average of $\lim_{x\to 0^+} g(x)$ and $\lim_{x\to 0^-} g(x)$.

The partial sum, $S_N(g)$ can be re-expressed as:

$$S_N(g;x) = \sum_{k=1}^{N} \frac{\sin kx}{k} = \int_0^x \sum_{k=1}^{N} \cos ky dy = \int_0^x \frac{1}{2}\left[\frac{\sin(N+\frac{1}{2})y}{\sin \frac{1}{2}y} - 1\right] dy \tag{5.45}$$

since

$$\frac{\sin kx}{k} = \int_0^x \cos ky dy.$$

We are looking for the maximum of the difference $S_N(g;x)-g(x)$. From elementary calculus we know that, at a point where the maximum occurs,

$$\frac{d}{dx}[S_N(g;x) - g(x)] = 0. \tag{5.46}$$

Away from its jumps, g is a linear function of slope $-\frac{1}{2}$, hence we are looking for points where

$$S'_N(g;x) = -\frac{1}{2}. \tag{5.47}$$

Let x_N denote the smallest, positive x where this holds. This is a reasonable place to look for the worst behavior as it is the local maximum error closest to the jump. Evaluating $S_N(g;x_N) - g(x_N)$ then gives a lower bound for the maximum difference.

From equation (5.45), we have

$$S'_N(g;x) = \frac{1}{2}\left[\frac{\sin(N+\frac{1}{2})x}{\sin\frac{1}{2}x} - 1\right].$$

Equation (5.47) holds if

$$\sin(N+\frac{1}{2})x = 0.$$

The number

$$x_N = \frac{\pi}{N+\frac{1}{2}}$$

is the smallest positive solution of this equation. The partial sum at x_N is given by

$$\begin{aligned} S_N(g;\frac{\pi}{N+\frac{1}{2}}) &= \frac{1}{2}\int_0^{\frac{\pi}{N+\frac{1}{2}}}\left[\frac{\sin(N+\frac{1}{2})y}{\sin\frac{1}{2}y} - 1\right]dy \\ &= \frac{1}{2}\int_0^{\frac{\pi}{N+\frac{1}{2}}}\frac{\sin(N+\frac{1}{2})y}{\sin\frac{1}{2}y}dy - \frac{1}{2}\frac{\pi}{N+\frac{1}{2}} \\ &= \frac{1}{2}\int_0^1 \frac{\sin(t\pi)}{\sin\left(\frac{1}{2}\frac{t\pi}{N+\frac{1}{2}}\right)}\frac{\pi dt}{N+\frac{1}{2}} - \frac{1}{2}\frac{\pi}{N+\frac{1}{2}} \end{aligned} \tag{5.48}$$

In the last line we used the change of variable, $y = \frac{t\pi}{N+\frac{1}{2}}$. Using the Taylor expansion for $\sin x$ gives

$$(N+\frac{1}{2})\sin\frac{t\pi}{2(N+\frac{1}{2})} = (N+\frac{1}{2})\left[\frac{t\pi}{2(N+\frac{1}{2})} - \frac{1}{6}\left(\frac{t\pi}{2(N+\frac{1}{2})}\right)^3 + \cdots\right].$$

Hence, as $N\to\infty$, the denominator of the integrand converges to $\frac{t\pi}{2}$ and the numerator converges to $\sin t\pi$. Therefore, we have

$$\lim_{N\to\infty} S_N(g; \frac{\pi}{N+\frac{1}{2}}) = \frac{1}{2}\int_0^1 \frac{\sin t\pi}{\frac{t\pi}{2}}\pi dt = \int_0^1 \frac{\sin t\pi}{t} dt. \tag{5.49}$$

From the definition of g,

$$\lim_{N\to\infty} g(\frac{\pi}{N+\frac{1}{2}}) = \frac{\pi}{2}$$

Evaluating the above integral numerically we obtain that

$$\lim_{N\to\infty} S_N(g; \frac{\pi}{N+\frac{1}{2}}) = \left(\frac{\pi}{2}\right) 1.178979744\cdots \tag{5.50}$$

This implies that

$$\lim_{n\to\infty} [S_N(g; \frac{\pi}{N+\frac{1}{2}}) - g(\frac{\pi}{N+\frac{1}{2}})] = \frac{\pi}{2} 0.178979744\cdots . \tag{5.51}$$

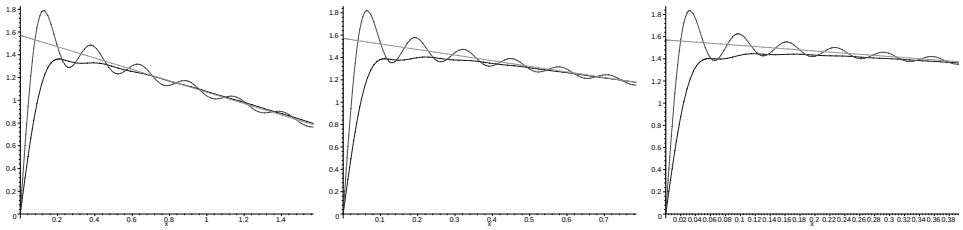
From the graphs in figure 5.3 it is clear that the oscillations in the partial sums become more and more concentrated near to the jump as N increases. The discussion above shows that the local maxima and minima of $S_N(g;x) - g(x)$ occur at the set of points $\{x_{N,k}\}$ which satisfy

$$\sin(N+\frac{1}{2})x_{N,k} = 0.$$

These are simply

$$x_{N,k} = \frac{k\pi}{N+\frac{1}{2}} \text{ for } k \in \mathbb{Z}.$$

The number of oscillations of a given size is essentially independent of N but the region in which they occur scales with N. The oscillations in $S_N(g;x)$ are concentrated in a region of size N^{-1} around the jump. The graphs in figure 5.4 show the original function, its partial sums and its "Fejer means." These are the less oscillatory curves lying below the graphs of g, and are explained in the next section. In these graphs we have rescaled the x-axis to illustrate that the Gibbs oscillations near the discontinuity remain of a constant size. This is a universal phenomenon for partial sums of the Fourier series of a function with simple jump discontinuities.



(a) N=24, magnification=4

(b) N=48, magnification=8

(c) N=96, magnification=16

Figure 5.4: Detail showing equi-oscillation property in Gibbs phenomenon

5.5.1 The general Gibbs phenomenon

Now suppose that f is piecewise differentiable function with a jump discontinuity at the point x_0. This means that the left and right limits

$$\lim_{x \to x_0^-} f(x) = L \text{ and } \lim_{x \to x_0^+} f(x) = R,$$

both exist but $L \neq R$. Suppose that $L < R$, fixing any sufficiently small $\epsilon > 0$ we show below that

$$\lim_{N \to \infty} \max_{0 < x - x_0 < \epsilon} (S_N(f;x) - f(x)) = (G-1)\frac{R-L}{2}.$$

The coefficient G is a universal constant defined by

$$G = \frac{2}{\pi} \int_0^1 \frac{\sin \pi t}{\sin t} dt = 1.178979744\ldots \tag{5.52}$$

In fact there is a sequence of points $< x_N >$ which converge to x_0 so that $|S_N(f;x_N) - f(x_N)|$ is about %9 of the size of the jump. The general result is the following.

Theorem 5.5.1. *Let f be a piecewise C^1-function with jump discontinuities at $\{x_1, \ldots, x_k\}$ of sizes $\{h_1, \ldots, h_k\}$ that is*

$$h_j = \lim_{x \to x_j^+} f(x) - \lim_{x \to x_j^-} f(x)$$

and set

$$\eta_j = \operatorname{sign} h_j.$$

For any sufficiently small $\epsilon > 0$ we have that

$$\lim_{N\to\infty} \max_{0<\eta_j(x-x_j)<\epsilon} (S_N(f;x) - f(x)) = (G-1)\frac{|h_j|}{2} \text{ for each } j \in \{1,\dots,k\}.$$

The constant G is given by (5.52).

Proof. First we consider $g_j(x) = \frac{h_j}{\pi} g(x - x_j)$ which is a scaled and translated version of g. From the analysis above of g, we have that

$$\lim_{N\to\infty} \max_{0<\eta_j(x-x_j)<\epsilon} (S_N(g_j;x) - g_j(x)) = (G-1)\frac{|h_j|}{2} \tag{5.53}$$

In section 5.6 it is shown that for $i \neq j$ we have

$$\lim_{N\to\infty} \max_{0<|x-x_j|<\epsilon} |S_N(g_i;x) - g_i(x)| = 0.$$

Rewrite f as

$$f(x) = f_1(x) + \sum_{j=1}^{k} \frac{h_j}{\pi} g(x - x_j), \text{ where } f_1(x) = f(x) - \sum_{j=1}^{k} \frac{h_j}{\pi} g(x - x_j).$$

Note that f_1 is a continuous and piecewise C^1-function. It is not difficult to show that f has an L^2-derivative and it therefore follows from (5.33) that the Fourier series of $f_1(x)$ converges uniformly to f_1. The jumps in the function f have been "transferred" to the sum of the g_js:

$$S_N(f;x) - f(x) = S_N(f_1;x) - f_1(x) + \sum_{j=1}^{k} S_N(g_j;x) - \sum_{j=1}^{k} g_j(x)$$

Since $\lim_{N\to\infty} S_N(f_1;x) = f_1(x)$ for every x it follows from (5.53) that

$$\lim_{N\to\infty} \max_{0<\eta_j(x-x_j)<\epsilon} (S_N(f;x) - f(x)) = \lim_{N\to\infty} \max_{0<\eta_j(x-x_j)<\epsilon} (S_N(g_j;x) - g_j(x)) = (G-1)\frac{|h_j|}{2}.$$

This completes the proof of the theorem. □

Remark 5.5.1. The proof of the theorem shows that, near the jumps, the partial sums $S_N(f)$ look like the partial sums of $S_N(g)$ near its jump. In (5.34) we showed that $|S_N(f_1;x) - f_1(x)|$ behaves like $N^{-\frac{1}{2}}$ for all x. This implies that $S_N(f;x)$ has the same oscillatory artifacts near the jump points as are present in $S_N(g)$. The Gibbs phenomenon places an inherent limitation on the utility of the Fourier transform when working with discontinuous data. In imaging applications such function often arise. Taking higher and higher partial sums does not lead, in and of itself, to a better reconstructed image near such points. In the next section we describe a different approximate inverse for the Fourier series which eliminates the Gibbs phenomenon.

Exercise 5.5.1. In section 5.6 it is shown that for $x \neq 0$

$$\lim_{N\to\infty} S_N(g;x) = g(x).$$

Assuming this, use partial summation to show that there is a constant M, which depends on x so that

$$|S_N(g;x) - g(x)| \leq \frac{M}{N}.$$

Explain why M *must* depend on x.

Exercise 5.5.2. Show that a piecewise differentiable, continuous periodic function has a bounded weak derivative.

5.5.2 Fejer means

The partial sums of the Fourier series of a function f defined on $[0,1]$ are given by

$$S_N(f,x) = \sum_{n=-N}^{N} \hat{f}(n) e^{2\pi i n x}.$$

The partial sum is expressible as the convolution of f with the Dirichlet kernel:

$$S_N(f;x) = D_N * f(x).$$

What makes the convergence of the partial sums so delicate is the fact that the Dirichlet kernel assumes both positive and negative values. This means that the convergence (or non-convergence) of the partial sums relies on subtle cancelations (or their absence). There is a general technique to obtain a more stable pointwise approximation for functions by finite trigonometric sums which does not sacrifice too much of the very important L^2-approximation properties of the partial sums.

Definition 5.5.1. The N^{th} *Fejer mean* of the partial sums is just the average of the first $N+1$ partial sums,

$$C_N(f;x) = \frac{S_0(f;x) + \cdots + S_N(f;x)}{N+1}.$$

Definition 5.5.2. Define the N^{th} *Fejer kernel* to be the average of the first $N+1$ Dirichlet kernels.

$$F_N(x) = \frac{D_0(x) + \cdots + D_N(x)}{N+1}.$$

A calculation shows that

$$F_N(x) = \frac{2\pi}{N+1} \left[\frac{\sin\left(\pi(N+1)x\right)}{\sin(\pi x)} \right]^2. \tag{5.54}$$

It follows from (5.42) that

$$C_N(f;x) = F_N * f(x).$$

The important difference between the Fejer kernel and the Dirichlet kernel is that the Fejer kernel only assumes non-negative values. Fejer's theorem is a consequence of this fact.

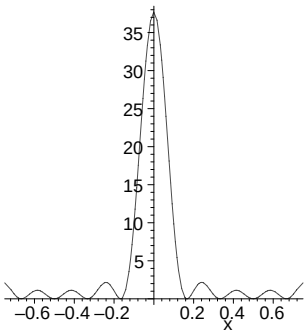


Figure 5.5: Graph of the Fejer kernel, $F_5(x)$

Theorem 5.5.2 (Fejer's Theorem). *If f is an absolutely integrable function which is continuous at x then*

$$\lim_{N\to\infty} C_N(f;x) = f(x).$$

Remark 5.5.2. As remarked above, the analogous statement for the partial sums is false.

Proof. The proof is very similar to the proof of the Fourier inversion formula. Note that the Fejer kernel shares three properties with $P_r(x)$:

(1). The Fejer kernel is non-negative.

(2). For every N, $\int_0^1 F_N(x)dx = 1$.

(3). Given $\epsilon > 0$ there is an M so that if $N > M$ then

$$F_N(x) < \epsilon \text{ for } \epsilon < x < 1-\epsilon. \tag{5.55}$$

The proof of Fejer's Theorem follows from these properties.

The first two properties imply that

$$\begin{aligned} |C_N(f;x) - f(x)| &= \left| \int_0^1 F_N(x-y)(f(y)-f(x))dy \right| \\ &\leq \int_0^1 F_N(x-y)|f(y)-f(x)|dy. \end{aligned} \tag{5.56}$$

As f is continuous at x, given $\epsilon > 0$ there is a $\delta > 0$ so that

$$|y-x| < \delta \Rightarrow |f(x)-f(y)| < \epsilon.$$

Using the third property of the Fejer kernel, there is an M so that if $N > M$ then

$$F_N(x) < \epsilon \text{ provided } \delta < x < 1-\delta.$$

We split the integral into two parts:

$$\begin{aligned} |C_N(f;x) - f(x)| &\leq \int_{|x-y|<\delta} F_N(x-y)|f(y)-f(x)|dy + \int_{|x-y|\geq\delta} F_N(x-y)|f(y)-f(x)|dy \\ &\leq \epsilon \int_{|x-y|<\delta} F_N(x-y)dy + \int_{|x-y|\geq\delta} \epsilon(|f(y)|+|f(x)|)dy \\ &\leq \epsilon(1+|f(x)|+\|f\|_{L^1}). \end{aligned} \tag{5.57}$$

As $\epsilon > 0$ is arbitrary, this completes the proof of the theorem. □

Remark 5.5.3. The proof shows that if f is a continuous, periodic function then $C_N(f;x)$ provides a uniformly accurate approximation. That is given $\epsilon > 0$ there is an N such that $|C_N(f;x) - f(x)| < \epsilon$ for all $x \in [0,1]$.

A further computation shows that the Fourier coefficients of $C_N(f;x)$ are given by

$$\widehat{C_N}(f;n) = \left[1 - \frac{|n|}{N+1}\right] \hat{f}(n). \tag{5.58}$$

From this is not hard to see that $\|C_N(f) - f\|_2$ goes to zero as $N \to \infty$, at worst, half as fast as $\|S_N(f) - f\|_2$. The smoother curves in figure 5.4 are the Fejer means of g. They are not nearly as oscillatory, near the jump points, as the partial sums. Moreover if f satisfies

$$m \leq f(x) \leq M \text{ for } x \in [0,1]$$

then for every N

$$m \leq C_N(f;x) \leq M \text{ for } x \in [0,1]$$

as well.

Using the Fejer kernel we can now prove the lemma used in the proof of the Parseval formula.

Theorem 5.5.3. *Let $f \in L^2([0,1])$ and fix an $\epsilon > 0$. There is a trigonometric polynomial g such that*

$$\|f - g\|_{L^2} < \epsilon.$$

Proof. In Corollary A.6.1 it is established that there is a continuous, periodic function g_0 so that

$$\|f - g_0\|_{L^2} < \frac{\epsilon}{2}.$$

Fejer's theorem implies that there is an N so that, for all $x \in [0,1]$

$$|g - C_N(g;x)| \leq \frac{\epsilon}{2}.$$

For this N we use the triangle inequality to obtain

$$\|f - C_N(g)\|_{L^2} \leq \|f - g\|_{L^2} + \|C_N(g) - g\|_{L^2} < \frac{\epsilon}{2} + \frac{\epsilon}{2}.$$

This completes the proof of the theorem. □

Example 5.5.2. We close this section by quantitatively comparing the Fejer means and the partial sums of the Fourier series for the function g defined in (5.44). Measuring the difference in the L^2-norm, we find that $\|f - S_N(f)\|_2$ is about half the size of $\|f - F_N(f)\|_2$. Table 5.1 compares the mean square errors of the partial sums and Fejer means.

Exercise 5.5.3. Prove that if $m \leq f(x) \leq M$ for $x \in [0,1]$ then $m \leq C_N(f;x) \leq M$ for $x \in [0,1]$ is as well.

Exercise 5.5.4. Prove the closed form expression for the Fejer kernel (5.54).

N	Partial sum	Fejer mean
4	.695306572	.942107038
8	.369174884	.570026118
12	.251193672	.411732444
16	.190341352	.323158176
20	.153218060	.266320768
24	.128210482	.226669996
28	.110220018	.197397242
32	.096656766	.174881126

Table 5.1: Comparison of the mean square errors

Exercise 5.5.5. Prove the formulæ for the Fourier coefficients of F_N, (5.58).

Exercise 5.5.6. Show that the Fourier transform also suffers from the Gibbs phenomenon by analyzing $S_R(\chi_{[-1,1]})$ as $R \to \infty$. The partial inverse S_R is defined in (3.38).

Exercise 5.5.7. Explain why the Fejer kernel is the Fourier series analogue of the kernel $\mathcal{F}^{-1}(\chi_{2,B})$ defined in (3.48).

5.5.3 Resolution in the partial sums of the Fourier series

The Fejer means produce visually more appealing images near to jump points. However this is at the expense of reducing the overall resolution. Convolution with the Fejer kernel progressively attenuates the high frequencies eventually cutting them off entirely at $n = \pm N$. It is important to remember that the Fourier coefficients are global, each coefficient is a (complex) weighted average of the function over it entire interval of definition. Employing Fejer means to reduce the Gibbs effect near to jumps, inevitably results in an overall decrease in the available resolution. A calculation using Taylor's formula shows that as N tends to infinity

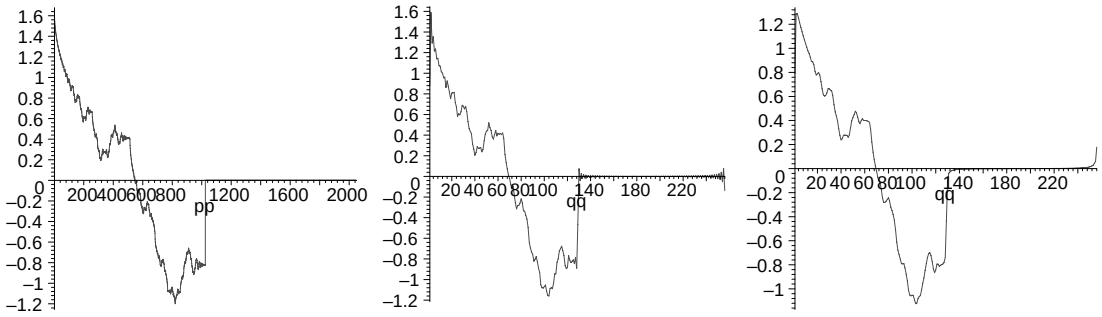
$$\text{FWHM}(D_N) \simeq \frac{.86}{\pi N} \tag{5.59}$$

whereas

$$\text{FWHM}(F_N) \simeq \frac{1.33}{\pi N}. \tag{5.60}$$

Here we are measuring the FWHM of the central peaks. Using the same number of Fourier coefficients the partial sum has about $\frac{3}{2}$-times as much FWHM resolution as the Fejer mean.

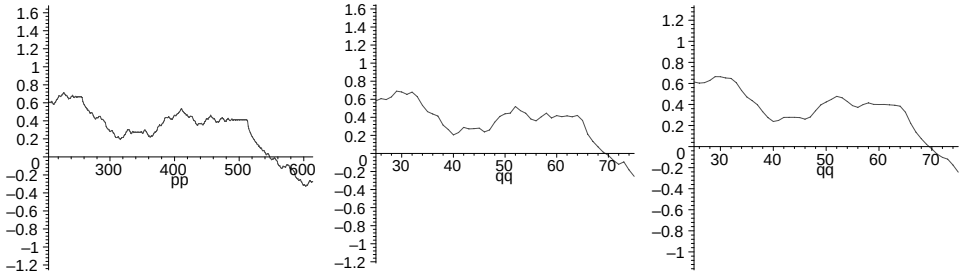
Example 5.5.3. In the graphs we have an "interesting" function which we reconstruct first using a partial sum of its Fourier series and then with Fejer means. It is clear that the oscillatory artifact has disappeared in the Fejer means reconstruction. The resolution of the latter reconstruction is also evidently lower.



(a) A function (b) Partial sum (c) Fejer mean

Figure 5.6: Graphs comparing the partial sums and Fejer means.

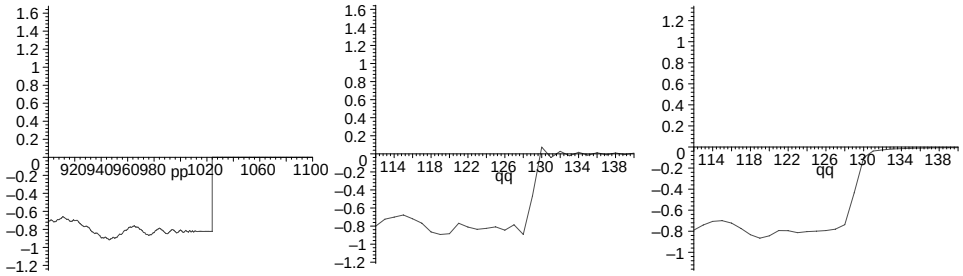
The graphs in figure 5.7 are expanded views of these functions between .1 and .3. Here the loss of resolution in the Fejer means, at points away from the jump is quite evident.



(a) A function (b) Partial sum (c) Fejer mean

Figure 5.7: Expanded view showing the loss of resolution in the Fejer means.

Finally we compare the behavior of these approaches near the jump discontinuity. Both the Gibbs ringing and higher resolution are again quite evident in the partial sums.



(a) The jump (b) Partial sum (c) Fejer mean

Figure 5.8: Expanded view showing Gibbs phenomenon in the partial sums.

Exercise 5.5.8. Derive (5.59) and (5.60).

5.6 The localization principle*

Each Fourier coefficient is defined by an integral of f over its whole domain of definition and therefore depends on the value of the function everywhere. Like the Fourier transform, the Fourier series is very sensitive to the local behavior of a function. This is the content of the following theorem.

Theorem 5.6.1 (The localization principle). *Let f and g be L-periodic and absolutely integrable over $[0, L]$. Suppose that for some x, $S_N(g; x)$ converges to $g(x)$ as $N \to \infty$. If $f(t) = g(t)$ in an interval $[x - \epsilon, x + \epsilon]$ for an $\epsilon > 0$ then*

$$\lim_{N\to\infty} S_N(f; x) = f(x)$$

as well.

Proof. In the computations below recall that x is a fixed point. Let $s_N(x) = D_N * f(x)$, and $t_N(x) = D_N * g(x)$. By linearity, $(s_N(x) - t_N(x)) = D_N * (f - g)(x)$, and $D_N * (f - g)(x)$

$$\begin{aligned} D_N * (f-g)(x) &= \int_0^L \sin\left(\frac{\pi(2N+1)(x-y)}{L}\right)\left[\frac{f(y)-g(y)}{L\sin\left(\frac{\pi(x-y)}{L}\right)}\right]dy \\ &= \operatorname{Im}\left\{\int_0^L e^{\frac{2\pi i N(x-y)}{L}} e^{\frac{i\pi(x-y)}{L}}\left[\frac{f(y)-g(y)}{L\sin\left(\frac{\pi(x-y)}{L}\right)}\right]dy\right\} \\ &= \operatorname{Im}\left\{e^{\frac{\pi i(2N+1)x}{L}}\int_0^L e^{-\frac{2\pi i N y}{L}}\left[\frac{f(y)-g(y)}{L\sin\left(\frac{\pi(x-y)}{L}\right)}\right]e^{-\frac{i\pi y}{L}}dy\right\}. \end{aligned}$$

Since $f(y) = g(y)$ for $y \in [x - \epsilon, x + \epsilon]$, the last integrand is absolutely integrable:

$$\int\left|\frac{f(y)-g(y)}{\sin\left(\frac{2\pi(x-y)}{L}\right)}e^{-\frac{i\pi y}{L}}\right|dy < \infty.$$

Therefore, the last integral is the N^{th} Fourier coefficient of an integrable function. By the Riemann-Lebesgue lemma, it goes to zero as $N \to \infty$. By the hypothesis, we know that $\lim_{N\to\infty} t_N(x) = g(x) = f(x)$. Rewrite this as

$$\lim_{N\to\infty} s_N(x) = \lim_{N\to\infty} (s_N(x) - t_N(x)) + \lim_{N\to\infty} t_N(x).$$

Since $s_N(x) - t_N(x)$ goes to zero as $N \to \infty$, we are done. □

This result shows that if an absolutely integrable function $f(x)$ is well behaved in the neighborhood of a point x_0, then $S_N(f; x_0)$ converges to $f(x_0)$. Suppose that f is continuously differentiable in an interval $(x_0 - \epsilon, x_0 + \epsilon)$. Let $\varphi(x)$ be an infinitely differentiable function which satisfies

$$\varphi(x) = \begin{cases} 1 & \text{for } |x - x_0| < \frac{\epsilon}{2}, \\ 0 & \text{for } |x - x_0| > \frac{3\epsilon}{4}. \end{cases}$$

If we set $g(x) = f(x)\varphi(x)$ then g is a continuously differentiable, periodic function and

$$f(x) = g(x) \text{ for } x \in (x_0 - \frac{\epsilon}{2}, x_0 + \frac{\epsilon}{2}).$$

Because $g(x)$ is continuously differentiable it follows easily that

$$\sum_{n=-\infty}^{\infty} |\hat{g}(n)| < \infty$$

therefore we can apply (5.6) to conclude that

$$g(x) = \sum_{n=-\infty}^{\infty} \hat{g}(n) e^{2\pi i n x}$$

for any x. The localization principle states that

$$\lim_{N\to\infty} \sum_{n=-N}^{N} \hat{f}(n) e^{2\pi i n x_0} = f(x_0),$$

no matter how wildly f behaves outside the interval $(x_0 - \epsilon, x_0 + \epsilon)$. Notice that the asymptotic behavior of the sequence $\{\hat{f}(n)\}$ will in general be completely different from that of $\{\hat{g}(n)\}$. This just makes the localization principle all the more remarkable.

Exercise 5.6.1. In the proof of Theorem 5.6, show that it is not necessary to assume that $f(t) = g(t)$ in an interval containing x by explaining why it suffices to assume that

$$h(y) = \frac{f(y) - g(y)}{x - y}$$

is an integrable function.

5.7 Higher dimensional Fourier series

The theory of Fourier series extends without difficulty to functions defined on the unit cube in $\mathbb{R}^n$. For completeness we include statements of the basic results. Proofs can be found in [16] or [72]. The Fourier coefficients are now labeled by vectors $\mathbf{k} \in \mathbb{Z}^n$, that is vectors of the form

$$\mathbf{k} = (k_1, \ldots, k_n) \text{ where } k_j \in \mathbb{Z} \text{ for } j = 1, \ldots, n.$$

If $f(x_1, \ldots, x_n)$ is an absolutely integrable function defined on

$$[0,1]^n = [0,1] \times \underbrace{\cdots}_{n-\text{times}} \times [0,1]$$

then its Fourier coefficients are defined by

$$\hat{f}(\mathbf{k}) = \int_{\mathbb{R}^n} f(\mathbf{x}) e^{-2\pi i \langle \mathbf{k}, \mathbf{x} \rangle} d\mathbf{x}.$$

Many aspects of the theory are quite similar in higher dimensions, however the theory of pointwise convergence is much more involved and has not yet been completely worked out.

If the Fourier coefficients of f tend to zero rapidly enough then we have an inversion formula:

Proposition 5.7.1. *Suppose that f is an absolutely integrable function on $[0,1]^n$ such that*

$$\sum_{\mathbf{k} \in \mathbb{Z}^n} |\hat{f}(\mathbf{k})| < \infty \tag{5.61}$$

then

$$f(x) = \sum_{\mathbf{k} \in \mathbb{Z}^n} \hat{f}(\mathbf{k}) e^{2\pi i \langle \mathbf{k}, \mathbf{x} \rangle}.$$

In general the Fourier coefficients of an absolutely integrable function may not satisfy (5.61). Indeed as the dimension increases this gets to be a more and more restrictive condition. In order for the infinite sum

$$\sum_{\mathbf{k} \in \mathbb{Z}^n} \frac{1}{(1 + \|\mathbf{k}\|)^\alpha}$$

to converge it is necessary to take $\alpha > n$. However the Riemann Lebesgue generalizes.

Proposition 5.7.2 (Riemann Lebesgue Lemma). *If f is an absolutely integrable function on $[0,1]^n$ then*

$$\lim_{\|\mathbf{k}\| \to \infty} \hat{f}(\mathbf{k}) = 0.$$

Once again the proof is by approximating L^1-functions by the n-dimensional analogue of step functions.

In this generality there is, as before, no estimate on the rate at which the Fourier coefficients go to zero. As in the one dimensional case, when working with Fourier series we need to consider f as a periodic function of period 1 in each variable. That is we extend f to all of $\mathbb{R}^n$ by using the condition

$$f(\mathbf{x}) = f(\mathbf{x} + \mathbf{k}) \text{ for every } \mathbf{k} \in \mathbb{Z}^n.$$

The inversion formula defines a function on all of $\mathbb{R}^n$ with this property. As before, in the context of Fourier series, a function is considered continuous if its periodic extension to $\mathbb{R}^n$ is continuous and differentiable if its periodic extension to $\mathbb{R}^n$ is differentiable, etc.

If the Fourier coefficients do not satisfy (5.61) then the problem of summing the Fourier series can be quite subtle. The question of the pointwise convergence for the partial sums is considerably more complicated in higher dimensions than in one dimension. In the one dimensional case there is, in essence only one reasonable way to define partial sums. In n-dimensions there are many different possible choices. The simplest way is to define the N^{th}-partial sum to be

$$S_N(f;\mathbf{x}) = \sum_{k_1=-N}^{N} \cdots \sum_{k_n=-N}^{N} \hat{f}(\mathbf{k}) e^{2\pi i \langle \mathbf{k}, \mathbf{x} \rangle}.$$

Because there is a very fast algorithm to do this calculation (at least for N a power of 2) this is the usual meaning of "partial sums" of the Fourier series in applications. However it is by no means the only way to partially invert the higher dimensional Fourier series. We could equally well consider the sum over all vectors $\mathbf{k}$ such that $\|\mathbf{k}\| \leq R$. Let

$$\Sigma_R(f;\mathbf{x}) = \sum_{\{\mathbf{k} \,:\, \|\mathbf{k}\| < R\}} \hat{f}(\mathbf{k}) e^{2\pi i \langle \mathbf{k}, \mathbf{x} \rangle},$$

denote this sum. While not as useful in applications, this form of the partial inverse is easier to analyze. From this analysis, it is known that the localization principle fails in higher dimensions. The convergence of the Fourier series at $\mathbf{x}$ is sensitive to the behavior of f at points distant from $\mathbf{x}$. The relationship between $S_N(f;\mathbf{x})$ and $\Sigma_R(f;\mathbf{x})$ has, so far, not been completely elucidated. An analysis of $\Sigma_R(f)$ is given in [57].

The Gibbs phenomenon also persists in higher dimensions but is, as expected more complicated to analyze. If a piecewise smooth function f has a simple jump along a smooth hypersurface S then the behavior of the partial sums near $\mathbf{x} \in S$ is determined in part by the size of the jump at $\mathbf{x}$ as well as the *curvature* of S at $\mathbf{x}$. Asymptotic formulæ for $\Sigma_R(f;\mathbf{x})$ are given in [57]. If S itself is not smooth then even more complicated phenomena arise. As the techniques involved are far beyond the scope of this text we content ourselves with giving as an example a partial sum (of S_N-type) for the Fourier series of $\chi_{[-1,1]}(x)\chi_{[-1,1]}(y)$. Note the Gibbs oscillations parallel to the edges of the square and the "Gibbs shadow" near the corner.

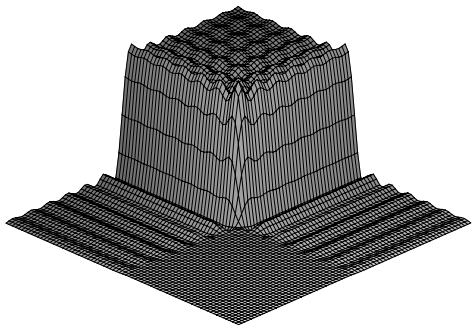


Figure 5.9: Illustration of the $2d$-Gibbs phenomenon

There is an obvious generalization of the notion convolution for periodic functions on $\mathbb{R}^n$ given by

$$f * g(\mathbf{x}) = \int_{[0,1]^n} f(\mathbf{x}-\mathbf{y})g(\mathbf{y})d\mathbf{y}.$$

It is connected to the Fourier series just as in one dimension:

$$\widehat{f * g}(\mathbf{k}) = \hat{f}(\mathbf{k})\hat{g}(\mathbf{k}) \text{ for all } \mathbf{k} \in \mathbb{Z}^n. \tag{5.62}$$

We can also define the convolution of two sequences $A =< a_{\mathbf{k}} >, B =< b_{\mathbf{k}} >$ indexed by $\mathbb{Z}^n$ by setting

$$(A \star B)_{\mathbf{k}} = \sum_{\mathbf{j}\in\mathbb{Z}^n} a_{\mathbf{k}-\mathbf{j}} b_{\mathbf{j}}.$$

The Fourier series of a pointwise product is then given by

$$\widehat{fg}(\mathbf{k}) = \hat{f} \star \hat{g}(\mathbf{k}). \tag{5.63}$$

Exercise 5.7.1. Let $r_1, \ldots, r_n$ be numbers between 0 and 1. Compute the Fourier coefficients of

$$f(x_1, \ldots, x_n) = \chi_{[-r_1,r_1]}(x_1) \cdots \chi_{[-r_n,r_n]}(x_n).$$

Exercise 5.7.2. Show that

$$\sum_{\mathbf{k}\in\mathbb{Z}^n} \frac{1}{(1+\|\mathbf{k}\|)^\alpha}$$

converges if $\alpha > n$ and diverges if $\alpha \leq n$. Hint: Compare this sum to an integral.

Exercise 5.7.3. Find a periodic function, $D_{N,n}$ on $\mathbb{R}^n$ for which

$$S_N(f) = D_{N,n} * f.$$

Exercise 5.7.4. Find a periodic function, $D'_{R,n}$ on $\mathbb{R}^n$ for which

$$\Sigma_R(f) = D'_{R,n} * f.$$

5.7.1 L^2-theory

See: A.4.7.

As in the one dimensional case, a much more complete theory is available for square integrable functions. The basic result is the Parseval formula.

Proposition 5.7.3 (Parseval Formula). *A function $f \in L^2([0,1]^n)$ if and only if its Fourier coefficients are square summable, in this case we have*

$$\int_{[0,1]^n} |f(\mathbf{x})|^2 d\mathbf{x} = \sum_{\mathbf{k}\in\mathbb{Z}^n} |\hat{f}(\mathbf{k})|^2.$$

The sense in which the Fourier series converges can also be made precise in this case.

Proposition 5.7.4 (L^2-Inversion formula). *If $f \in L^2([0,1]^n)$ then the partial sums of the Fourier series of f converge to f in the L^2-norm. That is*

$$\lim_{N\to\infty} \|f - S_N(f)\|_{L^2} = 0. \tag{5.64}$$

As in one dimension the partial sums may fail to converge pointwise.

By defining L^2-partial derivatives we generalize the theory of L^2-derivatives to higher dimensions. In this case we use the integration by parts formulation.

Definition 5.7.1. A function $f \in L^2([0,1]^n)$ has an L^2-partial derivative in the x_j-direction if there is a function $f_j \in L^2([0,1]^n)$ such that for every *periodic*, once differentiable function φ we have

$$\int_{[0,1]^n} f(\mathbf{x})\partial_{x_j}\varphi(\mathbf{x})d\mathbf{x} = -\int_{[0,1]^n} f_j(\mathbf{x})\varphi(\mathbf{x})d\mathbf{x}.$$

The restriction to *periodic* test functions φ is very important. More generally we can define higher L^2-derivatives.

Definition 5.7.2. A function $f \in L^2([0,1]^n)$ has m L^2-derivatives if for each multi-index $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}| \le m$ there is a function $f_{\boldsymbol{\alpha}} \in L^2([0,1]^n)$ such that for every m-times differentiable *periodic* test function φ we have

$$\int_{[0,1]^n} f(\mathbf{x})\partial_{\mathbf{x}}^{\boldsymbol{\alpha}}\varphi(\mathbf{x})d\mathbf{x} = (-1)^{|\boldsymbol{\alpha}|}\int_{[0,1]^n} f_{\boldsymbol{\alpha}}(\mathbf{x})\varphi(\mathbf{x})d\mathbf{x}.$$

As before the standard notations are used to denote the L^2-derivatives, i.e. $\partial_{\mathbf{x}}^{\boldsymbol{\alpha}} f$.

The existence of L^2-derivatives is intimately tied to the rate of decay of the Fourier coefficients.

Proposition 5.7.5. *A function $f \in L^2([0,1]^n)$ has m L^2-derivatives if and only if*

$$\sum_{\mathbf{k}\in\mathbb{Z}^n} \|\mathbf{k}\|^{2m} |\hat{f}(\mathbf{k})|^2 < \infty.$$

In this case

$$\int_{\mathbb{R}^n} |\partial_{\mathbf{x}}^{\boldsymbol{\alpha}} f(x)|^2 dx = \sum_{\mathbf{k}\in\mathbb{Z}^n} |\mathbf{k}^{\alpha} \hat{f}(\mathbf{k})|^2$$

and

$$\widehat{\partial_{\mathbf{x}}^{\alpha} f}(\mathbf{k}) = (i\mathbf{k})^{\boldsymbol{\alpha}} \hat{f}(\mathbf{k}),$$

for every multi-index $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}| \leq m$.

As noted above, a faster rate of decay is needed in higher dimensions to be able to conclude that the Fourier coefficients are absolutely summable. In one dimension we showed that a function with one L^2-derivative is continuous. In dimension n, slightly more that $(n/2)$ derivatives are required for this conclusion.

Functions that are defined on products of intervals $[a_1, b_1] \times \cdots \times [a_n, b_n]$ can be rescaled to be defined on $[0,1]^n$ and can therefore be expanded in Fourier series as well. We leave the details of this discussion to the interested reader. While intervals are the only connected subsets of the real line, higher dimensional spaces have a rich array of such subsets. The Fourier series in higher dimensions is, of course only defined for functions that are defined in products of intervals. The analysis of functions defined in other sorts of regions requires more sophisticated mathematical techniques. For example we cannot directly apply Fourier series to study functions defined in the unit disk. The interested reader is referred to [17].

Exercise 5.7.5. Prove (5.62) and (5.63)

Exercise 5.7.6. Let $[n/2]$ be the largest integer smaller than $n/2$. Show that a periodic function with $[n/2] + 1$ L^2-derivatives is a continuous function. Hint: Use the Cauchy-Schwarz inequality.

Chapter 6

Poisson summation, Sampling and Nyquist's theorem

See: A.1, A.7.1, A.6.2.

In Chapters 3-5 we developed the mathematical tools needed to describe images and methods employed to analyze and reconstruct them. This chapter serves as a bridge between the world of pure mathematics and its applications to problems of image reconstruction. In the first sections of this chapter we imagine that our "image" is a function of a single variable $f(x)$. In a purely mathematical context x is a real number which can assume any value along a continuum of numbers. The function also takes values in a continuum, either in $\mathbb{R}, \mathbb{C}$ or perhaps $\mathbb{R}^n$. In practical applications we can only evaluate f at a finite set of points $< x_j >$. This is called *sampling*. As most of the processing takes place in digital computers, both the points $< x_j >$ and the measured values $< f(x_j) >$ are forced to lie in a preassigned, finite subset of numbers known to the computer. This is called *quantization*. The reader is urged to review section A.1, where these ideas are discussed in some detail.

But for a brief discussion of quantization, this chapter is about the consequences of sampling. We examine the fundamental question: how much information about a function is contained in a finite or infinite set of samples? Central to this analysis is the *Poisson summation formula*. This formula is a bridge between the Fourier transform and the Fourier series. While the Fourier transform is well suited to an abstract analysis of image or signal processing, it is the Fourier series that is actually used to do the work. The reason is quite simple: the Fourier transform and its inverse require integrals over the entire real line, whereas the Fourier series is phrased in terms of sums and integrals *over finite intervals*. The latter integrals are approximated by finite sums. This chapter covers the first step from the abstract world of the infinite and the infinitesimal to the real world of the finite.

6.1 Sampling and Nyquist's theorem

See: A.1.4, A.6.2, B.1.

Let f be a real or complex valued function of a real variable x. A set of points $\{x_j\}$ contained in an interval (a, b) is discrete if no subsequence converges to a point in (a, b). Recall that our basic model for a measurement is the evaluation of a function at a point. Evaluating a function on a discrete set of points is called *sampling.*

Definition 6.1.1. Suppose that $f(x)$ is a function defined in an interval (a, b) and $\{x_j\}$ is a discrete set of points in (a, b). The points $\{x_j\}$ are called the *sample points.* The values $\{f(x_j)\}$ are called the *samples* of f at the points $\{x_j\}$.

Samples provide a simple mathematical model for a set of measurements. In most applications the discrete set is of the form $\{x_0 + jl \mid j \in \mathbb{Z}\}$ where l is a fixed positive number. These are called *equally spaced samples*, the number l is called the *sample spacing.* The reciprocal of l, l^{-1} is called the *sampling rate.* Sampling theory studies the problem of reconstructing functions of a continuous variable from a set of samples and the relationship between these reconstructions and the idealized data.

Because a function cannot, in general be evaluated at a point, actual measurements involve some sort of averaging. A more accurate model for measurement is the evaluation of the convolution $\{f * \varphi(x_j)\}$, where φ is an absolutely integrable, weight function which models the measuring apparatus. The Fourier transform of $f * \varphi$ is $\hat{f}\hat{\varphi}$. The Riemann-Lebesgue Lemma, Theorem 3.2.1, implies that $\hat{\varphi}(\xi)$ tends to zero as $|\xi|$ tends to infinity and this means that the measuring apparatus attenuates the high frequency information in f. In applications one often makes the assumption that there is "no high frequency information," or that is has been filtered out.

Definition 6.1.2. A function, f whose Fourier transform is zero for $|\xi| > L$ is called *L-bandlimited* or bandlimited to $[-L, L]$.

6.1.1 Nyquist's theorem

A bandlimited function can be completely determined from its samples provided that the sample spacing is sufficiently small. This is the content of Nyquist's theorem.

Theorem 6.1.1 (Nyquist's Theorem). *If f is a square integrable function and*

$$\hat{f}(\xi) = 0 \text{ for } |\xi| > L$$

then f can be reconstructed from the samples $\{f(\frac{\pi n}{L}) \ : \ n \in \mathbb{Z}\}$.

Proof. Because $\hat{f}(\xi) = 0$ for $|\xi| > L$ the Fourier inversion formula implies that

$$f(x) = \frac{1}{2\pi} \int_{-L}^{L} \hat{f}(\xi) e^{ix\xi} d\xi, \tag{6.1}$$

see exercise 3.2.11. If we think of $\hat{f}(\xi)$ as a function *defined* on the interval $[-L, L]$ then it follows from (6.1) that the numbers $\{2\pi f(\frac{\pi n}{L})\}$ are the Fourier *coefficients* of $\hat{f}(\xi)$. The inversion formula for Fourier series then applies to give

$$\hat{f}(\xi) = \left(\frac{\pi}{L}\right) \underset{N\to\infty}{LIM} \left[\sum_{n=-N}^{N} f(\frac{n\pi}{L}) e^{-\frac{n\pi i\xi}{L}} \right] \text{ if } |\xi| < L. \tag{6.2}$$

For the remainder of the proof we use the notation

$$\left(\frac{\pi}{L}\right) \sum_{n=-\infty}^{\infty} f(\frac{n\pi}{L}) e^{-\frac{n\pi i\xi}{L}}$$

to denote this LIM. The function defined by this infinite sum is periodic of period $2L$; we can use it to express $\hat{f}(\xi)$ in the form

$$\hat{f}(\xi) = \left(\frac{\pi}{L}\right) \left[\sum_{n=-\infty}^{\infty} f(\frac{n\pi}{L}) e^{-\frac{n\pi i\xi}{L}} \right] \chi_{[-L,L]}(\xi). \tag{6.3}$$

This proves Nyquist's theorem, for a function in $L^2(\mathbb{R})$ is completely determined by its Fourier transform. □

Remark 6.1.1. The Fourier transform of a bandlimited function is absolutely integrable. The Riemann Lebesgue lemma therefore implies that such a function is continuous and tends to zero as $|x| \to \infty$, moreover absolute integrability implies square integrability.

If this were as far as we could go, Nyquist's theorem would be an interesting result of little practical use. However the original function f can be explicitly reconstructed using (6.3) in the Fourier inversion formula. To justify our manipulations we assume that f tends to zero rapidly enough so that

$$\sum_{n-\infty}^{\infty} |f(\frac{n\pi}{L})| < \infty. \tag{6.4}$$

With this understood we obtain

$$\begin{aligned} f(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \chi_{[-L,L]}(\xi) \hat{f}(\xi) e^{i\xi x} d\xi \\ &= \frac{1}{2\pi} \frac{\pi}{L} \int_{-L}^{L} \sum_{n=-\infty}^{\infty} f(\frac{n\pi}{L}) e^{ix\xi - \frac{n\pi i\xi}{L}} d\xi \\ &= \frac{1}{2L} \sum_{n=-\infty}^{\infty} f(\frac{n\pi}{L}) \int_{-L}^{L} e^{ix\xi - \frac{n\pi i\xi}{L}} d\xi \\ &= \sum_{n=-\infty}^{\infty} f(\frac{n\pi}{L}) \frac{\sin(Lx - n\pi)}{(Lx - n\pi)}. \end{aligned}$$

This gives a formula to determine the value of $f(x)$ for every $x \in \mathbb{R}$, from the samples of $\{f(\frac{n\pi}{L})\}$. If f is in $L^2(\mathbb{R})$ but does not satisfy (6.4) then the most we can assert is that

$$f(x) = \underset{N\to\infty}{LIM} \sum_{n=-N}^{N} f(\frac{n\pi}{L}) \frac{\sin(Lx - n\pi)}{(Lx - n\pi)}.$$

The exponentials $e^{\pm iLx}$ have period $\frac{2\pi}{L}$ and frequency $\frac{L}{2\pi}$. If a function is L-bandlimited then $\frac{L}{2\pi}$ is the highest frequency appearing in its Fourier representation. Nyquist's theorem states that we must sample such a function at the rate $\frac{L}{\pi}$, that is at twice its highest frequency. As we shall see, sampling at a lower rate does not provide enough information to completely determine f.

Definition 6.1.3. The optimal sampling rate for an L-bandlimited function, $\frac{L}{\pi}$, is called the *Nyquist rate.* Sampling at a lower rate is called *undersampling* and sampling at a higher rate *oversampling.*

6.1.2 Shannon-Whittaker Interpolation

See: A.6.2.

The explicit interpolation formula for $f(x)$ in terms of its samples at $\{\frac{n\pi}{L} \mid n \in \mathbb{Z}\}$:

$$f(x) = \sum_{n=-\infty}^{\infty} f(\frac{n\pi}{L}) \operatorname{sinc}(Lx - n\pi). \tag{6.5}$$

is sometimes called the Shannon-Whittaker interpolation formula. In section A.6.2 we consider other methods for interpolating a function from sampled values. These formulæ involve finite sums and only give exact reconstructions for a finite dimensional family of functions. The Shannon-Whittaker formula gives an exact reconstruction for all L-bandlimited functions. Since it requires an infinite sum, it is mostly of theoretical significance. In a practical applications only a finite part of this sum can be used. That is, we would set

$$f(x) \approx \sum_{n=-N}^{N} f(\frac{n\pi}{L}) \operatorname{sinc}(Lx - n\pi).$$

Because

$$\operatorname{sinc}(Lx - n\pi) \simeq \frac{1}{n}$$

and $\sum n^{-1} = \infty$ the partial sums of this series may converge to f very slowly. In order to get a good approximation to $f(x)$ one would therefore need to take N very large.

Formula (6.5) is only one of an infinite family of similar interpolation formulæ . Suppose that $f(x)$ is an $(L-\eta)$-bandlimited function for an $\eta > 0$. Then it is also an L-bandlimited function. This makes it possible to use oversampling to obtain more rapidly convergent interpolation formulæ . To find other similar formulæ select a function $\varphi(x)$ such that

(1). $\hat{\varphi}(\xi) = 1$ for $|\xi| \leq L - \eta$,

(2). $\hat{\varphi}(\xi) = 0$ for $|\xi| > L$.

From (6.2) it follows that

$$\hat{f}(\xi) = \left(\frac{\pi}{L}\right) \sum_{n=-\infty}^{\infty} f(\frac{\pi n}{L}) e^{-\frac{n\pi i\xi}{L}} \text{ for } |\xi| < L. \tag{6.6}$$

Since $\hat{f}(\xi)$ is supported in $[\eta - L, L - \eta]$ and $\hat{\varphi}(\xi)$ satisfies the condition (1), above it follows that

$$\hat{f}(\xi) = \hat{f}(\xi)\hat{\varphi}(\xi).$$

Using this observation and (6.2) in the Fourier inversion formula gives

$$\begin{aligned} f(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi)\hat{\varphi}(\xi) e^{i\xi x} d\xi \\ &= \frac{1}{2\pi} \frac{\pi}{L} \sum_{n=-\infty}^{\infty} f(\frac{n\pi}{L}) \int_{-L}^{L} \hat{\varphi}(\xi) e^{ix\xi - \frac{n\pi i\xi}{L}} d\xi \\ &= \frac{1}{2L} \sum_{n=-\infty}^{\infty} f(\frac{n\pi}{L}) \varphi(x - \frac{n\pi}{L}). \end{aligned} \tag{6.7}$$

This is a different interpolation formula for f; the sinc-function is replaced by $[2L]^{-1}\varphi(x)$. The Shannon-Whittaker formula, (6.5) corresponds to the choice $\hat{\varphi}(\xi) = \chi_{[-L,L]}(\xi)$.

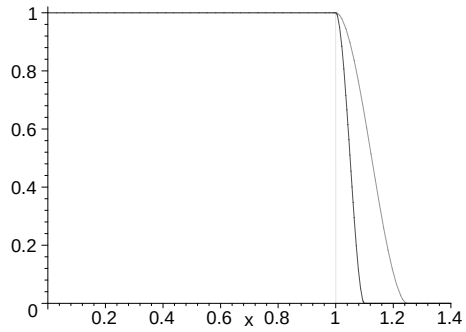
Recall that more *smoothness* in the Fourier transform of a function is reflected in faster decay of the function itself. Using a smoother function for $\hat{\varphi}(\xi)$ therefore leads to a more rapidly convergent interpolation formula for $f(x)$. There is a price to pay for using a different choice of $\hat{\varphi}(\xi)$. The first issue is that

$$\varphi(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{\varphi}(\xi) e^{ix\xi} d\xi$$

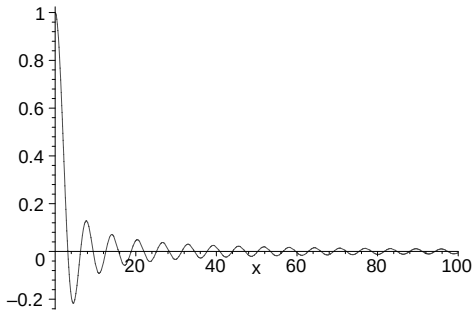
may be more difficult to accurately compute than the sinc-function. The second is that we need to sample f above the Nyquist rate. In this calculation, f is an $(L - \eta)$-bandlimited function, but we need to use a sample spacing

$$\frac{\pi}{L} < \frac{\pi}{L - \eta}.$$

On the other hand, a little oversampling and additional computational overhead often leads to far superior results. Figure 6.1(a) shows graphs of the characteristic function of an interval and two (second order) smoothed versions of this function. Figure 6.1(b) shows the ordinary sinc-pulse used in the standard Shannon-Whittaker interpolation formula. In figure 6.2 are graphs of possible interpolation functions, available if the data is oversampled. Figure 6.2(a) shows %10 oversampling, while figure 6.2(b) shows %25 oversampling. Notice how rapidly the interpolating functions, with smoother Fourier transforms decays.

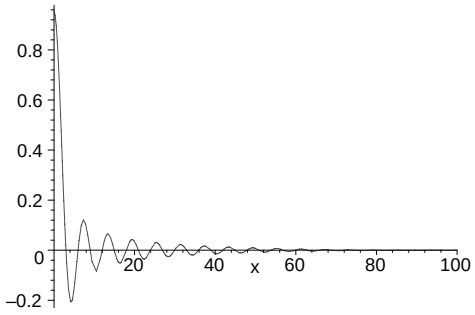


(a) Window functions in Fourier space.

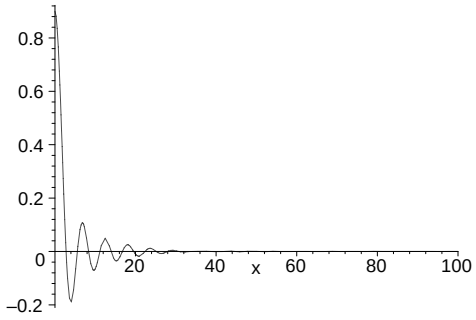


(b) Ordinary sinc-function.

Figure 6.1: Window functions in Fourier space and the ordinary sinc-pulse.



(a) %10-oversampling.



(b) %25-oversampling.

Figure 6.2: Shannon-Whittaker interpolation functions with second order smoothed windows.

Exercise 6.1.1. Use the Shannon-Whittaker formula to reconstruct the function

$$f(x) = \frac{\sin(Lx)}{\pi x}$$

from the samples $\{f(\frac{n\pi}{L})\}$.

Exercise 6.1.2. Show that for each $n \in \mathbb{N}$, function, $\text{sinc}(Lx - n\pi)$ is L-bandlimited. The Shannon-Whittaker formula therefore expresses an L-bandlimited function as a sum of such functions.

Exercise 6.1.3. The Fourier transform of

$$f(x) = \frac{1 - \cos(x)}{2\pi x}$$

is $\hat{f}(\xi) = i \operatorname{sign} \xi \chi_{[-1,1]}(\xi)$. Use the Shannon-Whittaker formula to reconstruct f from the samples $\{f(\frac{n}{\pi})\}$.

Exercise 6.1.4. Re-express the Shannon-Whittaker formula in terms of the sample spacing.

6.2 The Poisson Summation Formula

What if we do not have enough samples to satisfy the Nyquist criterion? For example, what if our signal is not band limited? Functions which describe images in medical applications generally have bounded support, so they cannot be bandlimited and therefore we are always undersampling (see Chapter 3, Proposition 3.2.10). To analyze the effects of undersampling we introduce the Poisson summation formula. It gives a relationship between the Fourier transform and the Fourier series.

6.2.1 The Poisson summation formula

Assume that f is a continuous function which decays reasonably fast as $|x| \to \infty$. We construct a periodic function out of f by summing the values of f at its integer translates. Define $f_p(x)$ by

$$f_p(x) = \sum_{n=-\infty}^{\infty} f(x+n). \tag{6.8}$$

This is a periodic function of period 1, $f_p(x+1) = f_p(x)$. If f is absolutely integrable on $\mathbb{R}$ then it follows from Fubini's theorem that f_p is absolutely integrable on $[0,1]$.

The Fourier *coefficients* of f_p are closely related to the Fourier *transform* of f :

$$\begin{aligned}
\hat{f}_p(m) &= \int_0^1 f_p(x) e^{-2\pi i m x} dx \\
&= \int_0^1 \sum_{n=-\infty}^{\infty} f(x+n) e^{-2\pi i m x} dx = \sum_{n=-\infty}^{\infty} \int_0^1 f(x+n) e^{-2\pi i m x} dx \\
&= \int_{-\infty}^{\infty} f(x) e^{-2\pi i m x} dx = \hat{f}(2\pi m).
\end{aligned}$$

The interchange of the integral and summation is easily justified if f is absolutely integrable on $\mathbb{R}$.

Proceeding formally we use the Fourier inversion formula for periodic functions, Theorem 5.1.2 to find Fourier series representation for f_p,

$$f_p(x) = \sum_{n=-\infty}^{\infty} \hat{f}_p(n) e^{2\pi i n x} = \sum_{n=-\infty}^{\infty} \hat{f}(2\pi n) e^{2\pi i n x}.$$

Note that $\{\hat{f}_p(n)\}$ are the Fourier coefficients of the 1-periodic function $f_p(x)$ whereas $\hat{f}(\xi)$ is the Fourier transform of the absolutely integrable function $f(x)$ defined on all of $\mathbb{R}$. To justify these computations it is necessary to assume that the coefficients $\hat{f}(2\pi n)$ go to zero sufficiently rapidly. If f is a sufficiently smooth function then this will be true. The *Poisson summation formula* is a precise formulation of these observations.

Theorem 6.2.1 (Poisson summation formula). *If $f(x)$ is an absolutely integrable function such that*

$$\sum_{n=-\infty}^{\infty} |\hat{f}(2\pi n)| < \infty$$

then, at points of continuity of $f_p(x)$, we have

$$\sum_{n=-\infty}^{\infty} f(x+n) = \sum_{n=-\infty}^{\infty} \hat{f}(2\pi n) e^{2\pi i n x}. \tag{6.9}$$

Remark 6.2.1. The hypotheses in the theorem are not quite optimal. Some hypotheses are required as there examples of absolutely integrable functions f such that both

$$\sum_{n=-\infty}^{\infty} |f(x+n)| \text{ and } \sum_{n=-\infty}^{\infty} |\hat{f}(2\pi n)|$$

converge but (6.9) does *not* hold. A more detailed discussion can be found in [40].

Using the argument above and rescaling one easily finds an Poisson summation formula for $2L$-periodic functions:

$$\sum_{n=-\infty}^{\infty} f(x+2nL) = \frac{1}{2L} \sum_{n=-\infty}^{\infty} \hat{f}(\frac{\pi n}{L}) e^{\frac{\pi i n x}{L}}. \tag{6.10}$$

The hypotheses are the same as those in the theorem.

As an application of (6.9) we can prove an x-space version of Nyquist's theorem. Suppose that $f(x) = 0$ outside the interval $[-L, L]$, i.e. f is a *space limited function*. For each $x \in [-L, L]$, only the $n = 0$ term on the left hand side of (6.9) is non-zero. The Poisson summation formula states that

$$f(x) = \sum_{n=-\infty}^{\infty} \hat{f}(\frac{\pi n}{L}) e^{\frac{\pi i n x}{L}} \text{ for } x \in [-L, L].$$

Therefore, if f is supported in $[-L, L]$ then it can be reconstructed from the samples of its Fourier transform

$$\{\hat{f}(\frac{\pi n}{L}) \mid n \in \mathbb{Z}\}.$$

This situation arises in magnetic resonance imaging (MRI). In this modality one directly measures samples of the Fourier transform of the image function. That is we measure

$\{\hat{f}(n\Delta\xi)\}$. On the other hand the function is known, *a priori* to be supported in a fixed bounded set $[-L, L]$. In order to reconstruct f exactly we need to take

$$\Delta\xi \leq \frac{\pi}{L}. \tag{6.11}$$

Thus, if we measure samples of the Fourier transform then Nyquist's theorem places a constraint on the sample spacing in the *Fourier domain*. An extensive discussion of sampling in MRI can be found in [80]. Figure 6.3 shows the result, in MRI, of undersampling the Fourier transform.

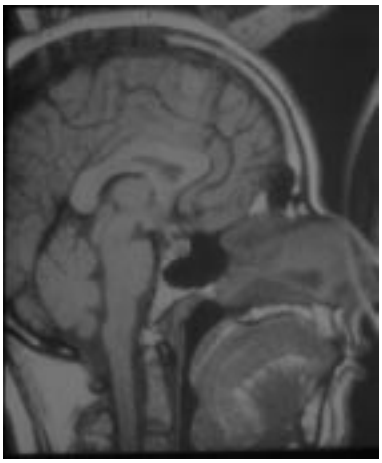


Figure 6.3: Aliasing in MRI

It most applications one samples a function rather than its Fourier transform. The analysis of undersampling in this situation requires the *dual Poisson summation formula*. Let f be a function such that the sum

$$\sum_{-\infty}^{\infty} \hat{f}(\xi + 2nL)$$

converges. Considering

$$\hat{f}_p(\xi) = \sum_{-\infty}^{\infty} \hat{f}(\xi + 2nL)$$

and its Fourier coefficients in the same manner as above we obtain:

Theorem 6.2.2 (The dual Poisson summation formula). *If f is a function such that $\hat{f}(\xi)$ is absolutely integrable and*

$$\sum_{n=-\infty}^{\infty} |f(\frac{\pi n}{L})| < \infty$$

then at, a point of continuity of $\hat{f}_p(\xi)$,

$$\sum_{n=-\infty}^{\infty} \hat{f}(\xi + 2nL) = \left(\frac{\pi}{L}\right) \sum_{n=-\infty}^{\infty} f(\frac{\pi n}{L}) e^{-\frac{n\pi i\xi}{L}}. \tag{6.12}$$

Exercise 6.2.1. Explain formula (6.11).

Exercise 6.2.2. * This exercise requires a knowledge of the Fourier transform for generalized functions, see section 3.2.13. Suppose that f is a periodic function of period 1. The generalized function l_f has a Fourier transform which is a generalized function. Using the dual Poisson summation formula, show that

$$\widehat{l_f} = 2\pi \sum_{n=-\infty}^{\infty} \hat{f}(n)\delta(2\pi n - \xi), \tag{6.13}$$

here $\{\hat{f}(n)\}$ are the Fourier coefficients defined in (5.1).

Exercise 6.2.3. * What is the analogue of formula (6.13) for a $2L$-periodic function?

6.2.2 Undersampling and aliasing

Using the Poisson summation formula we analyze the errors introduced by undersampling. Whether or not f is an L-bandlimited function, the samples $\{f(\frac{n\pi}{L}) | n \in \mathbb{Z}\}$ are interpolated by an L-bandlimited function, given by the Shannon-Whittaker formula:

$$F_L(x) = \sum_{n=-\infty}^{\infty} f(\frac{n\pi}{L}) \operatorname{sinc}(Lx - n\pi).$$

The properties of the sinc-function show that F_L interpolates f at the sample points,

$$F_L(\frac{n\pi}{L}) = f(\frac{n\pi}{L}) \text{ for } n \in \mathbb{Z}.$$

Reversing the steps in the derivation of the Shannon-Whittaker formula, we see that Fourier transform of F_L is given by

$$\widehat{F_L}(\xi) = \sum_{n=-\infty}^{\infty} \hat{f}(\xi + 2nL)\chi_{[-L,L]}(\xi). \tag{6.14}$$

If f is L-bandlimited then for all ξ we have

$$\hat{f}(\xi) = \widehat{F_L}(\xi),$$

if f is not L-bandlimited then

$$\hat{f}(\xi) - \widehat{F_L}(\xi) = \begin{cases} \hat{f}(\xi) & \text{if } |\xi| > L, \\ -\sum_{n\neq 0} \hat{f}(\xi + 2nL) & \text{if } |\xi| \leq L. \end{cases} \tag{6.15}$$

The function F_L or its Fourier transform $\hat{F}_L$ encodes all the information present in the sequence of samples. Formula (6.15) shows that there are two distinct sources of error in F_L. The first is "truncation error," as F_L is L-bandlimited the high frequency information in f is no longer available in F_L. The second source of error arises from the fact that the high frequency information in $f(x)$ *reappears* at low frequencies in the function $F_L(x)$. This

latter type of distortion is called *aliasing*. The high frequency information in the original signal is not only "lost" but resurfaces, corrupting the low frequencies. Hence F_L faithfully reproduces neither the high frequency nor the low frequency information in f.

Aliasing is familiar in everyday life: If one observes the rotation of the wheels of a fast moving car in movie, it appears that the wheels rotate very slowly. A movie image is actually a sequence of samples (24 frames/second). This sampling rate is below the Nyquist rate needed to accurately reproduce the motion of the rotating wheel.

Example 6.2.1. If a car is moving at 60mph and the tires are 3ft in diameter then the angular velocity of the wheels is

$$\omega = 58\frac{1}{3}\frac{\text{rotations}}{\text{second}}.$$

We can model the motion of a point on the wheel as $(r\cos((58\frac{1}{3})2\pi t), r\sin((58\frac{1}{3})2\pi t)$. The Nyquist rate is therefore

$$2 \cdot 58\frac{1}{3}\frac{\text{frames}}{\text{second}} \simeq 117\frac{\text{frames}}{\text{second}}.$$

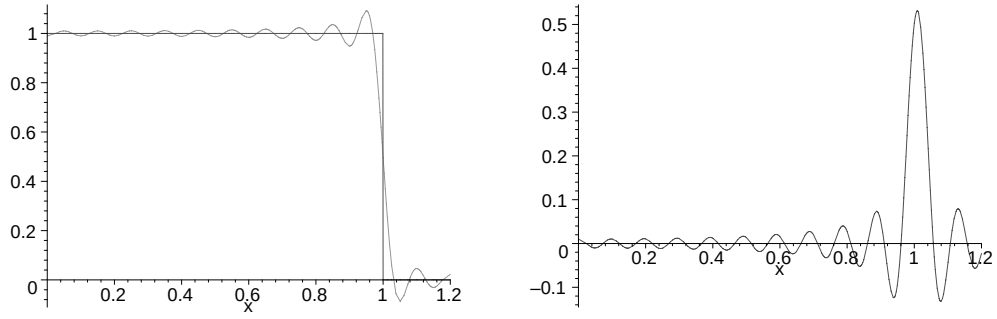
Sampling only 24 times-per-second leads to aliasing with the aliased frequencies given by

$$\pm(10\frac{1}{3}) = \pm(10\frac{1}{3} + 2 * 24).$$

The previous example is useful to conceptualize the phenomenon of aliasing but has little direct bearing on imaging. To better understand the role of aliasing in imaging we rewrite F_L in terms of its Fourier transform,

$$F_L(x) = \frac{1}{2\pi}\int_{-L}^{L} \hat{f}(\xi)e^{ix\xi}d\xi + \frac{1}{2\pi}\int_{-L}^{L}\sum_{n\neq 0}\hat{f}(\xi + 2nL)e^{ix\xi}d\xi.$$

The first term is the partial Fourier inverse of f. For a function with jump discontinuities this term produces Gibbs oscillations. The second term is the aliasing error itself. In many examples of interest in imaging, the Gibbs artifacts are as pronounced as the aliasing itself. What makes either term troublesome is slow decay of the Fourier transform.



(a) Partial Fourier inverse.

(b) Pure aliasing contribution.

Figure 6.4: The two faces of aliasing, $d = .05$.

Example 6.2.2. In figure 6.4 the two contributions to $f - F_L$ are shown separately, for the rectangle function $f(x) = \chi_{[-1,1]}(x)$. Figure 6.4(a) shows the Gibbs contribution, figure 6.4(b) shows the "pure aliasing" part. Figure 6.5 shows the original function, its partial Fourier inverse and its Shannon-Whittaker interpolant. The partial Fourier inverse is the solid line, the dotted line is the Shannon-Whittaker interpolant. In this example, the contributions of the Gibbs artifact and the pure aliasing error are of about the same size and have same general character. It is evident that the Shannon-Whittaker interpolant is more distorted than the partial inverse of the Fourier transform, though visually they are quite similar.

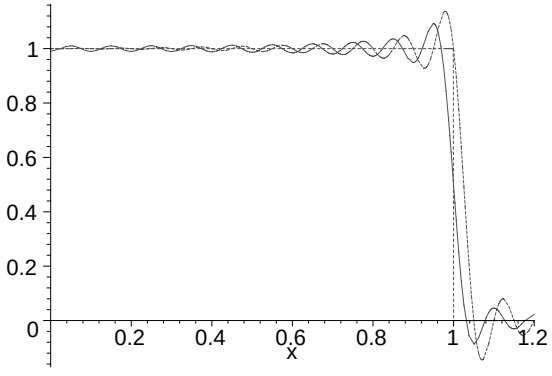
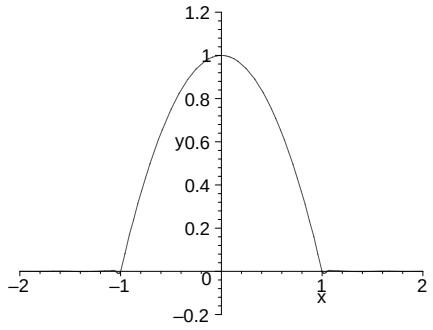


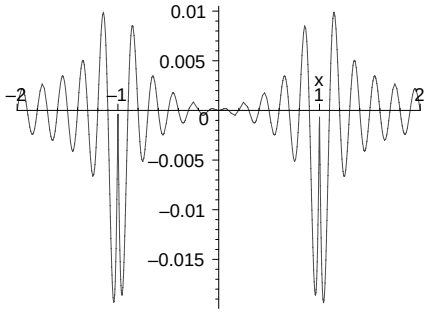
Figure 6.5: Partial Fourier inverse and Shannon-Whittaker interpolant.

Example 6.2.3. For comparison consider the continuous function $g(x) = \chi_{[-1,1]}(x)(1 - x^2)$ and its reconstruction using the sample spacing $d = .1$. In figure 6.6(a) it is just barely possible to distinguish the original function from its approximate reconstruction. The worst

errors occur near the points where g is finitely differentiable. Figure 6.6(b) shows the graph of the difference, $g - G_L$, note the scale along the y-axis.



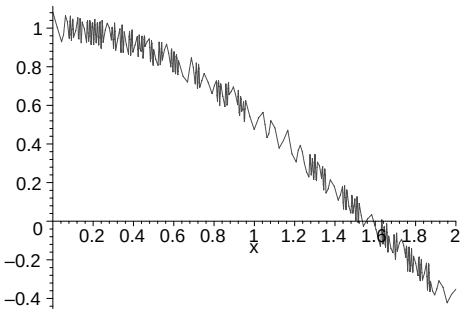
(a) The Shannon-Whittaker interpolation.



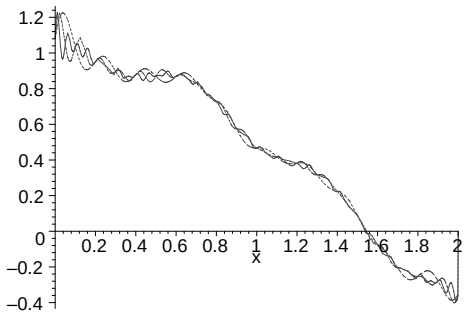
(b) The difference.

Figure 6.6: What aliasing looks like for a smoother function, $d = .1$.

Example 6.2.4. As a final example we consider the effect of sampling on a "furry function." Here we use a function of the sort introduced in example 3.2.5. These are continuous functions with "sparse," but slowly decaying Fourier transforms. Figure 6.7(a) is the graph of such a function and figure 6.7(b) shows the Shannon-Whittaker interpolants with $d = .1, .05$ and $.025$. For a function of this sort, Shannon-Whittaker interpolation appears to produce *smoothing*.



(a) A furry function.



(b) Shannon-Whittaker interpolants.

Figure 6.7: What aliasing looks like for a furry function, $d = .1, .05, .025$.

The functions encountered in imaging applications are usually spatially limited and therefore cannot be bandlimited. However, if the function f is smooth enough then its

Fourier transform decays rapidly and therefore, by choosing L sufficiently large, the difference, $\hat{f}(\xi) - \hat{F}_L(\xi)$ can be made small. One says that such a function is *effectively bandlimited*. Though this concept does not have a precise definition it is very important in imaging. In most applications it is not enough to have $\hat{f}(\xi)$ itself small outside on $[-L, L]$. Examples 6.2.2 and 6.2.3 illustrate what is meant by effective bandlimiting. As both functions have bounded support, neither is actually bandlimited. Because of the Gibbs phenomenon a Shannon-Whittaker interpolant for f displays large oscillatory artifacts, not matter how large L is taken. On the other hand, away from the jumps, the Shannon-Whittaker interpolant does a good job reconstructing f.

In applications one needs to select the sampling rate to be sufficiently large so that the aliasing error,

$$\sum_{n\neq 0} \hat{f}(\xi + 2nL)$$

is under control. Whether or not this is possible depends upon whether there exists an L so that f is effectively L-bandlimited. To diminish the effects of aliasing, an analogue signal may be passed through a "low pass filter" *before* it is sampled. An *ideal* low pass filter removes the high frequency content in the signal. In this way the sampled data accurately represents the low frequency information present in the original signal without corruption from the high frequencies. An ideal low pass filter would replace $f(x)$ with the signal $f_L(x)$ defined by the following properties:

$$\begin{aligned} &\hat{f}_L(\xi) = \hat{f}(\xi) \text{ if } |\xi| \leq L, \\ &\hat{f}_L(\xi) = 0 \text{ if } |\xi| \geq L. \end{aligned} \tag{6.16}$$

The samples $\{f_L(\frac{n\pi}{L})\}$ contain all the low frequency information in f *without* the aliasing errors. Using the Shannon-Whittaker formula to reconstruct a function, with these samples, gives $f_L(x)$ for all x. This function is just the partial Fourier inverse of f,

$$f_L(x) = \frac{1}{2\pi} \int_{-L}^{L} \hat{f}(\xi) e^{ix\xi} d\xi$$

and is still subject to unpleasant artifacts like the Gibbs phenomenon.

Realistic measurement processes lead to some sort of approximate low pass filtering. More concretely suppose that $\psi(x)$ is a function such that

(1). $\psi(x) \geq 0$ and $\operatorname{supp} \psi \subset [-\eta, \eta]$ for some positive number η.

(2).

$$\int_{-\eta}^{\eta} \psi(x) dx = 1.$$

A mathematical model for sampling the function $f(x)$ at the points $\frac{n\pi}{L}$ is the computation of the averages

$$\psi * f(\frac{n\pi}{L}) = \int_{-\eta}^{\eta} f(\frac{n\pi}{L} - x)\psi(x) dx.$$

That is, "measuring" the function f at $x = \frac{n\pi}{L}$ is the same thing as sampling the convolution $f * \psi(x)$ at $\frac{n\pi}{L}$. The Fourier transform of $\psi(x)$ goes to zero as the frequency goes to infinity; the smoother ψ is, the faster this occurs. As the Fourier transform of $\psi * f$ is

$$\widehat{\psi * f}(\xi) = \hat{\psi}(\xi)\hat{f}(\xi),$$

the measurement process itself attenuates the high frequency content of f. On the other hand

$$\hat{\psi}(0) = \int_{-\infty}^{\infty} \psi(x)dx = 1$$

and therefore $\widehat{\psi * f}(\xi)$ resembles $\hat{f}(\xi)$ for sufficiently low frequencies.

The more sharply peaked ψ is, the larger the interval over which the "measurement error,"

$$\widehat{\psi * f}(\xi) - \hat{f}(\xi) = (1 - \hat{\psi}(\xi))\hat{f}(\xi)$$

can be controlled. The aliasing error in the measured samples is

$$\sum_{n \neq 0} \hat{\psi}(\xi + 2nL)\hat{f}(\xi + 2nL).$$

By choosing ψ to be smoother, this can be made as small as one likes. If ψ is selected so that $\hat{\psi}(nL) = 0$ for $n \in \mathbb{Z} \setminus \{0\}$ then the Gibbs-like artifacts which result from truncating the Fourier transform to the interval $[-L, L]$ can also be eliminated.

Remark 6.2.2. A detailed introduction to *wavelets* which includes very interesting generalizations of the Poisson formula and the Shannon-Whittaker formula can be found in [27].

Exercise 6.2.4. Derive formula (6.14) for $\widehat{F_L}$.

Exercise 6.2.5. Compute the Fourier transform of $g(x) = \chi_{[-1,1]}(x)(1 - x^2)$.

Exercise 6.2.6. What forward velocity of the car in example 6.2.1 corresponds to the apparent rotational velocity of the wheels? What if the car is going 40mph?

Exercise 6.2.7. Sometimes in a motion picture or television image the wheels of a car appear to be going counterclockwise, even though the car is moving forward. Explain this by giving an example.

Exercise 6.2.8. Explain why the artifact produced by aliasing looks like the Gibbs phenomenon. For the function $\chi_{[-1,1]}(x)$ does the size of the pointwise error,in the Shannon-Whittaker interpolant go to zero as the sample spacing goes to zero, or not?

Exercise 6.2.9. Experiment with the family of functions

$$f_\alpha(x) = \chi_{[-1,1]}(x)(1 - x^2)^\alpha,$$

to understand effective bandlimiting. For a collection of $\alpha \in [0, 2]$ see whether there is a Gibbs like artifact in the Shannon-Whittaker interpolants and if not, at what sample spacing is the Shannon-Whittaker interpolant visually indistinguishable from the original function (over $[-2, 2]$).

Exercise 6.2.10. The ideal low pass filtered function, $f_L(x)$ can be expressed as a convolution

$$f_L(x) = f * k_L(x).$$

Find the function k_L. If the variable x is "time" explain the difficulty in implementing an ideal low pass filter.

Exercise 6.2.11. Suppose that ψ is an even function which assumes its maximum at $x = 0$ then explain why the interval over which $\widehat{\psi * f}(\xi) - \hat{f}(\xi)$ is small is controlled by

$$\int_{-\infty}^{\infty} x^2 \psi(x) dx.$$

Exercise 6.2.12. Show that if

$$\psi(x) = \varphi * \chi_{[-\frac{\pi}{L}, \frac{\pi}{L}]}(x)$$

then $\hat{\psi}(nL) = 0$ for all $n \in \mathbb{Z} \setminus \{0\}$.

6.2.3 Sub-sampling

Sub-sampling is a way to take advantage of aliasing to "demodulate" a band limited signal whose Fourier transform is supported in a set of the form $[-\omega - B, -\omega + B]$ or $[\omega - B, \omega + B]$. In this context ω is called the *carrier frequency* and $2B$ the bandwidth of the signal. This situation arises in FM-radio as well as in MR imaging. For simplicity suppose that there is a positive integer N so that

$$\omega = NB$$

Let $f(x)$ be a signal whose Fourier transform is supported in $[\omega - B, \omega + B]$. If we sample this signal at the points $\{\frac{n\pi}{B}, : n \in \mathbb{Z}\}$ and use formula (6.5) to obtain $F_L(x)$, then (6.12) implies that

$$\widehat{F}_L(\xi) = \hat{f}(\omega + \xi). \tag{6.17}$$

The function F_L is called the demodulated version of f; the two signal are very simply related

$$f(x) = e^{-i\omega x} F_L(x).$$

From the formula relating f and F_L it is clear that if F_L is real valued then, in general, the measured signal, f is not. A similar analysis applies to a signal with Fourier transform supported in $[-\omega - B, -\omega + B]$.

Exercise 6.2.13. Suppose that ω is not an integer multiple of B and that f is a signal whose Fourier transform is supported in $[\omega - B, \omega + B]$. If F_L is constructed as above from the samples $\{f(\frac{n\pi}{B})\}$ then determine $\widehat{F}_L$. What should be done to get a faithfully demodulated signal by sampling. Keep in mind that normally $\omega >> B$.

Exercise 6.2.14. Suppose that $f(x)$ is a real valued function whose Fourier transform is supported in $[-\omega - B, -\omega + B] \cup [\omega - B, \omega + B]$. Assuming that $\omega = NB$, and f is sampled at $\{\frac{n\pi}{B}\}$, how is $\widehat{F}_L$ related to $\hat{f}$?

6.2.4 Sampling periodic functions

See: A.2.10.

We now adapt the discussion of sampling to periodic functions. Let f be a function defined on the real line satisfying

$$f(x+L) = f(x).$$

The Fourier coefficients and the inversion formula are given by

$$\begin{aligned} \hat{f}(n) &= \int_0^L f(x) e^{-\frac{2\pi i n x}{L}} dx, \\ f(x) &= \frac{1}{L} \sum_{n=-\infty}^{\infty} \hat{f}(n) e^{\frac{2\pi i n x}{L}}. \end{aligned}$$

A periodic function is called band limited if there exists an N such that $\hat{f}(n) = 0$ for all n with $|n| > N$. In this case, from the inversion formula, we have

$$f(x) = \frac{1}{L} \sum_{n=-N}^{N} \hat{f}(n) e^{\frac{2\pi i n x}{L}}.$$

This is conceptually simpler than the previous case since f already lies in the finite dimensional space of functions spanned by

$$\{e^{\frac{2\pi i n x}{L}} \ : \ -N \le n \le N\}.$$

Suppose that f is sampled at $\{\frac{jL}{2N+1} : j = 0, \ldots, 2N\}$. Substituting the Fourier inversion formula into the sum gives:

$$\begin{aligned} \sum_{j=0}^{2N} f(\frac{jL}{2N+1}) e^{-\frac{2\pi i m j}{2N+1}} &= \sum_{j=0}^{2N} \frac{1}{L} \sum_{n=-N}^{N} \hat{f}(n) \exp\left(\frac{2\pi i n j L}{(2N+1)L} - \frac{2\pi i m j}{2N+1} \right) \\ &= \frac{1}{L} \sum_{n=-N}^{N} \hat{f}(n) \sum_{j=0}^{2N} \exp\left(\frac{2\pi i j}{2N+1} (n-m) \right) \\ &= \frac{2N+1}{L} \hat{f}(m). \end{aligned} \tag{6.18}$$

Going from the second to the third line uses the fact that

$$\sum_{j=0}^{2N} \exp\left(\frac{2\pi i j}{2N+1} (n-m) \right) = \begin{cases} 2N+1 & n = m, \\ 0 & n \neq m. \end{cases} \tag{6.19}$$

In this case the non-zero Fourier coefficients are easily obtained from the samples. The formulæ in (6.19) have a nice geometric interpretation: the set of vectors

$$\{(1, e^{\frac{2\pi ij}{2N+1}}, e^{\frac{4\pi ij}{2N+1}}, \ldots, e^{\frac{4N\pi ij}{2N+1}}) \mid j = 0, \ldots, 2N\}$$

are an orthogonal basis for $\mathbb{C}^{2N+1}$. These vectors are obtained by sampling the functions $\{e^{\frac{2\pi ijx}{L}} \mid \ \ j = 0, \ldots, 2N\}$ at the points $\{\frac{nL}{2N+1} \mid n = 0, 1, \ldots, 2N\}$. Formula (6.19) implies the following theorem.

Theorem 6.2.3 (Nyquist's theorem for periodic functions). *If $f(x + L) = f(x)$ and $\hat{f}(n) = 0$ for $|n| > N$, then f can be reconstructed from the samples $\{f(\frac{jL}{2N+1}),\ j = 0, 1, \ldots, 2N\}$.*

From equation (6.18) and the inversion formula, we have

$$\begin{aligned} f(x) &= \frac{1}{L} \sum_{n=-N}^{N} \hat{f}(n) e^{\frac{2\pi inx}{L}} \\ &= \frac{1}{L} \sum_{n=-N}^{N} \frac{L}{2N+1} \sum_{j=0}^{2N} f(\frac{jL}{2N+1}) e^{-\frac{2\pi inj}{2N+1}} e^{\frac{2\pi inx}{L}} \\ &= \frac{1}{2N+1} \sum_{j=0}^{2N} f(\frac{jL}{2N+1}) \sum_{n=-N}^{N} e^{-\frac{2\pi inj}{2N+1}} e^{\frac{2\pi inx}{L}} \\ &= \frac{1}{2N+1} \sum_{j=0}^{2N} f(\frac{jL}{2N+1}) \frac{\sin \pi(2N+1)(\frac{x}{L} - \frac{j}{2N+1})}{\sin \pi(\frac{x}{L} - \frac{j}{2N+1})} \end{aligned} \tag{6.20}$$

This is an interpolation formula similar to (6.5).

Even if f is not bandlimited then (6.20) defines an N-bandlimited function,

$$F_N(x) = \frac{1}{2N+1} \sum_{j=0}^{2N} f(\frac{jL}{2N+1}) \frac{\sin \pi(2N+1)(\frac{x}{L} - \frac{j}{2N+1})}{\sin \pi(\frac{x}{L} - \frac{j}{2N+1})}.$$

As before this function interpolates f at the sample points

$$F_N(\frac{jL}{2N+1}) = f(\frac{jL}{2N+1}), j = 0, 1, \ldots, 2N.$$

The Fourier coefficients of F_N are related to those of f by

$$\hat{F}_N(m) = \sum_{n=-\infty}^{\infty} \hat{f}(m + n(2N+1)) = \hat{f}(m) + \sum_{n \neq 0} \hat{f}(m + n(2N+1)) \quad -N \leq m \leq N.$$

If f is not N-bandlimited then F_N has aliasing distortion: high frequency data in f distorts the low frequencies in F_N. Of course if f is discontinuous then F_N also displays Gibbs oscillations.

Exercise 6.2.15. Prove (6.19), remember to use the Hermitian inner product!

Exercise 6.2.16. Suppose that f is an N-bandlimited, L-periodic function. Let

$$\{x_1, \ldots, x_{2N+1}\} \subset [0, L)$$

such that $x_j \neq x_k$ if $j \neq k$. Show that f can be reconstructed from the samples

$$\{f(x_j) \, : \, j = 1, \ldots, 2N+1\}.$$

From the point of view of computation, explain why equally spaced samples are preferable?

Exercise 6.2.17. Prove that

$$F_N(\frac{jL}{2N+1}) = f(\frac{jL}{2N+1}), j = 0, 1, \ldots, 2N.$$

Exercise 6.2.18. Find analogues of the generalized Shannon-Whittaker formula in the periodic case.

6.2.5 Quantization errors

See: A.1.

In the foregoing sections it is implicitly assumed that the we have a continuum of numbers at our disposal to make measurements and do computations. As digital computers are used to implement the various filters this is not the case. In section A.1.2 we briefly discuss how numbers are actually stored and represented in a computer. For simplicity we consider a base 2, fixed point representation of numbers. Suppose that we have $(n+1)$ bits and let the binary sequence $(b_0, b_1, \ldots, b_n)$ correspond to the number

$$(b_0, b_1, \ldots, b_n) \leftrightarrow (-1)^{b_0} \frac{\sum_{j=0}^{n-1} b_{j+1} 2^j}{2^n}. \tag{6.21}$$

This allows us to represent numbers between -1 and $+1$ with a maximum error of 2^{-n}. There are several ways to map the continuum onto $(n+1)$-bit binary sequences. Such a correspondence is called a *quantization* map. In essentially any approach, numbers greater than or equal to 1 are mapped to $(0, 1, \ldots, 1)$ and those less than or equal to -1 are mapped to $(1, 1, \ldots, 1)$. This is called clipping and is very undesirable in applications. To avoid clipping the data is usually scaled before it is quantized.

The two principal quantization schemes are called *rounding* and *truncation*. For a number x between -1 and $+1$ its rounding is defined to be the number of the form in (6.21) closest to x. If we denote this by $Q_r(x)$ then clearly

$$|Q_r(x) - x| \leq \frac{1}{2^{n+1}}.$$

There exist finitely many numbers which are equally close to two such numbers, for these a choice simply has to be made. If

$$x = (-1)^{b_0} \frac{\sum_{j=-\infty}^{n-1} b_{j+1} 2^j}{2^n} \text{ where } b_j \in \{0, 1\},$$

then its $(n+1)$-bit truncation corresponds to the binary sequence $(b_0, b_1, \ldots, b_n)$. If we denote this quantization map by $Q_t(x)$ then

$$0 \leq x - Q_t(x) \leq \frac{1}{2^n}.$$

We use the notation $Q(x)$ to refer to either quantization scheme.

Not only are measurements quantized but arithmetic operations are as well. The usual arithmetic operations must be followed by a quantization step in order for the result of an addition or multiplication to fit into the same number of bits. The machine uses

$$Q(Q(x) + Q(y)) \text{ and } Q(Q(x) \cdot Q(y))$$

for addition and multiplication respectively. The details of these operations depend on both the quantization scheme and the representation of numbers, i.e. fixed point or floating point. We consider only the fixed point representation. If $Q(x), Q(y)$ and $Q(x) + Q(y)$ all lie between -1 and $+1$ then no further truncation is needed to compute the sum. If $Q(x) + Q(y)$ is greater than $+1$ we have an *overflow* and if the sum is less than -1 an *underflow*. In either case the value of the sum is clipped. On the other hand if

$$Q(x) = (-1)^{b_0} \frac{\sum_{j=0}^{n-1} b_{j+1} 2^j}{2^n} \text{ and } Q(y) = (-1)^{c_0} \frac{\sum_{j=0}^{n-1} c_{j+1} 2^j}{2^n}$$

then

$$Q(x)Q(y) = (-1)^{b_0+c_0} \frac{\sum_{j,k=1}^{n-1} b_{j+1} c_{k+1} 2^{j+k}}{2^{2n}}.$$

This is essentially a $(2n+1)$-bit binary representation and therefore must be re-quantized to obtain an $(n+1)$-bit representation. Overflows and underflows cannot occur in fixed point multiplication.

It is not difficult to find numbers x and y between -1 and 1 so that $x+y$ is also between -1 and 1 but

$$Q(x + y) \neq Q(x) + Q(y).$$

This is a consequence of the fact that quantization is not a linear map! Because it is *non-linear,* quantization is difficult to analyze. An *exact* analysis requires entirely new techniques. Another approach is to regard the error $e(x) = x - Q(x)$ as *quantization noise.* If $\{x_j\}$ is a sequence of samples then $\{e_j = x_j - Q(x_j)\}$ is the quantization noise sequence. For this approach to be useful one needs to assume that the sequence $\{e_j\}$ is "random." This means that it has good statistical properties, e.g. it is of mean zero and the successive values are not highly correlated. If the original signal is sufficiently complex then this is a good approximation. However, if the original signal is too slowly varying then these assumptions may not hold. This approach is useful because it allows an analysis of the affect on the signal-to-noise ratio of the number of bits used in the quantization scheme. It is beyond the scope of this text to consider these problems in detail, a thorough treatment and references to the literature can be found in Chapter 9 of [53].

6.3 Higher dimensional sampling

In imaging applications one usually works with functions of two or three variables. Let f be a function defined on $\mathbb{R}^n$ and $\{\mathbf{x}_k\} \subset \mathbb{R}^n$, a discrete set of points. As before, the values $\{f(\mathbf{x}_k)\}$ are the *samples* of f at the *sample points* $\{\mathbf{x}_k\}$. Parts of the theory of sampling in higher dimensions exactly parallels the one dimensional theory though the problems of sampling and reconstruction are considerably more complicated. Suppose that f is a function defined on $\mathbb{R}^n$.

As in the one dimensional case samples are usually collected on a uniform grid. In this case it is more convenient to label the sample points using vectors with integer coordinates. To avoid confusion bold face letters are used to denote such vectors, i.e.

$$\mathbf{j} = (j_1, \ldots, j_n) \text{ where } j_i \in \mathbb{Z}, \quad i = 1, \ldots, n.$$

Definition 6.3.1. The *sample spacing* for a set of uniformly spaced samples in $\mathbb{R}^n$ is a vector $\mathbf{h} = (h_1, \ldots, h_n)$ with positive entries. The index $\mathbf{j}$ corresponds to the sample point

$$\mathbf{x}_\mathbf{j} = (j_1 h_1, \ldots, j_n h_n).$$

A values of a function, $\{f(\mathbf{x}_\mathbf{j})\}$ at these points is a uniform sample set.

A somewhat more general definition of uniform sampling is sometimes useful: fix n orthogonal vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. For each $\mathbf{j} = (j_1, \ldots, j_n) \in \mathbb{Z}^n$ define the point

$$\mathbf{x}_\mathbf{j} = j_1 \mathbf{v}_1 + \cdots j_n \mathbf{v}_n. \tag{6.22}$$

The set of points $\{\mathbf{x}_\mathbf{j} : \mathbf{j} \in \mathbb{Z}^n\}$ defines a uniform sample set. This sample set is the result of applying a rotation to a uniform sample set with sample spacing $(\|\mathbf{v}_1\|, \ldots, \|\mathbf{v}_n\|)$. As in the one dimensional case, the definitions of sample spacing and uniform sampling depend on the choice of coordinate system. A complication in several variables is that there are many different coordinate systems that naturally arise.

Example 6.3.1. Let $(h_1, \ldots, h_n)$ be a vector with positive coordinates. The set of points,

$$\{(j_1 h_1, \ldots, j_n h_n) : (j_1, \ldots, j_n \in \mathbb{Z}^n\}$$

is a uniform sample set.

Example 6.3.2. Let (r, θ) denote polar coordinates for $\mathbb{R}^2$; they are related to rectangular coordinates by

$$x = r\cos\theta, \quad y = r\sin\theta.$$

In CT-imaging we often encounter functions which are uniformly sampled on a *polar grid*. Let $f(r, \theta)$ be a function on $\mathbb{R}^2$ in terms of polar coordinates and let $\rho > 0$ and $M \in \mathbb{N}$ be fixed. The set of values

$$\{f(j\rho, \frac{2k\pi}{M}) : j \in \mathbb{Z}, k = 1, \ldots M\}$$

are uniform samples of f, in polar coordinates, however the points

$$\{(j\rho\cos\left(\frac{2k\pi}{M}\right), \quad j\rho\sin\left(\frac{2k\pi}{M}\right))\}$$

are *not* a uniform sample set as defined above.

In more than one dimension there are several different reasonable notions of "bandlimited" data.

Definition 6.3.2. A function f defined in $\mathbb{R}^n$ is $\mathbf{B}$-bandlimited, with $\mathbf{B} = (B_1, \ldots, B_n)$ an n-tuple of positive numbers if

$$\hat{f}(\xi_1, \ldots, \xi_n) = 0 \text{ if } |\xi_j| > B_j \text{ for } j = 1, \ldots, n. \tag{6.23}$$

Definition 6.3.3. A function f defined in $\mathbb{R}^n$ is R-bandlimited if

$$\hat{f}(\xi_1, \ldots, \xi_n) = 0 \text{ if } \|\xi\| > R \tag{6.24}$$

There are other reasonable choices as well.

The results proved above carry over easily to $\mathbf{B}$-bandlimited functions. However these generalizations are often inadequate to handle problems which arise in practice. Nyquist's theorem has an obvious generalization.

Theorem 6.3.1 (Higher dimensional Nyquist Theorem). *Let $\mathbf{B} = (B_1, \ldots, B_n)$ be an n-tuple of positive numbers. If f is a square integrable function which is $\mathbf{B}$-bandlimited then f can be reconstructed from the samples*

$$\{f(\frac{j_1\pi}{B_1}, \ldots, \frac{j_n\pi}{B_n}) \,:\, (j_1, \ldots, j_n) \in \mathbb{Z}^n\}.$$

This result is "optimal."

In order to apply this result to an R-bandlimited function one would need to collect the samples

$$\{f(\frac{j_1\pi}{R}, \ldots, \frac{j_n\pi}{R}) \,:\, (j_1, \ldots, j_m) \in \mathbb{Z}^n\}.$$

As $\hat{f}$ is known to vanish in a large part of $[-R, R]^n$ this would appear to be some sort of oversampling.

Neither Theorem 6.1.1 nor Theorem 6.3.1 say anything about non-uniform sampling. It is less of an issue in one dimension. If the vectors, $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ are linearly independent but not orthogonal, then formula (6.22) defines a set of sample points $\{\mathbf{x_j}\}$. Unfortunately Nyquist's theorem is not directly applicable to decide whether or not the set of samples $\{f(\mathbf{x_j}) \,:\, \mathbf{j} \in \mathbb{Z}^n\}$ suffices to determine f. There are many results in the mathematics literature which state that a function whose Fourier transform has certain support properties is determined by its samples on appropriate subsets, though few results give an explicit interpolation formula like (6.5). The interested reader is referred to [64] and [55].

The Poisson summation formula also has higher dimensional generalizations. If $f(\mathbf{x})$ is a rapidly decreasing function then

$$f_p(\mathbf{x}) = \sum_{\mathbf{j} \in \mathbb{Z}^n} f(\mathbf{x} + \mathbf{j})$$

is a periodic function. The Fourier coefficients of f_p are related to the Fourier transform of f in much the same way as in one dimension:

$$\begin{aligned}\widehat{f_p}(\mathbf{k}) &= \int\limits_{[0,1]^n} f_p(\mathbf{x}) e^{-2\pi i \langle \mathbf{x},\mathbf{k}\rangle} d\mathbf{x} \\ &= \int\limits_{\mathbb{R}^n} f(\mathbf{x}) e^{-2\pi i \langle \mathbf{x},\mathbf{k}\rangle} d\mathbf{x} \\ &= \hat{f}(2\pi i \mathbf{k}).\end{aligned} \tag{6.25}$$

Applying the Fourier series inversion formula with a function that is smooth enough and decays rapidly enough shows that

$$\sum_{\mathbf{j}\in\mathbb{Z}^n} f(\mathbf{x}+\mathbf{j}) = \sum_{\mathbf{k}\in\mathbb{Z}^n} \hat{f}(2\pi i\mathbf{k}) e^{2\pi i\langle \mathbf{x},\mathbf{k}\rangle}. \tag{6.26}$$

This is the n-dimensional Poisson summation formula.

The set of samples points is sometimes determined by the physical apparatus used to make the measurements. As such, one often has samples of a function, f on a non-uniform grid: $\{f(\mathbf{y}_k)\}$. To use computationally efficient methods it is often important to have samples on a uniform grid $\{\mathbf{x}_\mathbf{j}\}$. To that end *approximate* values for f, at these points, are obtained by interpolation. Most interpolation schemes involve averaging the known values at nearby points. This sort of averaging does not usually lead to smoothing and introduces new sources of error, beyond aliasing. An efficient method for multi-variable interpolation is discussed in section 8.7. Another approach is to find a computational scheme adapted to the non-uniform grid. An example of this is presented in section 8.4

Exercise 6.3.1. Prove Theorem 6.3.1.

Exercise 6.3.2. Find an n-dimensional generalization of the Shannon-Whittaker interpolation formula (6.5).

Exercise 6.3.3. Give a definition of oversampling and a generalization of formula (6.7) for the n-dimensional case.

Exercise 6.3.4. For a set of linearly independent vectors $\{\mathbf{v}_1,\ldots,\mathbf{v}_n\}$ find a notion of $\mathbf{V}$-bandlimited so that a $\mathbf{V}$-bandlimited function is determined by the samples $\{f(\mathbf{x}_\mathbf{j}) : \mathbf{j}\in\mathbb{Z}^n\}$ and this result is optimal.

Exercise 6.3.5. Using the results proved earlier about Fourier series give hypotheses on the smoothness and decay of f which are sufficient for (6.26) to be true.

Chapter 7

Filters

This chapter discusses basic concepts in filtering theory. A *filter* is the engineering term for any process that maps an *input* or collection of inputs to an *output* or collection of outputs. As inputs and outputs are generally functions of a variable (or variables), in mathematical terms, a filter is a map from one space of functions to another space of functions. Most of our discussion is devoted to linear filters, recasting our treatment of the Fourier transform in the language of linear filtering theory.

The beginning of the chapter is mostly linguistic, introducing engineering vocabulary for concepts already presented from a mathematical standpoint. In imaging applications the functions of interest usually depend on two or three *spatial* variables. The measurements themselves are, of necessity also functions of time though this dependence is often suppressed or ignored. In most of this chapter we consider filters acting on inputs which are functions of a single variable. There are three reasons for doing this: first the discussion is simpler, second, it reflects the origins of filtering theory in radio and communications, and third, filters acting on functions of several variables are usually implemented "one variable at a time." Most linear filters are expressed, at least formally as integrals. The fact that higher dimensional filters are implemented one variable at a time reflects the fact that higher dimensional integrals are actually computed as iterated, one dimensional integrals, that is

$$\iint\limits_{[a_1,b_1]\times[a_2,b_2]} f(x,y)dxdy = \int\limits_{a_1}^{b_1}\left[\int\limits_{a_2}^{b_2} f(x,y)dy\right]dx.$$

In the second part we consider the implementation of shift invariant filters on sampled data. Section7.6 presents the basic concepts of image processing.

7.1 Basic definitions

See: A.3, A.4.1, A.4.4.

In what follows the input to a filter is often referred to as a "signal." As noted above it is usually a function of one variable which is sometimes thought of as "time." For

example, an input could be the sound produced by an orchestra. The output could be a tape recording of that sound. Evidently such an input is a function of time but it is also a function of the point in space where it is measured. With this in mind the output is really the recording made by a microphone at a *fixed location.* An input could also be a function of spatial parameters such as the density of a photographic image as a function of location in the film plane or the X-ray absorption coefficient of an object as a function of a point in space. In the first case, the output might be a drawing which locates the sharp edges in the photographic image (an edge enhancing filter). For the second case the Radon transform could be considered as a filter which maps the absorption coefficient to its integrals over lines in space.

Much of the terminology in filtering theory is connected with the intended application, so a given mathematical concept, when connected to a filter has one name if the filter is used to process radio signals and a different name if the filter is used in imaging. Functional notation, similar to that used for linear transformations, is often used to denote the action of a filter, that is a filter $\mathcal{A}$ takes an input, x to the output $\mathcal{A}\,x$.

7.1.1 Examples of filters

In applications one rarely considers "arbitrary" filters. Before beginning a careful analysis of filters we consider typical examples of filters acting on functions of a single variable.

Example 7.1.1. The operation of scaling defines a filter

$$Ax(t) = ax(t),$$

here a is a positive number often called the amplification factor.

Example 7.1.2. Shifting a signal in time defines a filter. Let τ denote a constant and define

$$A_\tau x(t) = x(t - \tau).$$

Example 7.1.3. Multiplying a signal by a function defines a filter, let $\psi(t)$ denote a function and define

$$M_\psi x(t) = \psi(t)x(t).$$

Example 7.1.4. Convolution with a function defines a filter. Let $\varphi(t)$ be a function and define

$$C_\varphi x(t) = \int_{-\infty}^{\infty} \varphi(t - s)x(s)ds.$$

Example 7.1.5. If $x(t)$ is a signal depending on a single variable then differentiation defines a filter

$$Dx(t) = \frac{dx}{dt}(t).$$

Example 7.1.6. The Fourier transform, $\mathcal{F} : x \to \hat{x}$ is a filter, as is its inverse.

From the last three examples it is clear that some filters can only be applied to certain kinds of signals, e.g. signals that are sufficiently regular for D and signals that decay sufficiently rapidly for C_φ and $\mathcal{F}$. An important difference between the mathematical and

engineering approaches to filtering lies in the treatment of the spaces of inputs and outputs. Before a mathematician starts to discuss a map from a space of functions to another space of functions he or she likes to have well defined domain and target spaces, often equipped with norms. By contrast, engineers often describe a process or write down a formula without stipulating the exact nature of the inputs or the expected properties of the outputs. Of course, the implementation of the filter requires that the actual inputs produce meaningful outputs. This inevitably entails approximations and the precise relationship between the implemented filter and its theoretical description is rarely made explicit.

All the examples considered so far are linear filters, in mathematical language these are linear transformations or linear operators.

Definition 7.1.1. A *linear filter* $\mathcal{A}$ is an operation mapping inputs to outputs which satisfies the conditions:

(1). If x_1 and x_2 are a pair of inputs then

$$\mathcal{A}(x_1 + x_2) = \mathcal{A}(x_1) + \mathcal{A}(x_2).$$

(2). If x is an input and α is a constant then

$$\mathcal{A}(\alpha x) = \alpha\, \mathcal{A}(x).$$

In order to be clear about this distinction we consider some examples of *non-linear* filters.

Example 7.1.7. The squaring operation, $Sx(t) = (x(t))^2$ is a non-linear filter. This filter is the basis of FM radio.

Example 7.1.8. Suppose that the input is a pair of signals, $(x_1(t), x_2(t))$ and the output is their product

$$P(x_1, x_2)(t) = x_1(t)x_2(t).$$

If we think of one signal as fixed so the filter really acts only on the other, e.g. $x_1 \mapsto P(x_1, x_2)$, then this is a linear operation. However, as a filter acting on a pair of inputs it is non-linear

$$\begin{aligned} &P(x_1 + y_1, x_2 + y_2)(t) = (x_1(t) + y_1(t))(x_2(t) + y_2(t)) \neq \\ &x_1(t)y_1(t) + x_2(t)y_2(t) = P(x_1, x_2)(t) + P(y_1, y_2)(t). \end{aligned} \tag{7.1}$$

Example 7.1.9. An electrical diode is circuit element which only passes current moving in the positive direction. Its action is modeled by the formula

$$Rx(t) = \chi_{[0,\infty)}(x(t))x(t),$$

In electrical engineering this filter is called a *rectifier*.

Example 7.1.10. The process of quantization is defined by a non-linear filter. Suppose that for each time t the binary representation of $x(t)$ is given by

$$x(t) = \operatorname{sign} x(t) \sum_{j=-\infty}^{\infty} b_j(x(t))2^j.$$

One scheme used for quantization is called truncation, we let

$$Q_{N,M}x(t) = \operatorname{sign} x(t) \sum_{j=-M}^{N} b_j(x(t))2^j.$$

Example 7.1.11. The function $x(t)$ might represent a signal which we would like to measure and $\mathcal{A}\,x(t)$ is the result of our measurement. The measurement process itself defines the filter $\mathcal{A}$, which can either be linear or non-linear. A simple model for measurement is evaluation of a weighted average,

$$\mathcal{A}_l\,x(t) = \int_{-\infty}^{\infty} \psi(t-s)x(s)ds.$$

On the other hand many "detectors" become saturated when the signal is too strong. A model for such a device might be

$$\mathcal{A}_{nl}\,x(t) = \int_{-\infty}^{\infty} \psi(t-s)G[x(s)]ds.$$

Here $G(x)$ is a non-linear function which models the saturation of the detector, e.g.

$$G(x) = \begin{cases} x & \text{if } |x| < T, \\ T & \text{if } x \geq T, \\ -T & \text{if } x \leq -T. \end{cases}$$

A slightly different model is given by $\mathcal{A}'_{nl}x(t) = G[(\mathcal{A}_l x)(t)]$.

Exercise 7.1.1. Show that the filters in examples 7.1.1- 7.1.6 above are linear.

Exercise 7.1.2. Show that examples 7.1.7- 7.1.10 are non-linear.

7.1.2 Linear filters

See: A.4.6, A.4.7.

A filter is a map from a space of functions to a space of functions. From our finite dimensional experience it should come as no surprise that *linear* filters are much easier to analyze than *non-linear* filters. This analysis often begins by expressing the action of the filter as an integral:

$$\mathcal{A}\,x(t) = \int_{-\infty}^{\infty} a(t,s)x(s)ds.$$

The function $a(t,s)$ is called the *kernel function;* it completely describes the action of the filter on an 'arbitrary' input. For example, the linear filter $\mathcal{A}$ which assigns to an input, $x(t)$ its anti-derivative, $X(t)$ is expressed as an integral by setting

$$X(t) = \mathcal{A}\,x(t) = \int_0^\infty a(t,s)x(s)ds,$$

where

$$a(t,s) = \begin{cases} 1 & \text{if } 0 \le s \le t, \\ 0 & \text{if } s > t. \end{cases}$$

Sometimes the action of a filter is expressed as an integral even though the integral does not, strictly speaking make sense. This is the case if the kernel function is a 'generalized function.' The Hilbert transform is often "defined" by the expression

$$\mathcal{H}x(t) = \frac{i}{\pi}\int_{-\infty}^\infty \frac{x(s)}{t-s}ds.$$

The 'kernel function' for the Hilbert transform would appear to be $i/[\pi(t-s)]$. Since this function is not locally integrable, this expression requires interpretation. The Hilbert transform is actually as the Cauchy principal value of this integral. In this case the kernel is really the generalized function on $\mathbb{R}^2$ given by $\mathrm{P.V.}[i/\pi(t-s)]$.

When a filter is described by a formula, one needs to be careful that the formula makes sense for the signals one has in mind. For example if a signal $x(t)$ is non-zero for large times then an integral of the form

$$\mathcal{A}\,x = \int_{-\infty}^\infty a(t,s)x(s)ds$$

only makes unambiguous sense if $a(t,s)x(s)$ is an integrable function of s. A meaning can often be assigned to such expressions even when the integrals involved are not absolutely convergent. The Fourier transform provides a good illustration. The operation $x \mapsto \hat{x}$ is Initially defined for absolutely integrable functions by the formula

$$\hat{x}(t) = \int_{-\infty}^\infty x(s)e^{-ist}ds,$$

the kernel function is $a(t,s) = e^{-ist}$. The Parseval formula allows an extension of the Fourier transform to L^2-functions. The extended operation is thought of as a map from $L^2(\mathbb{R})$ to itself; though the integral may not exist, even as an improper Riemann integral. Using the Parseval formula and duality the Fourier transform can be further extended to generalized functions. For these extensions of the Fourier transform the integral above is a *formal* expression, it is *not* defined as a limit of Riemann sums.

Perhaps the simplest general class of filters are the *multiplication* filters. If $\psi(t)$ is a function then the operation:

$$M_\psi : x(t) \mapsto \psi(t)x(t)$$

defines a filter. This operation makes sense for very general types of inputs. In applications the multiplier ψ is frequently taken to equal one for t in an interval and zero for t outside a larger interval. In this case M_ψ *windows* the input. If ψ_1 and ψ_2 are two functions then we have the relations

$$M_{\psi_1}(M_{\psi_2}(x)) = M_{\psi_1\psi_2}(x) = M_{\psi_2}(M_{\psi_1}(x)).$$

In other words, the order in which multiplication filters are applied does not affect the outcome. Mathematically one says that multiplication filters *commute*. The next section treats another class of filters which have this property.

7.1.3 Shift invariant filters

In applications, the most important class of filters is called the *shift invariant filters.* If the input is shifted in time and such a filter is applied then the result is just shifted in time. In other words, the response of the filter to an input does not depend on the time that the input arrives.

Definition 7.1.2. A linear filter $\mathcal{A}$ is called *shift invariant* if for all real numbers τ we have the identity

$$\mathcal{A}(x_\tau) = (\mathcal{A}\,x)_\tau \text{ where } x_\tau(t) = x(t-\tau).$$

A linear shift invariant filter has a simpler kernel function. Suppose that $\mathcal{A}$ is such a filter with kernel function $a(t,s)$. Changing variables gives following equalities:

$$\mathcal{A}(x_\tau)(t) = \int_{-\infty}^{\infty} a(t,s)x_\tau(s)ds = \int_{-\infty}^{\infty} a(t,s)x(s-\tau)ds = \int_{-\infty}^{\infty} a(t,\sigma+\tau)x(\sigma)d\sigma,$$

on the other hand

$$(\mathcal{A}\,x)_\tau(t) = \mathcal{A}(x)(t-\tau) = \int_{-\infty}^{\infty} a(t-\tau,\sigma)x(\sigma)d\sigma.$$

Comparing these results shows that if $a(t,s)$ defines a shift invariant filter then

$$a(t,\sigma+\tau) = a(t-\tau,\sigma) \text{ for all } \sigma \in \mathbb{R}. \tag{7.2}$$

Setting $\sigma = 0$ gives

$$a(t,\tau) = a(t-\tau,0).$$

In other words, the kernel function $a(t,s)$, which describes the action of $\mathcal{A}$, depends only on the difference $t-s$. Setting $k(t) = a(t,0)$ gives

$$\mathcal{A}\,x(t) = \int_{-\infty}^{\infty} k(t-s)x(s)ds = k * x(t). \tag{7.3}$$

Proposition 7.1.1. *A linear filter is shift invariant if and only if it can represented as a convolution.*

The exercise completes the proof of the proposition.

Exercise 7.1.3. Suppose that a filter $\mathcal{A}$ is given by convolution with a function

$$\mathcal{A}\,x(t) = \int_{-\infty}^{\infty} \varphi(t-s)x(s)ds,$$

show that $\mathcal{A}$ is a linear shift invariant filter.

7.1.4 Harmonic components

One often assumes that the input signal has a Fourier transform, $\hat{x}(\xi)$. If the Fourier transform is written in terms of polar coordinates in the complex plane,

$$\hat{x}(\xi) = |\hat{x}(\xi)|e^{i\phi(\xi)}$$

then $|\hat{x}(\xi)|$ is called the amplitude of the signal at frequency ξ and $\phi(\xi)$ is called the phase. Because the complex exponential is 2π-periodic, the phase is not unambiguously defined. That is

$$e^{i\phi} = e^{i(\phi+2n\pi)} \text{ for any } n \in \mathbb{Z}.$$

How to choose the phase depends on the context. Often one fixes an interval of length 2π, for example $[-\pi, \pi)$ or $[0, 2\pi)$, and then insists that $\phi(\xi)$ belong to this interval. A different choice is to take the phase to be a *continuous* function of ξ.

Example 7.1.12. Suppose that $x(t) = e^{it^2}$. Using the first approach the phase $\phi_1(t)$ is computed by finding an integer n so that

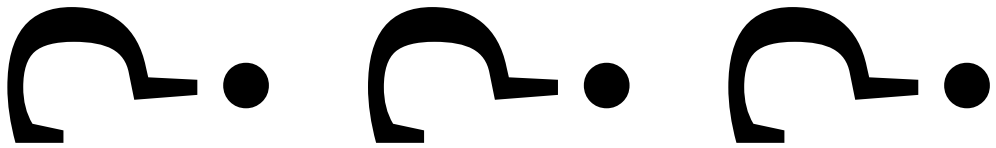
$$0 \le t^2 - 2\pi n < 2\pi;$$

the phase is then $\phi_1(t) = t^2 - 2\pi n$. In the second approach the phase is simply $\phi_2(t) = t^2$.

It is reasonable to enquire where the "information" in the Fourier transform lies. Is it more important to get the amplitude or phase correct? The answer depends on the intended application. In image processing it turns out that the *phase* is more important than the amplitude. Random errors in the amplitude produce little distortion in the reconstructed image whereas random errors in the phase produce serious artifacts. Intuitively this is reasonable as the phase of the Fourier transform encodes the relative positions of objects. Translating a function f by a vector $\boldsymbol{\tau}$, produces an overall shift in the phase of the Fourier transform,

$$\widehat{f_{\boldsymbol{\tau}}}(\boldsymbol{\xi}) = e^{i\langle \boldsymbol{\xi}, \boldsymbol{\tau}\rangle}\hat{f}(\boldsymbol{\xi}).$$

Figure 7.1(a) shows a grey scale image defined by a density function, f, figure 7.1(b) shows the results of random errors in the amplitude of $\hat{f}$ and figure 7.1(c) shows the results of random errors in the phase of $\hat{f}$. By contrast, in audio signal processing, it is often asserted that the phase carries no "useful information."



(a) A natural scene. (b) Errors in the amplitude of $\hat{f}$.. (c) Errors in the phase of $\hat{f}$..

Figure 7.1: The effects of errors in the amplitude and phase of the Fourier transform on a reconstructed image.

Physically one thinks of the Fourier transform as giving a decomposition of a signal into *harmonic components*. The actual operation of Fourier transform is somewhat at variance with an intuitive understanding of this concept. To determine the "amount" of a signal $x(t)$ at a frequency ξ, that is $\hat{x}(\xi)$, a knowledge of the signal for *all* times is required. This is because

$$\hat{x}(\xi) = \int_{-\infty}^{\infty} x(t)e^{-it\xi}dt.$$

Intuitively one thinks of a signal as having an "instantaneous frequency." For example if $x(t) = \cos(\omega t)$ for t in an interval $[t_1, t_2]$ then one would probably say that the frequency of $x(t)$, in that time interval, is $\frac{\omega}{2\pi}$. The idealized signal $x_i(t) = \cos(\omega t)$ for *all* t, its Fourier transform is a generalized function

$$\hat{x}_i(\xi) = \pi[\delta(\xi - \omega) + \delta(\xi + \omega)].$$

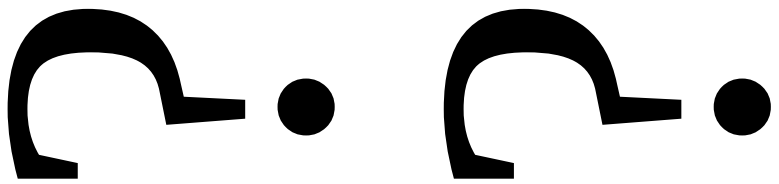
This formula can be heuristically justified by putting $\hat{x}_i(\xi)$ into the Fourier inversion formula. A real signal "at frequency $\frac{\omega}{2\pi}$" would be of the form $x_r(t) = \psi(t)\cos(\omega t)$, where $\psi(t) = 1$ over an interval $[t_1, t_2]$ and is zero outside a larger interval. The Fourier transform of the real signal is

$$\hat{x}_r(\xi) = \frac{1}{2}[\hat{\psi}(\xi - \omega) + \hat{\psi}(\xi + \omega)].$$

If $\psi(t) = 1$ over a long interval and vanishes smoothly outside it then $\hat{\psi}(\xi)$ is sharply peaked at zero and decreases rapidly as $|\xi| \to \infty$. In this case the Fourier transform of x_r is a good approximation to that of x_i.

Thus far we have used the Fourier transform and Fourier series as tools. The Fourier transform of a function is viewed as an indicator of the qualitative features of the original function, without much significance of its own or as a convenient way to represent convolution operators. In many applications it is not the function itself but its Fourier transform which contains the information of interest. This is the case if $x(t)$ describes the

state of a system which is composed of collection of resonant modes. In magnetic resonance spectroscopy a complicated molecule is caused to vibrate and emit a radio frequency signal, $x(t)$. This signal is composed of a collection of exponentially damped vibrations. Figure 7.2(a) shows a typical time series measurement. The useful information in $x(t)$ is extracted by taking the Fourier transform as shown in figure 7.2(a). The locations of the peaks determine the frequencies of the different vibrational modes and their widths give a measure of the damping. This information can in turn be used to deduce the structure of the molecule.



(a) The time series signal $x(t)$. (b) The amplitude of $\hat{x}$.

Figure 7.2: Using magnetic resonance to determine the vibrational modes of a molecule.

Spectral analysis of this sort is used throughout science and engineering; it provides a different perspective on the meaning of the Fourier transform. In magnetic resonance spectroscopy the signal decays exponentially, so little error results from cutting it off after a finite time and computing the Fourier transform of the time limited signal. In other applications, the signal does not decay, in any reasonable sense and so is regarded instead as a periodic signal. Such signals are analyzed using the Fourier series. In the world of real measurements and computation, where everything is done with finite data sets, the practical distinctions between the Fourier transform and Fourier series disappear. These distinctions remain important in the design of algorithms and the interpretation of results.

Exercise 7.1.4. Compare the results of these two approaches to the spectral analysis of a function $x(t)$ assuming that $x(t)$ *is* periodic but with a period much larger than T.

Exercise 7.1.5. Why might it be preferable to use a window function which goes smoothly from 1 to 0?

Exercise 7.1.6. Is the Fourier transform a shift invariant filter?

7.1.5 The transfer function

See: A.2.3.

The action of a shift invariant filter is expressed as a convolution in equation (7.3). If the input has a Fourier transform and the kernel function does as well then the Fourier transform of the output of such a filter is simply the product the Fourier transforms of the input and the Fourier transform of $k(t)$:

$$\widehat{k * x}(\xi) = \hat{k}(\xi)\hat{x}(\xi).$$

Definition 7.1.3. If $\mathcal{A}$ is a shift invariant filter defined by convolution with k then the Fourier transform $\hat{k}$ of k is called the *transfer function* of the filter. In imaging applications $\hat{k}(\xi)$ is called the *modulation transfer function.*

If the input to the filter is the exponential $x_\xi(t) = e^{it\xi}$ then (at least formally) the output is $\hat{k}(\xi)e^{it\xi}$. The action of the filter can be described in terms of its transfer function and the inverse Fourier transform by

$$\mathcal{A}\,x(t) = \mathcal{F}^{-1}(\hat{k}\hat{x})(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{k}(\xi)\hat{x}(\xi)e^{ix\xi}d\xi. \tag{7.4}$$

The description of a shift invariant filter in terms of its transfer function is called the *frequency space description.* This representation is a reason that linear, shift invariant filters play such a large role in applications. The Fourier transform and its inverse have very efficient approximate implementations which lead to very efficient implementations of linear, shift invariant filters. The Fourier transform converts a shift invariant filter into a multiplication filter. Both types of filters are very simple to implement though the relationship between these two classes is rather complicated.

For the moment, let us take for granted that the Fourier transform can be efficiently approximated and consider the relationship between the computations required to implement a shift invariant filter and a general linear filter. A general linear filter is represented as an integral:

$$\mathcal{A}\,x = \int_{-\infty}^{\infty} a(t,s)x(s)ds.$$

As discussed in Chapter 6, a signal is usually sampled at a finite set of times $\{s_1, \ldots, s_N\}$. that is we measure $\{x(s_j)\}$. Suppose that we would like to approximate $\mathcal{A}\,x(t_i)$ for N times $\{t_1, \ldots, t_N\}$. A reasonable way to do this is to approximate the integral defining $\mathcal{A}$ by a Riemann sum

$$\mathcal{A}\,x(t_i) \approx \sum_{j=1}^{N} a(t_i, s_j)x(s_j)(s_j - s_{j-1}).$$

Examining this formula shows that the action of the filter has been approximated by an $N \times N$ matrix multiplying an N-vector. If most of the entries of the matrix $a_{ij} = a(t_i, s_j)$ are non-zero, then this computation requires $O(N^2)$ arithmetic operations to perform.

The analogous computation for a shift invariant filter, *in the frequency space description*, is the approximate determination of the pointwise product, $\hat{k}(\xi)\hat{x}(\xi)$. This is done by evaluating $\hat{k}(\xi_j)$ and $\hat{x}(\xi_j)$ for N values of ξ and then computing the products $\{\hat{k}(\xi_j)\hat{x}(\xi_j)\}$. The matrix analogue is multiplying the vector $(\hat{x}(\xi_1), \ldots, \hat{x}(\xi_N))$ by the *diagonal* matrix

$$k_{ij} = \hat{k}(\xi_i)\delta_{ij}.$$

This requires $O(N)$ arithmetic operations. As we shall see, the approximate computation of the Fourier transform requires $O(N \log_2 N)$ operations, provided that N is a power of 2. In applications $N = 2^{10}$ is not unusual, in this case

$$N^2 = 2^{20}, \qquad N\log_2(N) \approx 2^{14}, \quad \frac{N^2}{N\log_2 N} \approx 64.$$

The impulse response of a filter can be a generalized function. Even when this is the case, the transfer function may be an ordinary function. In this case, the frequency space description of the filter is much simpler to use than its time domain description. As an example consider the filter which associates to a signal its first derivative, $Dx(t) = x'(t)$. The impulse response of this filter is the first derivative of the δ-function. It is difficult to approximate this filter as an integral, however it has a very simple frequency space description:

$$Dx(t) = \mathcal{F}^{-1}(i\xi\hat{x}(\xi))(t).$$

Of course such a filter can only be applied to a signal which has, in some reasonable sense, a first derivative. The Hilbert transform is another example of a filter whose impulse response is a generalized function but whose transfer function is an ordinary function:

$$\mathcal{H}(\delta) = \mathrm{P.V.}\,\frac{i}{\pi t} \text{ but } \mathcal{F}(\mathrm{P.V.}(\frac{i}{\pi t})) = \mathrm{sign}(\xi).$$

If we write the transfer function in polar coordinates,

$$\hat{k}(\xi) = A(\xi)e^{i\theta(\xi)},$$

where $A(\xi)$ and $\theta(\xi)$ are real numbers, then $A(\xi)$ is called the *amplitude* of the response and $e^{i\theta(\xi)}$ is called the *phase shift,*

$$\mathcal{F}(k * x) = A(\xi)|\hat{x}(\xi)|e^{i(\theta(\xi)+\phi(\xi))}.$$

If $A(\xi) > 1$ then the filter amplifies the component of the signal at frequency ξ and if $A(\xi) < 1$ it attenuates the component of the signal at this frequency. The phase of the output at frequency ξ is shifted by $\theta(\xi)$,

$$\phi(\xi) \mapsto \phi(\xi) + \theta(\xi).$$

Exercise 7.1.7. A multiplication filter has a frequency space description as a convolution. If ψ is a function such that $\hat{\psi}$ is absolutely integrable show that

$$M_\psi x(t) = \mathcal{F}^{-1}(\hat{\psi} * \hat{x})(t).$$

Exercise 7.1.8. Let ψ be a smooth function and let D denote the shift invariant filter $Dx = \partial_t x$. Compute the difference

$$DM_\psi - M_\psi D.$$

7.1.6 The δ-function revisited

See: A.4.6.

The δ-function is an idealization for a very short intense input. Formally, it is defined by the condition

$$\int_{-\infty}^{\infty} \delta(s)f(s)ds = f(0),$$

for any continuous function $f(x)$. We stress the very important fact that

$$\boxed{\text{The } \delta\text{-'function' is } \textbf{not} \text{ a function.}} \tag{7.5}$$

In the mathematics literature it is called a *distribution* and in the engineering literature a *generalized function.* One needs to exercise care when working with it, for example, its square $\delta^2(t)$ has no well defined meaning. On the other hand, one can derive in many different ways, that the Fourier transform of the delta function must be the constant function $\hat{\delta}(\xi) = 1$. Using the definition of $\delta(x)$ gives

$$\hat{\delta}(\xi) = \int e^{-i\xi x}\delta(x)dx = e^{-i\xi 0} = 1.$$

Often $\delta(s)$ is thought of as the input to a linear system. In this context it is called a *unit impulse* or *point source.* If the kernel function of $\mathcal{A}$ is $k(t-s)$ then

$$\mathcal{A}\,\delta(t) = k(t);$$

the response of a linear, shift invariant filter to the unit impulse completely determines the action of the filter. Because it is the response of the filter when the input is a unit impulse, the function $k(t)$ is called the *impulse response* of the filter $\mathcal{A}$. In imaging applications k is sometimes called the *point spread function.* In this context we usually use a spatial variable, say x, then $\delta(x)$ models a point source and the output, $A\delta(x) = k(x)$ describes how the filter spreads out a point source, hence the name. In actual applications the unit impulse is approximated by a highly localized, energetic input. For example in an acoustical problem a unit impulse might be approximated as a very short duration spark or explosion. In imaging a point source may be a bright, tiny light source.

There are many ways to mathematically approximate a δ-function. Let $\varphi(t)$ be a non-negative function of a single variable which satisfies the conditions:

$$\begin{aligned} &\varphi(0) = 1, \\ &\varphi(t) = 0, \text{ if } |t| > 1, \\ &\int_{-\infty}^{\infty} \varphi(t)dt = 1. \end{aligned} \tag{7.6}$$

The family of functions

$$\varphi_\epsilon(t) = \frac{1}{\epsilon}\varphi\left(\frac{t}{\epsilon}\right), \quad \epsilon > 0$$

gives an approximation to the δ-function in the sense that for any continuous function $f(t)$ it is true that

$$\lim_{\epsilon \downarrow 0} \int_{-\infty}^{\infty} \varphi_\epsilon(t) f(t) dt = f(0).$$

If φ_ϵ can be used as an input to $\mathcal{A}$ then, for small ϵ,

$$k_\epsilon(t) = \mathcal{A}\,\varphi_\epsilon(t)$$

provides an approximation to the impulse response of $\mathcal{A}$. Whether or not φ_ϵ can be used as an input to $\mathcal{A}$ usually depends on its smoothness. If for example $\mathcal{A} f = \partial_t f$ then the function $\chi_{-\frac{1}{2},\frac{1}{2}}(t)$ is not a good input and the corresponding scaled family does not provide a usable approximation for the impulse response.

Closely related to the δ-function is the *Heaviside function*. It is defined by

$$H(x) = \begin{cases} 0 \text{ if } x < 0 \\ 1 \text{ if } 0 \leq x. \end{cases}$$

This is of course just $\chi_{[0,\infty)}(x)$. Formally it is the indefinite integral of the δ-function

$$H(x) = \int_{-\infty}^{x} \delta(y) dy.$$

With this interpretation it is clear that the Fourier transform of H should be

$$\mathcal{F}(H)(\xi) = \frac{1}{i\xi}.$$

Because ξ^{-1} is not locally integrable, this requires interpretation. To find the correct interpretation we think of

$$H(x) = \lim_{\epsilon \downarrow 0} H(x) e^{-\epsilon x}.$$

This shows that

$$\mathcal{F}(H)(\xi) = \lim_{\epsilon \downarrow 0} \frac{1}{i\xi + \epsilon}.$$

This regularization of ξ^{-1} differs from $\mathrm{P.V.}[t^{-1}]$ used in the definition of the Hilbert transform, see section 4.3. Often the Heaviside function is used to window data, that is we define the filter

$$\mathcal{A}\,x(t) = H(t)x(t).$$

This is a multiplication filter; in the Fourier domain it is represented as a convolution by

$$\mathcal{A}\,x(t) = \hat{H} * \hat{x}(\xi) \stackrel{d}{=} \lim_{\epsilon \downarrow 0} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\hat{x}(\eta) e^{it\eta} d\eta}{\epsilon + i(\xi - \eta)}.$$

Exercise 7.1.9. The *identity filter* is defined as the filter which takes an input to itself,

$$\mathrm{Id}\, x(t) = x(t).$$

Show that the impulse response of the identity filter is $\delta(t)$. What is its transfer function.

7.1.7 Causal filters

In the context of time dependent signals, there is a special subclass of filters called *causal filters*.

Definition 7.1.4. A filter is causal if the output, at a given time, depends only upon the behavior of the signal at earlier times. For a linear filter this means that

$$\mathcal{A}\, x(t) = \int_{-\infty}^{t} a(t,s)x(s)ds$$

A linear, shift invariant filter is causal if and only if its impulse response $k(t)$ vanishes for $t < 0$.

This condition is important when working with time dependent signals if a filter must be implemented in "real time." In the context of image processing this distinction it often less important because an image is represented as a function of spatial variables. To avoid aliasing in the data acquisition step, it is useful to attenuate the high frequency components before the signal is sampled. This is called "low pass filtering" and must often be realized by a causal filter. The transfer function of a causal filter has an important analytic property.

Proposition 7.1.2. *If the filter $\mathcal{A}$ defined by $\mathcal{A}\, x = k * x$ is causal then the $\hat{k}(\xi)$ has a complex analytic extension to the lower half plane.*

Proof. The hypothesis that $\mathcal{A}$ is causal implies that $k(t) = 0$ for $t < 0$ and therefore

$$\hat{k}(\xi) = \int_{0}^{\infty} k(t)e^{-it\xi}dt.$$

If we replace ξ by $z = \xi + i\sigma$, with $\sigma < 0$ then

$$\mathrm{Re}[-i(\xi + i\sigma)] = \sigma t < 0$$

in the domain of the integration. Thus of

$$\begin{aligned} \hat{k}(z) &= \int_{0}^{\infty} k(t)e^{-itz}dt \\ &= \int_{0}^{\infty} k(t)e^{\sigma t}e^{-it\xi}dt, \end{aligned} \tag{7.7}$$

the real exponential in the integrand is decaying. Differentiating under the integral sign shows that $\partial_{\bar{z}}\hat{k} = 0$ in the lower half plane. □

Exercise 7.1.10. Prove that a linear, shift invariant filter is causal if and only if its impulse response $k(t)$ vanishes for $t < 0$.

Exercise 7.1.11. Among the examples 7.1.1–7.1.10 which are causal and which are not?

7.1.8 Bandpass filters

In filtering theory there are certain idealized filters which are frequently employed. Define the rectangle function rect(t) by

$$\text{rect}(t) = \begin{cases} 1 & |t| \leq 1/2, \\ 0 & |t| > 1/2 \end{cases}$$

and the function $\text{rect}_{[\alpha,\beta]}$ by

$$\text{rect}_{[\alpha,\beta]}(\xi) = \begin{cases} 1 & \alpha \leq \xi \leq \beta, \\ 0 & \text{otherwise.} \end{cases}$$

A *bandpass filter* is defined in the Fourier representation by

$$B_{[\alpha,\beta]}x = \mathcal{F}^{-1}[\text{rect}_{[\alpha,\beta]}(|\xi|)\hat{x}(\xi)],$$

with $0 \leq \alpha < \beta$.

The filtered signal contains only the part of the input signal with frequencies in the band $[\alpha, \beta]$, this is called the *passband*. Computing the inverse Fourier transform we see that $B_{[\alpha,\beta]}$ is represented by convolution with

$$b_{[\alpha,\beta]}(t) = 2\,\text{Re}\left[e^{-i\frac{t(\alpha+\beta)}{2}}\frac{\sin\frac{t(\beta-\alpha)}{2}}{\pi t}\right].$$

Because the sinc-function is non-zero for positive and negative times, a bandpass filter is *never* causal! If the passband is of the form $[0, B]$ the the filter is usually called a *low pass filter*. A ideal *high pass filter* has a transfer function of the form $1 - \chi_{[-B,B]}(\xi)$.

Example 7.1.13. Let $x(t)$ denote a time signal and $\hat{x}(\xi)$ its Fourier transform. The action of an ideal low pass filter H_B with passband $[-B, B]$ is given by

$$H_B x(t) = \frac{1}{2\pi}\int_{-B}^{B} \hat{x}(\xi)e^{it\xi}d\xi.$$

This is a "partial inverse" for the Fourier transform; its pointspread function is given by

$$h_B(t) = \frac{\sin(Bt)}{\pi t}.$$

In section 5.5, we describe the Gibbs phenomenon. This phenomenon also occurs for the partial inverse of the Fourier integral. If the function $x(t)$ has a jump discontinuity at t_0 and is otherwise smooth then the low pass filtered functions $\{H_B x(t)\}$ oscillate for t near to t_0 and the size of these oscillations do not decrease as B tends to infinity. On the other hand, the oscillations are concentrated in a region of size B^{-1} around the jump in $x(t)$. The underlying cause of the these oscillations is the fact that the pointspread functions $\{h_B(t)\}$ are not absolutely integrable. As in the case of the Fourier series, these oscillations can be

damped by using a smoother approximation to $\chi_{[-B,B]}$ to define the transfer function of an *approximate* low pass filter. In imaging applications a filter which attenuates the high frequencies and passes low frequencies with little change is called an *apodizing filter*. Its transfer function is called an apodizing function. We consider two examples of such filters.

Example 7.1.14. A simple example is an analogue of the Fejer mean. Instead of $\chi_{[-B,B]}$ we use the "tent"-function

$$\hat{t}_B(\xi) = \frac{1}{B}\chi_{[-\frac{B}{2},\frac{B}{2}]} * \chi_{[-\frac{B}{2},\frac{B}{2}]}(\xi).$$

This function is continuous, piecewise linear and satisfies

$$\hat{t}_B(0) = 1, \quad \hat{t}_B(\xi) = 0 \text{ if } |\xi| \geq B.$$

Its pointspread function is easily computed using the convolution theorem for the Fourier transform:

$$t_B(t) = \frac{1}{B}\left[\frac{\sin\left(\frac{Bt}{2}\right)}{\pi t}\right]^2.$$

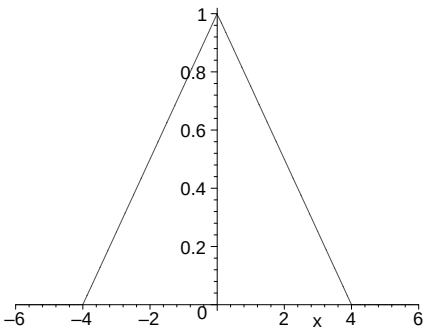


Figure 7.3: Transfer function for a tent filter.

Example 7.1.15. A second example is called the "Hanning window." Its transfer function is

$$\hat{h}_B(\xi) = \begin{cases} \cos^2\left(\frac{\pi\xi}{2B}\right) & \text{if } |\xi| < B, \\ 0 & \text{if } |\xi| > B. \end{cases}$$

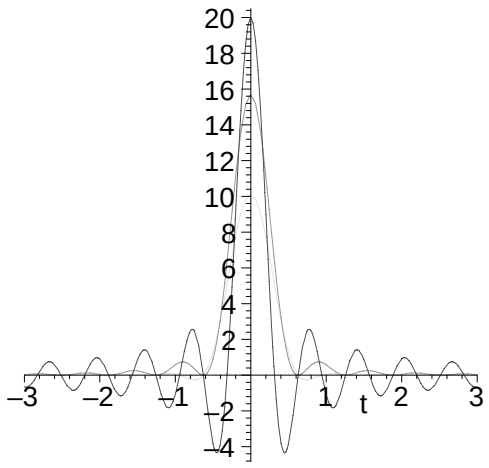


Figure 7.4: Pointspread functions for lowpass filters

This function is even smoother than the tent function, having a continuous first derivative. To compute its pointspread function we use the identity

$$\cos^2(x) = \frac{\cos(2x)+1}{2},$$

obtaining

$$h_B(t) = \left[\frac{\pi}{2B^2}\right] \frac{\sin(tB)}{t\left[\left(\frac{\pi}{B}\right)^2 - t^2\right]}.$$

This function decays like t^{-3} as t tends to infinity and is therefore absolutely integrable. The pointspread functions for the three low pass filters are shown in figure 7.4. The tallest peak corresponds to the ideal low pass filter, the middle is the tent function and the smallest comes from the Hanning window.

It is useful to generalize the notion of a bandpass filter.

Definition 7.1.5. If E is a subset of the real numbers, we define the generalized bandpass filter with passband E by defining

$$\mathcal{B}_E x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \chi_E(\xi)\hat{x}(\xi)d\xi,$$

where

$$\chi_E(\xi) = \begin{cases} 1 & \text{if } x \in E, \\ 0 & \text{if } x \notin E. \end{cases}$$

The following is an important property of a bandpass filter which is a consequence of the fact that $\chi_E^2(\xi) = \chi_E(\xi)$.

Proposition 7.1.3. *If $E \subset \mathbb{R}$ and the convolution $h * x$ is defined then*

$$\mathcal{B}_E(h * x) = (\mathcal{B}_E h) * x = h * (\mathcal{B}_E x) = (\mathcal{B}_E h) * (\mathcal{B}_E x). \tag{7.8}$$

Exercise 7.1.12. Show that a high pass filter cannot be causal.

Exercise 7.1.13. The function $\chi_{[0,\infty)}(\xi)$ defines the transfer function of a filter P which removes all negative frequency components. Show that

$$P = \frac{1}{2}(\mathrm{Id} + \mathcal{H}),$$

here $\mathcal{H}$ is the Hilbert transform.

Exercise 7.1.14. Prove Proposition 7.1.3.

7.1.9 Resolution*

In applications it is very important to have a notion of resolution. This is not a purely mathematical concept and may be defined in a variety of ways. Let $\mathcal{A}$ denote a filter. The resolution in the output of $\mathcal{A}$ is given as a length, $R_{\mathcal{A}}$. It can be variously interpreted as the size of the smallest feature which is discernible in the output, or as the minimum separation between just discernible features, or finally as the extent to which a point-like object is spread out by the filter. Whichever definition is used, the resolution *increases* as $R_{\mathcal{A}}$ *decreases*. In section 3.2.9 we discussed the "*full width half maximum*" definition of resolution. In this section we consider several other definitions applicable to linear shift invariant filters.

Suppose that $x(t)$ is a signal and $\mathcal{A}$ is a linear, shift invariant filter with pointspread function $k(t)$ and transfer function $\hat{k}(\xi)$. If $|\hat{k}(\xi)|$ is non-vanishing and stays uniformly away from zero as $|\xi| \to \infty$ then the output $Ax = k * x$ has the same resolution as the input. The input can be reconstructed by performing a bounded operation

$$x = \mathcal{F}^{-1}\left[\frac{\mathcal{F}(\mathcal{A}\,x)}{\hat{k}}\right].$$

For the remainder of this section we therefore suppose that both $k(t)$ and $\hat{k}(\xi)$ are ordinary functions which tend to zero as their arguments go to infinity.

Full width κ-maximum:

This is a family of definitions which apply to filters whose pointspread functions assume their maxima at zero and decay as $|t| \to \infty$. The case $\kappa = \frac{1}{2}$ is just the full width half maximum definition considered previously. Let $M = k(0)$ denote the maximum value attained by $k(t)$. For a number $0 < \kappa < 1$ let $t_-(\kappa) < 0 < t_+(\kappa)$ denote the largest and smallest values (respectively) where $k(t) = \kappa M$. The *full width κ-maximum* of the filter $\mathcal{A}$ is defined to be

$$\Delta_{\mathcal{A},\kappa} = t_+(\kappa) - t_-(\kappa).$$

The numbers, $\Delta_{\mathcal{A},\kappa}$ increase as κ decreases, hence this definition of resolution is more stringent for smaller values of κ.

Example 7.1.16. An extreme case is the rectangle function $(2d)^{-1}\chi_{[-d,d]}(t)$. The corresponding filters $\mathcal{A}_d$ average the signal over an interval of length of $2d$. In these cases

$$\Delta_{\mathcal{A}_d,\kappa} = 2d$$

for all values of κ.

Example 7.1.17. A less extreme case is provided by the tent functions

$$t_d(t) = \begin{cases} 0 & \text{if } |t| \geq d, \\ t+d & \text{if } -d < t < 0, \\ d-t & \text{if } 0 \geq t < d. \end{cases}$$

Letting T_d denote the corresponding filters we see that

$$\Delta_{T_d,\kappa} = 2(1-\kappa)d.$$

First zero:

If the pointspread function of a filter vanishes then the locations of the first positive and negative zeros can be used to give another definition of resolution. This is just the previous definition with $\kappa = 0$. Suppose that $k(t)$ is the pointspread function of a filter $\mathcal{A}$ and it vanishes at positive and negative values. Let $t_- < 0$ be the largest negative zero and $t_+ > 0$ be the smallest positive zero, define

$$\Delta_{\mathcal{A},0} = t_+ - t_-.$$

Example 7.1.18. Let F_B denote the Fejer mean filter with transfer function $\hat{t}_B(\xi)$ and pointspread function

$$t_B(t) = \frac{1}{B}\left[\frac{\sin\left(\frac{Bt}{2}\right)}{\pi t}\right]^2,$$

see example 7.1.14. The ideal low pass filter with passband $[-B, B]$ has pointspread function

$$d_B(t) = \frac{\sin(Bt)}{\pi t}.$$

We see that

$$\Delta_{F_B,0} = \frac{2\pi}{B} \text{ and } \Delta_{D_B,0} = \frac{\pi}{B}.$$

By this measure, the output of D_B has twice the resolution of the output of F_B. This should be compared with the computations of the full width half maxima given in section 3.2.9.

Equivalent width:

Suppose that $\mathcal{A}$ is a filter with pointspread function $k(t)$ which assumes its maximum value at 0 and is an integrable function. We then define the *equivalent width* resolution of $\mathcal{A}$ to be

$$\Delta_{\mathcal{A},\text{ew}} = \frac{\int_{-\infty}^{\infty} k(t)dt}{k(0)} = \frac{\hat{k}(0)}{k(0)}.$$

This is a measure of how much the filter smears out a point source. The number $\Delta_{\mathcal{A},\text{ew}}$ is the width of the rectangle function enclosing the same *signed area* as the graph of $k(t)/k(0)$. The fact that we use signed area has a significant effect on the value and interpretation of this number. It is not difficult to construct an example of a function $k(t)$ so that $k(0) = 1$ and

$$\int_{-\infty}^{\infty} k(t)dt = 0.$$

For the filter $\mathcal{A}$, defined by this example, $\Delta_{\mathcal{A},\text{ew}} = 0$, in other words, by this measure there is no loss in resolution.

More pertinent examples are provided by D_B and F_B; for these filters we have

$$\Delta_{F_B,\text{ew}} = \frac{4\pi^2}{B} \text{ and } \Delta_{D_B,\text{ew}} = \frac{\pi}{B}.$$

Using the previous definition, D_B has twice the resolution of F_B whereas, using the equivalent width definition, D_B has $4\pi \approx 12$ times the resolution of F_B. This is a reflection of the fact that d_B assumes both positive and negative values while F_B has a non-negative pointspread function. The equivalent width definition rewards filters with negative sidelobes. In light of this, it might be more meaningful to modify this definition by instead using the integral of $|k(t)|$. With the latter definition, the equivalent width of D_B would be infinite! Equivalent absolute width only gives a useful measure of resolution for filters with absolutely integrable pointspread functions.

ϵ-Nyquist width:

Suppose that the transfer function, $\hat{k}$ of a filter $\mathcal{A}$ is non-vanishing in an interval $[-B, B]$ and zero outside a slightly larger interval. The Nyquist criterion gives a heuristic for measuring the resolution in the output of such a filter. To perfectly reconstruct a B-bandlimited signal from uniformly spaced samples requires that the sample spacing d satisfy $d \leq \pi B^{-1}$. On the other hand, if a signal is uniformly sampled at points, separated by a distance d then it is quite clear that the resolution in the sampled data is at most d. With this is mind we define the *Nyquist width* of the filter $\mathcal{A}$ to be

$$\Delta_{\mathcal{A},\text{ny}} = \frac{\pi}{B}.$$

With this definition D_B and F_B satisfy

$$\Delta_{D_B,\text{ny}} = \frac{\pi}{B} = \Delta_{F_B,\text{ny}}.$$

This definition, evidently leaves something to be desired. If we suppose that k is real and that $|\hat{k}(\xi)|$ assumes its maximum at $\xi = 0$ then we can get a more satisfactory definition by adding a parameter $0 < \epsilon < 1$. Let ξ_- be the largest negative value of ξ so that $|\hat{k}(\xi)| < \epsilon|\hat{k}(0)|$ and let ξ_+ be the smallest positive value of ξ so that $|\hat{k}(\xi)| \leq \epsilon|\hat{k}(0)|$. We define the *$\epsilon$-Nyquist width* to be

$$\Delta_{\mathcal{A},\text{ny},\epsilon} = \frac{\pi}{\min\{|\xi_+|, |\xi_-|\}}.$$

With this definition we see that

$$\Delta_{D_B,\mathrm{ny},\frac{1}{2}} = \frac{\pi}{B} \text{ while } \Delta_{F_B,\mathrm{ny},\frac{1}{2}} = \frac{2\pi}{B},$$

which is in better agreement with our intuitive notion of resolution.

Exercise 7.1.15. Determine the resolution, according to each definition for the low pass filters defined in section 7.1.6.

Exercise 7.1.16. Suppose that $k(t)$ is the point spread function of a filter, for $0 < B$ define $k_B(t) = B^{-1}k(B^{-1}t)$. For each definition of resolution, how is the resolution of the filter defined by k_B related to that defined by k?

7.1.10 Cascaded filters

See: A.2.3, A.2.6.

Suppose that $\{\mathcal{A}_1, \ldots, \mathcal{A}_k\}$ are linear filters. The output of one can be used as the input to another, this is the filtering analogue of the composition of linear maps,

$$x \to \mathcal{A}_1\, x \to \mathcal{A}_2(\mathcal{A}_1\, x) \to \cdots \to \mathcal{A}_k(\mathcal{A}_{k-1}(\ldots(\mathcal{A}_1\, x)\ldots)) = \mathcal{A}_k \circ \cdots \circ \mathcal{A}_1\, x.$$

In this way, complex filtering operations are built up out of simpler pieces. A filter built in this way is called a *cascade* of filters. For general linear filters the order in which the operations are performed is quite important. In section 7.1.5 we saw that the action of a linear filter is analogous to multiplying a vector by a matrix. Cascading filters is then analogous to matrix multiplication. It is a familiar fact from linear algebra that if A and B are non-diagonal matrices then generally

$$AB \neq BA.$$

For general linear filters $\mathcal{A}_1$ and $\mathcal{A}_2$ it is also the case that

$$\mathcal{A}_1 \circ \mathcal{A}_2 \neq \mathcal{A}_2 \circ \mathcal{A}_1.$$

The shift invariant case, which is analogous to multiplication of a vector by a *diagonal matrix*, is much simpler. If the filters $\{\mathcal{A}_j\}$ have impulses responses $\{h_j\}$ then the cascade $\mathcal{A}_k \circ \cdots \circ \mathcal{A}_1$ is given by

$$x \mapsto (h_k * \cdots * h_1) * x.$$

This is again a shift invariant filter with impulse response $h_k * \cdots * h_1$. From the commutativity property of the convolution product,

$$f * g = g * f,$$

it follows that the result of applying a cascade of shift invariant, linear filters is, *in principle*, independent of the order in which the component filters are applied. In actual practice, where the actions of the filters can only be approximated, this is often not the case. Different orders of processes can produce substantially different outputs.

Example 7.1.19. Suppose that

$$\mathcal{A}_1\, x(t) = \partial_t x(t)$$

and

$$\mathcal{A}_2\, x(t) = \int_{-\infty}^{\infty} \varphi(t-s)x(s)ds,$$

where $\varphi(t)$ is a differentiable function, vanishing outside a bounded interval. The two possible compositions are

$$\mathcal{A}_1 \circ \mathcal{A}_2\, x(t) = \int_{-\infty}^{\infty} \partial_t\varphi(t-s)x(s)ds,$$

and

$$\mathcal{A}_2 \circ \mathcal{A}_1\, x(t) = \int_{-\infty}^{\infty} \varphi(t-s)\partial_s x(s)ds.$$

To implement the first case we need to approximate the convolution $\varphi_t * x$ whereas in the second case we first need to approximate $\partial_t x$ and then the convolution $\varphi * x_t$. Because of the difficulties in approximating differentiation, the composition $\mathcal{A}_1 \circ \mathcal{A}_2$ is much easier to implement than $\mathcal{A}_2 \circ \mathcal{A}_1$.

Example 7.1.20. If $\mathcal{A}_1$ is shift invariant and $\mathcal{A}_2$ is not then generally $\mathcal{A}_1 \circ \mathcal{A}_2 \neq \mathcal{A}_2 \circ \mathcal{A}_1$. As an example let $\mathcal{A}_1\, x = \partial_t x$ and

$$\mathcal{A}_2\, x(t) = \int_{-\infty}^{\infty} (s+t)x(s)ds.$$

A direct computation shows that

$$\mathcal{A}_1 \circ \mathcal{A}_2\, x(t) = \int_{-\infty}^{\infty} x(s)ds$$

whereas, integrating by parts gives

$$\mathcal{A}_2 \circ \mathcal{A}_1\, x(t) = -\int_{-\infty}^{\infty} x(s)ds.$$

The transfer function for the cascaded filter defined by impulse response $h = h_k * \cdots * h_1$ is the product of the transfer functions

$$\hat{h}(\xi) = \hat{h}_k(\xi) \cdots \hat{h}_1(\xi).$$

In the implementation of a cascade, using the Fourier representation, it is important to account for the limitations of finite precision arithmetic when selecting the order in which to multiply the terms in the transfer function. By grouping terms carefully one can take advantage of cancelations between large and small factors, thereby avoiding overflows or underflows.

Exercise 7.1.17. Assuming that $x(t)$ and $x_t(t)$ are absolutely integrable, prove the formulæin example 7.1.20.

7.1.11 The resolution of a cascade of filters*

In section 7.1.9 we discussed a variety of definitions for the resolution available in the output of a linear, shift invariant filter. If $\mathcal{A}_1$ and $\mathcal{A}_2$ are such filters it is reasonable to enquire how the resolution of $\mathcal{A}_1 \circ \mathcal{A}_2$ is related to the resolution of the components. The answers depend on the type of filter and the choice of definition.

Full width κ-maximum:

For values of κ between 0 and 1 and general filters it is difficult to relate $\Delta_{\mathcal{A}_1 \circ \mathcal{A}_2,\kappa}$ to $\Delta_{\mathcal{A}_1,\kappa}$ and $\Delta_{\mathcal{A}_2,\kappa}$. For the special case of filters with Gaussian pointspread functions there is a simple relation. For each $a > 0$, set

$$g_a(t) = e^{-at^2}$$

and let G_a denote the filter with this pointspread function. A simple calculation shows that

$$\Delta_{G_a,\kappa} = \sqrt{\frac{-\log \kappa}{a}}.$$

Using the fact that

$$\mathcal{F}(g_a)(\xi) = \sqrt{\frac{\pi}{a}} e^{-\frac{\xi^2}{4a}}$$

we obtain the relation

$$\Delta_{G_a \circ G_b,\kappa} = \sqrt{\Delta^2_{G_a,\kappa} + \Delta^2_{G_b,\kappa}}. \tag{7.9}$$

First zero:

If k_1 and k_2, the pointspread functions of filters $\mathcal{A}_1$ and $\mathcal{A}_2$, have the following additional properties

- k_1 and k_2 are even functions, i.e. $k_j(-t) = k_j(t)$,
- Each function is positive in an interval $(-t_j, t_j)$ and vanishes outside this interval

then

$$\Delta_{\mathcal{A}_1 \circ \mathcal{A}_2,0} = \Delta_{\mathcal{A}_1,0} + \Delta_{\mathcal{A}_2,0}. \tag{7.10}$$

This follows from the fact that the pointspread function of $\mathcal{A}_1 \circ \mathcal{A}_2$ is $k_1 * k_2$ and the support properties of convolutions given in Lemma 3.3.2.

Equivalent width:

If k_1 and k_2, the pointspread functions of filters $\mathcal{A}_1$ and $\mathcal{A}_2$, are non-negative and assume their maximum values are 0 then

$$\Delta_{\mathcal{A}_1 \circ \mathcal{A}_2,\mathrm{ew}} \geq \sqrt{\Delta_{\mathcal{A}_1,\mathrm{ew}} \Delta_{\mathcal{A}_2,\mathrm{ew}}}.$$

The proof of this estimate uses the mean value theorem for integrals, Theorem B.8.4. We can suppose that $k_1(0) = k_2(0) = 1$ as this does not affect the equivalent width. Because both functions are non-negative, the MVT for integrals applies to give constants t_1 and t_2 such that

$$k_1 * k_2(0) = \int_{-\infty}^{\infty} k_1(t)k_2(-t)dt = k_1(t_1) \int_{-\infty}^{\infty} k_2(-t)dt = k_2(t_2) \int_{-\infty}^{\infty} k_1(t)dt.$$

On the other hand

$$\int_{-\infty}^{\infty} k_1 * k_2(t)dt = \int_{-\infty}^{\infty} k_1(t)dt \int_{-\infty}^{\infty} k_2(t)dt.$$

Thus

$$\begin{aligned} \Delta_{\mathcal{A}_1 \circ \mathcal{A}_2,\mathrm{ew}} &= \frac{\int_{-\infty}^{\infty} k_1 * k_2(t)dt}{k_1 * k_2(0)} \\ &= \frac{\int_{-\infty}^{\infty} k_1(t)dt \int_{-\infty}^{\infty} k_2(t)dt}{\sqrt{k_1(t_1)k_2(t_2) \int k_1 dt \int k_2 dt}} \\ &= \sqrt{\frac{\Delta_{\mathcal{A}_1,\mathrm{ew}}\Delta_{\mathcal{A}_2,\mathrm{ew}}}{k_1(t_1)k_2(t_2)}}. \end{aligned} \tag{7.11}$$

The proof is completed by observing that $k_1(t_1)k_2(t_2) \leq 1$.

ϵ-Nyquist width:

The simplest and most general relation holds for the 0-Nyquist width. If $\hat{k}_1$ and $\hat{k}_2$ are the transfer functions for filters $\mathcal{A}_1$ and $\mathcal{A}_2$, then the transfer function of $\mathcal{A}_1 \circ \mathcal{A}_2$ is $\hat{k}_1\hat{k}_2$. Since $\hat{k}_1(\xi)\hat{k}_2(\xi) \neq 0$ if and only if each factor is non-vanishing it follows that

$$\Delta_{\mathcal{A}_1 \circ \mathcal{A}_2,\mathrm{ny},0} = \max\{\Delta_{\mathcal{A}_1,\mathrm{ny},0}, \Delta_{\mathcal{A}_2,\mathrm{ny},0}\}. \tag{7.12}$$

For $\epsilon > 0$ there is a somewhat less precise result:

$$\max\{\Delta_{\mathcal{A}_1,\mathrm{ny},\epsilon^2}, \Delta_{\mathcal{A}_2,\mathrm{ny},\epsilon^2}\} \leq \Delta_{\mathcal{A}_1 \circ \mathcal{A}_2,\mathrm{ny},\epsilon^2} \leq \max\{\Delta_{\mathcal{A}_1,\mathrm{ny},\epsilon}, \Delta_{\mathcal{A}_2,\mathrm{ny},\epsilon}\}. \tag{7.13}$$

Exercise 7.1.18. Prove (7.9).

Exercise 7.1.19. Provide the details of the proof (7.10).

Exercise 7.1.20. Suppose that for $j = 1, 2$ the pointspread functions $k_j(t)$ are positive in an interval $(-t_{j-}, t_{j+})$ where $t_{j-} < 0 < t_{j+}$ and vanish otherwise. Show that

$$\Delta_{\mathcal{A}_1 \circ \mathcal{A}_2,0} \geq \Delta_{\mathcal{A}_1,0} + \Delta_{\mathcal{A}_2,0}.$$

Exercise 7.1.21. Prove (7.12) and (7.13).

7.1.12 Filters and RLC-circuits*

In many applications it is important to filter signals in real time. For example, before sampling a time series, it is preferable to pass the analogue signal through a low pass filter. This reduces the effects of aliasing. A simple way to do this is to use RLC-circuits. These implement *passive linear filters* and are built out of three basic components known as *resistors, capacitors and inductors.* These are "ideal" circuit elements, characterized by the relationship between the current $I(t)$, through the device and the voltage $V(t)$, across the device:

(1). A *resistor* is characterized by its *impedence* which is a positive real number R. The voltage and current then satisfy the relationship

$$V(t) = RI(t). \tag{7.14}$$

(2). A *capacitor* is characterized by its *capacitance* which is a positive number C. The voltage and current then satisfy the relationship

$$I(t) = C\frac{dV}{dt}(t). \tag{7.15}$$

(3). An *inductor* is characterized by its *inductance* which is a positive number L. The voltage and current then satisfy the relationship

$$V(t) = L\frac{dI}{dt}(t). \tag{7.16}$$

A measurement of the current through a capacitor gives the derivative of the voltage, while the change in the current through an inductor over a time interval computes the definite integral of the voltage.

Impedence is a measure of how much an electrical device impedes the flow of current, as such, only a resistor has a well defined impedence. The extent to which a capacitor or inductor impedes the flow of current depends on the frequency of the input. Taking the Fourier transforms of the relations above we obtain

$$\begin{aligned} \text{Resistor} \qquad & \hat{V}(\xi) = R\hat{I}(\xi), \\ \text{Capacitor} \qquad & \hat{V}(\xi) = \frac{1}{i\xi C}\hat{I}(\xi), \\ \text{Inductor} \qquad & \hat{V}(\xi) = i\xi L\hat{I}(\xi). \end{aligned} \tag{7.17}$$

This shows that the *effective impedence* of a capacitor decreases as the frequency increases while that of an inductor increases with frequency.

Kirchoff's laws

Using these components, circuits can be built which approximate low pass, high pass and bandpass filters *and* operate in real time. *Kirchoff's laws* allow such circuits to be analyzed. We briefly describe them and then analyze some very simple circuits. To describe Kirchoff's laws, a circuit is considered to be a directed network consisting of nodes and edges. Each edge is oriented to define the positive direction of current flow.

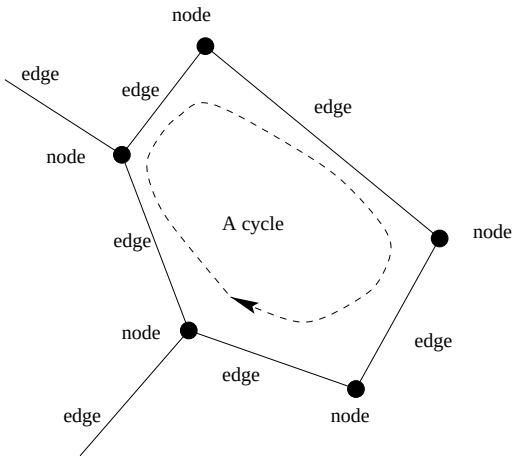


Figure 7.5: A network

One of the ideal circuit elements is placed along each edge, In circuit diagrams resistors, capacitors and inductors are represented by the symbols shown in figure 7.6.

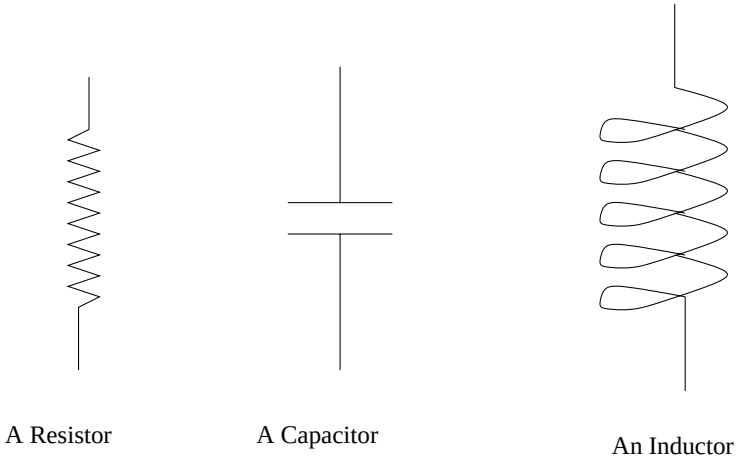


Figure 7.6: Standard symbols for passive circuit elements

Figure 7.7 shows two simple circuit diagrams. The circle indicates a *voltage source* which is used to model the input to the filter defined by the circuit. Equations (7.14)-(7.16), give the relationships between the current and voltage along each edge. There is a current flowing *through* each edge and a voltage drop *across* each circuit element. Kirchoff's

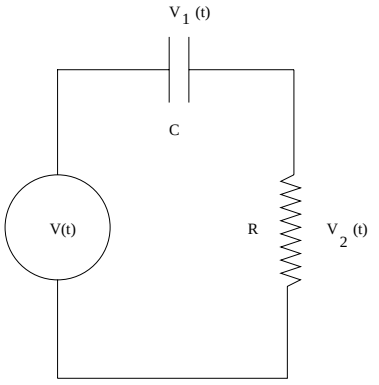
laws relate the currents flowing into a node and the voltage drops around closed paths or *cycles* in the circuit:

- At each node the sum of the currents entering and leaving is zero. Thinking of the current as a flow of electrons, this is just the "conservation of charge."
- A cycle is a closed path in a network. The sum of the voltage drops around any cycle is zero. This is the "conservation of energy:" a charged particle which travels through the circuit and returns to its starting point should experience no net change in its energy.

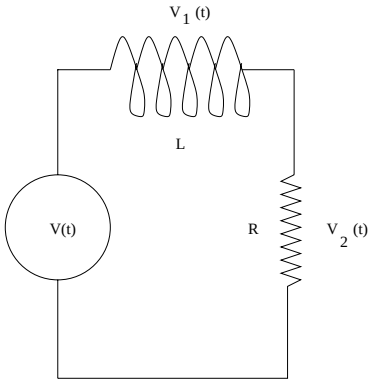
Using the defining relations, (7.14), (7.15) and (7.16) and Kirchoff's laws one can obtain a system of differential equations which relate the voltages and currents for the edges of an RLC circuit. To use an RLC-circuit as a filter one imagines that the input is a voltage source connected to a pair of nodes and the output is the voltage measured between two other nodes in the circuit. We now consider some very simple examples.

Approximate high and low pass filters*

High and low pass filters can be crudely approximated using just two circuit elements as shown in figure 7.7. The input is a voltage source $V(t)$ and the output is the voltage $V_2(t)$, measured across the resistor.



(a) RC-circuit, a "high pass filter."



(b) RL-circuit, a "low pass" filter.

Figure 7.7: Simple RLC-circuits

Example 7.1.21. Circuit (7.7)(a) is a capacitor "in series" with the voltage source. The circuit consists of a single loop so the current is the same everywhere. Kirchoff's law for voltage drops gives

$$V_1 + V_2 = V.$$

Using the defining relations (7.14) and (7.15) this implies that the current $I(t)$ satisfies

$$C\frac{dV}{dt} = I + RC\frac{dI}{dt}. \tag{7.18}$$

The causal solution of this equation is given by

$$I(t) = C \int_{-\infty}^{t} e^{RC(t-s)} \frac{dV}{dt}(s) ds.$$

The voltage across the resistor is therefore

$$V_2(t) = RC \int_{-\infty}^{t} e^{RC(t-s)} \frac{dV}{dt}(s) ds. \tag{7.19}$$

To see how this filters affects different frequency components of the input voltage we take the Fourier transform of (7.18) to obtain

$$\hat{I}(\xi) = \frac{i\xi C \hat{V}(\xi)}{1 + iRC\xi}.$$

The Fourier transform of $V_2(t)$ is therefore

$$\hat{V}_2(\xi) = \frac{i\xi RC \hat{V}(\xi)}{1 + iRC\xi}.$$

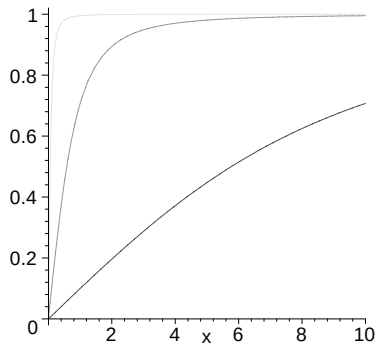
The transfer function of this filter is

$$\hat{h}(\xi) = \frac{i\xi RC}{1 + iRC\xi}.$$

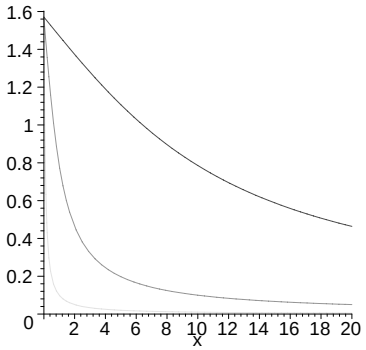
Its modulus and phase for three different values of RC are plotted below. As $h(0) = 0$ and

$$\lim_{\xi \to \pm\infty} \hat{h}(\xi) = 1$$

this circuit gives a crude approximation to a high pass filter.



(a) Filter amplitude, $|\hat{h}(\xi)|$



(b) Filter phase shift, $\phi(\xi)$

Figure 7.8: The amplitude and phase of the transfer function of an RC-filter.

Example 7.1.22. Now we analyze the affect of an inductor in series with a voltage source. Again the circuit consists of a single loop so the current is everywhere the same. Kirchoff's law for voltage drops gives

$$V_1 + V_2 = V.$$

Using the defining relations (7.14) and (7.16) this implies that the current $I(t)$ satisfies

$$V(t) = RI + L\frac{dI}{dt}. \tag{7.20}$$

The causal solution of this equation is given by

$$I(t) = \frac{1}{R}\int_{-\infty}^{t} e^{\frac{L}{R}(t-s)}V(s)ds.$$

The voltage across the resistor is therefore

$$V_2(t) = \int_{-\infty}^{t} e^{\frac{L}{R}(t-s)}V(s)ds. \tag{7.21}$$

To see how this filters affects different frequency components of the input voltage we take the Fourier transform of (7.20) to obtain

$$\hat{I}(\xi) = \frac{\hat{V}(\xi)}{R + iL\xi}.$$

The Fourier transform of $V_2(t)$ is therefore

$$\hat{V}_2(\xi) = \frac{R\hat{V}(\xi)}{R + iL\xi}.$$

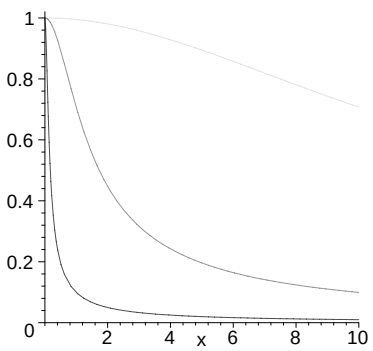
As a map from $V(t)$ to $V_2(t)$ the transfer function of this filter is

$$\hat{l}(\xi) = \frac{R}{R + iL\xi}.$$

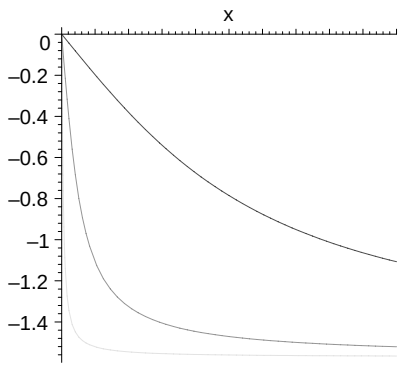
Its modulus and phase for three different values of L/R are plotted below. As $\hat{l}(0) = 1$ and

$$\lim_{\xi\to\pm\infty} \hat{l}(\xi) = 0$$

this circuit gives a crude approximation to a low pass filter.



(a) Filter amplitude, $|\hat{l}(\xi)|$



(b) Filter phase shift, $\phi(\xi)$

Figure 7.9: The amplitude and phase of the transfer function of an RL-filter

Resonant circuits

In magnetic resonance imaging a third RLC-circuit plays an important role. This is a circuit that has a resonant frequency. There is a basic difference between resistors, on the one hand and capacitors and inductors on the other. Resistors only dissipate energy, whereas capacitors and inductors can store it. The dissipative quality of resistors is clearly seen by considering the outputs of the circuits above when the voltage source is replaced by a wire. For the RC-circuit the "un-forced" solution is $I(t) = ke^{\frac{-t}{RC}}$ and for the RL-circuit, the solution is $I(t) = ke^{\frac{-Rt}{L}}$. These solutions tend to zero as $t \to \infty$. The analogous LC-circuit is shown in figure 7.10.

The current $I(t)$ is constant throughout the circuit, relations (7.15) and (7.16) lead to a differential equation for I:

$$L\frac{d^2I}{dt^2} + \frac{1}{C}I = 0. \tag{7.22}$$

The general solution of this equation is

$$I(t) = \alpha e^{it\omega} + \beta e^{-it\omega},$$

where $\omega^{-1} = \sqrt{LC}$. These solutions are periodic, in particular, they do not decay. This is an example of a *resonant* circuit with frequency ω. At normal temperatures, real circuit elements have some resistance. This is modeled by putting a resistor into the above circuit.

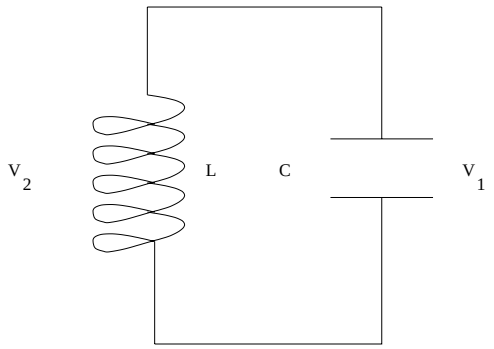


Figure 7.10: A resonant circuit.

As a final example we examine the affect of a resonant circuit on a voltage source in series with a resistor.

Example 7.1.23. Consider the circuit in figure 7.11. The input to the filter is a voltage $V(t)$, the output is the voltage measured across the inductor. Kirchoff's laws give the relations

$$\begin{aligned} V &= V_1 + V_2, \qquad & V_2 &= V_3, \\ V &= V_1 + V_3, \qquad & I_1 &= I_2 + I_3. \end{aligned} \tag{7.23}$$

Combining these equations with the defining relations, (7.14)- (7.16) gives a differential equation relating $I_3(t)$ and $V(t)$:

$$\frac{d^2 I_3}{dt^2} + \frac{1}{RC}\frac{dI_3}{dt} + \frac{1}{LC} I_3 = \frac{1}{RLC} V(t). \tag{7.24}$$

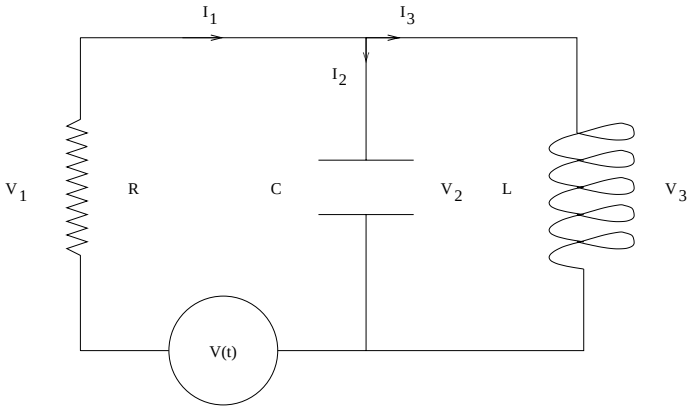


Figure 7.11: An RLC-circuit.

To find the unforced solutions the voltage source, V is set to zero, physically this amounts to replacing the voltage source with a wire. If the characteristic "frequencies" are

defined to be

$$\omega_{\pm} = \frac{1}{2}\left[\frac{1}{RC} \pm \sqrt{\frac{1}{R^2C^2} - \frac{4}{LC}}\right], \tag{7.25}$$

then the unforced solutions are linear combinations of $e^{\omega_{\pm}t}$. If

$$4R^2C < L,$$

making the expression under the square root is positive, then both of these solutions are decaying exponentials. The dissipative character of the resistor is dominating the behavior of the circuit. If $4R^2C > L$ then the square root is a purely imaginary number and the unforced solutions are damped oscillations at the frequencies,

$$\pm\sqrt{\frac{1}{LC} - \frac{1}{4R^2C^2}}, \tag{7.26}$$

the larger the resistance, the slower the oscillations are damped.

Exercise 7.1.22. Explain, physically why it is reasonable that in example 7.1.23 a larger value of R produces less damping.

Regarding the voltage source $V(t)$ as the input and the voltage $V_3(t)$, across the inductor as the output, the transfer function for this filter is

$$H(\xi) = \frac{iL\xi}{-RLC\xi^2 + iL\xi + R}.$$

The amplification factor,

$$|H(\xi)| = \frac{|\xi|}{\sqrt{R^2C^2\xi^4 + \xi^2(1 - 2R^2CL^{-1}) + R^2L^{-2}}}$$

assumes its maximum value at $\xi_{\pm} = \pm[LC]^{-\frac{1}{2}}$, the resonant frequencies of the undamped LC-circuit. At the resonant frequencies $|H(\xi_{\pm})| = 1$. That $H(0) = 0$ and $\lim H(\xi) = 0$ as $|\xi|$ tends to infinity, shows that this circuit gives an approximation to a bandpass filter with pass band centered on $\xi_{\pm}$. Figure 7.12 shows $|H(\xi)|$ with $4R^2C/L = .1, 1, 80$, the graph becomes more sharply peaked as this parameter increases.

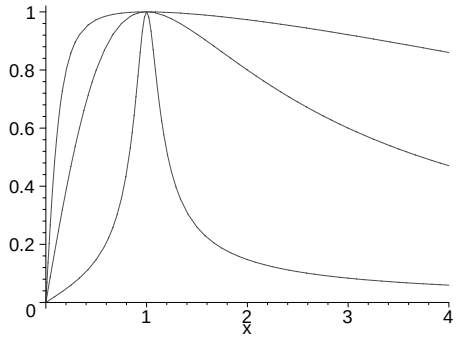


Figure 7.12: The amplitude of the transfer function.

Good basic references for electrical networks are [6] and [60].

Exercise 7.1.23. Find the transfer functions in the preceding example with $V(t)$ the input and V_1 or V_2 as the output.

Exercise 7.1.24. Analyze the RLC-circuit shown in figure 7.13. What happens to the transfer function at $[LC]^{-\frac{1}{2}}$ as R tends to zero? How should this be understood?

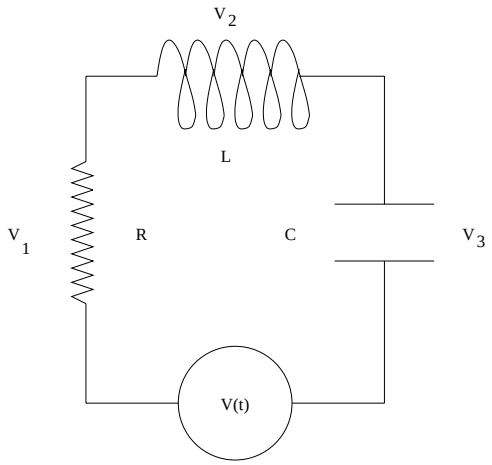


Figure 7.13: A second RLC-circuit.

Exercise 7.1.25. Sometimes it is desirable to have $4R^2C < L$ and sometimes it is desirable to have $4R^2C > L$. Give physical examples of each case and explain why the other case would be problematic.

7.2 Filtering periodic signals

In the previous section we considered the fundamentals of filtering for signals which are represented as functions of a real variable. To use the Fourier representation one needs to assume that the signals under consideration have Fourier transforms. For example, that they are bounded and supported in a bounded interval or perhaps absolutely integrable or in $L^2(\mathbb{R})$. For many applications these hypotheses are not appropriate. A case of particular interest is that of periodic signals. A signal $x(t)$, defined for $t \in \mathbb{R}$ is L-periodic if $x(t+L) = x(t)$. If, on the other hand, $x(t)$ is only defined in the interval $[0, L)$ then it can be extended to $\mathbb{R}$ as an L-periodic function by setting

$$x(t + nL) \stackrel{d}{=} x(t) \text{ for } t \in [0, L) \text{ and } n \in \mathbb{Z}.$$

In this section we briefly consider the modifications needed to analyze linear, shift invariant filters acting on periodic functions. A filter $\mathcal{A}$ is "L-periodic" if it carries L-periodic functions to L-periodic functions. For $\tau \in \mathbb{R}$, the shift operation,

$$x \mapsto x(t - \tau) = x_\tau(t)$$

is evidently an L-periodic filter. An L-periodic filter, $\mathcal{A}$ is shift invariant if $\mathcal{A}(x_\tau) = (\mathcal{A}\,x)_\tau$.

Recall that if f and g are L-periodic then their periodic convolution, defined by

$$f * g(t) = \int_0^L f(t-s)g(s)ds$$

is also L-periodic. The "L-periodic" unit impulse, δ_L is given formally as the sum

$$\delta_L = \sum_{j=-\infty}^{\infty} \delta(t + jL).$$

If $\mathcal{A}$ is an L-periodic, shift invariant, linear filter then it is determined by its impulse response, k which is given by

$$k(t) = \mathcal{A}(\delta_L).$$

The impulse response can be either an ordinary function or a generalized function. As before $\mathcal{A}$ has a representation as convolution with its impulse response,

$$\mathcal{A}\,f(t) = \int_0^L k(t-s)f(s)ds.$$

Instead of the Fourier transform, the Fourier series now provides a spectral representation for a shift invariant filter. In this case the "transfer function" $\hat{k}(n)$, is defined for $n \in \mathbb{Z}$ by applying the filter directly to the complex exponentials

$$\hat{k}(n) = \mathcal{A}(e^{-\frac{2\pi int}{L}}).$$

If $k(t)$ is an ordinary function then this agrees with the usual definition for its Fourier coefficients. If k is a *periodic* generalized function then this formula still makes sense. Formula (5.35) in Chapter 5 implies that the Fourier coefficients of $\mathcal{A}\,x = k * x$ are given by $\{\hat{k}(n)\hat{x}(n)\}$ and therefore the Fourier space representation of $\mathcal{A}$ is

$$\mathcal{A}\,x(t) = \frac{1}{L}\sum_{n=-\infty}^{\infty} \hat{k}(n)\hat{x}(n)e^{\frac{2\pi int}{L}}. \tag{7.27}$$

As in the case of the real line, a composition of linear, shift invariant, L-periodic filters is again a filter of the same type. It is also true that linear, shift invariant, L-periodic filters commute. As before, one needs to exercise caution when implementing cascades and choose an ordering which avoids numerical difficulties.

Example 7.2.1. A periodic function of period L has an expansion as a Fourier series

$$f(t) = \frac{1}{L}\sum_{j=-\infty}^{\infty} \hat{f}(j)e^{\frac{2\pi ijt}{L}}.$$

For each N the N^{th}-partial sum operator is defined by

$$S_N(f;t) = \frac{1}{L}\sum_{j=-N}^{N} \hat{f}(j)e^{\frac{2\pi ijt}{L}}.$$

The N^{th}-partial sum is a shift invariant filter given by convolution with

$$d_N(t) = \frac{\sin\left(\frac{2\pi(2N+1)t}{L}\right)}{L\sin\left(\frac{\pi t}{L}\right)}.$$

The impulse response of S_N is $d_N(t)$.

Example 7.2.2. An L-periodic analogue of the Hilbert transform is defined by

$$\begin{aligned}\mathcal{H}_L x(t) &= \lim_{\epsilon\downarrow 0}\left[\int_{-\frac{L}{2}}^{-\epsilon} + \int_{\epsilon}^{\frac{L}{2}}\right]\frac{\cos\left(\frac{\pi s}{L}\right)}{2\sin\left(\frac{\pi s}{L}\right)}f(t-s)ds \\ &= \text{P. V.}\int_{-\frac{L}{2}}^{\frac{L}{2}}\frac{\cos\left(\frac{\pi s}{L}\right)}{2\sin\left(\frac{\pi s}{L}\right)}f(t-s)ds\end{aligned} \tag{7.28}$$

Example 7.2.3. The transfer function for the Hilbert transform is given by

$$\hat{h}(n) = \begin{cases} 1 & n > 0, \\ 0 & n = 0, . \\ -1 & n < 0 \end{cases}$$

By convention $\hat{h}(0) = 0$. The Hilbert transform, in the Fourier representation, is given by

$$\mathcal{H}x(t) = \frac{1}{L}\left[\sum_{n=1}^{\infty}\hat{x}(n)e^{\frac{2\pi int}{L}} - \sum_{n=-\infty}^{-1}\hat{x}(n)e^{\frac{2\pi int}{L}}\right].$$

Example 7.2.4. A bandpass filter is defined in the Fourier representation by

$$\hat{k}(n) = \begin{cases} 1 & |n| \in [a,b], \\ 0 & k \notin [a,b]. \end{cases}$$

Example 7.2.5. As noted above, for each N the N^{th}-partial sum of the Fourier series $S_N(f)$ defines a shift invariant, linear filter. It is a bandpass filter with passband $[0,N]$. Its transfer function is therefore

$$\hat{d}_N(n) = \begin{cases} 1 & \text{for } |n| \leq N, \\ 0 & \text{for } |n| \geq N. \end{cases}$$

Exercise 7.2.1. Suppose that k_1 and k_2 are the impulses responses of a pair of L-periodic, shift invariant filters. Show that the impulse response of the composition is $k_1 * k_2$. What is the transfer function?

Exercise 7.2.2. * Give a definition for a periodic, generalized function.

7.2.1 Resolution of periodic filters*

Let $\mathcal{A}$ denote a shift invariant, linear L-periodic filter with pointspread function $k(t)$. Some of the definitions of resolution given in section 7.1.9 can be adapted to the context of periodic filters and signals. The *full width κ-maximum* definitions carry over in an obvious way, at least for pointspread functions with a well defined maximum at zero. We can also use the *first zero* definition for pointspread functions which vanish. The *equivalent width* definition can be adapted if we use the integral of $k(t)$ over a single period:

$$\Delta_{\mathcal{A},\text{ew}} = \frac{\int_0^L k(t)dt}{k(0)} = \frac{\hat{k}(0)}{k(0)}.$$

Applying the Nyquist criterion for periodic functions we can also define the ϵ-Nyquist width for an L-periodic filter. Let $< \hat{k}(n) >$ denote the Fourier coefficients of k and suppose that $0 < \epsilon < 1$. If

$$|\hat{k}(j)| \geq \epsilon |\hat{k}(0)| \text{ for } |j| \leq N$$

but either $|\hat{k}(N+1)| < \epsilon|\hat{k}(0)|$ or $|\hat{k}(-(N+1))| < \epsilon|\hat{k}(0)|$ then the ϵ-Nyquist width of $\mathcal{A}$ is defined to be

$$\Delta_{\mathcal{A},\text{ny},\epsilon} = \frac{L}{2N+1}.$$

Exercise 7.2.3. Show that the first derivative $x \mapsto \partial_t x$ defines an L-periodic shift invariant filter. What is its transfer function?

Exercise 7.2.4. Show that the shift $\mathcal{A}_\tau\, x(t) = x(t-\tau)$ defines an L-periodic shift invariant filter. What is its transfer function?

Exercise 7.2.5. The time reversal filter is defined by $\mathcal{A}\, x(t) = \mathcal{A}\, x(-t)$. It certainly carries L-periodic functions to L-periodic functions. Is it shift invariant?

Exercise 7.2.6. The "periodizing" map defined by

$$x \mapsto \sum_{j=-\infty}^{\infty} x(t + jL)$$

defines a filter which maps functions on $\mathbb{R}$ to L-periodic functions. Show that, in an appropriate sense, this is a shift invariant filter. What are its impulse response and transfer function?

Exercise 7.2.7. Show that the definitions of the Hilbert transform in examples 7.2.2 and 7.2.3 agree.

Exercise 7.2.8. Show that the Fejer means $C_N(f)$, see definition 5.5.1, are periodic, linear shift invariant filters. For each N, what are the impulse response and transfer function?

7.2.2 The comb filter and Poisson summation

See: A.4.6.

In Chapter 6 we saw that the result of sampling a signal defined on $\mathbb{R}$ is to pass from the realm of the Fourier transform to the realm of the Fourier series. Sampling is not, in any reasonable sense, a shift invariant operation. It is rather a generalized multiplication filter which can be analyzed using a formalism quite similar (actually dual) to that used to analyze shift invariant filters. To that end recall the properties of the delta function:

- $x * \delta(L) = \int x(t)\delta(L-t)dt = f(L)$,
- $\hat{\delta}(\xi) = 1$.
- $\hat{\delta}_L(\xi) = \int \delta(t-L)e^{i\xi t}dt = e^{i\xi L}$.

Multiplying the delta function by a continuous function x gives

$$x(t)\delta(t-L) = x(L)\delta(t-L).$$

Repeating this for a sum of shifted delta functions gives

$$x(t)\sum_{-\infty}^{\infty}\delta(t-nL) = \sum x(nL)\delta(t-nL).$$

This gives a model for the sequence of samples as a train of impulses located at the sample points. Sampling is defined as multiplication by

$$C_L(t) = \sum_{n=-\infty}^{\infty}\delta(t-nL).$$

This generalized function is sometimes called a *Comb filter.*

Integrating the output of the comb filter gives

$$\int_{-\infty}^{\infty} x(t)\sum_{n=-\infty}^{\infty}\delta(t-nL) = \sum_{n=-\infty}^{\infty} x(nL).$$

Parseval's formula, for L^2 functions f and g, is

$$\int f(t)\overline{g(t)}dt = \frac{1}{2\pi}\int \hat{f}(\xi)\overline{\hat{g}(\xi)}d\xi. \tag{7.29}$$

On the other hand the Poisson summation formula states that

$$\sum f(nL) = \frac{1}{L}\sum \hat{f}(\frac{2\pi n}{L}). \tag{7.30}$$

The delta function is not a function, but arguing by analogy and comparing (7.29) to (7.30) gives

$$\int_{-\infty}^{\infty} x(t)C_L(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{x}(\xi)\overline{\hat{C}_L(\xi)}d\xi.$$

Thus the Fourier transform of the comb filter is also a generalized function:

$$\mathcal{F}[\sum_{n=-\infty}^{\infty} \delta(t-nL)] = \frac{1}{L}\sum_{n=-\infty}^{\infty} \delta(\xi - \frac{2\pi n}{L}).$$

As the comb filter is defined as the product of x and C_L it has a Fourier representation as a convolution

$$\mathcal{F}(x\cdot C_L)(\xi) = \hat{x} * \hat{C}_L(\xi) = \frac{1}{L}\sum_{n=-\infty}^{\infty} \hat{x}(\xi - \frac{2\pi n}{L}).$$

This formula is another way to write the Poisson summation formula.

The operation of sampling, which is done in the time domain is frequently followed by windowing in the frequency domain and then reconstruction or interpolation, done in the time domain. As this is a composition of a multiplication filter and a shift invariant filter, it is not itself shift invariant, nonetheless there is a very simple formula for the kernel function of the composite operation. Let L denote the time domain sample spacing and $\hat{\varphi}$ the frequency domain windowing function. The sampling step takes

$$S : x \longrightarrow \sum_{n=-\infty}^{\infty} x(nl)\delta(t-nL),$$

the output of the windowing and reconstruction steps is

$$WR : S(x) \longrightarrow \mathcal{F}^{-1}\left[\hat{\varphi}(\xi)\mathcal{F}\left(\sum_{n=-\infty}^{\infty} x(nl)\delta(t-nL)\right)\right].$$

Using the formula for the inverse Fourier transform of a product gives

$$WR\circ S(x) = \varphi * \left[\sum_{n=-\infty}^{\infty} x(nl)\delta(t-nL)\right] = \sum_{n=-\infty}^{\infty} x(nl)\varphi(t-nL). \tag{7.31}$$

This is the generalized Shannon-Whitaker formula, (6.7). Formula (7.31) shows that the kernel function for $WR\circ S$ is the generalized function

$$a_{WRS}(t,s) = \sum_{n=-\infty}^{\infty} \varphi(t-nL)\delta(nL-s).$$

In a more realistic model for measurement, the samples $\{x(nL)\}$ are replaced by the samples of an average $\{\psi * x(nL)\}$. Here ψ models the sampling device. The output of the composite filter becomes

$$\sum_{n=-\infty}^{\infty} \varphi(t-nL)\psi * x(nL).$$

This is easily incorporated into the kernel function for $WR \circ S$. Letting C_ψ denote convolution with ψ, the kernel function for $WR \circ S \circ C_\psi$ is

$$a_{WRS\psi} = \sum_{n=-\infty}^{\infty} \varphi(t - nL)\psi(nL - s). \tag{7.32}$$

The kernel function for $WR \circ S \circ C_\psi$ is an ordinary function. This is a reflection of the more realistic model for sampling incorporated into this filter.

Exercise 7.2.9. Show that

$$\int_{-\infty}^{\infty} a_{WRS}(t, s)x(s)ds = (WR \circ Sf)(x).$$

Exercise 7.2.10. While the filter $WR \circ S \circ C_\psi$ is fairly realistic it still involves an infinite sum. A further refinement is to collect only a finite number of samples. Let $\chi(t)$ denote a function with bounded support.

(1). Find the kernel function for the filter

$$x \mapsto WR \circ S \circ C_\psi(\chi x).$$

(2). Find the kernel function for the filter

$$x \mapsto WR \circ S \circ (\chi C_\psi x).$$

7.3 The inverse filter

Let k denote the impulse response of a shift invariant filter $\mathcal{A}$,

$$\mathcal{A}(x) = k * x.$$

Suppose that $x(t)$ is a signal that we would like to determine and $k * x$ is the output of a measurement device. How can $x(t)$ be reconstructed given only the available measurements? What is required is a filter which *undoes* the action of $\mathcal{A}$. Such a filter is called an *inverse filter*. For some types of filters it is very clear that it is not possible to recover the original signal from the filtered output. For example, if we apply the bandpass filter $B_{[\alpha,\beta]}$ to x then all the information about the signal at frequencies outside the passband is irrevocably lost. In other cases the Fourier transform suggests a way to try to recover x from the knowledge of the filtered output $k * x$. In the Fourier representation the *inverse filter* should be given by

$$\mathcal{A}^{-1} : x \longrightarrow \mathcal{F}^{-1}\left[\frac{\widehat{k * x}}{\hat{k}}\right]. \tag{7.33}$$

From the examples we have studied it is clear that this formula often does not define a useful operation.

If k is function which goes to zero as $|t| \to \infty$ in a reasonable way, for example

$$\int |k(t)|dt < \infty$$

then $\hat{k}(\xi)$ goes to 0 as $|\xi| \to \infty$. This means that the process of dividing by $\hat{k}(\xi)$ takes the measured data and increasingly amplifies the high frequency components. If the measured data behaved like the convolution, $k * x$ then this would not be a problem: the high frequencies in the original signal will have been attenuated. In a real situation there is noise; the measurement is then modeled as $k * x + n$ where n is the noise. The noise part is **not** the result of sending a signal through the measurement device. In this case

$$\frac{\mathcal{F}(k * x + n)}{\hat{k}}(\xi) = \hat{x}(\xi) + \frac{\hat{n}(\xi)}{\hat{k}(\xi)}.$$

The high frequency content in the noise is amplified by this attempt to reverse the measurement process. One way to try to avoid this problem is to cut off the transfer function of the inverse filter outside a bounded interval. If $\hat{k}(\xi) \neq 0$ for $\xi \in [-a, a]$ then an approximate inverse filter is given by

$$\mathcal{F}^{-1}[\frac{\text{rect}_{[-a,a]}(\xi)}{\hat{k}(\xi)}\hat{x}]. \tag{7.34}$$

This gives a perfect reconstruction for data whose Fourier transform vanishes outside of $[-a, a]$ and otherwise suppresses the amplification of high frequency noise.

Though real signals are rarely bandlimited, they are usually considered to be *effectively bandlimited*. This means that all the "useful information" in the signal is contained in a finite frequency band $[-a, a]$. This is called the *effective bandwidth* in the measurement. Frequency components in the measurement from outside this band are regarded as coming from noise. By having estimates for the spectral properties of the data, the measuring apparatus and the noise one can formulate quantitative criteria for the effective bandwidth of the data, see section 6.2.2.

A related problem is that $\hat{k}(\xi)$ might vanish at finite frequencies. Recall that

$$\text{rect}_n(x) = n\chi_{[-\frac{1}{2n},\frac{1}{2n}]}(x);$$

the filter defined by rect_n averages a signal over an interval of length n^{-1}. The Fourier transform of rect_n is:

$$\widehat{\text{rect}_n}(\xi) = \int_{-1/(2n)}^{1/(2n)} ne^{-i\xi x}dx = -\frac{n}{i\xi}[e^{-\frac{i\xi}{2n}} - e^{\frac{i\xi}{2n}}] = \frac{2n\sin\left(\frac{\xi}{2n}\right)}{\xi}.$$

This function vanishes at the points $\{4\pi nm\}$ where $m \in \mathbb{Z} \setminus \{0\}$. If there were no noise, then the Fourier transform of $\text{rect}_n * x$ would also vanish at the zeros of $\widehat{\text{rect}_n}$ and dividing $\widehat{\text{rect}_n * x}$ by $\widehat{\text{rect}_n}$ would reproduce the original signal. In practice, this is not a good idea as division by $\widehat{\text{rect}_n}$ infinitely amplifies anything supported on its zero set. One approach would be to simply cut-off $[\text{rect}_n(\xi)]^{-1}$ outside of an interval $[-a, a]$ contained in $(-4\pi n, 4\pi n)$. If

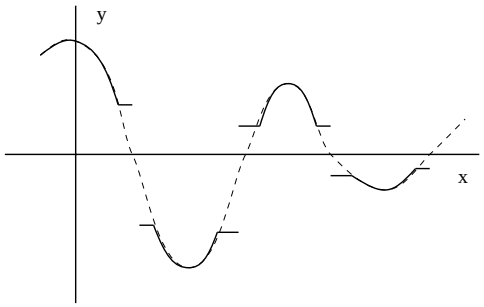


Figure 7.14: Modified Fourier Transform of the rectangle function

the effective bandwidth of the data is larger than $[-4\pi n, 4\pi n]$ then a less drastic approach would be to modify $[\mathcal{F}(\text{rect}_n)(\xi)]^{-1}$ in intervals containing the zeros of $\widehat{\text{rect}_n}$, for example one could let

$$\mathcal{F}(\text{rect}_n)_\epsilon(\xi) = \begin{cases} \mathcal{F}(\text{rect}_n)(\xi) & \text{if } |\mathcal{F}(\text{rect}_n)(\xi)| > \epsilon, \\ \epsilon & \text{if } 0 < \mathcal{F}(\text{rect}_n)(\xi) \leq \epsilon, \\ -\epsilon & \text{if } -\epsilon \leq \mathcal{F}(\text{rect}_n)(\xi) \leq 0. \end{cases}$$

An approximate inverse filter is then given by

$$\mathcal{F}^{-1}\left[\frac{\hat{x}}{\mathcal{F}(\text{rect}_n)_\epsilon}\right].$$

An even better idea is to combine the two approaches, repairing the transfer function near its zeros and cutting it off entirely outside the effective bandwidth of the data.

Designing an inverse filter is largely an engineering problem. A formula such as (7.33) provides a starting point. As the model is never exact and the measured data always contains noise, this is generally not a bounded operation and *must* therefore be approximated. The fine detail in a signal is contained in the high frequency components of its Fourier transform. A measurement process usually involves averaging which suppresses this information. This is reflected in the various definitions of the resolution available in the output of a filter considered in section 7.1.9. The implementation of an inverse filter is constrained on the one hand by a desire to retain as much of this high frequency information as possible and on the other hand by the presence of noise. A characteristic feature of noise is its irregularity which is reflected in the slow decay of its Fourier transform, see example 3.2.5. Finding an "optimal" approximation for the inverse filter begins by modeling the noise in the signal and the measurement process but ultimately requires empirical adjustment of the parameters.

Example 7.3.1. Suppose that $x(t)$ is an L-bandlimited signal. Nyquist's theorem says that in order to perfectly reconstruct the signal we must sample $x(t)$ at equally spaced points that are no further apart than $\frac{\pi}{L}$. Of course a real measurement is an average and not a point evaluation. Let $\epsilon > 0$ and define the averaging function

$$h_\epsilon(t) = \frac{1}{\epsilon}\,\text{rect}(\frac{t}{\epsilon}).$$

Observe that if x is L-bandlimited then so is $h_\epsilon * x$, to see this we compute the Fourier transform:

$$\widehat{h_\epsilon * x} = \hat{h}_\epsilon(\xi)\hat{x}(\xi). \tag{7.35}$$

If we sample the filtered function at the points $\{\frac{n\pi}{L} \mid n \in \mathbb{Z}\}$ then we can reconstruct $\widehat{h_\epsilon * x}$ using (6.2):

$$\widehat{h_\epsilon * x}(\xi) = \chi_{[-L,L]}(\xi) \sum_{n=-\infty}^{\infty} h_\epsilon * x(\frac{n\pi}{L}) e^{-\frac{n\pi i\xi}{L}} \tag{7.36}$$

On the other hand

$$\hat{h}_\epsilon(\xi) = \frac{2\sin(\frac{\epsilon\xi}{2})}{\epsilon\xi}$$

has its first zero at $\xi_0 = 2\pi/\epsilon$. If ϵ is chosen so that

$$\frac{2\pi}{\epsilon} \geq L$$

then it follows from (7.35) that $\widehat{h_\epsilon * x}(\xi)$ can be divided by $\hat{h}_\epsilon(\xi)$ leading to an exact reconstruction of $\hat{x}(\xi)$. The estimate for ϵ can be re-written

$$\frac{\epsilon}{2} < \frac{\pi}{L}.$$

For an exactly L-bandlimited function, $\epsilon = \frac{2\pi}{L}$ works but does not give a stable method for reconstructing the original signal from the measurements. This is because the function $\hat{h}_\epsilon(\xi)$ vanishes at $\xi = \pm L$. Notice also that the consecutive intervals

$$[\frac{n\pi}{L} - \frac{\epsilon}{2}, \frac{n\pi}{L} + \frac{\epsilon}{2}] \text{ and } [\frac{(n+1)\pi}{L} - \frac{\epsilon}{2}, \frac{(n+1)\pi}{L} + \frac{\epsilon}{2}]$$

overlap. A more stable algorithm results if we take $\epsilon = \frac{\pi}{L}$. In this case the consecutive intervals do not overlap and therefore the values of $x(t)$ which are averaged to determine the consecutive measurements $h_\epsilon * x(\frac{n\pi}{L})$ and $h_\epsilon * x(\frac{(n+1)\pi}{L})$ do not overlap. The smallest value that $\hat{h}_{\frac{\pi}{L}}(\xi)$ attains on $[-L, L]$ is

$$\hat{h}_{\frac{\pi}{L}}(L) = \frac{2}{\pi} \simeq 0.63661977\ldots$$

This example shows that a bandlimited function can be exactly and stably reconstructed from "realistic" measurements, provided the resolution of the measuring device is sufficiently high.

Example 7.3.2. The Hilbert transform, $\mathcal{H}$ is a filter which has a well behaved inverse. In fact $\mathcal{H}$ is its own inverse. In the Fourier representation

$$\mathcal{H}x = \frac{1}{2\pi} \int_{-\infty}^{\infty} \operatorname{sign}\xi \hat{x}(\xi) e^{ix\xi} d\xi.$$

The assertion that $\mathcal{H} = \mathcal{H}^{-1}$ follows from the equation $\operatorname{sign}\xi \cdot \operatorname{sign}\xi = 1$.

Exercise 7.3.1. Keeping the Gibbs phenomenon in mind, explain why (7.34) might be a poor choice for an approximate inverse filter. Suggest a modification likely to produce better results.

7.4 Higher dimensional filters

See: A.4.7.

In imaging applications the data is not usually a function of a time variable but rather a function of several spatial variables. Typically they are functions defined on $\mathbb{R}^2$ and $\mathbb{R}^3$. The theory of filtering functions of several variables is formally quite similar to that for functions of a single variable, though concepts like causality have no reasonable analogues. In this section boldface letters, e.g. $\mathbf{x}$ and $\boldsymbol{\xi}$ are used to denote points in $\mathbb{R}^n$ (for an appropriate n) and f denotes a function on this space.

A linear filter acting on functions of n-variables is usually represented as an integral

$$\mathcal{A}\, f(\mathbf{x}) = \int_{\mathbb{R}^n} a(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\mathbf{y}. \tag{7.37}$$

As before $a(\mathbf{x}, \mathbf{y})$ is called the *kernel function* defining the filter $\mathcal{A}$; it may be an ordinary function or a generalized function.

Example 7.4.1. The *identity filter* is the filter which carries a function to itself. It is usually denoted by Id, so that $(\mathrm{Id}\, f)(\mathbf{x}) = f(\mathbf{x})$. The kernel function for the identity acting on functions of n-variables is $\delta(\mathbf{x}-\mathbf{y})$, where δ is the n-dimensional delta function, see (3.100).

Example 7.4.2. Suppose that $f(\mathbf{x})$ is a function of n-variables and $\boldsymbol{\tau} \in \mathbb{R}^n$ is a fixed vector, the "shift by $\boldsymbol{\tau}$" is the filter defined by

$$\mathcal{A}_{\boldsymbol{\tau}}\, f(\mathbf{x}) = f(\mathbf{x} - \boldsymbol{\tau}) = f_{\boldsymbol{\tau}}(\mathbf{x}).$$

The kernel function for $\mathcal{A}_{\boldsymbol{\tau}}$ is $\delta(\mathbf{x} - \mathbf{y} - \boldsymbol{\tau})$.

Definition 7.4.1. A filter $\mathcal{A}$ acting on functions of n-variables is shift invariant if

$$\mathcal{A}\, f_{\boldsymbol{\tau}} = (\mathcal{A}\, f)_{\boldsymbol{\tau}}$$

for all inputs f and vectors $\boldsymbol{\tau} \in \mathbb{R}^n$.

A shift invariant filter is expressible as convolution.

Proposition 7.4.1. *Let $\mathcal{A}$ be a shift invariant filter acting on functions defined on $\mathbb{R}^n$. If $k(\mathbf{x}) = \mathcal{A}\,\delta(\mathbf{x})$ is the impulse response and f is an "arbitrary" input then*

$$\mathcal{A}\, f(\mathbf{x}) = \int_{\mathbb{R}^n} k(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\mathbf{y}.$$

As in the one dimensional case, the impulse response may be a generalized function, in which case this formula requires careful interpretation. The Fourier transform of the impulse response, $\hat{k}(\boldsymbol{\xi})$ is called the transfer function (in electrical engineering) or modulation

transfer function (in imaging). It provides a frequency space description for the action of a shift invariant filter:

$$\mathcal{A}\, f(\mathbf{x}) = \frac{1}{[2\pi]^n} \int_{\mathbb{R}^n} \hat{k}(\boldsymbol{\xi}) \hat{f}(\boldsymbol{\xi}) e^{i\langle \mathbf{x}, \boldsymbol{\xi} \rangle} d\boldsymbol{\xi}.$$

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be coordinates for $\mathbb{R}^n$. The simplest functions on $\mathbb{R}^n$ are functions which can be expressed as products of functions on $\mathbb{R}^1$. In filtering theory such a function is called *separable*. If

$$k(\mathbf{x}) = k_1(x_1) \cdots k_n(x_n)$$

is the impulse response of a filter $\mathcal{A}$, then $\mathcal{A}$ is said to be a *separable filter*. The transfer function of a separable filter is also a product

$$\hat{k}(\boldsymbol{\xi}) = \hat{k}_1(\xi_1) \cdots \hat{k}_n(\xi_n).$$

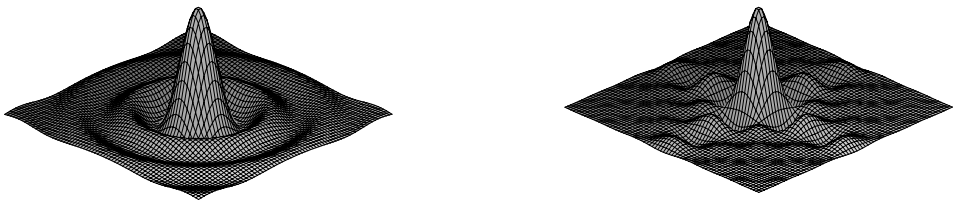
Example 7.4.3. Because the frequency, in n-dimensions is a vector, a "low pass" filter in n-dimensions can be defined in many different ways. The transfer function $s_R(\boldsymbol{\xi}) = \chi_{[0,R]}(\|\boldsymbol{\xi}\|)$ defines a filter which removes all harmonic components whose frequencies have *length* greater than R. Another possibility is to use

$$m_R(\boldsymbol{\xi}) = \prod_{j=1}^{n} \chi_{[0,R]}(|\xi_j|).$$

This filter removes harmonic components whose frequency *in any coordinate direction* exceeds R. These functions define shift invariant filters

$$S_R(f) = \mathcal{F}^{-1}(s_R(\boldsymbol{\xi}) \hat{f}(\boldsymbol{\xi})), \quad M_R(f) = \mathcal{F}^{-1}(m_R(\boldsymbol{\xi}) \hat{f}(\boldsymbol{\xi})).$$

Their impulse responses (in 2-dimensions) are shown in the figure 7.15. The filter M_R is separable whereas the filter S_R is not.



(a) Impulse response for S_R.

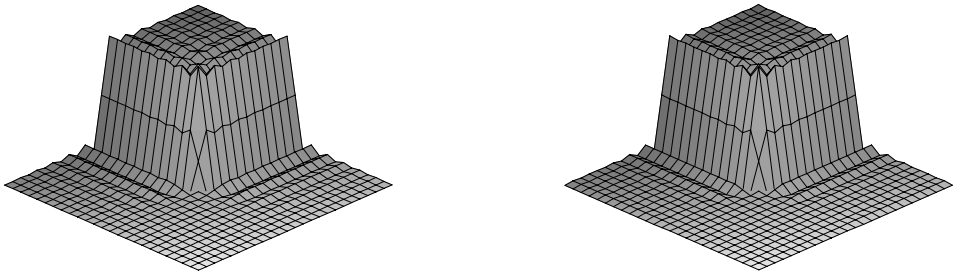
(b) Impulse response for M_R.

Figure 7.15: Impulse responses for 2-dimensional low pass filters.

Each of the filters considered above defines a partial inverse to the Fourier transform, in the sense that either $S_R f$ or $M_R f$ converges to f as R tends to infinity. Figure 7.16 shows that result of applying these filters to the characteristic function of a square,

$$\chi_{[-1,1]^2}(\mathbf{x}) = \chi_{[-1,1]}(x_1)\chi_{[-1,1]}(x_2).$$

As the data has jump discontinuities the filtered image exhibits "Gibbs artifacts." Note the ringing artifact parallel to the edges of the square and also its absence in the "Gibbs shadow" formed by the vertex. This is an indication of the fact that the detailed analysis of the Gibbs phenomenon is more complicated in higher dimensions than it is in one dimension. Note also that the Gibbs artifact is more pronounced in (a).



(a) S_R applied to $\chi_{[-1,1]^2}$. (b) M_R applied to $\chi_{[-1,1]^2}$.

Figure 7.16: Low pass filters in two dimensions.

Example 7.4.4. A filter can also act selectively in different directions. For example a low pass filter in the x_1-direction is defined by the transfer function $\chi_{[0,R]}(|\xi_1|)$. Given a unit vector $\omega \in S^{n-1}$ the transfer function $\chi_{[0,R]}(|\langle \omega, \boldsymbol{\xi}\rangle|)$ removes all harmonic components whose frequency in the ω-direction exceeds R.

A linear transformation $U : \mathbb{R}^n \to \mathbb{R}^n$ is a rigid rotation if

$$\|U\mathbf{x}\|_2 = \|\mathbf{x}\|_2 \tag{7.38}$$

for all vectors $\mathbf{x} \in \mathbb{R}^n$. Fixing an orthogonal coordinate system for $\mathbb{R}^n$ the rigid rotations are defined by matrices U which satisfy the matrix equation

$$U^t U = \mathrm{Id} = U U^t.$$

This collection of matrices is denoted $O(n)$. The rigid rotations of $\mathbb{R}^2$ are given by the matrices

$$O(2) = \left\{ \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}, \quad \begin{pmatrix} -\sin\theta & \cos\theta \\ \cos\theta & \sin\theta \end{pmatrix} \text{ for } \theta \in [0, 2\pi) \right\}.$$

A linear transformation, U defines an action on functions by setting

$$f_U(\mathbf{x}) = f(U\mathbf{x}).$$

Definition 7.4.2. A filter $\mathcal{A}$ acting on functions on $\mathbb{R}^n$ is *isotropic* if it commutes with all rotations, that is

$$\mathcal{A}\, f_U(\mathbf{x}) = (\mathcal{A}\, f)_U(\mathbf{x}) = \mathcal{A}\, f(U\mathbf{x}).$$

An isotropic filter can be linear or non-linear. For example, the filter which takes a function f to its absolute value, $|f|$ is a non-linear isotropic filter. Linear, shift invariant isotropic filters have a very simple characterization.

Proposition 7.4.2. *A linear, shift invariant filter $\mathcal{A}$ is isotropic if and only if its impulse response (or transfer function) is a radial function.*

Proof. Let $k = \mathcal{A}\,\delta$ denote the impulse response of $\mathcal{A}$. From the results in section 3.3.4 it follows that k is a radial function if and only if $\hat{k}$ is a radial function. If k is a radial function then there is a function κ such that

$$k(\mathbf{x}) = \kappa(\|\mathbf{x}\|).$$

Let U be a rigid rotation, then

$$\begin{aligned} \mathcal{A}\, f_U(\mathbf{x}) &= \int_{\mathbb{R}^n} \kappa(\|\mathbf{x}-\mathbf{y}\|) f(U\mathbf{y}) d\mathbf{y} \\ &= \int_{\mathbb{R}^n} \kappa(\|U(\mathbf{x}-\mathbf{y})\|) f(U\mathbf{y}) d\mathbf{y} \\ &= \mathcal{A}\, f(U(\mathbf{x})). \end{aligned} \tag{7.39}$$

The substitution $\mathbf{y}' = U\mathbf{y}$ is used to go from the second line to the last line. This shows that a radial impulse response defines an isotropic filter.

The converse statement is even easier because $\delta_U(\mathbf{x}) = \delta(\mathbf{x})$, this follows by formally changing variables in the integral defining δ_U :

$$\int_{\mathbb{R}^n} \delta(U\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \int_{\mathbb{R}^n} \delta(\mathbf{x}) f(U^{-1}\mathbf{x}) d\mathbf{x} = f(0).$$

The definition of an isotropic filter implies that

$$k_U(\mathbf{x}) = \mathcal{A}\,\delta_U(\mathbf{x}) = \mathcal{A}\,\delta(\mathbf{x}) = k(\mathbf{x}) \text{ for all } U \in O(n).$$

This shows that $k(\mathbf{x})$ only depends on $\|\mathbf{x}\|$. □

Example 7.4.5. If $\psi(\mathbf{x})$ is a smooth, non-negative function with support contained in the ball of radius ϵ and total integral 1 then the formula

$$\mathcal{A}_\psi\, f(\mathbf{x}) = \int_{\mathbb{R}^n} \psi(\mathbf{x}-\mathbf{y}) f(\mathbf{y}) d\mathbf{y}$$

defines a smoothing filter. Its transfer function is $\hat{\psi}(\boldsymbol{\xi})$. As

$$\int_{\mathbb{R}^n} \psi(\mathbf{x}) d\mathbf{x} = 1$$

it follows that $\hat{\psi}(0) = 1$. As ψ is smooth and vanishes outside a bounded set, its Fourier transform tends to zero as $\|\boldsymbol{\xi}\| \to \infty$. Thus $\mathcal{A}_\psi$ is an approximate low pass filter. As in the one-dimension, filters like $\mathcal{A}_\psi$ provide models for measuring devices. If ψ is a radial function then the filter $\mathcal{A}_\psi$ is isotropic.

Complex, higher dimensional filtering operation are often assembled out of simpler pieces. If $\mathcal{A}_1$ and $\mathcal{A}_2$ are linear filters then their composite, $\mathcal{A}_1 \circ \mathcal{A}_2$ is as well. While in general $\mathcal{A}_1 \circ A_2 \neq \mathcal{A}_2 \circ \mathcal{A}_1$, shift invariant filters do commute. If k_i, $i = 1, 2$ are the impulse responses of shift invariant filters $\mathcal{A}_i$, $i = 1, 2$ then the impulse response of the cascade $\mathcal{A}_1 \circ \mathcal{A}_2$ is

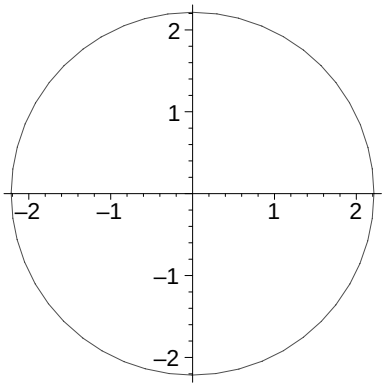
$$k_1 * k_2 = k_2 * k_1.$$

The transfer function is the product $\hat{k}_1(\boldsymbol{\xi})\hat{k}_2(\boldsymbol{\xi})$.

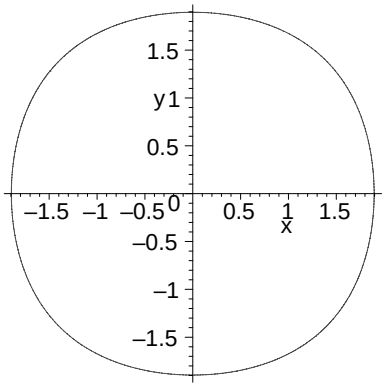
The discussion of resolution for one dimensional filters in section 7.1.9 can be repeated almost verbatim for higher dimensional, isotropic filters. If the filter is not isotropic then the situation is more complicated. For example, let $\mathcal{A}$ be a filter acting on functions of n variables with impulse response, $a(\mathbf{x})$. Suppose that a achieves its maximum at 0 and decays to zero as $\|\mathbf{x}\|$ tends to infinity. The set of points

$$\text{FWHM}(a) = \{\mathbf{x} \,:\, a(\mathbf{x}) = \frac{1}{2}|a(0)|\}$$

where a assumes half its maximum value is a hypersurface, the *half maximum hypersurface*. If $n = 2$ then this is curve, figure 7.17 shows level curves for the impulse responses of the filters S_R and M_R. If a is not a radial function then there are many possible ways to assign a single number which captures the resolution in the output of $\mathcal{A}$. A conservative assessment of the resolution in a non-isotropic filter is to use the largest distance between two points on the half maximum hypersurface. This is called the *diameter* of this hypersurface.



(a) Half maximum curve for S_R.



(b) Half maximum curve for M_R.

Figure 7.17: Half maximum curves for 2d low pass filters.

Definition 7.4.3. Let $\mathcal{A}$ be a linear, shift invariant filter with impulse response $a(\mathbf{x})$. Suppose that $|a|$ assumes it maximum at 0 and tends to zero as $\|\mathbf{x}\|$ tends to infinity. The hypersurface FWHM(a) may have several components, let FWHM$(a)_0$ be the component bounding a region which contains 0. The *full width half maximum* of $\mathcal{A}$ is defined to be the diameter of FWHM$(a)_0$.

Generalizing the other definitions is left as an exercise for the curious reader.

Exercise 7.4.1. Prove Proposition 7.4.1 assuming the f is a smooth function with bounded support.

Exercise 7.4.2. Prove that if $\mathcal{A}_1$ and $\mathcal{A}_2$ are shift invariant filters then $\mathcal{A}_1 \circ \mathcal{A}_2 = \mathcal{A}_2 \circ \mathcal{A}_1$.

Exercise 7.4.3. Suppose that $\mathcal{A}_1$ and $\mathcal{A}_2$ are isotropic filters, show that $\mathcal{A}_1 \circ \mathcal{A}_2$ is as well.

Exercise 7.4.4. Show that squared length of the gradient

$$\mathcal{A}\, f = \sum_{j=1}^{n} \left(\frac{\partial f}{\partial x_j} \right)^2$$

is an isotropic filter.

Exercise 7.4.5. Show that the following filters are *not* isotropic

$$\begin{aligned} \mathcal{A}_\infty\, f &= \max\{ \left| \frac{\partial f}{\partial x_j} \right| \,:\, j = 1, \ldots, n\}, \\ \mathcal{A}_1\, f &= \sum_{j=1}^{n} \left| \frac{\partial f}{\partial x_j} \right|. \end{aligned} \tag{7.40}$$

7.5 Implementing shift invariant filters

See: A.6.2, A.7.1.

In practical applications one cannot measure a time signal $x(t)$ continuously, rather the signal is sampled at a discrete sequence of times. Frequently the sample times are equally spaced and of the form $\{t_0 + j\tau \,|\, j \in \mathbb{Z}\}$, τ is called the sample spacing. In certain situations the physical measuring apparatus makes it difficult to collect equally spaced samples. In such situations the data is often interpolated to create an equally spaced set of samples for further analysis. This is done, even though it introduces a new source of error, because algorithms for equally spaced samples are so much simpler than those for unequal sample spacing. In this section we use the tools developed above to understand the transition from continuous time signals to sampled data and the implementation of shift invariant linear filters using the finite Fourier transform. The one dimensional case is treated in detail, higher dimensional filters are briefly considered.

In terms of the continuous parameters t and ξ there are two different representations for a shift invariant filter H with impulse response $h(t)$ and transfer function $\hat{h}(\xi)$: the time domain representation as a convolution

$$Hx(t) = h * x(t) = \int_{-\infty}^{\infty} h(t-s)x(s)ds, \tag{7.41}$$

and the frequency domain representation as a Fourier integral

$$Hx(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{h}(\xi)\hat{x}(\xi)e^{it\xi} d\xi. \tag{7.42}$$

Each representation leads to a different discretization scheme. For several reasons, the frequency domain representation is usually employed: certain filters, like the Hilbert transform or differentiation are difficult to represent as convolutions, since the impulse response is a generalized function. For most shift invariant filters, using the "fast Fourier transform" algorithm makes the frequency domain computation vastly more efficient than the time domain computation.

7.5.1 Sampled data

Sampling a function entails evaluating it at points. From the point of view of measurements, a function only has a well defined value at points of continuity and therefore it is assumed throughout this analysis that both the signal x and the impulse response of the filter, h are continuous functions. With τ the sample spacing in the time domain let

$$h_s(j) = h(j\tau) \text{ for } j \in \mathbb{Z},$$

denote samples of the impulse response and

$$x_s(j) = x(j\tau) \text{ for } j \in \mathbb{Z},$$

samples of the input signal. The time domain representation of the filter is approximated by using a Riemann sum

$$\begin{aligned} Hx(j\tau) &= h * x(j\tau) \\ &\approx \sum_{k=-\infty}^{\infty} h_s(j-k)x_s(k)\tau = \tau h_s \star x_s(j). \end{aligned} \tag{7.43}$$

In the second line of (7.43), $h_s \star x_s$ denotes the discrete convolution operation, defined in section 5.4. A Riemann sum is just one possible approximation for the convolution, if the integrand is smooth, a higher order, numerical integration method might give superior results, see section A.7.1.

For the first step of the analysis it is assumed that $x(t)$ is absolutely integrable and that an infinite sequence of samples is collected. In imaging applications the data is usually supported in a bounded interval, so this is not an unreasonable assumption. The sequence of samples has a *sample Fourier transform*

$$\hat{x}_s(\xi) = \sum_{j=-\infty}^{\infty} x_s(j)e^{-ij\tau\xi}\tau, \tag{7.44}$$

which is a $\frac{2\pi}{\tau}$-periodic function.

The sample Fourier transform is connected to discrete convolution in a simple way:

$$\begin{aligned}\widehat{\tau h_s \star x_s}(\xi) &= \sum_{j=-\infty}^{\infty} h_s \star x_s(j\tau) e^{-ij\tau\xi}\tau \\ &= \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} h((j-k)\tau) e^{-i(j-k)\tau\xi} x(k\tau) e^{-ik\tau\xi}\tau^2 \\ &= \hat{h}_s(\xi)\hat{x}_s(\xi).\end{aligned} \tag{7.45}$$

In other words, the sample Fourier transform of $\tau h_s \star x_s$ is simply $\hat{h}_s(\xi)\hat{x}_s(\xi)$.

The dual Poisson summation formula, (6.12) relates $\hat{x}_s$ to $\hat{x}$:

$$\hat{x}_s(\xi) = \sum_{j=-\infty}^{\infty} \hat{x}(\xi + \frac{2\pi j}{\tau}). \tag{7.46}$$

Formula (7.44) is a Riemann sum for the integral defining $\hat{x}(\xi)$. From (7.46) it is apparent that careful consideration is needed to understand in what sense $\hat{x}_s(\xi)$ is an approximation to $\hat{x}(\xi)$.

When the input is $\frac{\pi}{\tau}$-bandlimited Nyquist's theorem implies that

$$\hat{x}(\xi) = \hat{x}_s(\xi)\chi_{[-\frac{\pi}{\tau},\frac{\pi}{\tau}]}(\xi).$$

If the transfer function for the filter H is also supported in this interval then

$$Hx(t) = \frac{1}{2\pi}\int_{-\frac{\pi}{\tau}}^{\frac{\pi}{\tau}} \hat{x}_s(\xi)\hat{h}_s(\xi)e^{it\xi}d\xi.$$

This indicates that, for this application, $\hat{x}_s(\xi)\chi_{[-\frac{\pi}{\tau},\frac{\pi}{\tau}]}(\xi)$ should be regarded as an approximation for $\hat{x}(\xi)$. The accuracy of this approximation depends on the high frequency behavior of $\hat{x}(\xi)$ and the sample spacing.

From the Fourier inversion formula it follows that

$$h_s \star x_s(l)\tau = \frac{1}{2\pi}\int_{-\frac{\pi}{\tau}}^{\frac{\pi}{\tau}} \hat{h}_s(\xi)\hat{x}_s(\xi)e^{il\tau\xi}d\xi. \tag{7.47}$$

Thus if x **and** h are $\frac{\pi}{\tau}$-bandlimited then

$$h * x(l\tau) = h_s \star x_s(l)\tau \text{ for all } j \in \mathbb{Z},$$

the discrete convolution of the sampled sequences consists of samples of the continuous convolution. In any case, the integral on the right hand side of (7.47) gives an approximation to $h * x(j\tau)$:

$$h * x(l\tau) \approx \frac{1}{2\pi}\int_{-\frac{\pi}{\tau}}^{\frac{\pi}{\tau}} \hat{h}_s(\xi)\hat{x}_s(\xi)e^{il\tau\xi}d\xi \tag{7.48}$$

An important variant on (7.48) is to use the exact transfer function $\hat{h}(\xi)$ for the filter (which is often known) instead of the sample Fourier transform, $\hat{h}_s(\xi)$; this gives

$$h * x(l\tau) \approx \frac{1}{2\pi} \int_{-\frac{\pi}{\tau}}^{\frac{\pi}{\tau}} \hat{h}(\xi)\hat{x}_s(\xi)e^{il\tau\xi}d\xi. \tag{7.49}$$

Usually the approximations to $h * x$ provided by (7.48) and (7.49) are different. Equations (7.43), (7.48) and (7.49) form the foundation for the implementation of shift invariant filters on sampled data.

In imaging, the signal $x(t)$ is usually zero outside a finite interval. In section 3.2.14 it is shown that this prevents the signal from also being bandlimited. Heuristically, the sample spacing is chosen to attain a certain degree of resolution in the final result. From (7.46) we see that in order for $\hat{x}_s(\xi)$ to be a good approximation to $\hat{x}(\xi)$ over $[-\frac{\pi}{\tau}, \frac{\pi}{\tau}]$ it is necessary for $\hat{x}(\xi)$ to be small outside of this interval. In other words, $x(t)$ must be effectively bandlimited to $[-\frac{\pi}{\tau}, \frac{\pi}{\tau}]$. In the present application this means that in order for $\hat{x}_s$ to give a good approximation to $\hat{x}$ for $\xi \in [-\frac{\pi}{\tau}, \frac{\pi}{\tau}]$ the sum

$$\sum_{j\neq 0} \hat{x}(\xi - \frac{2\pi j}{\tau}) \text{ for } \xi \in [-\frac{\pi}{\tau}, \frac{\pi}{\tau}]$$

must be uniformly small. To insure that this is so, a signal is often passed through a low pass filter *before it is sampled.* Once the signal is sampled, the sample spacing *fixes* the effective bandwidth of the sampled data to be $\frac{2\pi}{\tau}$.

7.5.2 The finite Fourier transform

See: A.2.5, A.2.10.

A real measurement consists of a finite set of numbers. In this section we define the finite Fourier transform.

Definition 7.5.1. Let $(x_0, \ldots, x_{N-1})$ be a sequence of (real or complex) numbers. The sequence $(\hat{x}_0, \ldots, \hat{x}_{N-1})$ defined by

$$\hat{x}_k = \frac{1}{N} \sum_{j=0}^{N-1} x_j e^{-\frac{2\pi ijk}{N}} \tag{7.50}$$

is called the *finite Fourier transform* of the original sequence. Sometimes it is denoted by

$$\mathcal{F}_N(x_0, \ldots, x_{N-1}) = (\hat{x}_0, \ldots, \hat{x}_{N-1}).$$

The inverse of the finite Fourier transformation is given by

$$x_j = \sum_{k=0}^{N-1} \hat{x}_k e^{\frac{2\pi ijk}{N}}. \tag{7.51}$$

Notice that the summation in this formula is quite similar to (7.50), the exponential multipliers have been replaced by their complex conjugates. This means that a fast algorithm for computing $\mathcal{F}_N$ automatically provides a fast algorithm for computing $\mathcal{F}_N^{-1}$.

Definition 7.5.2. Let $(x_0, \ldots, x_{N-1})$ be a sequence of length N. Its N-periodic extension is defined by

$$x_{j+lN} = x_j \text{ for } 0 \leq j \leq N-1, l \in \mathbb{Z}.$$

Periodic sequences can be convolved.

Definition 7.5.3. Given two N-periodic sequences, $(x_j)_0^{N-1}, (y_j)_0^{N-1}$ their *periodic convolution* is defined by

$$(x \star y)_k = \sum_{j=0}^{N-1} x_j y_{k-j}.$$

For example, the 0th component of $x \star y$ is

$$(x \star y)_0 = \sum_{i=0}^{N-1} x_i y_{-i}.$$

Observe that

$$\begin{aligned} \widehat{x \star y}_k &= \frac{1}{N} \sum_{j=0}^{N-1} (x \star y)_j e^{-\frac{2\pi i j k}{N}} \\ &= \frac{1}{N} \sum_{j=0}^{N-1} (\sum_{l=0}^{N-1} x_l y_{j-l}) e^{-\frac{2\pi i j k}{N}} \\ &= N \hat{x}_k \hat{y}_k \end{aligned} \tag{7.52}$$

The periodic convolution can therefore be computed using the finite Fourier transform and its inverse.

$$x \star y = N \mathcal{F}_N^{-1}(\hat{x}_0 \hat{y}_0, \ldots, \hat{x}_{N-1} \hat{y}_{N-1}). \tag{7.53}$$

If $N = 2^k$ then this is the most efficient way to do such calculations.

Exercise 7.5.1. Explain why line 2 equals line 3 in (7.52).

Exercise 7.5.2. To directly compute the sums defining $\mathcal{F}_N$ requires $O(N^2)$ arithmetic operations. An algorithm for computing $\mathcal{F}_N$ is "fast" if it uses $O(N^\alpha)$ operations for an $\alpha < 2$. Explain why a fast algorithm for computing $\mathcal{F}_N$ also gives a fast algorithm for computing $\mathcal{F}_N^{-1}$.

7.5.3 Approximation of Fourier coefficients

Let f be a function on $[0,1]$; define $x_j = f(j/N), j = 0, \ldots, N-1$. The sequence $(\hat{x}_j)$ denotes its finite Fourier transform. It is reasonable to expect that $\hat{x}_k$ is an approximation for the k^{th} Fourier coefficient of f, as it is defined by a Riemann sum for the integral:

$$\hat{f}(k) = \int_0^1 f(x)e^{-2\pi ikx}dx \approx \frac{1}{N}\sum_{j=0}^{N-1} x_j e^{-\frac{2\pi ijk}{N}} = \hat{x}_k.$$

The error in replacing the integral with the sum gets worse as k becomes larger. This can be explained by a crude estimate of the error in replacing the integral for $\hat{f}(\xi)$ by this Riemann sum:

$$\text{The error} \approx \max(\text{derivative of the integrand}) \times (\text{mesh size}) \sim k\tfrac{1}{N}.$$

The estimate for the maximum of the derivative, k comes entirely from the exponential factor and ignores the contribution of the function f. This shows that, even for a very smooth function, if $k > N/2$ then $\hat{x}_k$ is a poor approximation to $\hat{f}(k)$. For k in this range a different interpretation for $\hat{x}_k$ is therefore needed. Formula (7.50) defines $\hat{x}_k$ for all values of k. Observe that $\hat{x}_{k+lN} = \hat{x}_k$, in particular

$$\hat{x}_{-k} = \hat{x}_{N-k} \text{ for } k = \frac{N}{2}, \frac{N}{2}+1, \ldots, N-1.$$

As

$$e^{-\frac{2\pi ijk}{N}} = e^{-\frac{2\pi ij(k-N)}{N}},$$

$\hat{x}_k$ could equally well be interpreted as a Riemann sum for the integral

$$\int_0^1 f(x)e^{-2\pi i(k-N)x}dx = \hat{f}(k-N).$$

For $\frac{N}{2} < k \leq N-1$ this turns out to be a more useful interpretation.

In light of this interpretation the approximate partial sums of the Fourier series of $f(x)$ should be

$$f(x) \approx \sum_{j=0}^{N/2} \hat{x}_j e^{2\pi ijx} + \sum_{j=N/2+1}^{N-1} \hat{x}_j e^{2\pi i(j-N)x}. \tag{7.54}$$

If f is real valued one could also try to use

$$\text{Re}\Big[\sum_{j=0}^{N-1} \hat{x}_j e^{\frac{2\pi ijk}{N}}\Big] \tag{7.55}$$

as an approximate partial sum for the Fourier series of f. This function agrees with f at the sample points. If f is sufficiently smooth and N is large enough, formula (7.54) can be used to accurately interpolate values of $f(x)$ between the sample points, whereas formula (7.55)

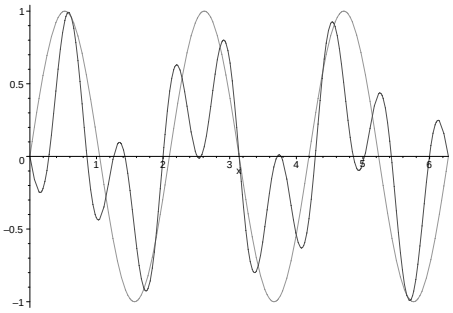


Figure 7.18: Bad interpolation using formula 7.55.

usually gives poor results. Figure 7.18 shows the result of using formula (7.55) to interpolate $\sin(3x)$ using 11 sample points. Using the formula (7.54) with this many sample points gives an exact reconstruction.

To summarize: the output of the finite Fourier transform, when indexed by frequency, should be interpreted to be

$$(\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_{\frac{N-1}{2}}, \hat{x}_{-\frac{N}{2}}, \hat{x}_{1-\frac{N}{2}}, \ldots, \hat{x}_{-1}).$$

If f is a function defined on $[0, L]$ and we collect N-equally spaced samples,

$$x_j = f\left(\frac{jL}{N}\right) \text{ for } j = 0, \ldots, N-1,$$

then

$$\begin{aligned} \hat{x}_k &= \frac{1}{N}\sum_{j=0}^{N-1} f\left(\frac{jL}{N}\right) e^{-\frac{2\pi i j k}{N}} \\ &= \frac{1}{L}\sum_{j=0}^{N-1} f\left(\frac{jL}{N}\right) e^{-\frac{2\pi i j L k}{NL}} \frac{L}{N} \\ &\approx \begin{cases} \frac{1}{L}\hat{f}(k) & \text{if } 0 \le k \le \frac{N}{2}, \\ \frac{1}{L}\hat{f}(k-N) & \text{if } \frac{N}{2} < k \le N-1. \end{cases} \end{aligned} \tag{7.56}$$

Exercise 7.5.3. Explain why (7.56) agrees with Nyquist's theorem for periodic functions.

Exercise 7.5.4. For f a once differentiable, periodic function compare the approximations

to $\hat{f}(k)$ obtained by sampling the integrands to approximate the following integrals

$$\begin{aligned}\hat{f}(k) &= \int_0^1 f(x)e^{-2\pi ikx}dx,\\ \hat{f}(k) &= \frac{1}{2\pi ik}\int_0^1 f'(x)e^{-2\pi ikx}dx.\end{aligned} \tag{7.57}$$

7.5.4 Implementing periodic convolutions on sampled data

The action of a shift invariant filter H on a 1-periodic function is defined in terms of its impulse response $h(t)$ (another 1-periodic function) by the periodic convolution

$$Hf(t) = \int_0^1 f(s)h(t-s)ds.$$

Let

$$\hat{h}_k = \int_0^1 h(t)e^{-2\pi ikt}dt$$

denote the Fourier coefficients of the impulse response. The Fourier representation of this filter is

$$\begin{aligned}H(f)(t) =& \mathcal{F}^{-1}(\hat{f}(k)\hat{h}_k)\\ =& \sum_{k=-\infty}^{\infty}\hat{f}(k)\hat{h}_k e^{2\pi ikt}.\end{aligned} \tag{7.58}$$

At sample points, a finite frequency approximation to such a filter is given by

$$(Hx)_j = \sum_{j=0}^{N/2}\hat{x}_k\hat{h}_k e^{\frac{2\pi ijk}{N}} + \sum_{j=N/2+1}^{N-1}\hat{x}_k\hat{h}_{k-N}e^{\frac{2\pi ijk}{N}}.$$

Example 7.5.1. The Hilbert transform is defined in the Fourier representation by the transfer function

$$\hat{h}_k = \begin{cases}1 & k>0,\\ 0 & k=0,.\\ -1 & k<0\end{cases}$$

A finite frequency approximation to the Hilbert transform is given by

$$(\mathcal{H}x)_j \approx \sum_{k=1}^{N/2}\hat{x}_k e^{\frac{2\pi ijk}{N}} - \sum_{k=N/2+1}^{N-1}\hat{x}_k e^{\frac{2\pi ijk}{N}}.$$

7.5.5 Implementing filters on finitely sampled data

We use the pieces developed so far to explain how the finite Fourier transform is used to approximate a shift invariant filter using finitely sampled data. The inversion formula, (7.51) defines the sequence, (x_k) for all values of k as an N-periodic sequence, $x_k = x_{k+lN}$. Once the finite Fourier transform is used to approximate a filter the data *must* be regarded as samples of a *periodic* function rather than a function with bounded support on the real line. This section treats the details of using the finite Fourier transform to approximately implement a shift invariant filter on finitely sampled data.

Let f be a function defined on $\mathbb{R}$ with support in $[0,1]$. Let h be the impulse response of a filter H so that

$$Hf(x) = h * f(x).$$

In general h is not assumed to have support in $[0,1]$. Suppose that f is sampled at the points $\{\frac{j}{N} : j = 0, \ldots, N-1\}$, it is implicit that the remaining "uncollected" samples are zero. A Riemann sum gives an approximate value for $f * h$ at the sample points:

$$f * h(\frac{k}{N}) \approx \sum_{j=0}^{N-1} h(\frac{k-j}{N}) f(\frac{j}{N}) \frac{1}{N}.$$

To do this calculation for $k \in \{0, \ldots, N-1\}$ requires a knowledge of $2N-1$ values of h,

$$h_s = (h(\frac{1-N}{N}), h(\frac{2-N}{N}), \ldots, h(\frac{0}{N}), \ldots, h(\frac{N-1}{N})).$$

In order to use the finite Fourier transform to compute this discrete convolution it must be interpreted as a *periodic convolution* of two sequences of length $2N-1$. This means that the vector of samples of f must be augmented to get a sequence of the same length. This is done by adding $(N-1)$-zeros to the end in a process called *zero padding*. It is consistent with the interpretation of $f(x)$ as a function defined on $\mathbb{R}$ with support contained in $[0,1]$. Let

$$f_s = (f(0), f(\frac{1}{N}), \cdots, f(\frac{N-1}{N}), \underbrace{0, 0, \cdots, 0}_{N-1}).$$

The zero padded sequence is, in effect, samples of a function defined on $[0, \frac{2N-1}{N}]$. Formula (7.56) implies that

$$\hat{f}_s(k) \approx \begin{cases} \frac{N}{2N-1} \hat{f}(k) & \text{for } 0 \le k \le N-1 \\ \frac{N}{2N-1} \hat{f}(k-2N+1) & \text{for } N \le k \le 2N-2. \end{cases} \tag{7.59}$$

Regarding h_s and f_s as $(2N-1)$-periodic sequences gives

$$\sum_{j=0}^{N-1} h(\frac{k-j}{N}) f(\frac{j}{N}) \frac{1}{N} = \sum_{j=0}^{2N-2} h_s(k-j) f_s(j) \frac{1}{N}. \tag{7.60}$$

Using (7.53) this can be rewritten in terms of the finite Fourier transform as

$$\sum_{j=0}^{N-1} h(\frac{k-j}{N}) f(\frac{j}{N}) \frac{1}{N} = \frac{2N-1}{N} \mathcal{F}_{2N-1}^{-1}(\hat{h}_s \hat{f}_s)(k). \tag{7.61}$$

Thinking of f and h as being defined on $\mathbb{R}$ with f effectively bandlimited to $[-N\pi, N\pi]$ leads to

$$\begin{aligned} h * f(\frac{k}{N}) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{h}(\xi)\hat{f}(\xi)e^{i\frac{k}{N}\xi}d\xi \\ &\approx \frac{1}{2\pi} \sum_{j=1-N}^{N-1} \hat{f}\left(\frac{2\pi jN}{2N-1}\cdot\frac{k}{N}\right)\hat{h}\left(\frac{2\pi jN}{2N-1}\cdot\frac{k}{N}\right)\exp\left(\frac{2\pi ijN}{2N-1}\cdot\frac{k}{N}\right)\left(\frac{2\pi N}{2N-1}\right) \\ &\approx \frac{2N-1}{N}\mathcal{F}^{-1}_{2N-1}(\hat{h}_s(0)\hat{f}_s(0),\ldots,\hat{h}_s(2N-2)\hat{f}_s(2N-2))(k). \end{aligned} \tag{7.62}$$

The last line is a consequence of (7.59); it provides a different derivation for the main result of this section that

$$h * f(\frac{k}{N}) \approx \frac{2N-1}{N}\mathcal{F}^{-1}_{2N-1}(\hat{h}_s(0)\hat{f}_s(0),\ldots,\hat{h}_s(2N-2)\hat{f}_s(2N-2))(k) \text{ for } k = 0,\ldots,N-1.$$

For latter applications it is useful to have formulæ for the approximate Fourier implementation of a linear, shift invariant filter H without scaling the data to be defined on $[0,1]$. Let $h(t)$ be the impulse response of H and sample $f(t)$ at a sequence of equally spaced points

$$x_j = f(t_0 + j\tau) \text{ for } j = 0,\ldots,N-1.$$

Let $(\hat{x}_0,\ldots,\hat{x}_{2N-2})$ be the finite Fourier transform of the zero padded sequence of length $2N-1$, $(x_0,\ldots,x_{N-1},0,\ldots,0)$ and $(\hat{h}_j)$ denote the finite Fourier transform of

$$h_s = (h((1-N)\tau),\ldots,h((N-1)\tau)).$$

The approximate values at the sample times, $\{t_0+j\tau\}$ given by the Riemann sum (computed using the finite Fourier transform) are

$$\begin{aligned} h * x(t_0 + j\tau) &\approx \sum_{k=0}^{N-1} x_k h(\tau(j-k))\tau \\ &= \tau(2N-1)\mathcal{F}^{-1}_{2N-1}(\hat{x}_0\hat{h}_0,\ldots,\hat{x}_{2N-2}\hat{h}_{2N-2})(j). \end{aligned} \tag{7.63}$$

In many applications the sequence $(\hat{h}_0,\ldots,\hat{h}_{2N-2})$ appearing in (7.63) is **not** computed as the finite Fourier transform of the samples h_s. Instead, it is obtained by directly sampling $\hat{h}(\xi)$. This is because the transfer function $\hat{h}(\xi)$ may be an ordinary function which can be computed exactly, even when the impulse response is a generalized function. From (7.63) it is clear that h should be regarded as a function defined on the finite interval $[(1-N)\tau, (N-$

$1)\tau]$. The discussion is section 7.5.3 shows that finite Fourier transform of h_s is given by

$$\begin{aligned}\hat{h}_s(k) &= \frac{1}{2N-1}\sum_{j=0}^{2N-2} h((1-N)\tau + j\tau)e^{-\frac{2\pi ijk}{2N-1}} \\ &= \frac{1}{\tau(2N-1)}\sum_{j=0}^{2N-2} h((1-N)\tau + j\tau)e^{-ij\tau\frac{2\pi k}{\tau(2N-1)}}\tau \qquad (7.64)\\ &\approx \begin{cases} \frac{1}{\tau(2N-1)}\hat{h}\left(\frac{2\pi k}{\tau(2N-1)}\right) \text{ if } 0 \le k \le N-1, \\ \frac{1}{\tau(2N-1)}\hat{h}\left(\frac{2\pi(k-2N+1)}{\tau(2N-1)}\right) \text{ if } N \le k \le 2N-2. \end{cases}\end{aligned}$$

Hence, the correct samples of $\hat{h}$ to use in (7.63) are

$$\hat{h}'_k = \frac{1}{\tau(2N-1)}\hat{h}\left(\frac{2\pi k}{\tau(2N-1)}\right), \text{ for } k \in \{-(N-1), \ldots, (N-1)\}, \qquad (7.65)$$

regarded as a $(2N-1)$-periodic sequence. With the sequence $< \hat{h}'_k >$ defined using these samples of $\hat{h}(\xi)$, formula (7.63) provides another approximation to $h * x(t_0 + j\tau)$. It is important to note that these two approaches generally give **different** results. Which result is preferable is often an empirical question.

The algorithm which provides the fast implementation of the finite Fourier transform is called the "fast Fourier transform" or FFT. This basics of this algorithm are described in section 7.5.8. The FFT only works for a sequence whose length equals a power of two. If $N = 2^k$ then the computation of $\mathcal{F}_N(x)$ from x requires about $CN\log_2 N$ computations. Here C is a constant that does not depend on N. The FFT computation of $h * f(j\tau)$ for $N/2$ values of j therefore requires about $2CN\log_2 N + N$ computations. To approximate this integral directly requires $O(N^2)$-computations.

The formulæ obtained above always lead to sequences of odd length. To use the FFT these sequences must be augmented to get even length sequences. If both the signal and the impulse response are sampled in the time domain, each sequence is padded with zeros until we reach a power of 2. If the transfer function is sampled in the Fourier domain then samples of its Fourier transform are added, symmetrically about zero frequency, until we reach a power of 2. To accomplish this we need to have either one more sample at a positive frequency than at negative frequency or vice versa. One simply needs to make a choice.

Exercise 7.5.5. From formula (7.62) it is clear that $\mathcal{F}^{-1}_{2N-1}$ provides an approximation to the inverse Fourier transform. Formula (7.62) is a Riemann sum approximation. By putting weight factors into the definition of $\hat{h}_s$ one can approximate other integration schemes such as the trapezoidal rule or Simpson's rule. How should the coefficients $\{\hat{h}_s(k)\}$ be modified so that

$$\frac{2N-1}{N}\mathcal{F}^{-1}_{2N-1}(\hat{h}_s(0)\hat{f}_s(0), \ldots, \hat{h}_s(2N-2)\hat{f}_s(2N-2))(k)$$

provides a trapezoidal rule or Simpson's rule approximation to

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{h}(\xi)\hat{f}(\xi)e^{i\frac{k}{N}\xi}d\xi\,?$$

7.5.6 Zero padding reconsidered

Padding the sequence of samples of f with $N-1$ zeros is a purely mechanical requirement for using the finite Fourier transform to evaluate the discrete convolution: the finite sum in (7.60) must be seen as a periodic convolution of two sequences of *equal* length. There is also an analytic interpretation for zero padding. For each positive integer m define the function

$$f_m(x) = \begin{cases} f(x) & \text{for } x \in [0,1], \\ 0 & \text{for } x \in (1,m]. \end{cases}$$

Let

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-ix\xi} dx$$

be the Fourier transform of f thought of as a function defined on the whole real line, supported in $[0,1]$.

Fix a sample spacing $\tau > 0$ in the time domain and collect $N_m = \frac{m}{\tau} + 1$ samples of f_m,

$$f_{m,s} = (f(0), f(\tau), \ldots, f((N_m - 1)\tau)).$$

Of course the samples $f_{m,s}(k)$ are zero for $k > \tau^{-1}$. The finite Fourier transform of this sequence is a sequence $\hat{f}_{m,s}$ of length N_m. Formula (7.56) implies that the correct interpretation of this sequence as an approximation to $\hat{f}$ is

$$\begin{aligned} \hat{f}_{m,s}(k) &\approx \frac{1}{m} \hat{f}_m(k) \text{ for } k \leq \frac{N_m}{2} \\ \hat{f}_{m,s}(k) &\approx \frac{1}{m} \hat{f}_m(k - N_m) \text{ for } k > \frac{N_m}{2}. \end{aligned} \tag{7.66}$$

On the other hand

$$\hat{f}_m(k) = \int_0^m f_m(x) e^{-\frac{2\pi i k x}{m}} dx = \hat{f}(\frac{2\pi k}{m})$$

and therefore

$$\begin{aligned} \hat{f}_{m,s}(k) &\approx \frac{1}{m} \hat{f}(\frac{2\pi k}{m}) \text{ for } k \leq \frac{N_m}{2}, \\ \hat{f}_{m,s}(k) &\approx \frac{1}{m} \hat{f}(\frac{2\pi(k - N_m)}{m}) \text{ for } k > \frac{N_m}{2}. \end{aligned} \tag{7.67}$$

As k varies from 0 to $\frac{m}{\tau}$ the sequence $\{\hat{f}_{m,s}(k)\}$ consists of approximate samples of $\hat{f}(\xi)$ for $\xi \in [-\frac{\pi}{\tau}, \frac{\pi}{\tau}]$. The effective bandwidth does **not** depend on m, whereas the sample spacing in the Fourier domain is $\frac{2\pi}{m}$. This shows that the effect of adding additional zeros to a sequence of samples of a function with bounded support is to decrease the effective mesh size in the Fourier domain.

7.5.7 Higher dimensional filters

Similar considerations apply to implement shift invariant, linear filters acting on inputs which depend on more than one variable. As in the one dimensional case the convolution is usually computed using a Fourier representation. As there is no essential difference between the 2-dimensional and n-dimensional cases we consider the general case. Boldface letters are used to denote points in $\mathbb{R}^n$, e.g.

$$\mathbf{t} = (t_1, \ldots, t_n), \quad \mathbf{x} = (x_1, \ldots, x_n) \text{ etc.}$$

Suppose that $f(\mathbf{t})$ is an input, with bounded support, depending continuously on n real variables. A uniform sample set in $\mathbb{R}^n$ is specified by a vector of positive numbers $\mathbf{h} = (h_1, \ldots, h_n)$, whose coordinates are the sample spacings in the corresponding coordinate directions. In one dimension the sample points and samples are labeled by an integer, in n dimensions it is more convenient to use n-tuples of integers. The integer vector $\mathbf{j} = (j_1, \ldots, j_n) \in \mathbb{Z}^n$ labels the sample point

$$\mathbf{x_j} = (j_1 h_1, \ldots, j_n h_n)$$

and the sample

$$f_{\mathbf{j}} = f(\mathbf{x_j}).$$

As it should cause no confusion, sets labeled by such integer vectors are usually called *sequences*.

Riemann sum approximations

Let $a(\mathbf{t})$ denote the impulse response of a shift invariant filter, $\mathcal{A}$ acting on a function of n variables,

$$\mathcal{A}\, f(\mathbf{t}) = \int_{\mathbb{R}^n} f(\mathbf{s}) a(\mathbf{t} - \mathbf{s}) d\mathbf{s}.$$

An n-dimensional integral is computed by re-writing it as iterated, 1-dimensional integrals,

$$\mathcal{A}\, f(\mathbf{t}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(s_1, \ldots, s_n) a(t_1 - s_1, \ldots, t_n - s_n) ds_1 \cdots ds_n.$$

The iterated integrals can, in turn be approximated by Riemann sums. Using the uniformly spaced samples defined by $\mathbf{h}$ to define a partition, the integral is approximated by

$$\mathcal{A}\, f(\mathbf{t}) \approx \sum_{j_1=-\infty}^{\infty} \cdots \sum_{j_n=-\infty}^{\infty} f(j_1 h_1, \ldots, j_n h_n) a(t_1 - j_1 h_1, \ldots, t_n - j_n h_n) h_1 \cdots h_n.$$

At sample points this can be re-written using the more economical notation, introduced above, as

$$\mathcal{A}\, f(\mathbf{x_k}) \approx h_1 \cdots h_n \sum_{\mathbf{j} \in \mathbb{Z}^n} f_{\mathbf{j}} a_{\mathbf{k}-\mathbf{j}}. \tag{7.68}$$

In the sequel the sample of the output $\mathcal{A}f(\mathbf{x_j})$ is denoted by $(\mathcal{A}f)_\mathbf{j}$.

Because the input, f is assumed to have bounded support, these sums can be replaced by finite sums. By translating the coordinates it can be assumed that $f_\mathbf{j}$ is only non-zero for $\mathbf{j}$ belonging to the set

$$\mathcal{J}_\mathbf{M} = \{\mathbf{j} \,:\, 1 \le j_i \le M_i, \quad i = 1, \ldots, n\}, \tag{7.69}$$

here $\mathbf{M} = (M_1, \ldots, M_n)$. Altogether there are $M_1 \cdots M_n$ potentially non-zero samples. The Riemann sum for $(\mathcal{A}f)_\mathbf{k}$ becomes

$$(\mathcal{A}f)_\mathbf{k} \approx [h_1 \cdots h_n] \sum_{j_1=1}^{M_1} \cdots \sum_{j_n=1}^{M_n} f_\mathbf{j} a_{\mathbf{k}-\mathbf{j}}. \tag{7.70}$$

If $a_\mathbf{k}$ is non-zero for most values of $\mathbf{k}$ and the numbers $\{M_j\}$ are powers of 2 then this is most efficiently computed using the Fourier representation. To compute the sum in (7.70), for all indices which satisfy (7.69), requires a knowledge of $a_\mathbf{k}$ for all indices in the set

$$\mathcal{J}_{2\mathbf{M}} = \{\mathbf{j} \,:\, 1 - M_i \le k_i \le M_i, \quad i = 1, \ldots, n\}. \tag{7.71}$$

To use the Fourier transform to compute (7.70) the set of samples $\{f_\mathbf{j}\}$, defined for $\mathbf{j} \in \mathcal{J}_\mathbf{M}$ must be augmented so that $f_\mathbf{j}$ is defined for all indices in $\mathcal{J}_{2\mathbf{M}}$. As f is assumed to vanish outside the sample set, this is done by adding the samples

$$\{f_\mathbf{j} = 0 \text{ for } \mathbf{j} \in \mathcal{J}_{2\mathbf{M}} \setminus \mathcal{J}_\mathbf{M}\}.$$

As in one dimension, this is called *zero padding*. In n-dimensions this amounts to adding about $(2^n - 1)M_1 \cdots M_n$ zero samples.

The finite Fourier transform

The n-dimensional finite Fourier transform is defined by iterating the one dimensional transform. Suppose that $f_\mathbf{j}$ is a collection of numbers parametrized by the set of indices

$$\mathcal{J}_\mathbf{N} = \{\mathbf{j} \,:\, 0 \le j_i \le N_i - 1, \quad i = 1, \ldots, n\}. \tag{7.72}$$

The finite Fourier transform of $(f_\mathbf{j})$ is the collection of numbers $(\hat{f}_\mathbf{k})$ defined by

$$\hat{f}_\mathbf{k} = \frac{1}{N_1 \cdots N_n} \sum_{j_1=0}^{N_1-1} \cdots \sum_{j_n=0}^{N_n-1} f_\mathbf{j} e^{-\frac{2\pi i j_1 k_1}{N_1}} \cdots e^{-\frac{2\pi i j_n k_n}{N_n}} \tag{7.73}$$

The formula defines $\hat{f}_\mathbf{k}$ for all $\mathbf{k} \in \mathbb{Z}^n$ as a periodic sequence, periodic of period N_j in the j^{th} index. Thus $\hat{f}_\mathbf{k}$ is also naturally parametrized by $\mathcal{J}_\mathbf{N}$. When it is important to emphasize the index set this, transform is denoted by $\mathcal{F}_\mathbf{N}$.

For $1 \le j \le n$, let $\mathcal{F}_l^1$ denote the one dimensional, finite Fourier transform acting only in the l^{th} index,

$$(\mathcal{F}_l^1 f)_\mathbf{k} = \frac{1}{N_l} \sum_{j=0}^{N_l-1} f_{k_1 \ldots k_{l-1} j k_{l+1} \ldots k_n} e^{-\frac{2\pi i j k_l}{N_l}}. \tag{7.74}$$

Suppose that $(a_{\mathbf{j}})$ and $(b_{\mathbf{j}})$ are sequences parametrized by $\mathcal{J}_{\mathbf{N}}$. These sequences can be extended to all of $\mathbb{Z}^n$ be requiring that they be periodic, of period N_i in the i^{th} index. A convolution operation is defined for such periodic sequences by setting

$$a \star b_{\mathbf{k}} = \sum_{\mathbf{j} \in \mathcal{J}_{\mathbf{N}}} a_{\mathbf{j}} b_{\mathbf{k}-\mathbf{j}}.$$

As in the one dimensional case, convolution is intimately connected to the finite Fourier transform:

$$\mathcal{F}_{\mathbf{N}}(a \star b)_{\mathbf{k}} = N_1 \cdots N_n \left[\hat{a} \cdot \hat{b}\right]. \tag{7.75}$$

Here $\hat{a} \cdot \hat{b}$ is the sequence whose $\mathbf{k}^{\text{th}}$-element is the ordinary product $\hat{a}_{\mathbf{k}} \hat{b}_{\mathbf{k}}$. This relation is the basis for computing convolutions using the Fourier transform,

$$a \star b = N_1 \cdots N_n \mathcal{F}_{\mathbf{N}}^{-1} \left[\hat{a}_{\mathbf{k}} \cdot \hat{b}_{\mathbf{k}}\right] \tag{7.76}$$

Exercise 7.5.6. If $l \neq m$ show that

$$\mathcal{F}_l^1 \mathcal{F}_m^1 = \mathcal{F}_m^1 \mathcal{F}_l^1.$$

The n-dimensional finite Fourier transform can be computed as an iterated sum. In the notation introduced above

$$\mathcal{F}_{\mathbf{N}}(f_{\mathbf{j}}) = \mathcal{F}_n^1 \circ \cdots \circ \mathcal{F}_1^1(f_{\mathbf{j}}) \tag{7.77}$$

Exercise 7.5.7. If M is a power of 2 then the length M, finite Fourier transform can be computed using $O(M \log_2 M)$ arithmetic operations. Show that if each $M_j, \quad j = 1, \ldots, n$ is a power of 2 then the "length $(M_1, \ldots, M_n)$," n-dimensional, finite Fourier transform can be computed using $O(M_1 \log_2 M_1 \cdots M_n \log_2 M_n)$ arithmetic operations. Can you give a better estimate?

Exercise 7.5.8. Using (7.77) find a formula for the inverse of the n-dimensional, finite Fourier transform.

Exercise 7.5.9. Prove the identity (7.75).

The Fourier representation for shift invariant filters

If the sequence $(f_{\mathbf{j}})$ is obtained as uniformly spaced samples of a function then the finite Fourier transform has an interpretation as an approximation to the Fourier transform of f. Let $\mathbf{h}$ denote the sample spacing and $\mathbf{M} = (M_1, \ldots, M_n)$ the number of samples in each coordinate direction. The sample set and its Fourier transform are parametrized by $\mathcal{J}_{\mathbf{M}}$. Because $\hat{f}_{\mathbf{k}}$ is periodic, it is also possible to parametrize it using the indices

$$\mathcal{J}'_{\mathbf{M}} = \{\mathbf{k} \, : \, -\frac{1 - M_i}{2} \leq k_i \leq \frac{M_i - 1}{2}\},$$

here we assume that the $\{M_i\}$ are odd numbers. For an index $\mathbf{k} \in \mathcal{J}'_{\mathbf{M}}$ the relationship between $\hat{f}_{\mathbf{k}}$ and $\hat{f}$ is easily expressed

$$\hat{f}_{\mathbf{k}} \approx \frac{1}{h_1 \cdots h_n} \hat{f}(\frac{2\pi k_1}{M_1 h_1}, \ldots, \frac{2\pi k_n}{M_n h_n}). \tag{7.78}$$

The effective bandwidth of the sample set,

$$[-\frac{\pi}{h_1},\frac{\pi}{h_1}]\times\cdots\times[-\frac{\pi}{h_n},\frac{\pi}{h_n}],$$

depends only on the sample spacing. The number of samples in each direction is then determined by the size of the support of f.

In light of (7.68) and (7.76) the finite Fourier transform can be used to approximate the output of a linear, shift invariant filter. Let $\mathcal{J}_{\mathbf{M}}$ denote the indices satisfying (7.69) and $\mathcal{J}_{2\mathbf{M}}$ the augmented index set satisfying (7.71). The samples of f are denoted $(f_s(\mathbf{j}) : \mathbf{j}\in\mathcal{J}_{\mathbf{M}})$ and the samples of a by $(a_s(\mathbf{j}) : \mathbf{j}\in\mathcal{J}_{2\mathbf{M}})$. To compute $a_s\star f_s$ the samples of f have to be zero padded to be defined on $\mathcal{J}_{2\mathbf{M}}$ and both sequences should be considered periodic with period $2M_i-1$ in the i^{th} index. If $\mathbf{h}$ is the vector of sample spacings then

$$\begin{aligned}\mathcal{A}\,f(j_1h_1,\ldots,j_nh_n)&\approx[h_1\cdots h_n](a_s\star f_s)_{\mathbf{j}}\\&=[h_1(2M_1-1)\cdots h_n(2M_n-1)]\mathcal{F}^{-1}_{2\mathbf{M}}(\hat{a}_s\hat{f}_s)_{\mathbf{j}}.\end{aligned}\tag{7.79}$$

As in the one dimensional case a slightly different way to approximate shift invariant filters is to bypass the impulse response and sample the transfer function directly. Again this is because the transfer function is often an ordinary function, even when the impulse response is not. Using (7.78), the discussion leading up to (7.65) can be adapted to show that the correct samples of $\hat{a}(\xi)$ to use in (7.79) are

$$\hat{a}_s(\mathbf{j})=\frac{1}{h_1(2M_1-1)\cdots h_n(2M_n-1)}\hat{a}\left(\frac{2\pi j_1}{h_1(2M_1-1)},\ldots,\frac{2\pi j_n}{h_1(2M_n-1)}\right),\quad\text{for } 1-M_i\le j_i\le M_i-1,\quad i=1,\ldots,n.\tag{7.80}$$

Exercise 7.5.10. Show that (7.79) gives a Riemann sum approximation to the Fourier representation of $a*f(j_1h_1,\ldots,j_nh_n)$.

Exercise 7.5.11. Give a detailed justification for (7.80).

Exercise 7.5.12. The Laplace operator is defined in two dimensions as

$$\Delta f=\partial^2_{x_1}f+\partial^2_{x_2}f.$$

The transfer function for this operator is $-(\xi_1^2+\xi_2^2)$. The Laplace operator can also be approximated by finite differences, for example

$$\begin{aligned}\Delta f(x_1,x_2)\approx&\frac{f(x_1+h,x_2)-2f(x_1,x_2)+f(x_1-h,x_2)}{h^2}+\\&\frac{f(x_1,x_2+h)-2f(x_1,x_2)+f(x_1,x_2-h)}{h^2}.\end{aligned}\tag{7.81}$$

Compare the approximations obtained from sampling the transfer function directly and using the Fourier representation of the finite difference formula.

Exercise 7.5.13. In the previous exercise what is the impulse response of the finite difference approximation to Δ?

7.5.8 Appendix: The Fast Fourier Transform

If $N = 2^q$ then there is a very efficient way to compute the finite Fourier transform of a sequence of length N. The fast algorithm for the finite Fourier transform is the Cooley-Tukey or fast Fourier transform algorithm, usually referred to as the "FFT." Let $\zeta = e^{\frac{2\pi i}{N}}$ be the primitive N^{th}-root of unity and let $\zeta_j = \zeta^j$. This makes the notation simpler in what follows. The finite Fourier transform of $(f(0), f(1), \ldots, f(N-1))$ is given by

$$\hat{f}(k) = \frac{1}{N}\sum_{j=1}^{N-1} f(j)e^{-\frac{2\pi ijk}{N}} = \frac{1}{N}\sum_{j=1}^{N-1} f(j)\bar{\zeta}_j^k,$$

which can be expressed as a matrix multiplying a vector:

$$\begin{pmatrix} \hat{f}(0) \\ \hat{f}(1) \\ \hat{f}(2) \\ \vdots \\ \hat{f}(N-1) \end{pmatrix} = \frac{1}{N}\begin{pmatrix} 1 & 1 & \cdots & \cdots & 1 \\ 1 & \bar{\zeta}_1 & \bar{\zeta}_2 & \cdots & \bar{\zeta}_{N-1} \\ 1 & \bar{\zeta}_1^2 & \bar{\zeta}_2^2 & \cdots & \bar{\zeta}_{N-1}^2 \\ 1 & \cdots & & \cdots & \\ 1 & \bar{\zeta}_1^{N-1} & \bar{\zeta}_2^{N-1} & \cdots & \bar{\zeta}_{N-1}^{N-1} \end{pmatrix}\begin{pmatrix} f(0) \\ f(1) \\ f(2) \\ \vdots \\ f(N-1) \end{pmatrix}$$

Denote this matrix by C_N. We now show that for $N = 2^q$, the number of calculations involved in multiplying C_N times a vector can be reduced to $3Nq$. For even modest values of q this is a much smaller number than $N(2N-1)$. This reduction comes from several observations about the structure of the matrix C_N.

If A is a square matrix with complex entries a_{ij} then the adjoint of A, denoted A^* is the matrix whose ij^{th}-entry is $\bar{a}_{ji}$. The matrix A is *unitary* if

$$A^{-1} = A^*.$$

The matrix

$$\sqrt{N}C_N$$

is a unitary matrix. This is a consequence of formulæ (7.50), (7.51). In matrix notation we have

$$\sqrt{N}C_N^*\sqrt{N}C_N = I.$$

The inverse of C_N therefore has essentially the same form as C_N^*. A fast algorithm for multiplication by C_N should also give a fast algorithm for multiplication by C_N^{-1}. In other words, if we can compute $\mathcal{F}_N$ efficiently then we can also compute $\mathcal{F}_N^{-1}$ efficiently.

The following identities among the N^{th} and $(2N)^{\text{th}}$ roots of unity lead to the fast Fourier transform algorithm. Let $\mu = e^{\frac{2\pi i}{2N}}$ be the primitive $(2N)^{\text{th}}$ root of unity; as above $\mu_j = \mu^j$. The following identities are elementary:

$$e^{\frac{2\pi i2kj}{2N}} = e^{\frac{2\pi ikj}{N}}, \quad e^{\frac{2\pi ik(j+N)}{N}} = e^{\frac{2\pi ikj}{N}} \text{ and } e^{\frac{2\pi i(2k+1)j}{2N}} = e^{\frac{2\pi ikj}{N}}e^{\frac{2\pi ij}{2N}}.$$

These identities can be rewritten

$$\mu_j^{2k} = \zeta_j^k, \quad \zeta_j^{k+N} = \zeta_j^k \text{ and } \mu_j^{2k+1} = \mu_j\zeta_j^k. \tag{7.82}$$

From the definition of C_N, the $(2k+1)^{\text{st}}$ and $(2k+2)^{\text{th}}$ rows of C_{2N} are given by

$$\begin{aligned}(2k+1)^{\text{st}} &: (1, \bar{\mu}_1^{2k}, \cdots, \bar{\mu}_{2N-1}^{2k})\\ (2k+2)^{\text{th}} &: (1, \bar{\mu}_1^{2k+1}, \cdots, \bar{\mu}_{2N-1}^{2k+1})\end{aligned}$$

Comparing these with the k^{th} row of C_N and using the relations in (7.82), rows of C_{2N} can be expressed in terms of the rows of C_N as follows:

$$(2k+1)^{\text{st}} \quad : \quad (1, \bar{\mu}_1^{2k}, \cdots, \bar{\mu}_{N-1}^{2k}, \bar{\mu}_N^{2k}, \cdots, \bar{\mu}_{2N-1}^{2k}) = (1, \bar{\zeta}_1^{k}, \cdots, \bar{\zeta}_{N-1}^{k}, 1, \bar{\zeta}_1^{k}, \cdots, \bar{\zeta}_{N-1}^{k}) \tag{7.83}$$

and,

$$\begin{aligned}(2k+2)^{\text{th}} \quad : \quad & (1, \bar{\mu}_1^{2k+1}, \cdots, \bar{\mu}_{N-1}^{2k+1}, \bar{\mu}_N^{2k+1}, \cdots, \bar{\mu}_{2N-1}^{2k+1})\\ & = (1, \bar{\zeta}_1^{k}\bar{\mu}_1, \bar{\zeta}_2^{k}\bar{\mu}_2 \cdots, \bar{\zeta}_{N-1}^{k}\bar{\mu}_{N-1}, \bar{\mu}_N, \bar{\zeta}_1^{k}\bar{\mu}_{N+1} \cdots, \bar{\zeta}_{N-1}^{k}\bar{\mu}_{2N-1})\end{aligned}$$

In terms matrices, C_{2N} is essentially obtained by multiplying 2 copies of C_N by another very simple matrix:

$$C_{2N} = C_N^{\#} U_N.$$

Define the $2N \times 2N$-matrix

$$C_N^{\#} = \begin{pmatrix} r_1 & \mathbf{0} \\ \mathbf{0} & r_1 \\ \vdots & \vdots \\ r_N & \mathbf{0} \\ \mathbf{0} & r_N \end{pmatrix},$$

where the $\{r_i\}$ are the rows of C_N and the vector $\mathbf{0} = (\underbrace{0, \ldots, 0}_{N})$,

$$C_N = \begin{pmatrix} r_1 \\ \vdots \\ r_N \end{pmatrix}.$$

The matrix U_N is defined by

$$U_N = \begin{pmatrix} I & I \\ D_N^1 & D_N^2 \end{pmatrix}$$

where

$$D_N^1 = \begin{pmatrix} 1 & & & & \\ & \bar{\mu}_1 & & 0 & \\ & & \bar{\mu}_2 & & \\ & 0 & & \ddots & \\ & & & & \bar{\mu}_{N-1} \end{pmatrix} \text{ and } D_N^2 = \begin{pmatrix} \bar{\mu}_N & & & & \\ & \bar{\mu}_{N+1} & & 0 & \\ & & \bar{\mu}_{N+2} & & \\ & 0 & & \ddots & \\ & & & & \bar{\mu}_{2N-1}. \end{pmatrix}$$

The very important feature of U_N is that it has exactly 2 non-zero entries per row. If $N = 2^q$ then this argument applies recursively to $C_N^{\#}$ to give a complete factorization.

Theorem 7.5.1. *If $N = 2^q$ then $C_N = E_1 E_2 \cdots E_q$ where each row of the $N \times N$ matrix E_i has 2 nonzero entries.*

It is not difficult to determine exactly which entries in each row of the matrices E_j are non-zero. For an arbitrary N-vector $\mathbf{v} = (v_1, \ldots, v_N)$, the computation of $E_j\mathbf{v}$ can be done using exactly $N(2\text{multiplications} + 1\text{addition})$. Using this factorization and the knowledge of which entries of the E_j are non-zero one can reduce the number of operations needed to compute the matrix product $C_N\mathbf{v}$ to $3qN = 3N\log_2 N$. Indeed the combinatorial structures of the matrices, $\{E_j\}$ are quite simple and this has led to very efficient implementations of this algorithm. Each *column* of E_j also has exactly two non-zero entries and therefore the factorization of C_N gives a factorization of C_N^* :

$$C_N^* = E_q^* E_{q-1}^* \ldots E_1^*.$$

Example 7.5.2. For example we can factor the matrix $2C_4$ as

$$4C_4 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & -i & 0 & i \end{pmatrix}. \tag{7.84}$$

For a more complete discussion of the fast Fourier transform see [53].

7.6 Image processing

See: A.4.7.

Image processing is a sub-discipline of higher dimensional filtering. In this context "images" are mathematical representations of "pictures," in the naive sense of the word. The Fourier transform of an image depends on the same number of variables but is *not* an image in this sense as it is not generally recognizable as a "picture" of anything, see figure 7.19. The basic goal of image processing is to make information in pictures more accessible to the human visual system. Another related aim is to help machines to "see." Image processing operations and the language used to describe them closely mirror these origins and intents. While the filters described below can be applied to any sufficiently regular function of the correct number of variables, the interpretation of the output is closely tied to these *a priori* assumptions about the inputs. Image processing is a very important component of medical imaging, though these operations are more important in the "post-processing" phase rather than in the basic measurement and image formation processes, which are our main topic. We present this material both because of its importance to medical imaging *per se* and because of the its rich mathematical content. It shows how particular aims shape the design and implementation of filters. Our presentation is adapted from the very extensive treatment of this subject in [32].

(a) An "image." (b) Its Fourier transform.

Figure 7.19: The Fourier transform of an image is not usually an image.

7.6.1 Basic concepts and operations

We first introduce basic image processing operations and examples of filters which implement them. Many image processing operations are non-linear. For simplicity we consider planar images which are represented by scalar functions of two variables. The value, $f(x_1, x_2)$, represents the "grey level" or density of the image at "(x_1, x_2)." With this interpretation it is reasonable to assume that f assumes only non-negative values. Very similar considerations apply to "color images" which are usually represented by a triple of functions $[r(x_1, x_2), g(x_1, x_2), b(x_1, x_2)]$. These functions represent the intensities of three "independent colors" at (x_1, x_2). Higher dimensional images are treated using similar methods.

A visual output device, such as a monitor, is required to pass from a functional description of an image, i.e. f to a picture, in the ordinary sense. Such a device has a fixed size and dynamic range. Mappings must be fixed between the coordinates used in the parameterization of f and coordinates in the output device as well as between the values f assumes and grey levels (or colors) available in the output. When it is necessary to distinguish between these two coordinate systems we speak of the "measurement plane" and the "image plane." Generally speaking f is defined on the measurement plane and the output device is the identified with the image plane. Sometimes the image plane refers to the original image itself. When this distinction is not needed we implicitly identify the image plane with the measurement plane. In the first part of this section we consider filtering operations defined on functions of continuous variables, at the end we briefly discuss the problems of sampling continuous images and implementation of the filters.

The basic operations of image processing fall into several classes:

Coordinate transformations

Images are sometimes distorted because of systematic modeling (or measurement) errors. In this connection it is useful to make a distinction between the "image plane" and the "measurement plane." In this paragraph the image plane is the plane in which the image lies; let (y_1, y_2) denote orthogonal coordinates in this plane. This means that if $f(y_1, y_2)$ is

displayed as a grey level image in the y_1y_2-plane then the image appears undistorted. The measurement plane refers to the coordinates defined by the apparatus used to *measure* the image; let (x_1, x_2) denote coordinates in this plane. Suppose that $f(x_1, x_2)$, for (x_1, x_2) lying in a subset D of $\mathbb{R}^2$, are measurements of an image. The parameters (x_1, x_2) in the measurement plane may differ from the coordinates in the image plane. Displaying $f(x_1, x_2)$ in the x_1x_2-plane would then result in a distorted image. This is called *geometric distortion*. The following example illustrates this point.

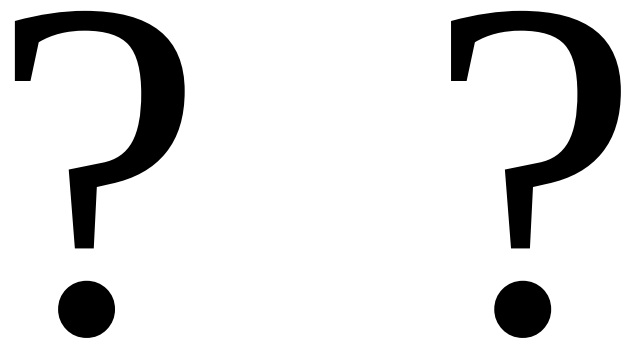
Example 7.6.1. Suppose that the measuring device is *calibrated* in polar coordinates with (x_1, x_2) corresponding to the point in the image plane with Cartesian coordinates

$$y_1 = x_1 \cos x_2, \quad y_2 = x_1 \sin x_2. \tag{7.85}$$

Displaying f in the x_1x_2-plane would result in a distorted image, see figure 7.20(a). Define a filter $\mathcal{A}_{pr}$ with

$$(\mathcal{A}_{pr}\, f)(y_1, y_2) = f(\sqrt{y_1^2 + y_2^2}, \tan^{-1}\left(\frac{y_2}{y_1}\right)).$$

The filtered image is shown in 7.20(b). This is a linear filter whose output is the image parametrized by Cartesian coordinates in the image plane. The actual calibration of the measuring equipment determines the choice of branch of $\tan^{-1}$. The value of $(\mathcal{A}_{pr}\, f)(y_1, y_2)$ is not defined if $(y_1, y_2) = (0, 0)$; notwithstanding that the inverse transformation, (7.85) is defined in the whole "polar coordinate plane." It is, however, not one-to-one.



(a) f displayed in the x_1x_2-plane.

(b) $\mathcal{A}\, f$ displayed in the image plane.

Figure 7.20: Removing geometric distortion.

The example illustrates most of the features of the general case. Let (x_1, x_2) be coordinates in the measurement plane and suppose that these differ from those in the image plane. Denote the latter coordinates by (y_1, y_2). The two coordinate systems are functionally related with

$$x_1 = g(y_1, y_2), x_2 = h(y_1, y_2),$$

defined for (y_1, y_2) belonging to a subset, D_i in the image plane. This defines a mapping Φ from D_i to $D_m = \Phi(D_i)$, a subset of the measurement plane. Let $f(x_1, x_2)$ for $(x_1, x_2) \in D$ denote the measurements of an image; then

$$(\mathcal{A}_\Phi f)(y_1, y_2) = f \circ \Phi(y_1, y_2) = f(g(y_1, y_2), h(y_1, y_2)) \tag{7.86}$$

defines a filter which maps the portion of the measured image lying over $D_m \cap D$ to an image defined in the corresponding part of D_i. As with the example of polar coordinates, this transformation may not be defined in the the entire image plane and there may be choices involved in the definition of the map $\Phi : (y_1, y_2) \mapsto (g(y_1, y_2), h(y_1, y_2))$. This operation defines a linear filter which is usually not translation invariant. The kernel function for $\mathcal{A}_\Phi$ is the generalized function

$$a_\Phi(y_1, y_2; x_1, x_2) = \delta(x_1 - g(y_1, y_2), x_2 - h(y_1, y_2)),$$

here δ is the two-dimensional delta function. Let $g(y_1, y_2), h(y_1, y_2)$ be a pair of functions defined in a subset $D \subset \mathbb{R}^2$. Setting

$$\Phi(y_1, y_2) = (g(y_1, y_2), h(y_1, y_2))$$

defines a map of D into a subset $D' = \Phi(D)$ of $\mathbb{R}^2$. Whether or not Φ is a change of coordinates, formula (7.86) defines a linear filter carrying functions defined on D' to functions defined on D.

Exercise 7.6.1. Find conditions on (g, h) which imply that $\mathcal{A}_\Phi$ is translation invariant.

Exercise 7.6.2. Suppose that the image is a transparency lying in the y_1y_2-plane, $f(y_1, y_2)$ describes the amount of incident light transmitted through the point (y_1, y_2). Suppose that the measurement is made by projecting the transparency onto a screen which lies in the plane $y_2 = y_3$ using a light source which produces light rays orthogonal to the y_1y_2-plane.

(1). Using $(x_1, x_2) = (y_1, \sqrt{2}y_3)$ as coordinates for the measurement plane, find an expression for the amount of light incident at each point of the measurement plane, see section 2.2.

(2). Why are these good coordinates for the measurement plane?

(3). Find a filter which removes the geometric distortion resulting from the projection.

Exercise 7.6.3. Let $\Phi = (g, h)$ be a pair of functions defined in $D \subset \mathbb{R}^2$ and let $D' = \Phi(D)$. What are necessary and sufficient conditions on Φ for the filter A_Φ to be invertible as a map from functions defined on D' to functions defined on D.

In the remainder of this section the measurement plane and image plane are assumed to agree.

Noise reduction

Images can be corrupted by noise. There are two main types of noise: *uniform* noise and *binary* noise. In the first case the noise is uniformly distributed and locally of mean zero, whereas binary noise consists of sparsely, but randomly, distributed large errors. It is often caused by sampling or transmission errors. In this section techniques are described for reducing the affects of uniform noise; binary noise is discussed in the section 7.6.2. Because uniform noise is locally of mean zero, replacing f by a weighted average generally has the effect of reducing the noise content.

A shift invariant filter of this type is defined by convolution with a weight function $\varphi(x_1, x_2)$ which satisfies the conditions

$$\begin{aligned} &\varphi(x_1, x_2) \geq 0, \\ &\int\limits_{\mathbb{R}^2} \varphi dx_1 dx_2 = 1. \end{aligned} \tag{7.87}$$

Let $\mathcal{A}_\varphi f(x_1, x_2) = \varphi * f(x_1, x_2)$ be the convolution filter defined by φ. The first condition ensures that a non-negative function is always carried to a non-negative functions while the second ensures that a constant function is mapped to a constant function. If φ has support in a small ball then the output of $\mathcal{A}_\varphi f$ at (x_1, x_2) only depends on values of $f(y_1, y_2)$ for (y_1, y_2), near to (x_1, x_2) and the filter acts in a localized manner. If the noise has a directional dependence then this can be incorporated into φ. If, for example, the noise is isotropic then it is reasonable to use a radial function.

The frequency representation of such a filter is

$$\mathcal{A}_\varphi f = \mathcal{F}^{-1}\left[\hat{\varphi}\hat{f}\right]. \tag{7.88}$$

As φ is integrable, its Fourier transform tends to zero as $\|\xi\|$ tends to infinity. This means that $\mathcal{A}_\varphi$ is an approximate low pass filter. If $\hat{\varphi}(\xi_1, \xi_2)$ is a function satisfying the conditions

$$\begin{aligned} \hat{\varphi}(0, 0) &= 1, \\ \lim_{\xi \to \infty} \hat{\varphi}(\xi) &= 0, \end{aligned} \tag{7.89}$$

then (7.88) defines an approximate low pass filter. Even if the inverse Fourier transform, φ assumes negative values, the effect of $\mathcal{A}_\varphi$ is to reduce uniform noise. In this generality there can be problems interpreting the output as representing an image. The filtered output, $\mathcal{A}_\varphi f$ may assume negative values, even if f is pointwise positive; to represent $\mathcal{A}_\varphi f$ as image it is necessary to remap its range to the allowable range of densities. This operation is discussed below in the paragraph on contrast enhancement. If $\hat{\varphi}$ vanishes outside a bounded set then φ cannot have bounded support, see Proposition 3.2.10. From the representation of $\mathcal{A}_\varphi f$ as a convolution, it follows that the value of $\mathcal{A}_\varphi f$ at a point (x_1, x_2) depends on values of f at points distant from (x_1, x_2).

As observed above, reduction of the available resolution is an undesirable side effect of low pass filtering. In an image this appears as blurring. Using statistical properties of the noise an "optimal filter" can be designed which provides an balance between noise reduction and blurring. This is discussed in Chapter 11. Noise might be locally uniform,

but non-uniform across the image plane. In this case a non-shift invariant filter might do a better job reducing the affects of noise. A heuristic, used in image processing to retain detail in a filtered image, is to represent an image as a linear combination of the low pass filtered output and the original image. That is, instead of using either $\mathcal{A}_\varphi f$ or f alone we use

$$\mu \mathcal{A}_\varphi f + (1-\mu) f, \tag{7.90}$$

where $0 \leq \mu \leq 1$. If the noise is non-uniform across the image plane then μ could be taken to depend on (x_1, x_2).

Sharpening

An image may be blurred during acquisition, it is sometimes possible to filter the image and recapture some of the fine detail. In essence this is an inverse filtering operation. The measured image is modeled as $\mathcal{A} f$ where f denotes the "original" image and $\mathcal{A}$ models the measurement process. If $\mathcal{A}$ is a shift invariant, linear filter then the methods discussed in section 7.3 can be applied to try to approximately invert $\mathcal{A}$ and restore details present in f which were "lost" in the measurement process. In general this leads to an amplification of the high frequencies which can, in turn, exacerbate problems with noise.

The fine detail in an image is also "high frequency information." A slightly different approach to the de-blurring problem is to simply remove, or attenuate the low frequency information. In the frequency domain representation, such a filter is given by

$$\mathcal{A}_\varphi f = \mathcal{F}^{-1}\left[\hat{\varphi}\hat{f}\right]$$

where, instead of 7.89, $\hat{\varphi}$ satisfies

$$\begin{aligned} \hat{\varphi}(0) &= 0, \\ \lim_{\xi \to \infty} \hat{\varphi}(\boldsymbol{\xi}) &= 1. \end{aligned} \tag{7.91}$$

If the function $\hat{\varphi} - 1$ has an inverse Fourier transform, ψ then this filter has spatial representation as

$$\mathcal{A}_\varphi f = f - \psi * f.$$

In other words A_φ is the difference between the identity filter and an approximate low pass filter, hence it is an approximate high pass filter.

Edge detection

Objects in an image are delimited by their edges. Edge detection is the separation of the boundaries of objects from other, more slowly varying features of an image. The rate at which a smooth function varies in direction ω is measured by its directional derivative

$$D_\omega f(\mathbf{x}) = \frac{d}{dt} f(\mathbf{x} + t\omega)\big|_{t=0} = \langle \nabla f, \omega \rangle.$$

Here

$$\nabla f = (\partial_{x_1} f, \partial_{x_2} f),$$

is the gradient of f. The Euclidean length of ∇f provides an isotropic measure of the variation of f, i.e. it is equally sensitive to variation in all directions. Thus points where $\|\nabla f\|$ is large should correspond to edges. An approach to separating the edges from other features is to set a threshold, t_{edge} so that points with $\|\nabla f(\mathbf{x})\| > t_{\text{edge}}$ are considered to belong to edges. A filtered image, showing only the edges, would then be represented by

$$\mathcal{A}_1 f(\mathbf{x}) = \chi_{(t_{\text{edge}},\infty)}(\|\nabla f(\mathbf{x})\|). \tag{7.92}$$

While this is a non-linear filter, it is shift invariant and the computation of ∇f can be done very efficiently using the Fourier transform.

This approach is not robust as the gradient of f is also large in highly textured or noisy regions. If D is a region in the plane with a smooth boundary and $f = \chi_D$ then $\nabla f(\mathbf{x}) = 0$ if $\mathbf{x} \notin bD$. From the point of view of sampling, it may be quite difficult to detect such a sharply defined edge. These problems can be handled by smoothing f before computing its gradient. Let φ denote a smooth function, with small support satisfying (7.87). In regions with a lot of texture or noise, but no boundaries, the gradient of f varies "randomly" so that cancelation in the integral defining

$$\nabla(\varphi * f) = \varphi * \nabla f$$

should lead to a relatively small result. On the other hand, along an edge, the gradient of f is dominated by a component in the direction orthogonal to the edge. Therefore the weighted average, $\varphi * \nabla f$ should also have a large component in that direction. Convolution smears the sharp edge in χ_D over a small region and therefore points where $\|\nabla(\varphi * \chi_D)\|$ is large are more likely to show up in a sample set. This give a second approach to edge detection implemented by the filter

$$\mathcal{A}_2 f(\mathbf{x}) = \chi_{(t_{\text{edge}},\infty)}(\|\nabla(\varphi * f)(\mathbf{x})\|). \tag{7.93}$$

Once again $f \mapsto \nabla\varphi * f$ is a linear shift invariant filter which can be computed efficiently using the the Fourier transform. In order not to introduce a preference for edges in certain directions, a radial function should be used to do the averaging.

The second filter helps find edges in regions with noise or texture but may miss edges where the adjoining grey levels are very close. In this case it might be useful to compare the size of the gradient to its average over a small region. In this approach the $\|\nabla f\|$ (or perhaps $\|\nabla\varphi * f\|$) is computed and is then convolved with a second averaging function to give $\psi * \|\nabla f\|$ (or $\psi * \|\nabla\varphi * f\|$). A point $\mathbf{x}$ then belongs to an edge if the ratio

$$\frac{\|\nabla f(\mathbf{x})\|}{(\psi * \|\nabla f\|)(\mathbf{x})}$$

exceeds a threshold τ_{edge}. A filter implementing this idea is given by

$$\mathcal{A}_3 f(\mathbf{x}) = \chi_{(\tau_{\text{edge}},\infty)}\left(\frac{\|\nabla f(\mathbf{x})\|}{\psi * \|\nabla f\|(\mathbf{x})}\right). \tag{7.94}$$

Once again the filter is non-linear but it major components can be efficiently implemented using the Fourier transform. The image produced by the output of these filters shows only the edges.

There is a related though somewhat more complicated approach to edge detection which entails the use of the Laplace operator,

$$\Delta f = \partial_{x_1}^2 f + \partial_{x_2}^2 f$$

as a way to measure the local variability of f. There are three reasons for this approach: (1) The Laplace operator is rotationally invariant. (2) The singularities of the function Δf are the same as those of f. (3) There is some evidence that animal optical tracts use a filtering operation of this sort to detect edges. The first statement is easily seen in the Fourier representation:

$$\mathcal{F}(\Delta f)(\xi_1, \xi_2) = -(\xi_1^2 + \xi_2^2)\hat{f}(\xi_1, \xi_2).$$

The second property requires more advanced techniques to prove, see [17]. The point of (2) is that the sharp edges of objects are discontinuities of the density function and will therefore remain discontinuities of Δf. For the reasons discussed above, the Laplace operator is often combined with a Gaussian smoothing operation,

$$G_\sigma f(\mathbf{x}) = \iint_{\mathbb{R}^2} e^{-\frac{|\mathbf{x}-\mathbf{y}|^2}{\sigma}} f(\mathbf{y}) d\mathbf{y}$$

to get

$$\mathcal{A}_\sigma f = \Delta G_\sigma f.$$

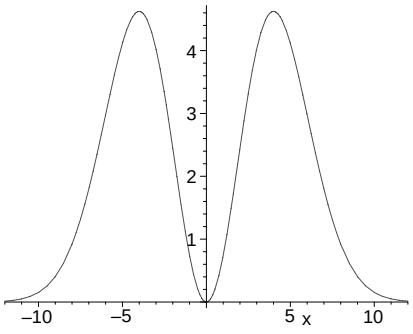
The transfer function of $\mathcal{A}_\sigma$ is

$$\hat{a}_\sigma(\xi_1, \xi_2) = -\pi\sigma(\xi_1^2 + \xi_2^2)e^{-\frac{\sigma(\xi_1^2+\xi_2^2)}{4}}$$

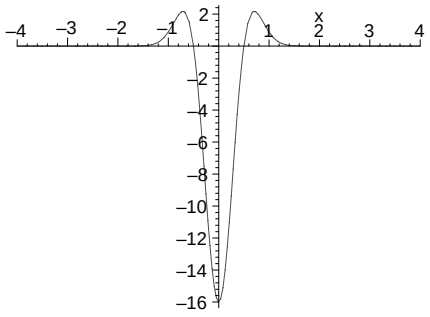
and its impulse response response is

$$a_\sigma(x_1, x_2) = \frac{4}{\sigma^2}(x_1^2 + x_2^2 - \sigma)e^{-\frac{x_1^2+x_2^2}{\sigma}}.$$

These are radial functions, the graphs of a radial section are shown in figure 7.21. Note that the impulse response assumes both positive and negative values.



(a) The transfer function.



(b) The impulse response.

Figure 7.21: The impulse response and transfer function for $\mathcal{A}_{.25}$.

It not immediately obvious how to use the output of $\mathcal{A}_\sigma$ to locate edges. This is clarified by considering the special case of a sharp edge. Applying $\mathcal{A}_\sigma$ to the function $\chi_{[0,\infty)}(x_1)$ gives

$$\mathcal{A}_\sigma\, \chi_{[0,\infty)}(x_1, x_2) = \frac{2cx_1}{\sqrt{\sigma}} e^{-\frac{x_1^2}{\sigma}},$$

here c is a positive constant. The zero crossing of $\mathcal{A}_\sigma\, \chi_{[0,\infty)}$ lies on the edge, nearby are two sharp peaks with opposite signs. Figure 7.22 shows a cross section in the output of $A_\sigma \chi_{[0,\infty)}$ orthogonal to the edge. The parameter σ takes the values $.01, .1, .25$ and 1; smaller values correspond to sharper peaks.

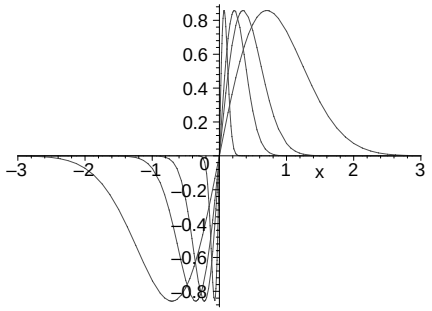


Figure 7.22: Output of Laplacian edge detection filters.

This example suggests that the absolute value of the output of filter $\mathcal{A}_\sigma$ can be used to locate edges. An edge shows up as a white arc (the zero crossings) flanked by two nearly parallel black arcs. A different approach is to define

$$B_\sigma f(x) = \frac{1}{2}\left[\text{sign}((\mathcal{A}_\sigma\, f)(x)) + 1\right].$$

The example above indicates that $B_\sigma f$ should be 0 on one side of an edge and 1 on the other. The image produced by $B_\sigma f$ is then a "schematic" showing the outlines of the objects with all the fine detail removed. An edge image can be combined with the original image, as in (7.90) to obtain a composite image with fine detail and slowly varying features as well as enhanced contrast between objects.

Exercise 7.6.4. Let U be a 2×2 matrix defining a rigid rotation. If f is a function defined in the plane the set

$$f_U(x_1, x_2) = f(U(x_1, x_2)).$$

Show that

$$\|\nabla f_U(0,0)\| = \|\nabla f(0,0)\|.$$

Explain the statement "∇f provides an isotropic measure of the variation of f."

Contrast enhancement

The equipment used to display an image has a fixed dynamic range. Suppose that the grey values that can be displayed are parametrized by the interval $[d_{\min}, d_{\max}]$. To output an image, a mapping needs to be fixed from the range of f to $[d_{\min}, d_{\max}]$. That is, the values of f need to be scaled to fit the dynamic range of the output device. Suppose that f assumes values in the interval $[m, M]$. Ordinarily, the scaling map is a monotone map $\gamma : [m, M] \to [d_{\min}, d_{\max}]$. This map is usually non-linear and needs to be adapted to the equipment being used. By choosing γ carefully, different aspects of the image can be emphasized or *enhanced*.

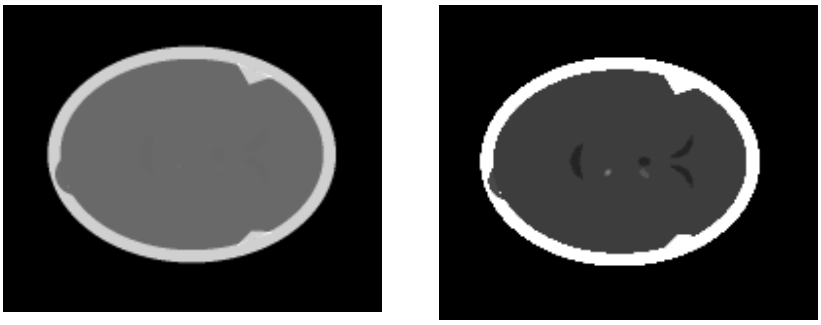
Suppose that there is a region R in the image where f varies over the range $[a, A]$. If $A - a$ is very small compared to $M - m$ then a linear scaling function,

$$\gamma(t) = d_{\max}\frac{t-m}{M-m} + d_{\min}\frac{M-t}{M-m},$$

would compress the information in R into a very small part of $[d_{\min}, d_{\max}]$. In the output this region would have very low contrast and so the detail present there might not be visible. The contrast in R can be enhanced by changing γ to emphasize values of f lying in $[a, A]$, though necessarily at the expense of values outside this interval. For example a piecewise linear scaling function

$$\gamma_{aA}(t) = \begin{cases} d_{\min} & \text{if } t < a, \\ d_{\max}\frac{t-a}{A-a} + d_{\min}\frac{A-t}{A-a} & \text{if } a \le t \le A, \\ d_{\max} & \text{if } t > A \end{cases}$$

would make the detail in R quite apparent. On the other hand, detail with grey values outside of $[a, A]$ is entirely lost. In figure 7.23 an example of a CT-phantom is shown with two different choices of γ. In medical imaging this is called windowing or thresholding. Another method used to make information present in an image more apparent in a visual display is to map the grey values to colors. In either case, no more information is actually "there" but, due to the physiology of human vision, it becomes more apparent.



(a) Low contrast display.

(b) High contrast display.

Figure 7.23: A CT-head phantom showing the effect of rescaling grey values.

7.6.2 Discretized images

In the previous section we considered an image to be a real valued function of a pair of continuous variables. As was the case with one dimensional signals, functions describing images are usually sampled on a discrete set of points and quantized to take values in a finite set. In this section we briefly discuss sampling of images and implementation of filters on the sampled data. As before we restrict our attention to two dimensional images. The sample set in imaging is usually a uniform rectangular grid,

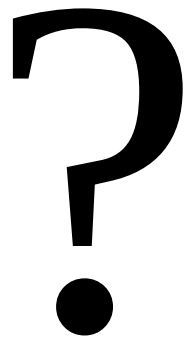
$$\{(jh_1, kh_2) \,:\, j, k \in \mathbb{Z}\},$$

where h_1 and h_2 are positive numbers. To simplify the discussion, we do not consider questions connected with quantization and work instead with the full range of real numbers.

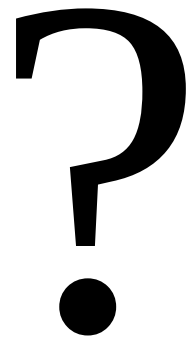
Suppose that the function $f(x_1, x_2)$ describing the image is supported in the unit square, $[0,1] \times [0,1]$. If the image has bounded support then this can always be arranged by scaling. The actual measurements of the image consist of a finite collection of samples, collected on a uniform rectangular grid. Let M and N denote the number of samples in the x_1 and x_2 directions respectively. The simplest model for measurements of an image is the set of samples

$$f_{jk} = f(\frac{j}{M}, \frac{k}{N}) \text{ for } 0 \leq j \leq M-1, \quad 0 \leq k \leq N-1.$$

The discussion of higher dimensional sampling presented in section 6.3 is directly applicable in this situation. In particular, Nyquist's theorem provides lower bounds on the sampling rates in terms of the bandwidth, needed to avoid aliasing. Because the bandwidth (or effective bandwidth) can vary with direction, aliasing can also have a directional component. The so called *moiré effect* is the result of undersampling a directional periodic structure, see figure 7.24.



(a) The original image.



(b) The result of undersampling.

Figure 7.24: The Moiré effect is directional aliasing.

As was the case with one dimensional signals, low pass filtering, prior to sampling, reduces aliasing artifacts due to undersampling. Most practical measurements involve some

sort of averaging and therefore the measurement process itself incorporates a low pass filter. A shift invariant measurement process is modeled as samples of a convolution $\varphi * f$. In either case the samples are the result of evaluating *some* function. For the remainder of this section we consider the sample set to be the values of a function at the sample points.

Suppose that $\{f_{jk}\}$ are samples of an image. To use these samples to reconstruct an image, the image plane is divided into rectangles. Each rectangle is called a *picture element* of *pixel*. The jk^{th} pixel, denoted p_{jk} is the rectangle

$$p_{jk} = \{(x_1, x_2) \,:\, \frac{j}{M} \leq x_1 < \frac{j+1}{M}, \quad \frac{k}{N} \leq x_2 < \frac{k+1}{N}\}.$$

A simple way to use the samples to define an image is to set

$$\tilde{f}(x_1, x_2) = \sum_{j=0}^{M-1} \sum_{k=0}^{N-1} f_{jk} \chi_{p_{jk}}(x_1, x_2). \tag{7.95}$$

In this image, each pixel has a constant grey level. More sophisticated methods, using interpolation, can be used to get a smoother looking result.

Coordinate transformations

Filters which employ changes of coordinate are difficult to implement on sampled data. Suppose that $\Phi = (g(x_1, x_2), h(x_1, x_2))$ is map defined on a domain D' whose image is contained in the unit square. As above, let $\{(\frac{j}{M}, \frac{k}{N})\}$ denote the sample points in $[0,1]^2$ and $\{(\frac{l}{M'}, \frac{m}{N'})\}$ uniformly spaced sample points in D'. If f is defined everywhere in $[0,1]^2$ then there is no difficulty sampling the values of $\mathcal{A}_\Phi f$ obtaining the samples

$$(\mathcal{A}_\Phi f)_{lm} = \{f(\Phi(\frac{l}{M'}, \frac{m}{N'})\}$$

If f has already been sampled, then a problem immediately presents itself: the images of sample points in D', $\{\Phi(\frac{l}{M'}, \frac{m}{N'})\}$ are unlikely to belong to the set $\{(\frac{j}{M}, \frac{k}{N})\}$. Several strategies can be used to approximate $\mathcal{A}_\Phi$ on sampled data. The easiest approach is to use the piecewise constant image function, $\tilde{f}$ defined in (7.95), setting

$$(\mathcal{A}_\Phi f)_{lm} \stackrel{d}{=} \tilde{f}(\Phi(\frac{l}{M'}, \frac{m}{N'})).$$

Another, more exact approach is to use some sort of "nearest neighbor interpolation." This is a three step process:

- For each $(\frac{l}{M'}, \frac{m}{N'})$ in D' find the sample points in $[0,1]^2$ "nearest" to $\Phi(\frac{l}{M'}, \frac{m}{N'})$. Denote these points $\{\mathbf{x}_1^{lm}, \ldots, \mathbf{x}_{j_{lm}}^{lm}\}$.

- Assign weights $\{w_q^{lm}\}$ to these point according to some measure of their distance to $\Phi(\frac{l}{M'}, \frac{m}{N'})$. The weights are usually non-negative and add up to 1.

- Define $(\mathcal{A}_\Phi f)_{lm}$ as a weighted average:

$$(\mathcal{A}_\Phi f)_{lm} = \sum_{q=1}^{j_{lm}} w_q^{lm} f(\mathbf{x}_q^{lm}).$$

A definition of "nearest neighbors" and a scheme for assigning weights is needed to implement this procedure.

Exercise 7.6.5. Explain how the first procedure used to define $(\mathcal{A}_\Phi f)_{lm}$ can be interpreted as an interpolation scheme.

Linear, shift invariant filters

Suppose that $a(x_1, x_2)$ is the impulse response of a linear shift invariant, two dimensional filter $\mathcal{A}$. Its action in the continuous domain is given by

$$\mathcal{A} f(x_1, x_2) = \iint_{\mathbb{R}^2} a(x_1 - y_1, x_2 - y_1) f(y_1, y_2) dy_1 dy_2.$$

A Riemann sum approximation for this filter, at sample points is

$$\mathcal{A} f(\frac{j}{M}, \frac{k}{N}) \approx \mathcal{A}_s f_{jk} \stackrel{d}{=} \sum_{l=0}^{M-1} \sum_{m=0}^{N-1} a_{(j-l)(k-m)} f_{lm} \frac{1}{MN}, \tag{7.96}$$

where $a_{jk} = a(\frac{j}{M}, \frac{k}{N})$. After zero padding the samples $\{f_{kj}\}$ by setting

$$f_{jk} = 0 \text{ if } M \leq j \leq 2M-1 \text{ or } N \leq k \leq 2N-1,$$

this discrete convolution can be computed using the finite Fourier transform. If $\hat{a}_{jk}$ and $\hat{f}_{jk}$ are the $(2N-1) \times (2M-1)$-finite Fourier transforms of $\{a_{jk}\}$ and $\{f_{jk}\}$ respectively then

$$(\mathcal{A}_s f)_{jk} = \frac{(2M-1)(2N-1)}{MN} [\mathcal{F}^{-1}(\hat{a} \cdot \hat{f})]_{jk},$$

here $(\hat{a} \cdot \hat{f})_{jk} = \hat{a}_{jk} \hat{f}_{jk}$. This gives an efficient way to compute if $\mathcal{A}_s f$ *provided* that a_{jk} is non-zero for most pairs (j, k).

In image processing, many operations on sampled data are naturally defined by sums like those on the right hand side (7.96) with only a small number of non-zero coefficients. Highly localized operations of this sort are directly implemented using this sum. A filter with this property is called a *local operation*. In this context the array of coefficients (a_{jk}) is called the *filter mask*. A local operation has a *small* filter mask. The mask can be represented as a (finite) matrix showing the non-zero elements, with a specification of the index corresponding to the "center" of the matrix.

Example 7.6.2. A standard method to reduce uniform noise in a uniformly sampled image is to average the "nearest neighbors." A simple example of such a filter is

$$(S^a f)_{jk} = \frac{1}{1+4a}[f_{jk} + a(f_{(j-1)k} + f_{(j+1)k} + f_{j(k-1)} + f_{j(k+1)})].$$

Its mask is nicely expressed as a matrix, showing only the non-zero elements, with the center of the matrix corresponding to $(0,0)$:

$$s^a_{jk} = \frac{1}{1+4a}\begin{pmatrix} 0 & a & 0 \\ a & 1 & a \\ 0 & a & 0 \end{pmatrix}.$$

Example 7.6.3. Partial derivatives can be approximated by finite differences. The x_1-partial derivative has three different, finite difference approximations

Forward Difference: $\partial_{x_1} f(\frac{j}{M}, \frac{k}{N}) \approx M[f(\frac{j+1}{M}, \frac{k}{N}) - f(\frac{j}{M}, \frac{k}{N})] \stackrel{d}{=} D_1^f f(j,k),$

Backward Difference: $\partial_{x_1} f(\frac{j}{M}, \frac{k}{N}) \approx M[f(\frac{j}{M}, \frac{k}{N}) - f(\frac{j-1}{M}, \frac{k}{N})] \stackrel{d}{=} D_1^b f(j,k),$

Symmetric Difference: $\partial_{x_1} f(\frac{j}{M}, \frac{k}{N}) \approx 2M[f(\frac{j+1}{M}, \frac{k}{N}) - f(\frac{j-1}{M}, \frac{k}{N})] \stackrel{d}{=} D_1^s f(j,k)$

Let $d^f_{jk}, d^b_{jk}, d^s_{jk}$ denote the corresponding filter masks. Each has only two non-zero entries in the $k=0$ row, as 1×3 matrices they are

$$\begin{aligned} d^f_{j0} &= (M, -M, 0), \\ d^b_{j0} &= (0, M, -M), \\ d^s_{j0} &= (-\frac{M}{2}, 0, \frac{-M}{2}). \end{aligned} \tag{7.97}$$

In each, the center entry corresponds to $(0,0)$. Using any of these approximations requires $O(MN)$ operations to approximate $\partial_{x_1} f$ as compared to $O(MN \log_2 M)$ operations, if M is a power of 2 and the FFT is used. For large M, the finite difference approximation is considerably more efficient.

Example 7.6.4. Suppose that $M = N$. A standard finite difference approximation for the Laplace operator is $\Delta \approx D_1^b \circ D_1^f + D_2^b \circ D_2^f$. Let Δ^s denote this finite difference operator, its filter mask is the 3×3 matrix

$$[\Delta^s]_{jk} = N^2 \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

The center of the matrix corresponds to $(0,0)$.

An important feature of the smoothing and edge detection filters defined in the previous section is that they are isotropic. Even if the sample spacings are equal in all directions, anisotropy is an unavoidable consequence of sampling. The coordinates axes become preferred directions in a uniformly sampled image. The direct implementation of an isotropic, shift invariant filter then leads to an an-isotropic filter. In Proposition 7.4.2 it is shown that a linear, shift invariant filter is isotropic if and only if its transfer function is radial, that is, can be expressed as a function of $\xi_1^2 + \xi_2^2$.

Example 7.6.5. The Laplace operator provides an example of this phenomenon. Let Δ^s denote the finite difference approximation to Δ defined in example 7.6.4. In the Fourier representation

$$(\Delta^s f)_{jk} = (\mathcal{F}_2^{-1}\hat{d}_{mn} \cdot \hat{f}_{mn})_{jk},$$

where $\mathcal{F}_2^{-1}$ is the inverse of the 2-dimensional finite Fourier transform and $\hat{d}_{mn}$ are the Fourier coefficients of Δ_s thought of as an $(N+2) \times (N+2)$-periodic matrix. A simple calculation shows that

$$\begin{aligned} \hat{d}_{mn} &= (2N^2)[\cos\left(\frac{2\pi l}{N+2}\right) + \cos\left(\frac{2\pi m}{N+2}\right) - 2] \\ &= (-4N^2)[\frac{\pi^2(l^2+m^2)}{(N+2)^2} - \frac{2\pi^4 l^4}{3(N+2)^4} - \frac{2\pi^4 m^4}{3(N+2)^4} + O(l^6+m^6)]. \end{aligned} \tag{7.98}$$

The fourth order terms are *not* radial. If the Laplace operator is implemented using the Fourier representation then the transfer function, $-4\pi^2(l^2+m^2)$ is radial though,for large N this requires considerably more computation.

Binary noise

Suppose that f_{jk} with $0 \leq j,k \leq N-1$ represents a discretized image. Due to transmission or discretization errors, for a sparse, "random" subset of indices, the values, $\{f_{j_1k_1}, \ldots, f_{j_mk_m}\}$ are dramatically wrong. This kind of noise is called binary noise; as it is not, in any sense, of "mean zero," it is not attenuated by low pass filtering. A different approach is needed to correct such errors. *Rank value* filtering is such a method. These are filters which compare the values that f_{jk} assumes at neighboring pixels.

A simple example is called the *median filter*. Fix a pair of indices jk. A pixel p_{lm} is a *neighbor* of p_{jk} if $\max\{|l-j|, |m-k|\} \leq 1$. With this definition of neighbor, each pixel has 9 neighbors, including itself. The values of f_{lm} for the neighboring pixels are listed in increasing order. The fifth number in this list is called the median value, denote it by m_{jk}. The median filter, $\mathcal{M}f$ replaces f_{jk} by m_{jk}. This removes wild oscillations from the image and otherwise produces little change in the local variability of f_{jk}. Other schemes replace f_{jk} by the maximum or the minimum value. There are many variants of this idea involving various sorts of averages to define $(\mathcal{M}f)_{jk}$. Any of these operations could be modified to leave the value of f_{jk} unchanged unless it differs dramatically, in the context of the neighboring pixels, from the median.

This concludes our very brief introduction to image processing. Complete treatments of this subject can be found in [42] or [32].

Exercise 7.6.6. Show that the median filter is shift invariant.

Exercise 7.6.7. Show that the median filter is non-linear.

Exercise 7.6.8. Suppose that the image $\{f_{jk}\}$ contains a sharp edge so $f_{jk} = 1$ for values on one side and 0 for values on the other side. A point is on the edge if it has neighbor whose value differs by ± 1. Considering the different possible orientations for an edge, determine the effect of the median filter at points on an edge.

7.7 General linear filters*

See: A.2.3.

Most of this chapter is devoted to the analysis of shift invariant, linear filters. Many actual filtering operations are not shift invariant, though for computational reasons, are approximated by such filters. The theory of shift invariant filters is very simple because, in the Fourier representation, a shift invariant filter becomes a multiplication filter. This has a variety of theoretical and practical consequences. Primary among them are the facts that (1) a pair of shift invariant filters commute, (2) an approximate inverse filter (if it exists) is easily constructed, (3) because the (approximate) Fourier transform and its inverse have efficient implementations so does any shift invariant filter. The main points of the discussion in section 7.5 are summarized as follows: A shift invariant filter $\mathcal{A}$ is approximated on a finite sample set consisting of N samples by a linear transformation of the form

$$\mathcal{A}\,x \approx \mathcal{F}_{2N}^{-1}\Lambda_a\mathcal{F}_{2N}(x_1,\ldots,x_N).$$

Here $\mathcal{F}_{2N}$ is the finite Fourier transform on sequences of length $2N$ and Λ_a is a *diagonal* matrix computed using samples of the transfer function of $\mathcal{A}$. Once the number of samples is fixed, *all* shift invariant, linear filters can be diagonalized by the the same change of basis. Furthermore, if N is a power of 2 then this particular change of basis is very economical to compute. This describes in a nutshell why shift invariant filters are so special and so important in applications.

A general linear filter usually does not have a simple representation as a multiplication filter. This means that the approximate implementation of such a filter on a sample sequence of length N requires $O(N^2)$ operations (as compared to $O(N\log_2 N)$ in the shift invariant case), see section 7.1.5. Two general linear filters rarely commute and it is almost always quite difficult to find an approximate formula for the inverse filter. Nonetheless many classes of general linear invariant filters have been analyzed in detail. In the mathematics literature these classes are called *pseudodifferential operators* and *Fourier integral operators*. Even for these classes, it is quite rare to have an efficient way to implement the filter or a good approximate inverse. The emphasis in this work is on the behavior of filters in the high frequency limit. From a mathematical standpoint this is quite important, however, as we have seen, the world of measurement and sampling is closer to the low frequency limit. The theory of pseudodifferential operators is beyond the scope of this text, an elementary treatment can be found in [76]; more thorough treatments are given in [29] and [75].

Suppose that $\mathcal{A}$ is a general linear filter, with kernel function $a(t,s)$, acting on functions of a single variable. The usual way to approximately implement $\mathcal{A}$ on sampled data is to find a Riemann sum approximation for the integral

$$\mathcal{A}\,x(t) = \int_{-\infty}^{\infty} a(t,s)x(s)ds.$$

If $\{t_1, \ldots, t_N\}$ are the sample points then

$$\mathcal{A}\, x(t_j) \approx \sum_{k=1}^{N-1} a(t_j, t_k) x(t_k)(t_{k+1} - t_k).$$

If the kernel function is a generalized function then it must first be approximated by an ordinary function before it is sampled. This calculation can only be done efficiently (faster than $O(N^2)$) if the kernel function has special structure. An approximate inverse can be found by solving the system of linear equations

$$\sum_{k=1}^{N} a(t_j, t_k) x(t_k)(t_{k+1} - t_k) = y(t_j).$$

If a finite matrix arises from approximating a linear transformation of a function space then, at least for large N, the properties of the approximating transformations mirror those of the infinite dimensional operator. The theory of numerically solving systems of linear equations is a very highly developed field in applied mathematics, see [19] or [78].

Exercise 7.7.1. Let $\{a_0(t), \ldots, a_n(t)\}$ be non-constant functions on $\mathbb{R}$. The differential operator

$$Df = \sum_{j=0}^{n} a_j(t) \frac{d^j f}{dt^j}$$

is a non-shift invariant filter. Compare implementations of this filter on sampled data obtained using the discrete Fourier transform and finite difference operators to approximate the derivatives. Which is more efficient.

Exercise 7.7.2. If $a \neq 0$ and $b \neq 1$ are real constants then the filter $f \mapsto Gf$ defined on functions on $\mathbb{R}$ by the conditions

$$\begin{aligned} &\frac{d^2(Gf)}{dt^2} + a\frac{d(Gf)}{dt} + b(Gf) = f, \\ &\lim_{t \to \pm\infty} (Gf)(t) = 0 \end{aligned} \tag{7.99}$$

is shift invariant, see example 3.2.6. Replacing the boundary condition with $G_0 f(0) = \partial_t G_0 f(0) = 0$ defines a non-shift invariant filter. Find an efficient way to implement G_0.

7.8 Linear filter analysis of imaging hardware*

A very important question in the design of an imaging device is its "resolution." There are two rather different limitations on the resolution, the first derives from the physical limitations of real measuring devices and the second arises from the sampling and processing done to reconstruct an image. From the point of view of signal processing the precise result is Nyquist's theorem which relates the sample spacing to the bandwidth of the sampled data. In imaging applications the data is spatially limited so it cannot be bandlimited. We therefore introduced the concept of effective bandlimiting as the frequency band where

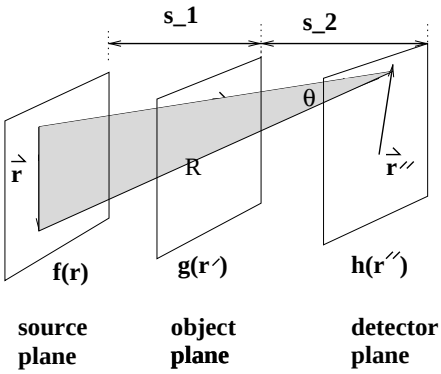


Figure 7.25: Arrangement of an imaging device with a source, object and detector.

"most" of the energy in the data lies. It is clear that, even with perfect measurements, one cannot expect to "resolve" objects that are smaller than the sample spacing. In this section we use linear filtering theory to analyze the distortion of the data that results from using real, physical measuring devices. These are limitations which are present in the measurements before any attempt is made to process the data and reconstruct the image. In this section we use geometric optics to model X-rays as a diverging flux of particles in much the same spirit as in section 2.2. This discussion is adapted from [4] where the reader can find, *inter alia* a careful discussion of γ-ray detectors, X-ray sources and collimators.

7.8.1 The transfer function of the scanner

In this section we consider a three dimensional situation, beginning with the simple setup in figure 7.25. A source of radiation is lying in the *source plane* with a distribution $f(\mathbf{r})$, i.e. $f(\mathbf{r})d\mathbf{r}$ is the number of photons per unit time emitted by an area on the source of size $d\mathbf{r}$ located at position $\mathbf{r}$. The source output is assumed to be independent of time.

In a *parallel* plane, at distance s_1 from the source plane is an object which is described by a transmittance function $g(\mathbf{r}')$. The fraction of the incident photons transmitted by an area $d\mathbf{r}'$ located at the point $\mathbf{r}'$ in the object plane is given by $g(\mathbf{r}')d\mathbf{r}'$. Usually $g(\mathbf{r}')$ takes values between 0 and 1. It is sometimes useful to think of g as the probability that a photon incident at $\mathbf{r}'$ will be transmitted. The object plane is usually thought of as a thin slice of a three dimensional object. If the width of the slice is ϵ then transmittance is related to absorption coefficient in Beer's law by

$$g(\mathbf{r}') = 1 - \exp[-\int_0^\epsilon \mu ds].$$

Here μ is an absorption coefficient and the integral is along the line perpendicular to object plane through the point $\mathbf{r}'$. Finally a detector lies in a second *parallel* plane, at distance s_2 from the object plane. For the moment assume that the detector is perfect, i.e. everything incident on the detector is measured. Later on a more realistic detector will be incorporated into the model. To analyze the source-object-detector geometry, first assume that the object is transparent, that is $g(\mathbf{r}') = 1$.

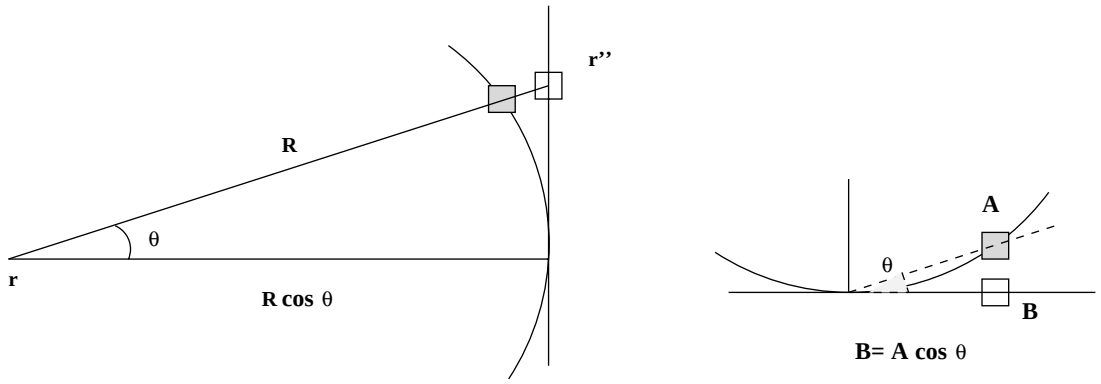


Figure 7.26: Computing the solid angle.

It is convenient to use a different systems of coordinates in each plane, $\mathbf{r}, \mathbf{r}', \mathbf{r}''$. As above $d\mathbf{r}, d\mathbf{r}'$ and $d\mathbf{r}''$ denote the corresponding area elements. A point source is isotropic if the flux through an area A on the sphere of radius ρ, centered on the source, is proportional to the ratio of A to the area of the whole sphere $4\pi\rho^2$. The constant of proportionality is the intensity of the source. The *solid angle* Ω subtended by a region D, relative to a point p is defined by projecting the region onto the sphere of radius 1 centered at p, along lines through the center of the sphere. If D' is the projected image of D then the solid angle it subtends is defined to be

$$\Omega = \text{area of the region } D'.$$

From the definition it is clear that the solid angle assumes values between 0 and 4π.

To find the flux through a region, D due to a planar distribution of isotropic sources with density $f(\mathbf{r})$ we need to find the contribution of each infinitesimal area element $d\mathbf{r}$. If $\Omega(D, \mathbf{r})$ is the solid angle subtended at $\mathbf{r}$ by D then the contribution of the area element centered at $\mathbf{r}$ to the flux through D is $f(\mathbf{r})d\mathbf{r}\Omega(D, \mathbf{r})/4\pi$. In our apparatus we are measuring the area element on a plane parallel to the source plane. As shown in figure 7.26, the solid angle subtended at a point $\mathbf{r}$ in the source plane by an infinitesimal area $d\mathbf{r}''$ at $\mathbf{r}''$ in the detector plane is

$$\frac{\cos\theta d\mathbf{r}''}{R^2}, \text{ where } R = \frac{s_1 + s_2}{\cos\theta}.$$

Therefore the infinitesimal solid angle is

$$d\Omega = \frac{\cos^3\theta}{(s_1 + s_2)^2} d\mathbf{r}''.$$

If no absorbing material is present, a detector at $\mathbf{r}''$ of planar area $d\mathbf{r}''$ absorbs $d\Omega/4\pi$ of the emitted radiation. If the intensity of the source at $\mathbf{r}$ is $f(\mathbf{r})d\mathbf{r}$ then we get

$$f(\mathbf{r})d\mathbf{r}\frac{d\Omega}{4\pi} = f(\mathbf{r})\frac{\cos^3\theta}{4\pi(s_1 + s_2)^2} d\mathbf{r}d\mathbf{r}''$$

as the measured flux. Notice that θ is a function of $\mathbf{r}$ and $\mathbf{r}''$.

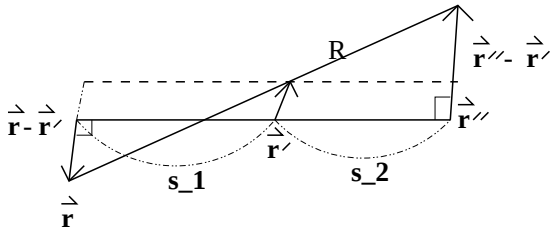


Figure 7.27: The similar triangle calculation.

Now we include the effect of an absorbing material. The measured flux at $\mathbf{r}''$ is an integral along the source plane given by

$$h(\mathbf{r}'') = \frac{1}{4\pi(s_1+s_2)^2}\int_{\text{source}} \cos^3\theta(\mathbf{r},\mathbf{r}'')f(\mathbf{r})g(\mathbf{r}'(\mathbf{r},\mathbf{r}''))d\mathbf{r}. \tag{7.100}$$

Here $h(\mathbf{r}'')$ is the measured photon flux density at the detector point $\mathbf{r}''$. We see that $\mathbf{r}'$ is a function of $\mathbf{r}$ and $\mathbf{r}''$. To obtain this relation, we can think of the 2-dimensional plane containing the three vectors $\mathbf{r}, \mathbf{r}'$ and $\mathbf{r}''$ as in figure 7.27. Since we have similar triangles, we see that

$$\frac{\mathbf{r}'-\mathbf{r}}{s_1} = \frac{\mathbf{r}''-\mathbf{r}'}{s_2}.$$

Or,

$$\mathbf{r}' = \frac{s_2}{s_1+s_2}\mathbf{r} + \frac{s_1}{s_1+s_2}\mathbf{r}'' = a\mathbf{r}'' + b\mathbf{r}$$

where

$$a = \frac{s_1}{s_1+s_2}, \quad b = \frac{s_2}{s_1+s_2} = 1-a.$$

Now equation (7.100) reads

$$h(\mathbf{r}'') = \frac{1}{4\pi(s_1+s_2)^2}\int \cos^3\theta f(\mathbf{r})g(a\mathbf{r}''+b\mathbf{r})d\mathbf{r}. \tag{7.101}$$

In a transmission imaging problem the source function $f(\mathbf{r})$ is assumed to be known. Relation (7.101) states that the output along the detector plane is a linear filter applied to f. It is more or less a convolution of the known source function $f(\mathbf{r})$ with the unknown transmittance function $g(\mathbf{r}')$. In a transmission imaging problem we are trying to determine $g(\mathbf{r}')$. As it stands, formula (7.101) is not quite in the form of a convolution for two reasons: (1) $\theta(\mathbf{r},\mathbf{r}'')$ is not a function of $\mathbf{r}-\mathbf{r}''$, (2)there are two coordinates systems, one in the source and one in the detector plane. Letting

$$\mathbf{r}_0'' = -\frac{b\mathbf{r}}{a},$$

we express everything in the coordinates of the detector plane. Let $\tilde{f}$ be the scaled source function and $\tilde{g}$ be the scaled transmission function given by:

$$\hat{f}(\mathbf{r}_0'') = f(-a\mathbf{r}_0''/b), \quad \tilde{g}(\mathbf{r}_0'') = g(a\mathbf{r}_0'')$$

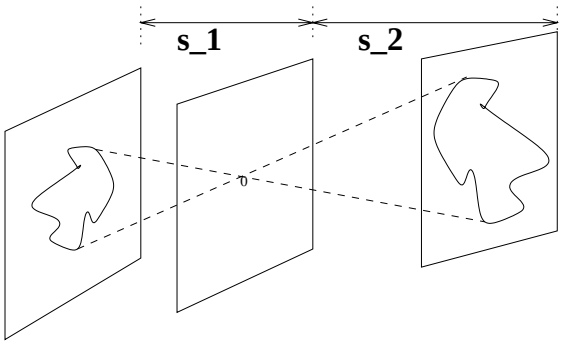


Figure 7.28: A pinhole camera.

Now, the measurement is expressed as

$$h(\mathbf{r}'') = \big(\frac{a}{b}\big)^2 \frac{1}{4\pi(s_1+s_2)^2} \int \tilde{f}(\mathbf{r}_0'')\tilde{g}(\mathbf{r}''-\mathbf{r}_0'')\cos^3\theta d\mathbf{r}_0''.$$

But for the $\cos^3\theta$-term this is a convolution. In many applications, the angle θ is close to 0 and therefore the cosine term can be approximated by 1. With this approximation, a shift invariant linear system relates the source and transmittance to the measured output,

$$h(\mathbf{r}'') = \big(\frac{a}{b}\big)^2 \frac{1}{4\pi(s_1+s_2)^2} \int \tilde{f}(\mathbf{r}_0'')\tilde{g}(\mathbf{r}''-\mathbf{r}_0'')d\mathbf{r}_0''. \tag{7.102}$$

This formula is not only useful for the analysis of transmission imaging, but is also useful in nuclear medicine. Instead of a known X-ray source we imagine that $f(\mathbf{r})$ describes an unknown distribution of radioactive sources. In this context we could use a pinhole camera to form an image of the source distribution. Mathematically this means that g is taken to equal 1 in a tiny disk and zero everywhere else. Let us consider this situation geometrically: only lines drawn from source points to detector points which pass through the support of g contribute to the image formed in the detector plane. If the support of g is a single point then the image formed would be a copy of f scaled by the ratio s_1/s_2. This is because each point $\mathbf{r}''$ in the detector plane is joined to a unique point $\mathbf{r}(\mathbf{r}'')$ in the source plane by a line passing through the support of g. Note however that the intensity of the image at a point $\mathbf{r}''$ is proportional to $|\mathbf{r}''-\mathbf{r}(\mathbf{r}'')|^{-2}$, see figure 7.28.

Now suppose that the support of g is the disk $B_0(d)$ for a very small d. The image formed at a point $\mathbf{r}''$ in the detector plane is the result of averaging the source intensities over the disk in the source plane visible from the point $\mathbf{r}''$ through the pinhole, $B_0(d)$. Of course the actual image is a weighted average. The result of a positive diameter pinhole is a blurred image.

Exercise 7.8.1. Find a more accurate model for the "output" of a pinhole camera by letting $g(\mathbf{r}') = \delta(\mathbf{r}')$ in (7.101).

Exercise 7.8.2. Find a formula for the output of a pinhole camera with

$$g(\mathbf{r}') = \chi_D(\mathbf{r}')$$

using the simplified formula (7.102).

7.8.2 The resolution of an imaging system

We now estimate the resolution of the transmission imaging system described in section 7.8.1 using the FWHM criterion. We consider the problem of resolving two spots in the object plane. The source is assumed to have constant intensity over a disk of diameter $d_{\rm fs}$; in this case f is the characteristic function of a disk $f(\mathbf{r}) = \chi_{D_1}(2|\mathbf{r}|/d_{\rm fs})$. In this section D_1 denotes the disk of radius 1 centered at $(0,0)$. The source is projected onto the detector plane, passing it through an object composed of two identical opaque spots separated by some distance. Mathematically it is equivalent to think of the object as being entirely opaque but for two disks of perfect transmittance separated by the same distance. It is clear that the outputs of the two configurations differ by a constant. For the calculation the object is modeled as the sum of two δ-functions,

$$g(\mathbf{r}') = \delta(\mathbf{r}' - \mathbf{r}'_1) + \delta(\mathbf{r}' - \mathbf{r}'_2).$$

Of course this is not a function taking values between 0 and 1. This situation is approximated by a transmittance given by

$$g_\epsilon = \chi_{D_1}(\frac{|\mathbf{r}' - \mathbf{r}'_1|}{\epsilon}) + \chi_{D_1}(\frac{|\mathbf{r}' - \mathbf{r}'_2|}{\epsilon}),$$

letting ϵ tend to zero **and** rescaling the output by ϵ^{-2} so that the total measured intensity is constant. The limiting case is mathematically equivalent to using the sum of δ-functions as the transmittance.

The image of a single tiny spot on the object plane located at $\mathbf{r}'_1$ is given by formula (7.102)

$$\begin{aligned} h(\mathbf{r}'', \mathbf{r}'_1) &= \frac{1}{4\pi(s_1+s_2)^2}\int \chi_{D_1}\left(\frac{2r}{d_{\rm fs}}\right)\delta(a\mathbf{r}'' + b\mathbf{r} - \mathbf{r}'_1)d\mathbf{r} \\ &= \frac{1}{4\pi(s_1+s_2)^2}\frac{1}{b^2}\chi_{D_1}\left(\frac{2|a\mathbf{r}'' - \mathbf{r}'_1|}{bd_{\rm fs}}\right) \\ &= \frac{1}{4\pi(s_1+s_2)^2}\frac{1}{b^2}\chi_{D_1}\left(\frac{2|\mathbf{r}'' - (\mathbf{r}'_1/a)|}{bd_{\rm fs}/a}\right) \end{aligned}$$

The image is a disk centered at $\mathbf{r}'_1/a$ with radius $bd_{\rm fs}/a$. This is the impulse response of the source-detector pair. The FWHM for the point spread function of this system is just the diameter of the image disk,

$$d''_{\rm fs} = \frac{b}{a}d_{\rm fs}.$$

According to this definition, two points sources in the object plane are resolvable if their images are completely non-overlapping, which might appear to be overly stringent.

We would actually like to know the minimum separation in the *object plane* for two point sources to be resolvable. This distance is found by projecting to the object plane,

$$\delta'_{\rm fs} = bd_{\rm fs} = \frac{s_2}{s_1+s_2}d_{\rm fs}.$$

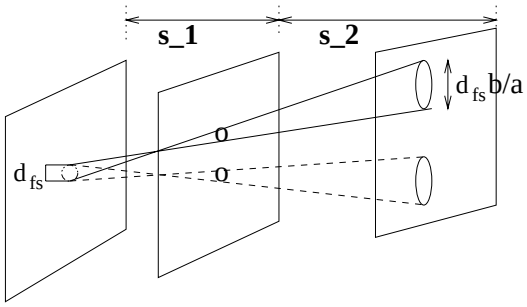


Figure 7.29: The image of two dots.

If $s_2 = 0$ then $\delta'_{\rm fs} = 0$. If the object is sitting on the detector then no matter how close the two point sources are, they can be distinguished. As $s_1 \to 0$, $\delta'_{\rm fs} \to d_{\rm fs}$, so the resolution is better for objects closer to the detector.

This simple model is completed by including a detector response function. Suppose that $k(\mathbf{r}'')$ is the impulse response of the detector. The complete imaging system with source f, transmittance g and detector response k is given by

$$h(\mathbf{r}'') = \left(\frac{a}{b}\right)^2 \frac{1}{4\pi(s_1+s_2)^2} \int\int \tilde{f}(\mathbf{r}''_0)\tilde{g}(\mathbf{r}''_1 - \mathbf{r}''_0)k(\mathbf{r}'' - \mathbf{r}''_1)\cos^3\theta d\mathbf{r}''_0 d\mathbf{r}''_1.$$

This models what is called an "imaging detector" such as photographic film or an array of scintillation counters and photo-multiplier tubes. Setting $\mathbf{r}'' = 0$ gives a model for a "counting detector." Such a detector only records the number of incident photons, making no attempt to record the location on the detector where the photon arrives. In this case $k(\mathbf{r}'')$ is the distribution function for the probability that the detector responds to a photon arriving at $\mathbf{r}''$.

We now consider the sensitivity of the detector. The bigger the area on the detector subtended by the image, the more photons we count. By taking $s_2 = 0$, the object is in contact with the detector. This gives the best resolution but, for a pair of infinitesimal objects, no measurable data. There is a trade off between the *resolution* of the image and the *sensitivity* of the detector. A larger detector captures more photons but gives less information about where they came from. In general, it is a very difficult problem to say which configuration is optimal. We consider the question of optimizing the placement of the object when both the source and detector are Gaussian. This means that

$$\hat{f} \text{ is proportional to } \exp(-\pi\left(\frac{|\xi|}{\rho_f}\right)^2),$$

$$\hat{d} \text{ is proportional to } \exp(-\pi\left(\frac{|\xi|}{\rho''_d}\right)^2),$$

Such a source is said to have a *Gaussian focal spot.*

When referred back to the object plane, we get the following modulation transfer function

$$\text{MTF} = \exp\left[-\pi|\xi|^2\left(\frac{1}{(\lambda\rho''_d)^2} + \frac{(\lambda-1)^2}{\lambda^2\rho_f^2}\right)\right]. \tag{7.103}$$

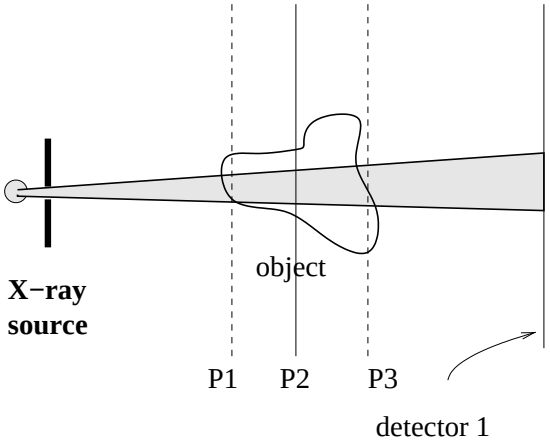


Figure 7.30: Beam spreading.

The parameter $\lambda = \frac{s_1+s_2}{s_1}$ is the magnification factor. In this context the optimal configuration is the one which distorts the object the least, this is the one for which the Gaussian in (7.103) decays as slowly as possible. For this family of functions this corresponds to

$$\frac{1}{(\lambda\rho_d'')^2} + \frac{(\lambda-1)^2}{\lambda^2\rho_f^2}$$

assuming its minimum value. Differentiating and setting the derivative equal to zero, we find that

$$\lambda^{opt} = 1 + \left(\rho_f/\rho_d''\right)^2.$$

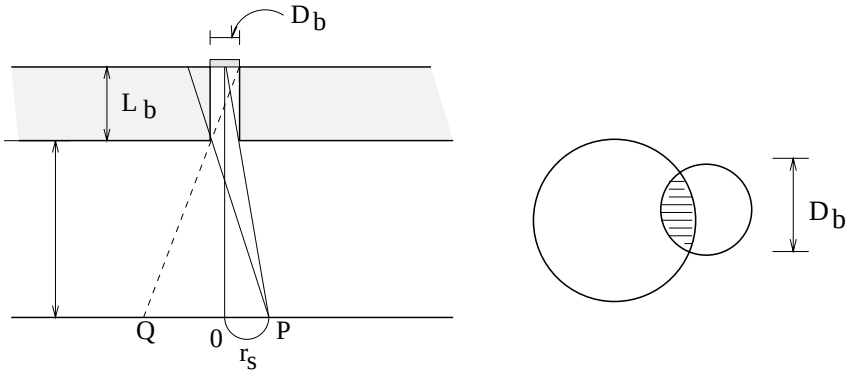
Note that a large spot corresponds to $\rho_f \to 0$ and a poor detector corresponds to $\rho_d'' \to 0$.

7.8.3 Collimators

In the previous section, we discussed a simple geometry for a source, object and detector. The object was simplified to be 2 dimensional. In medical imaging applications one can think of the object as being made of many slices. The analysis presented in the previous section indicates that the spreading of the X-ray beam causes distortion in the projected image which depends upon how far the object is from the source. As illustrated in Figure 7.30, the images of points in the plane P1 are spread out more than the images of points in P3. A diverging beam magnifies objects more in the slices closer to the source than those in slices further away. One could try reducing this effect by making the distance from the source to the detector much larger than the size of object. This has the undesirable effect of greatly attenuating the X-ray beam incident on the detector.

To control beam spreading distortion, we can reject X-rays arriving from certain directions by using a *collimator*. Collimators are an essential part of most imaging apparatus. In simple terms, a collimator is a cylindrical (or conical) hole bored through X-ray absorbing material. There are two physical parameters describing a collimator:

D_b: the diameter of the hole, L_b: the height of sides,



(a) (b)

Figure 7.31: The geometry of a collimator

called the *bore diameter* and *bore length*, see figure 7.31. Often collimators are used in arrays. In this case, each collimator is better modeled as narrow pipe made of X-ray absorbing material. In our discussion it is assumed that only photons which pass through the collimator bore reach the detector.

This is also a three dimensional analysis. The collimator is circularly symmetric about its axis, which implies that the response to a point source depends only on the distance, z from the point source to the front of the collimator and the distance, r_s from the source to the axis of the collimator.

The impulse response of a collimator can be deduced from the the analysis in section 7.8.1. The effect of the collimator on a X-ray beam is identical to the effect of two concentric pinholes of diameter D_b, cut into perfectly absorbing plates, lying in parallel planes a distance L_b apart. Let $f(\mathbf{r})$ model the X-ray source, $g_1(\mathbf{r}')$ the lower pinhole and $g_2(\mathbf{r}'')$ the upper pinhole. From section 7.8.1, we have that the photon flux incident on the upper plate is

$$h(\mathbf{r}'') = \frac{1}{4\pi(L_b+z)^2}\int f(r)g_1(a\mathbf{r}''+b\mathbf{r})d\mathbf{r}$$

where

$$a = \frac{z}{L_d+z}, \quad b = \frac{L_b}{L_b+z}.$$

The flux incident on the detector (through the upper pinhole) is therefore

$$\int h(\mathbf{r}'')g_2(\mathbf{r}'')d\mathbf{r}'' = \frac{1}{4\pi(L_b+z)^2}\int\int f(\mathbf{r})g_1(a\mathbf{r}''+b\mathbf{r})g_2(\mathbf{r}'')d\mathbf{r}''d\mathbf{r}.$$

It is assumed that the angle θ the beam makes with the detector plane is approximately 0 and therefore $\cos\theta \approx 1$.

Now suppose that the source is a point source, P located a distance r_s from the axis of the collimator i.e., $f(\mathbf{r}) = \delta(\mathbf{r}-\mathbf{r}_s)$ and g_1 is the characteristic function of a disk:

$$g_1(\mathbf{r}') = \chi_{D_1}\left(\frac{2|\mathbf{r}'|}{D_b}\right).$$

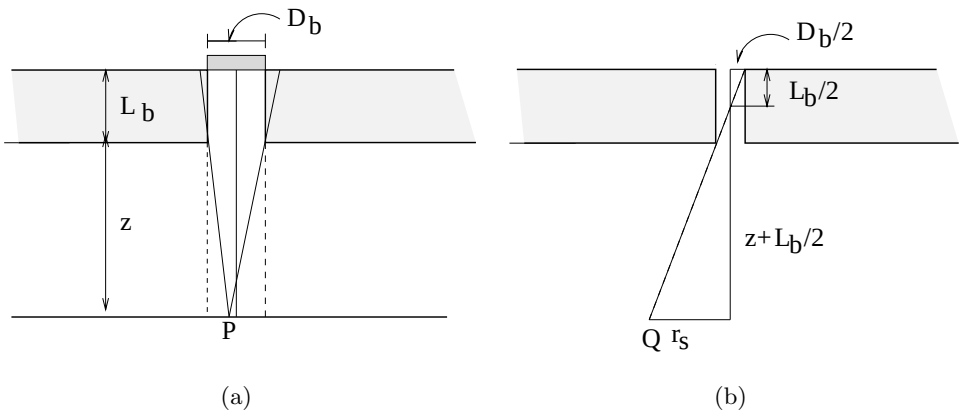


Figure 7.32: Evaluating the point spread function of a collimator.

For the sort of collimator considered here, g_2 is the same:

$$g_2(\mathbf{r}'') = \chi_{D_1}\left(\frac{2|\mathbf{r}''|}{D_b}\right).$$

The output of our detector is

$$p(\mathbf{r}_s; z) = \frac{1}{4\pi(L_b+z)^2}\int \chi_{D_1}\left(\frac{2|\mathbf{r}''|}{D_b}\right)\chi_{D_1}\left(\frac{2a|\mathbf{r}''+b\mathbf{r}_s/a|}{D_b}\right)d\mathbf{r}''.$$

This integral is exactly the area of the intersection of the two disks, as shown in Figure 7.31(b). The larger circle is the projection of the lower circle from the source point onto the upper plate. The $1/z^2$ scaling accounts for beam spreading.

In two special regions, $p(\mathbf{r}_s; z)$ is easy to determine. First, if $|\mathbf{r}_s| < \frac{1}{2}D_b$, as in figure 7.32(a) then the projected circle covers the upper circle, hence $p(\mathbf{r}_s; z)$ is the area of the smaller disk, that is

$$p(\mathbf{r}_s; z) = \pi\left(\frac{D_b}{2}\right)^2.$$

On the other hand, if P lies far enough from the origin then the two circles do not meet. The location of the first such point, Q is found using similar triangles:

$$\frac{L_b}{2} : \frac{D_b}{2} = \frac{L_b}{2} + z : \mathbf{r}_s \Rightarrow \mathbf{r}_s = \frac{D_b}{2}\left(\frac{L_b+2z}{L_b}\right),$$

see figure 7.32(b). Hence if $|\mathbf{r}_s| > \frac{D_b}{2}\left(\frac{L_b+2z}{L_b}\right)$, the area of the intersection is zero and $p(\mathbf{r}_s; z) = 0$. Figure (7.33) show the graph of $p(|\mathbf{r}_s|; z)$ for a fixed value of z.

To approximate the FWHM, we approximate p by a piecewise linear function. Let $\delta(z)$ be the resolution, defined to be

$$\delta(z) = \mathrm{FWHM}(p(\cdot; z)) = D_b\frac{(L_b+2z)}{L_b}.$$

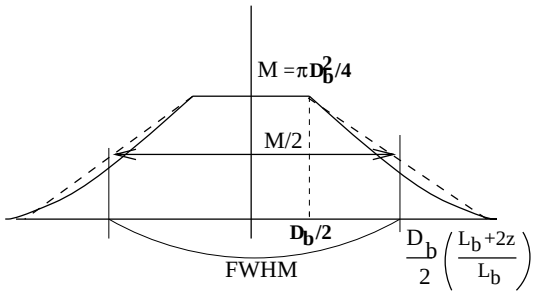


Figure 7.33: The graph of $p(\mathbf{r}_s; z)$, for a fixed z.

As z increases, the resolution gets worse, as observed in section 7.8.2. If the object is sitting right on the collimator face, i.e. $z = 0$ then $\delta(0) = D_b$ which is just the diameter of the collimator bore.

The other parameter of interest for a detector is its sensitivity. The sensitivity of a detector is important as it determines how intense a source is needed to make a usable image. To analyze the sensitivity we use a uniform planar source, instead of the point source used in the resolution analysis. Imagine that a uniform radioactive source is spread on a plane at distance z from the face of the collimator. The intensity of the "standard source" is $1\mu C/\,\mathrm{cm}^2$. For a typical radioactive material, this translates into a photon flux of $3.7 \times 10^4/\,\mathrm{cm}^2\,\mathrm{sec}$. The sensitivity of the measuring device is given by the number of photons arriving at the detector. For the collimator described above we obtain

$$\begin{aligned} S &= \frac{3.7 \times 10^4}{4\pi(L_b + z)^2} \iint \chi_{B_0}\left(\frac{2\mathbf{r}''}{D_b}\right) \chi_{B_0}\left(\frac{2|a\mathbf{r}'' + b\mathbf{r}|}{D_b}\right) d\mathbf{r} d\mathbf{r}'' \\ &= \frac{3.7 \times 10^4}{4\pi(L_b + z)^2} \frac{\pi D_b^2}{4b^2} \frac{\pi D_b^2}{4} \\ &= \frac{3.7 \times 10^4}{64} \frac{\pi D_b^4}{L_b^2} \end{aligned}$$

The sensitivity does not depend on z, this is consequence of using a uniform source of *infinite* extent.

With L_b and z fixed, the resolution is proportional to D_b, while the sensitivity is proportional to D_b^4. Thus we see that

$$S \text{ is proportional to } \delta^4.$$

To form a usable image the required dosage is roughly proportional to $1/S$. This shows that

$$\textit{dosage} \text{ is proportional to } \delta^{-4}.$$

To increase the resolution by a factor of 2, i.e., $\delta \to \frac{1}{2}\delta$, the dosage must be increased by a factor of 16.

Exercise 7.8.3. Find the transfer function for a collimator with a cylindrical bore. This can be modeled as two concentric, transparent disks of radii D_{b1} and D_{b2} lying in parallel planes a distance L apart.

Exercise 7.8.4. What is the effect of the collimator in the previous exercise if $D_{b1} < D_{b2}$? How about if $D_{b1} > D_{b2}$?

Chapter 8

Reconstruction in X-ray tomography

See: A.1.

At long last we are returning to the problem of reconstructing images in X-ray tomography. Recall that if f is a function defined on $\mathbb{R}^2$ which is bounded and has bounded support then its Radon transform, $\mathrm{R}f$ is a function on the space of oriented lines. The oriented lines in $\mathbb{R}^2$ are parametrized by pairs $(t, \omega) \in \mathbb{R} \times S^1$, with

$$l_{t,\omega} \leftrightarrow \{(x_1, x_2) \,:\, \langle (x_1, x_2), \omega \rangle = t\}.$$

The positive direction along $l_{t,\omega}$ is defined by the unit vector $\hat{\omega}$ orthogonal to ω with $\det(\omega\, \hat{\omega}) = +1$. The Radon transform of f is given by the line integrals:

$$\mathrm{R}f(t, \omega) = \int_{-\infty}^{\infty} f(t\omega + s\hat{\omega}) ds.$$

Exact inversion formulæ for the Radon transform are derived in Chapter 4. These formulæ assume that $\mathrm{R}f$ is known for *all* lines. In a real X-ray tomography machine $\mathrm{R}f$ is *approximately* sampled at a finite set of points. The goal is to construct a discrete image which is, to the extent possible, samples or averages of the original function f. In this chapter we see how the Radon inversion formulæ lead to methods for approximately reconstructing f from realistic measurements. We call such a method a *reconstruction algorithm.*

Before deriving and analyzing the reconstruction algorithms we review the setup in medical imaging and fix the notation for the remainder of the chapter. Beer's law is the basic principle underlying X-ray tomography. To an object D, in $\mathbb{R}^3$ there is associated an *absorption coefficient* $\mu(\mathbf{x})$. This is a non-negative function which describes the probability that an X-ray photon of a given energy which encounters the object at the point $\mathbf{x}$ is absorbed. Beer's law is phrased as a differential equation describing the change in the intensity of a (1-dimensional) "beam," composed of many photons, traveling along a line

L in $\mathbb{R}^3$. If $\Omega \in S^2$ is the direction of L and $\mathbf{x}_0$ is a point on L then the line is given parametrically by

$$L = \{s\Omega + \mathbf{x}_0 \ : \ s \in \mathbb{R}\}.$$

Let $I(s)$ denote the intensity of the photon beam at the point $s\Omega + \mathbf{x}_0$, Beer's law states that

$$\frac{dI}{ds} = -\mu(s\Omega + \mathbf{x}_0)I.$$

If D is a bounded object then $L \cap D$ is contained in an interval of parameter values: $s \in [a, b]$. In the simplest model for X-ray tomography the incoming intensity $I(a)$ is known and the outgoing intensity $I(b)$ is measured, integrating Beer's law gives the relation

$$\log\left[\frac{I(a)}{I(b)}\right] = \int_a^b \mu(s\Omega + \mathbf{x}_0)ds. \tag{8.1}$$

Tomography refers to a particular way of organizing this data. A coordinate system (x_1, x_2, x_3) for $\mathbb{R}^3$ is fixed. The data collected are the Radon transforms of the 2-dimensional "slices" of μ, in the x_3-direction. These are the functions

$$f_c(x_1, x_2) = \mu(x_1, x_2, c),$$

obtained by fixing the last coordinate. In the formulation above this corresponds to taking

$$\begin{aligned} \mathbf{x}_0 =&(t\omega, c) \text{ and } \Omega = (\hat{\omega}, 0) \text{ where } t, c \in \mathbb{R} \text{ and} \\ &\omega, \hat{\omega} \in S^1 = \{(x_1, x_2, 0) \ : \ x_1^2 + x_2^2 = 1\}. \end{aligned} \tag{8.2}$$

For a fixed c, the integrals in (8.1) are nothing but the Radon transform of f_c. With these measurements the function f_c can therefore be reconstructed.

A real X-ray CT machine only collects a finite number of projections. In a simple model for the actual measurements there is a finite set of values, $\{c_1, \ldots, c_n\}$ such that the Radon transforms of the functions, $\{f_{c_1}, \ldots, f_{c_n}\}$ are sampled along a finite set of lines $\{l_{t_j,\omega_j} \ : \ j = 1, \ldots, P\}$. The design of the machine determines which projections are measured. This chapter considers algorithms for reconstructing a single two dimensional slice. The problem of using the slices to re-assemble a 3-dimensional image is not treated in any detail. Note however the very important fact that the CT-machine itself determines a frame of reference, fixing, for example, the "x_3-direction." Once the machine is calibrated, the positions in space which correspond to the different slices and the projections, with-in a slice are known in advance. As we shall see, the design of the machine also singles out a particular coordinate system in the space of lines in $\mathbb{R}^2$.

The X-ray beam is not one dimensional but three dimensional. Let C denote the cross section of the beam at right angles to its direction. The cross section is often approximately rectangular

$$C \approx [-a, a] \times [-d, d],$$

where the second factor lies in the x_3-direction. The width of the second factor $2d$ is called the *slice thickness*; this parameter is usually adjustable when the measurements are made. The beam intensity also varies continuously within C falling off to zero at the edge.

As a practical matter, a larger cross section increases the energy in the beam which, in turn, improves the signal-to-noise ratio in the measurements. On the other hand, poorer spatial resolution is also a consequence of a larger cross section. In our initial discussion of the reconstruction algorithms we model the measurements as line integrals of a slice, $f_c(x_1, x_2) = \mu(x_1, x_2, c)$. In other words we assume that the X-ray beam *is* one dimensional. A linear model for the effect of a three dimensional beam is to replace these line integrals by weighted averages of such integrals. As usual, averaging with an integrable weight is a form of low pass filtering. This is very important because it reduces the effects of aliasing which result from sampling. These effects are easily analyzed using properties of the Radon transform and are considered in sections 8.5- 8.6. For the bulk of this chapter the third dimension is rarely mentioned explicitly.

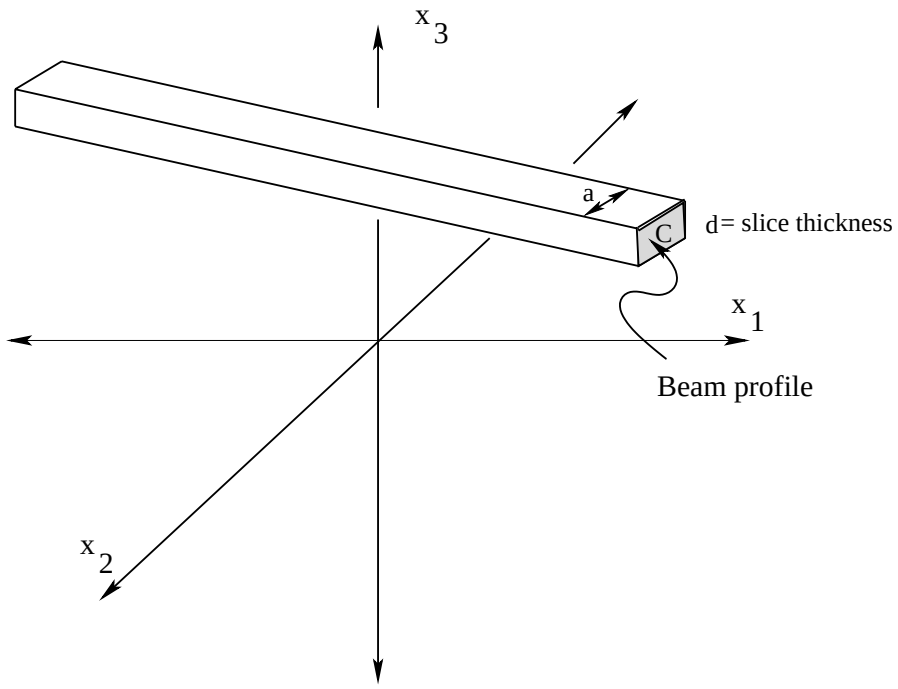


Figure 8.1: A 3-dimensional X-ray beam.

Let (x, y) denote Cartesian coordinates in the slice $x_3 = c$, which is heretofore fixed. The two dimensional object we would like to image lies in D_L, the disk of radius L, centered at $(0, 0)$ in this plane. In most of this chapter the X-ray source is assumed to be monochromatic of energy $\mathcal{E}$ and the object is described by its X-ray absorption coefficient f at this energy. As the object lies in D_L, f is assumed to vanish outside this set. Our goal is to use samples of $\mathrm{R}f$ to approximately determine the values of f on a uniform *reconstruction* grid,

$$\mathcal{R}_\tau = \{(x_j, y_k) = (j\tau, k\tau) \ : \ j, k \in \mathbb{Z}\},$$

in the (x, y)-plane. Here $\tau > 0$ denotes the sample spacing. It is selected to reflect the resolution available in the data. The reconstruction grid can also be thought of as dividing the plane into a grid of squares of side τ. Each square is called a *pixel*. As each slice is

really a three dimensional slab, and hence each square is really a cuboid, the elements in the reconstruction grid are often called *voxels*. The value reconstructed at a point $(x_j, y_k) \in \mathcal{R}_\tau$ should be thought of as a weighted average of f over the voxel containing this point. Of course f is only reconstructed at points of $\mathcal{R}_\tau$ lying in $[-L, L] \times [-L, L] \supset D_L$. We assume that f is bounded, and regular enough for its Radon transform to be sampled. As the actual measurements involve averaging μ with a continuous function, this is not a restrictive assumption.

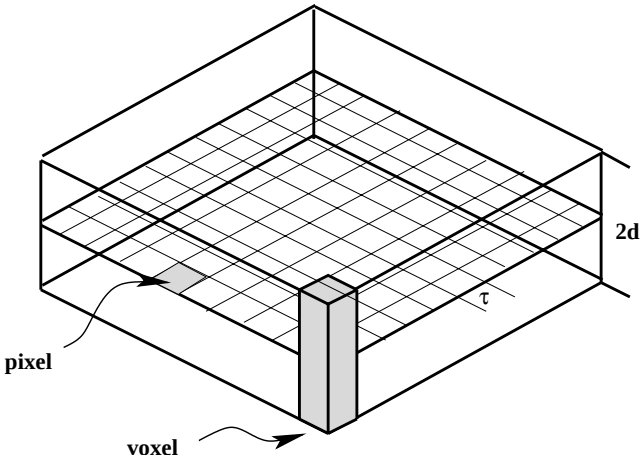


Figure 8.2: The reconstruction grid.

In most of this section it is assumed that measurements are made with infinite precision and that the full set of real numbers are available. Real measurements have errors and are quantized to fit into a finite number of bits. The accuracy and precision of the measurements are determined by the accuracy and sensitivity of the detectors *and* the stability of the X-ray source. The number of bits used in the quantization of the measurements is the ultimate mathematical constraint on the *contrast* available in the reconstructed image. This numerical resolution reflects the accuracy of the measurements themselves and the computations used to process them. It should not be confused with *spatial* resolution which is a function of the sample spacing and the point spread function of the reconstruction algorithm. A detailed discussion of X-ray sources and detectors can be found in [86] or [4].

8.1 Reconstruction formulæ

We now review the inversion formulæ derived earlier and consider how they might be approximated in a real computation. These formulæ are consequences of the Fourier inversion formula and central slice theorem, relating $\mathrm{R}f$ to $\hat{f}$,

$$\widetilde{\mathrm{R}f}(r, \omega) = \int_{-\infty}^{\infty} \mathrm{R}f(t, \omega) e^{-irt} dt = \hat{f}(r\omega).$$

Writing the Fourier inversion formula in polar coordinates gives

$$\begin{aligned} f(x,y) &= \frac{1}{[2\pi]^2}\int_0^{2\pi}\int_0^{\infty} \hat{f}(r\omega)e^{ir\langle(x,y),\omega\rangle} r\,dr\,d\omega \\ &= \frac{1}{[2\pi]^2}\int_0^{\pi}\int_{-\infty}^{\infty} \hat{f}(r\omega)e^{ir\langle(x,y),\omega\rangle} |r|\,dr\,d\omega. \end{aligned} \tag{8.3}$$

Note that the polar variable r is allowed to assume negative values with the understanding that if $r < 0$ then

$$r\omega = |r|(-\omega).$$

Using the central slice theorem gives

$$f(x,y) = \frac{1}{[2\pi]^2}\int_0^{\pi}\int_{-\infty}^{\infty} \widetilde{\mathrm{R}f}(r,\omega)|r|e^{ir\langle(x,y),\omega\rangle}\,dr\,d\omega. \tag{8.4}$$

Directly approximating this integral is one approach to reconstructing f.

The r integral is often interpreted as a linear, shift invariant filter acting in the t-variable,

$$\begin{aligned} \mathcal{G}(\mathrm{R}f)(t,\omega) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{\mathrm{R}f}(r,\omega)|r|e^{irt}\,dr \\ &= -i\mathcal{H}\partial_t\,\mathrm{R}f(t,\omega), \end{aligned} \tag{8.5}$$

leading to the filtered backprojection formula

$$f(x,y) = \frac{1}{2\pi}\int_0^{\pi} \mathcal{G}(\mathrm{R}f)(\langle(x,y),\omega\rangle,\omega)\,d\omega. \tag{8.6}$$

Of course these formulæ are entirely equivalent from the mathematical point of view. However, approximating mathematically equivalent formulæ can lead to very different algorithms. As above, τ is the uniform spacing in the reconstruction grid. The object is assumed to lie within the disk of radius L. The number of grid points on the x or y-axis is therefore $2K+1$ where

$$K = \left[\frac{L}{\tau}\right] + 1.$$

Our immediate goal is to approximately reconstruct the set of values

$$\{f(j\tau,k\tau)\ :\ -K \le j,k \le K\}.$$

In order to distinguish the reconstructed values from the actual samples, the reconstructed values are denoted by $\{\tilde{f}(j\tau,k\tau)\ :\ -K \le j,k \le K\}$. Of course f is known to be zero on grid points lying outside D_L. As remarked above real measurements have an interpretation as weighted averages, over the width of the X-ray beam, of values of $\mathrm{R}f$. In our initial

derivations of the reconstruction algorithms we overlook this point, the effects of the finite beam width are incorporated afterwards.

A good reconstruction algorithm is characterized by *accuracy*, *stability* and *efficiency*. An accurate algorithm is often found by starting with an exact inversion formula and making judicious approximations. In CT, stability is the result of low pass filtering and the continuity properties of the exact inversion formula. Whether an algorithm can be implemented efficiently depends on its overall structure as well as the hardware available to do the work. An X-ray CT-machine is *not* a general purpose computer, but rather a highly specialized machine designed to do one type of computation quickly and accurately. Such a machine may have many processors, allowing certain parts of the algorithm to be "parallelized." As we shall soon see, the measuring hardware naturally divides the measurements into a collection of *views*. The most efficient algorithms allow the data from a view to be processed as soon as it is collected.

8.2 Scanner geometries

The structure of a reconstruction algorithm is dictated by which samples of $\mathrm{R}f$ are available. Before discussing algorithms we therefore need to consider what kind of measurements are actually made. In broad terms there are two types of CT-machine:

(a) Parallel beam scanner, see figure 8.3(a).

(b) Divergent beam scanner, see figure 8.4.

The earliest scanner was a parallel beam scanner. This case is considered first because the geometry and algorithms are simpler to describe. Because the data can be collected much faster, most modern machines are divergent or *fan beam* scanners. Algorithms for these machines are a bit more involved and are treated later.

In a parallel beam scanner approximate samples of $\mathrm{R}f$ are measured in a finite set of directions,

$$\{\omega(k\Delta\theta) \text{ for } k = 0\ldots, M\}$$

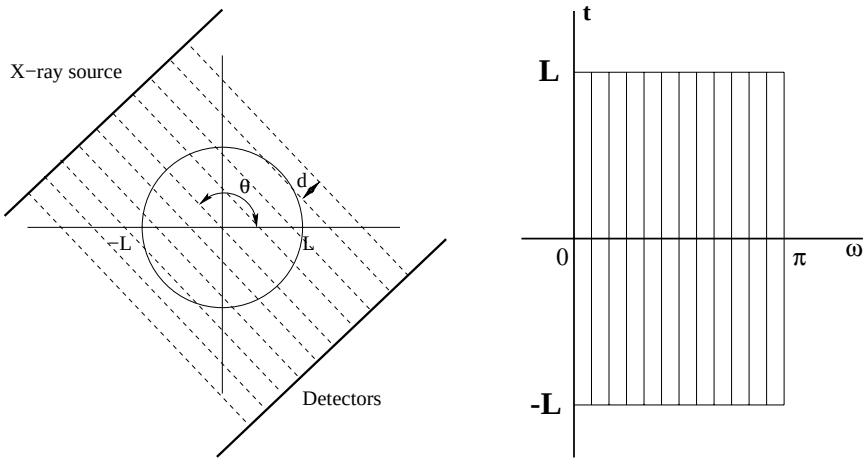
where

$$\Delta\theta = \frac{\pi}{M+1} \text{ and } \omega(k\Delta\theta) = (\cos(k\Delta\theta), \sin(k\Delta\theta)).$$

In terms of the angular variable, θ the samples are equally spaced. The measurements made in a given direction are then samples of $\mathrm{R}f$ at a set of equally spaced affine parameters

$$\{jd \,:\, j = -N, \ldots, N\},$$

here d is the sample spacing in the affine parameter and $N = Ld^{-1}$.



(a) A parallel beam scanner.

(b) The parallel beam sample space.

Figure 8.3: A parallel beam scanner and sample set.

Parallel beam data therefore consists of the samples

$$\{\mathrm{R}f(jd, \omega(k\Delta\theta)), \quad j = -N, \ldots, N, \quad k = 0, \ldots, M\}. \tag{8.7}$$

Because of the symmetry of the Radon transform,

$$\mathrm{R}f(-t, -\omega) = \mathrm{R}f(t, \omega), \tag{8.8}$$

measurements are only required for angles lying in $[0, \pi)$. Sometimes it is useful to make measurements over a full 360°, the extra measurements can then be averaged as a way to reduce the effects of noise and systematic errors.

The individual measurements are often called *rays*. A *view*, for a parallel beam machine, consists of the measurements of $\mathrm{R}f(t, \omega)$ for a fixed ω. These are the integrals of f along the collection of equally spaced parallel lines

$$\{l_{jd,\omega(k\Delta\theta)}, \quad j = -N, \ldots, N\}.$$

In (t, θ)-space, parallel beam data consists of equally spaced samples on the vertical lines shown in figure 8.3(b).

The other type of scanner in common use is called a *divergent beam* or *fan beam scanner*. A point source of X-rays is moved around a circle centered on the object being measured. The source is pulsed at a discrete sequence of angles, measurements of $\mathrm{R}f$ are collected for a finite family of lines passing through the source. In a machine of this type, data is collected by detectors which are usually placed on a circular arc. There are two different designs for fan beam machines which, for some purposes, need to be distinguished. In a *third generation* machine the detectors are placed on a circular arc centered on the source. The detectors and the source are rotated together. In a *fourth generation* machine the

detectors are on a fixed ring, centered on the object. Only the source is rotated, again on a circle centered on the object, within the ring of detectors. These designs are shown schematically in figure 8.4.

For a fan beam machine machine it is useful to single out the *central ray*. For a third generation machine this is the line that passes through the source and the center of rotation. The central ray is well defined no matter where the source is positioned. Since all the rays that are sampled pass through a source position, the natural angular parameter, for this geometry is the angle, ϕ between a given ray and the central ray. Suppose that the source is at distance R and the central ray makes an angle ψ with the positive x-axis. The affine parameter for the line passing through the source, making an angle ϕ with the central ray, is given by

$$t = R\sin(\phi), \tag{8.9}$$

see figure 8.5(a).

In a third generation machine the source is placed at a finite number of equally spaced angles

$$\psi_k \in \{\frac{2\pi k}{M+1} \ : \ k = 0, \ldots, M\},$$

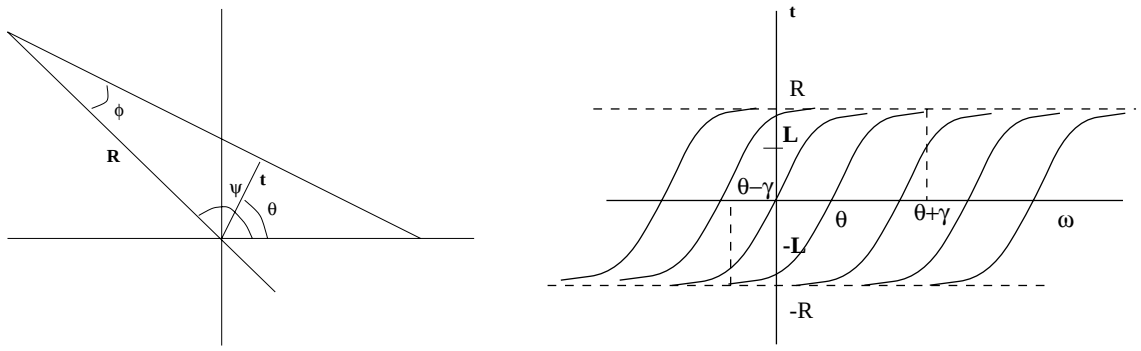
and data is collected along a set of lines through the source, equally spaced in the ϕ-parameter,

$$\phi_j \in \{\frac{j\pi}{N} \ : \ j = -P, \ldots, P\}.$$

Third generation, fan-beam data is the set of samples

$$\{\mathrm{R}f(\left(\frac{j\pi}{N} + \frac{2\pi k}{M+1} - \frac{\pi}{2}\right), R\sin(\frac{j\pi}{N})) \mid j = -P, \ldots, P, \quad k = 0, \ldots, M\}.$$

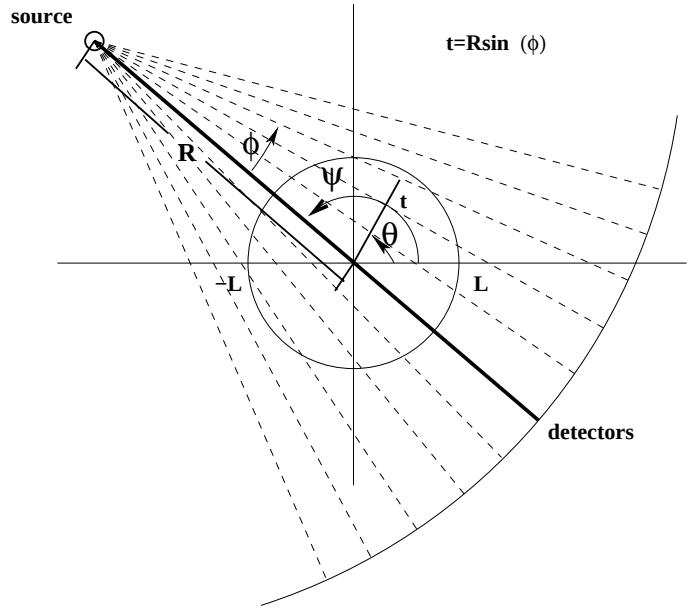
The maximum and minimum values, $\pm\frac{P\pi}{N}$ for the angle ϕ, are determined by the necessity of sampling all lines which meet an object lying in D_L, as shown in figure 8.4(a). The samples lie on the sine curves shown in figure 8.5(b). A *view*, for a third generation machine, is the set of samples from the rays passing through a given *source* position.



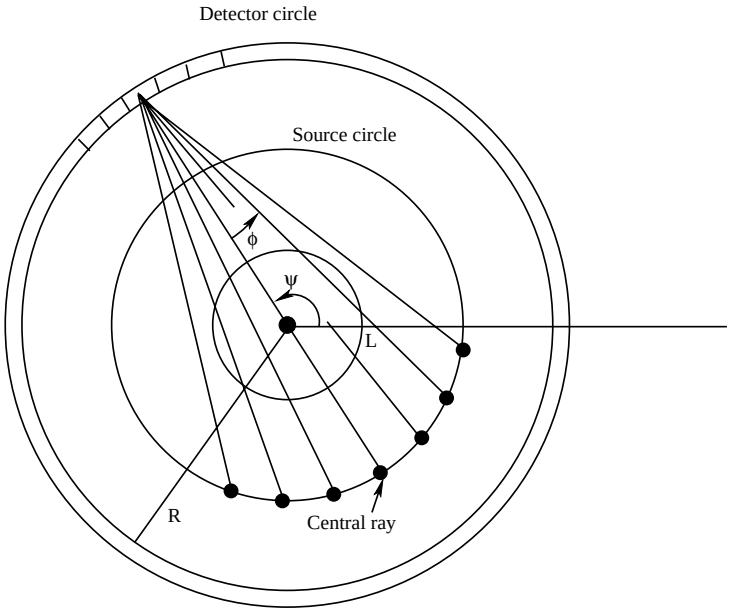
(a) Angular parameters for a fan beam machine.

(b) Sample set for a fan beam machine.

Figure 8.5: Parameters for a fan beam machine.



(a) A third generation scanner.



(b) A fourth generation scanner.

Figure 8.4: The two different divergent beam geometries.

For a fourth generation machine the source moves and the detectors remain fixed. Each time the source emits a pulse of X-rays, a sample, can in principle be collected by any detector which can "see" the source. In practice, data is collected for rays which pass through the target area, as shown in figure 8.4(b). In this way the data is grouped into views according to the detector position **not** the source position. For many purposes third and fourth generation machines can be treated together. Indeed we can still use the coordinates in figure 8.5(a) to parametrize the data, with the following small changes: the detectors now lie on the circle of radius R and ψ denotes the angle from the positive x-axis to the "central ray," which is now a line joining $(0, 0)$ to a fixed *detector*. The parameter ϕ measures the angle between lines through a detector and its central ray. With this understood, the sample set and data collected are essentially the same for third and fourth generation machines. A view for a fourth generation machine consists of all the rays passing through a given *detector*. At a basic level, third and fourth generation scanners are quite similar, though they differ in many subtle points. These differences are considered in greater detail in section 8.6.4.

In the engineering and medical literature one sometimes sees the raw measurements represented as density plots in (t, ω) or (ψ, ϕ)-space. Such a diagram is called a *sinogram*. These are difficult for a human observer to directly interpret, however as they contain *all* the information available in the data set, they may be preferable for machine based assessments.

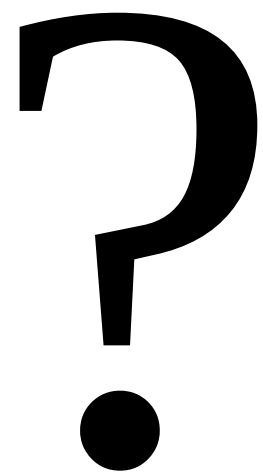


Figure 8.6: An example of a sinogram.

Exercise 8.2.1. For a third generation, fan beam machine, the set of angles $\{\phi_k\}$ could be selected so that the sample spacing in the affine parameter is constant. To get a sample spacing d what set of angles $\{\phi_k\}$ should be used?

8.3 Reconstruction algorithms for a parallel beam machine

See: A.7.1.

We now consider reconstruction algorithms for a parallel beam machine. A good starting point in the construction of an algorithm is to assume that we can measure *all* the data from a finite set of equally spaced directions. In this case the data would be

$$\{\mathrm{R}f(t,\omega(k\Delta\theta)),\quad k=0,\ldots,M,\quad t\in[-L,L]\}$$

where $\Delta\theta=\frac{\pi}{M+1}$. With this data we can apply the central slice theorem to compute angular samples of the 2-dimensional Fourier transform of f,

$$\hat{f}(r\omega(k\Delta\theta))=\int_{-\infty}^{\infty}\mathrm{R}f(t,\omega(k\Delta\theta))e^{-irt}dt.$$

Remark 8.3.1 *(***Important notational remark)**. Recall that in this chapter the polar variable r can be either positive or negative; if $r<0$ then

$$\hat{f}(r\omega(\theta))=\hat{f}(|r|\omega(\theta+\pi)). \tag{8.10}$$

8.3.1 Direct Fourier inversion

Formula (8.3) suggests using the 2 dimensional Fourier inversion formula directly, to reconstruct f. Using a Riemann sum in the angular direction gives

$$f(x,y)\approx\frac{1}{4\pi(M+1)}\sum_{k=0}^{M}\int_{-\infty}^{\infty}\hat{f}(r\omega(k\Delta\theta))e^{ir\langle(x,y),\omega(k\Delta\theta)\rangle}|r|dr.$$

Our actual measurements are the samples $\{\mathrm{R}f(jd,\omega(k\Delta\theta))\}$ of $\mathrm{R}f(t,\omega(k\Delta\theta))$. Since the sample spacing in the t-direction is d, the data is effectively bandlimited to $[-\frac{\pi}{d},\frac{\pi}{d}]$. In light of the results in section 7.5.3 these samples can be used to compute approximations to samples of the 2-dimensional Fourier transform of f,

$$\hat{f}(r_j\omega(k\Delta\theta))\quad r_j\in\{0,\pm\eta,\pm2\eta,\ldots,\pm N\eta\}\text{ where }\eta=\frac{1}{N}\frac{\pi}{d}=\frac{\pi}{L}.$$

This is a set of equally spaced samples of $\hat{f}$ in the *polar* coordinate system. Doing the computation directly in polar coordinates would entail $O((2M+2)(N+1))$ operations to compute the approximate inverse of the Fourier transform for a single point (x,y). As the number of points in the reconstruction grid is $O(K^2)$, the number of computations needed to reconstruct the image at every grid point is

$$O(MNK^2).$$

For realistic data this is a very large number. The size of this computation can be vastly reduced by using the fast Fourier transform algorithm.

The two dimensional FFT is only usable if samples of $\hat{f}$ are known on a uniform grid in a *rectangular* coordinate system. Our data $\{\hat{f}(r_j\omega(k\Delta\theta))\}$ are $\hat{f}$ sampled on a uniform grid in a polar coordinate system. To use the 2-dimensional FFT, the data must first be interpolated to get simulated measurements on a uniform, rectangular grid. Using nearest neighbor, linear interpolation, the amount of computation required to do this is a modest $O(K^2)$ calculations. Assuming that K is a power of 2, the FFT leading to $\tilde{f}(x_j, y_k)$ would require $O((K\log_2 K)^2)$ calculations. The full reconstruction of $\tilde{f}$, at the grid points, from the parallel beam data would therefore require $O(K^2(\log_2 K)^2 + 2MN\log_2 N)$ computations. The $2M\log_2 N$ term comes from using the 1-dimensional FFT to compute $< \hat{f}(r_j\omega(k\Delta\theta)) >$ from $< \mathrm{R}f(jd, \omega(k\Delta\theta)) >$. For realistic values of M and N this is a much smaller number than what was computed above for a direct inversion of the Fourier transform. Indeed, this algorithm is the fastest algorithm for implementation on a general purpose computer. However, with real data, simple linear interpolation leads to unacceptably large errors in the approximation of $\hat{f}$. This is due in part to the fact that $\hat{f}$ is complex valued and the extreme sensitivity of the reconstructed image to errors in the phase of the Fourier transform, see 7.1.4. A more sophisticated interpolation scheme is needed. A technique of this sort is presented in section 8.7.

Exercise 8.3.1. Explain how to use zero padding to obtain a approximation to $\hat{f}(r\omega)$ on a finer grid in the r-variable. Is this justified in the present instance? Is there any way to reduce the sample spacing in the angular direction?

8.3.2 Filtered backprojection

Formula (8.6) organizes the approximate inversion in a different way. The Radon transform is first filtered

$$\mathcal{G}\,\mathrm{R}f(t,\omega) = -i\mathcal{H}\partial_t\,\mathrm{R}f(t,\omega),$$

and then backprojected to find f at (x, y). The operation $\mathrm{R}f \mapsto \mathcal{G}\,\mathrm{R}f$ is a 1-dimensional, linear shift invariant filter. On a parallel beam scanner, the data for a given ω defines a single view and is collected with the source-detector array in a fixed position. Once the data from a view has been collected it can be filtered. In this way, a large part of the processing is done by the time all the data for a slice has been collected. Supposing, as before, that sampling only occurs in the angular variable, the data set for a parallel beam scanner would be the samples

$$\{\mathrm{R}f(t, \omega(k\Delta\theta)) \,:\, k = 0, \ldots, M\}.$$

In a filtered backprojection algorithm each view, $\mathrm{R}f(t, \omega(k\Delta\theta))$ is filtered immediately after it is measured, giving $\mathcal{G}\,\mathrm{R}f(t, \omega(k\Delta\theta))$. When all the data has been collected and filtered, the image is approximately reconstructed by using a Riemann sum approximation to the backprojection:

$$\tilde{f}(x,y) = \frac{1}{2(M+1)}\sum_{k=0}^{M} \mathcal{G}\,\mathrm{R}f(\langle (x,y), \omega(k\Delta\theta)\rangle, \omega(k\Delta\theta)); \tag{8.11}$$

here use is made of the symmetry (8.15). Assuming that all the necessary values of $\mathcal{G}\,\mathrm{R}f$ are known, the backprojection step requires $O(MK^2)$ operations to determine $\tilde{f}$ on the reconstruction grid, $\mathcal{R}_\tau$. This step is also *highly* parallelizable: with $O(K^2)$ processors the backprojection could be done simultaneously for all points in $\mathcal{R}_\tau$ in $O(M)$ cycles. The serious work of implementing this algorithm is in deciding how to approximate the filter $\mathcal{G}$ on data which is sampled in both t and ω.

In real applications the approximation to the impulse response of $\mathcal{G}$ is chosen to be the inverse Fourier transform of a function which satisfies certain properties. Denote the approximate impulse response by $\phi(t)$ and define

$$Q_\phi f(t,\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{\mathrm{R}f}(r,\omega)\hat{\phi}(r)e^{irt}dr.$$

In order for Q_ϕ to approximate $\mathcal{G}$, its modulation transfer function $\hat{\phi}(r)$ should provide an "approximation" to $|r|$ over the effective bandwidth of the data. Exactly what is meant by this statement is a subject of considerable discussion. For example, as $|r|$ is an even, real valued function $\hat{\phi}$ should also be. In order to have a stable algorithm and suppress noise, it is important for $\hat{\phi}(r)$ to tend to zero as $|r|$ tends to infinity. In the end, whether or not a given choice of ϕ provides a "good" approximation is a largely empirical question.

Once ϕ is chosen the filtered Radon transform is given by

$$\begin{aligned} Q_\phi f(t,\omega) &= \int_{-\infty}^{\infty} \mathrm{R}f(s,\omega)\phi(t-s)ds \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{\mathrm{R}f}(r,\omega)\hat{\phi}(r)e^{irt}dr. \end{aligned} \tag{8.12}$$

With "complete data" the approximate reconstruction, defined by ϕ, would be

$$f_\phi(x,y) = \frac{1}{2\pi}\int_0^{\pi} Q_\phi f(\langle(x,y),\omega\rangle,\omega)d\omega. \tag{8.13}$$

Approximating this integral, on the sampled data, in either the spatial or frequency representation and using a Riemann sum approximation for the backprojection gives an approximate reconstruction, $\tilde{f}_\phi$ at points $(x_m, y_l) \in \mathcal{R}_\tau$, e.g.

$$\tilde{f}_\phi(x_m,y_l) = \frac{d}{2(M+1)}\sum_{k=0}^{M}\sum_{j=-N}^{N} \mathrm{R}f(jd,\omega(k\Delta\theta))\phi(\langle(x_m,y_l),\omega(k\Delta\theta)\rangle - jd). \tag{8.14}$$

Under constraints like those enumerated above, one tries to choose the function $\phi(t)$ to optimize some aspect of this reconstruction. As always, there are trade-offs between efficiency, resolution and noise reduction. The function ϕ is regarded as a *parameter* which

can be adjusted to achieve certain aims. In the imaging literature, $\hat{\phi}$ is often expressed as a product

$$\hat{\phi}(r) = A(r)|r|.$$

Here $A(r)$ is a function which tends to zero as $|r| \to \infty$; it is called an *apodizing function.*

Remark 8.3.2 (**Important notational remark.**). In the sequel, the notation f_ϕ refers to an approximate reconstruction of f using an integral like that in (8.13) whereas $\tilde{f}_\phi$ refers to a discretization of this integral, as in (8.14).

This family of reconstruction formulæ has a very important feature which follows from the fact that the quantity

$$\langle (x,y), \omega(k\Delta\theta)\rangle - jd$$

is the signed distance from the line $l_{jd,\omega(k\Delta\theta)}$ to the point (x,y) The algorithm defined by ϕ can be expressed as

$$\tilde{f}_\phi(x,y) = \frac{d}{2(M+1)} \sum_{k=0}^{M} \sum_{j=-N}^{N} \mathrm{R}f(jd, \omega(k\Delta\theta))\phi(\mathrm{dist}[(x,y), l_{jd,\omega(k\Delta\theta)}]).$$

This is also a feature of the exact reconstruction formula. We now consider the details of implementing (8.14) on parallel beam data.

Exercise 8.3.2. Show that

$$\mathcal{G}\,\mathrm{R}f(-t,-\omega) = \mathcal{G}\,\mathrm{R}f(t,\omega). \tag{8.15}$$

8.3.3 Ram-Lak filters

From formula (8.14) it would appear that the computation of $\tilde{f}_\phi(x_m, y_l)$, for a point $(x_m, y_l) \in \mathcal{R}_\tau$, requires a knowledge of the values

$$\begin{aligned}\{\phi(\langle (x_m,y_l), \omega(k\Delta\theta)\rangle - jd) \text{ for} \\ -N \le j \le N,\ -K \le l,m \le K \text{ and } 0 \le k \le M\}.\end{aligned} \tag{8.16}$$

The filtering operation could still be done as soon as the data from a view, $\{\mathrm{R}f(jd, \omega(k\Delta\theta)) : -N \le j \le N\}$ is collected. But, from (8.16), it would appear that a *different* filter function is required for each point in $\mathcal{R}_\tau$. In other words the filter step would have to be repeated $O(K^2)$ times for each view! Ramachandran and Lakshminarayanan found a computational shortcut so that the filter operation would only have to be done once for each view. Their idea was fix values of ϕ at the sample points, $\{\phi(jd) : -N \le j \le N\}$ and then *linearly* interpolate $\phi(t)$ to the intermediate values.

We first consider what this means and why it reduces the computational load in the implementation of the filtered backprojection algorithm so dramatically. The function ϕ is defined to be linear between the sample points, for any $\alpha \in [0,1]$

$$\phi(\alpha(k+1)d + (1-\alpha)(kd)) \stackrel{d}{=} \alpha\phi((k+1)d) + (1-\alpha)\phi(kd).$$

Since

$$\alpha(k+1)d + (1-\alpha)(kd) - ld = \alpha(k+1-l)d + (1-\alpha)(k-l)d$$

this implies that

$$\phi(\alpha(k+1)d + (1-\alpha)(kd) - ld) = \alpha\phi((k+1-l)d) + (1-\alpha)\phi((k-l)d). \tag{8.17}$$

For each $(x_m, y_l) \in \mathcal{R}_\tau$ and direction $\omega(k\Delta\theta)$ there is an integer $n_{klm} \in [-N, N]$ such that

$$n_{klm}d < \langle (x_m, y_l), \omega(k\Delta\theta)\rangle \leq (n_{klm}+1)d.$$

Thus there is a real number $\alpha_{klm} \in [0,1]$ so that

$$\langle (x_m, y_l), \omega(k\Delta\theta)\rangle = \alpha_{klm}(n_{klm}+1)d + (1-\alpha_{klm})n_{klm}d.$$

The *trivial* but *crucial* observation is that (8.17) implies that, for any integer l,

$$\langle (x_m, y_l), \omega(k\Delta\theta)\rangle - ld = \alpha_{klm}[(n_{klm}+1)d - ld] + (1-\alpha_{klm})[n_{klm}d - ld]. \tag{8.18}$$

If ϕ is linearly interpolated between sample points then

$$\begin{aligned} Q_\phi\tilde{f}(\langle (x_m, y_l), \omega(k\Delta\theta)\rangle, \omega(k\Delta\theta)) &= \alpha_{klm} Q_\phi\tilde{f}((n_{klm}+1)d, \omega(k\Delta\theta)) \\ &\quad + (1-\alpha_{klm}) Q_\phi\tilde{f}(n_{klm}d, \omega(k\Delta\theta)). \end{aligned} \tag{8.19}$$

In other words $Q_\phi\tilde{f}(\langle (x_m, y_l), \omega(k\Delta\theta)\rangle, \omega(k\Delta\theta))$ is a weighted average of $Q_\phi\tilde{f}$ at *sample points*. In fact the interpolation step can be done as part of the backprojection

$$\begin{aligned} &\tilde{f}_\phi(x_m, y_l) = \\ &\frac{1}{2(M+1)} \sum_{j=0}^{M} \big[\alpha_{klm} Q_\phi\tilde{f}((n_{klm}+1)d, \omega(k\Delta\theta)) + (1-\alpha_{klm}) Q_\phi\tilde{f}(n_{klm}d, \omega(k\Delta\theta))\big]. \end{aligned} \tag{8.20}$$

As a practical matter, the sampling angles and reconstruction grid are essentially fixed and therefore the coefficients, $\{\alpha_{klm}, n_{klm}\}$ can be evaluated once and stored in tables.

The filtered backprojection algorithm using a *Ram-Lak* filter is the following sequence of steps:

(1) Approximately compute the filtered Radon transform, $Q_\phi f(jd, \omega((k\Delta\theta))$ at sample points. The Riemann sum approximation is

$$Q_\phi\tilde{f}(jd, \omega(k\Delta\theta)) = d \sum_{l=-\infty}^{\infty} \mathrm{R}(ld, \omega(k\Delta\theta))\phi((j-l)d).$$

(2) Backproject, with linearly interpolated values for $Q_\phi\tilde{f}(\langle (x_m, y_l), \omega(k\Delta\theta)\rangle, \omega(k\Delta\theta))$, to determine the values of $\tilde{f}_\phi(x_m, y_l)$ for $(x_m, y_l) \in \mathcal{R}_\tau$.

Step (1) can be done a view at a time, it requires a knowledge of $\phi(jd)$ for $-(2N-1) \leq j \leq (2N-1)$. The calculation of $Q_\phi f(jd, \omega)$ is a discrete convolution which is usually computed using the FFT.

As noted above, the coefficients used in the backprojection step can be computed in advance and this step can also be highly parallelized. The interpolation used in the filtered

backprojection is interpolation of a real valued function and is much less damaging than that used in the direct Fourier reconstruction. Empirically it leads to an overall blurring of the image but does not introduce complicated oscillatory artifacts or noise. Approximately MK^2 calculations are required once the values $\{Q_\phi \tilde{f}(jd, \omega(k\Delta\theta))\}$ are computed. Using an FFT the computation of Q_ϕ requires about $MN \log_2 N$-steps and using a direct convolution about MN^2-steps. The FFT is clearly faster but the backprojection step already requires a comparable number of calculations to that needed for a direct convolution.

Exercise 8.3.3. How should formula (8.14) be modified if sampling is done around the full circle, i.e. θ varies between 0 and 2π?

8.3.4 Shepp-Logan analysis of the Ram-Lak filters

See: B.7.

Ramachandran and Lakshminarayanan introduced the linear interpolation method described in the previous section as a way to reduce the computational burden of the filtered backprojection algorithm. Initially it was assumed that using a simple linear interpolation to define $\phi(t)$ would result in a significant loss in accuracy. However that turned out not to be the case. Shepp and Logan explained the surprisingly high quality of the Ram-Lak reconstructions by analyzing the following example:

$$\phi(0) = \frac{4}{\pi d^2}, \quad \phi(kd) = \frac{-4}{\pi d^2(4k^2 - 1)}, \tag{8.21}$$

and ϕ is linear otherwise. This function has a tall narrow peak at zero and a long *negative* tail. The figure shows ϕ with $d = .28$ along with the impulse response for an approximation to $\mathcal{G}$ obtained in section 4.4.

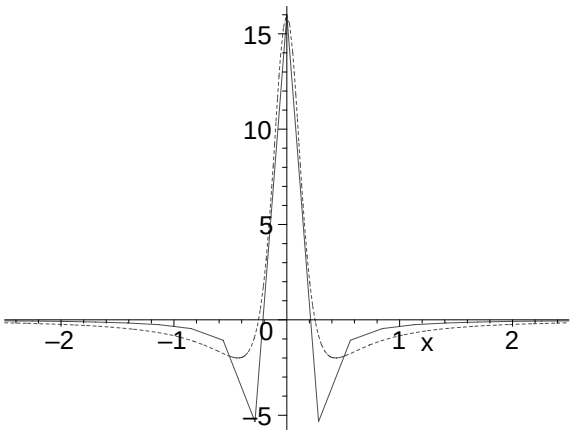


Figure 8.7: The impulse response for a RamLak filter (solid) and a continuous approximation (dotted).

With this choice

$$\hat{\phi}(\xi) = \left|\frac{2}{d}\sin\frac{d\xi}{2}\right|\left[\frac{\sin(d\xi/2)}{d\xi/2}\right]^2. \tag{8.22}$$

From the Taylor series of $\sin x$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots$$

we obtain that

$$\begin{aligned}\hat{\phi}(\xi) &= \frac{2}{d}\left|\frac{d\xi}{2} + O(\frac{d\xi}{2})^3\right|\left[\frac{d\xi/2 + O(d\xi/2)^3)}{d\xi/2}\right]^2 \\ &= |\xi|(1 + O(d^2\xi^3))[1 + O(\frac{d\xi}{2})^2]^2 \\ &= |\xi|(1 + O(d^2\xi^3 + |d\xi|^2)).\end{aligned}$$

By choosing d small enough, $\hat{\phi}(\xi)$ can be made quite close to $|\xi|$ over any desired interval. This approximation also has the desirable feature that $\phi(t)$ decays at the optimal rate for a function whose Fourier transform has a $|\xi|$-type singularity at $\xi = 0$:

$$\phi(t) = O(\frac{1}{t^2})$$

To see that this is the optimal rate suppose that $\hat{\phi}(\xi)$ has the same singularity as $|\xi|$ at $|\xi| = 0$ and is otherwise smooth with absolutely integrable derivatives. The inverse Fourier transform, $\phi(t)$ is computed by twice integrating by parts:

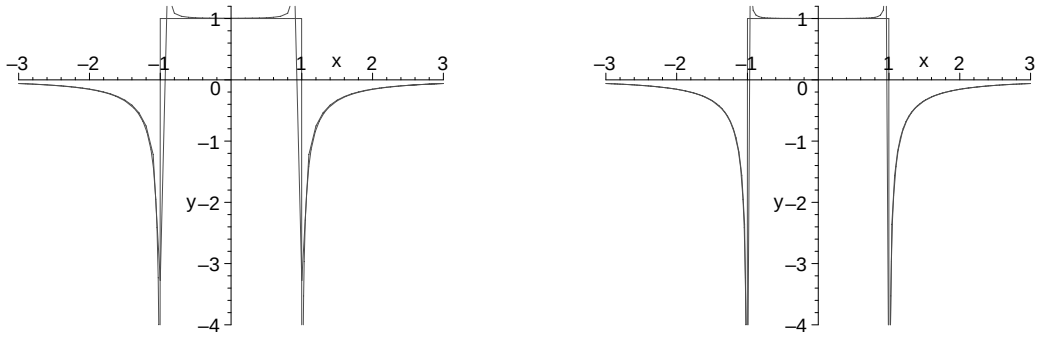
$$\begin{aligned}\int_{-\infty}^{\infty}\hat{\phi}(\xi)e^{it\xi}d\xi &= \left[\int_{-\infty}^{0} + \int_{0}^{\infty}\right]\hat{\phi}(\xi)e^{it\xi}d\xi \\ &= \frac{e^{it\xi}}{it}\hat{\phi}(\xi)\Big|_{-\infty}^{0} + \frac{e^{it\xi}}{it}\hat{\phi}(\xi)\Big|_{0}^{\infty} - \left[\int_{-\infty}^{0} + \int_{0}^{\infty}\right]\frac{e^{it\xi}}{it}\phi'(\xi)d\xi \\ &= \frac{e^{it\xi}}{t^2}\phi'(\xi)\Big|_{-\infty}^{0} + \frac{e^{it\xi}}{t^2}\phi'(\xi)\Big|_{0}^{\infty} - \left[\int_{-\infty}^{0} + \int_{0}^{\infty}\right]\frac{e^{it\xi}}{t^2}\phi''(\xi)d\xi \\ &= \frac{\phi'(0^-) - \phi'(0^+)}{t^2} + O(\frac{1}{t^2}).\end{aligned}$$

Since $\phi'(0^-) = -1$ and $\phi'(0^+) = 1$, the Fourier transform of a function with this sort of discontinuity cannot decay faster than $1/t^2$. The Shepp-Logan filter considered above is essentially optimal in this regard. Formally, the impulse response of the exact filter is a constant times $\partial_t \,\mathrm{P.V.}\, t^{-1}$, which also decays like t^{-2} at infinity.

Example 8.3.1. Let $f_r(x, y)$ be the characteristic function of the disk of radius r centered on $(0, 0)$, its Radon transform is

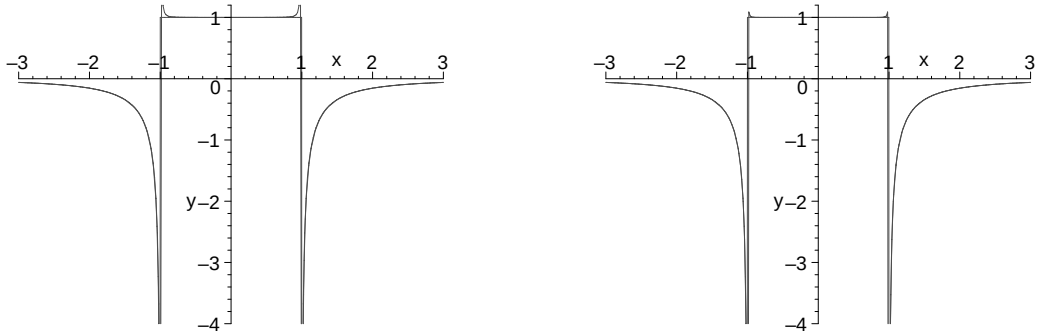
$$\mathrm{R}f_r(t, \omega) = 2\sqrt{r^2 - t^2}\chi_{[-r,r]}(t).$$

Graphs of $Q_\phi f_r$ are shown on figures 8.8-8.9 with $r = 1$ and $d = .1, .04, .02., .005$ For comparison, the exact filtered function, $-\frac{i}{2}\mathcal{H}\partial_t \mathrm{R}f_1$, computed in (4.40) is also shown.



(a) $d = .1$ (b) $d = .04$

Figure 8.8: Ram-Lak filters applied to $\mathrm{R}f_1$.



(a) $d = .02$ (b) $d = .005$

Figure 8.9: Ram-Lak filters applied to $\mathrm{R}f_1$.

The Nyquist theorem says that if the spacing in the t direction is d then the essential bandwidth of $\mathrm{R}f$ should be no more than πd^{-1}. In order for the reconstruction to be accurate $\hat{\phi}(\xi)$ should closely approximate $|\xi|$ over this range. In earlier work

$$\hat{\phi}(\xi) = \begin{cases} |\xi| & |\xi| < \frac{\pi}{d}, \\ 0 & |\xi| > \frac{\pi}{d} \end{cases}$$

so that

$$\phi(t) = \frac{\pi}{d}\frac{\sin(\pi t/d)}{\pi t} - \frac{1-\cos(\pi t/d)}{\pi t^2}.$$

Note that this function only decays like $1/t$ as $t \to \infty$. This is a reflection of the fact that $|\xi|\chi_{[-\frac{\pi}{d},\frac{\pi}{d}]}(\xi)$ has a jump discontinuity at $\pm\pi/d$. Using a sharp cutoff leads to Gibbs artifacts in the reconstruction. On the other hand, a Ram-Lak filter is *never* bandlimited: if the impulse response, ϕ is defined by linear interpolation then the modulation transfer function, $\hat{\phi}(\xi)$ cannot be supported in a finite interval. This is a simple consequence of the Paley-Wiener theorem, see section 3.2.14. If $\hat{\phi}(\xi)$ had bounded support then $\phi(t)$ would be an analytic function. If an analytic function agrees with a linear function on a open interval then it must equal that linear function *everywhere*. This is clearly not possible unless $\phi \equiv 0$.

In the present case both $\{\phi(jd)\}$ and the Fourier transform of its linear interpolant $\hat{\phi}(\xi)$ are known. In the FFT implementation, the modulation transfer function of the filter can be approximated either as the discrete Fourier transform of the sequence,

$$\phi_s(j) =< \phi(-2^q d), \phi(-(2^q-1)d), \ldots, \phi((2^q-1)d) >$$

or by sampling $\hat{\phi}(\xi)$. In general the two approaches give different answers.

Example 8.3.2. For the Shepp-Logan filter is is easy to show that these two approaches to implementing the Ram-Lak filter give different answers. While $\hat{\phi}(0) = 0$, the 0-frequency component in the finite Fourier transform of the sequence, $< \phi_s(j) >$ is

$$\begin{aligned}\mathcal{F}_{2^{q+1}}(\phi_s)(0) &= \frac{1}{2^{q-1}\pi d^2}\left[1 - \sum_{j=1}^{2^q-1}\left(\frac{1}{2j-1} - \frac{1}{2j+1}\right) - \frac{1}{2^{2(q+1)}-1}\right] \\ &= \frac{4}{\pi d^2}\cdot\frac{1}{2^{2(q+1)}-1}.\end{aligned} \tag{8.23}$$

Exercise 8.3.4. Derive the result in (8.23).

Exercise 8.3.5. Find an input to the Shepp-Logan filter for which the two implementations give different results. Do a numerical experiment to see if you can visually identify the difference.

How is the Fourier transform of the interpolated filter function found? This is a nice application of the theory of generalized functions, see sections A.4.6 and 3.2.13. We follow the derivation in [56]. Let the sample spacing be d and suppose that the samples $\phi_n = \phi(nd)$ are specified. The function $\phi(t)$ is defined at intermediate points by interpolation. The computation below allows for different interpolation schemes. To compute $\hat{\phi}(\xi)$, $\phi(t)$ is written as a convolution. Let P denote the generalized function

$$P(t) = d\sum_{n=-\infty}^{\infty}\phi_n\delta(t-nd)$$

and $I(t)$, a function with support in the interval $[-d, d]$ and total integral 1. The filter function is *defined* at intermediate points by setting

$$\phi(t) = I * P(t).$$

If

$$I(t) = \frac{1}{d}\left(1 - \frac{|t|}{d}\right)\chi_{[0,d]}(|t|)$$

then ϕ is the Shepp-Logan linearly interpolated filter, while

$$I(t) = \frac{1}{d}\chi_{[-\frac{d}{2},\frac{d}{2}]}(|t|)$$

produces a piecewise constant interpolant. The Fourier transform of P is the $\left(\frac{2\pi}{d}\right)$-periodic function

$$\hat{P}(\xi) = \sum_{n=-\infty}^{\infty} d\phi_n e^{-ind\xi},$$

see exercise 6.2.2. Using the convolution theorem for the Fourier transform gives

$$\hat{\phi}(\xi) = \hat{P}(\xi)\hat{I}(\xi).$$

For the Shepp-Logan filter

$$\hat{I}(\xi) = \operatorname{sinc}^2\left[\frac{\xi d}{2}\right],$$

piecewise constant interpolation gives

$$\hat{I}(\xi) = \operatorname{sinc}\left[\frac{\xi d}{2}\right].$$

In the Shepp-Logan example, the sinc^2-factor is therefore the result of using linear interpolation.

Exercise 8.3.6. For the Shepp-Logan example, (8.21) give the details of the computation of $P(\xi)$.

8.3.5 Sample spacing in a parallel beam machine

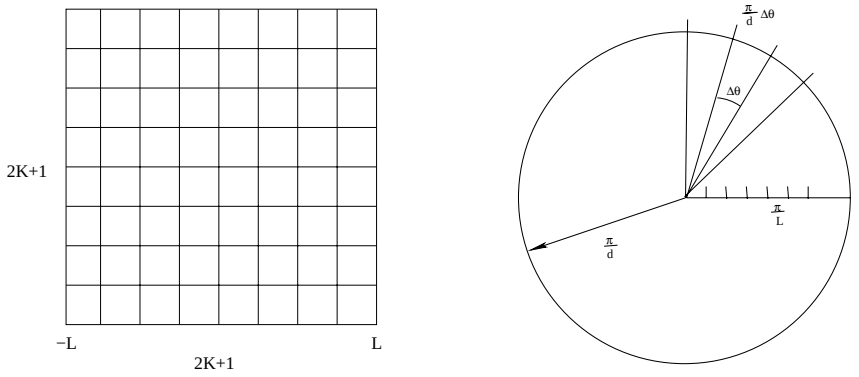
In a parallel beam scanner with filtered back projection as the reconstruction algorithm, there are three different sample spacings. We discuss how these parameters are related; our discussion is adapted from [39]. Assume that the absorption coefficient is supported in the disk of radius L and that the measurements are samples of $\mathrm{R}f$. The finite width of the X-ray beam has a significant effect on this analysis which we return to in section 8.6.2.

The sample spacings are:

(1) The reconstruction grid sample spacing, τ: The reconstruction is usually made in a square, say $[-L, L] \times [-L, L]$ which is divided into a $(2K + 1) \times (2K + 1)$ grid, $\tau = LK^{-1}$.

(2) Spacing between consecutive projection angles: $\Delta\theta = \frac{\pi}{M+1}$.

(3) The sample spacing in the affine parameter is $d = LN^{-1}$. There are $2N + 1$ samples of each projection.



(a) The reconstruction grid.

(b) Sampling in frequency space.

Figure 8.10: How to choose sample spacings.

A priori each of these numbers could be chosen independently, however once one is fixed, it is possible to determine reasonable values for the others. The data is effectively bandlimited and the sample spacings are chosen to reflect the essential bandwidth of the data. The sample spacing in the affine parameter is $d = \frac{L}{N}$. From Nyquist's theorem it follows that, in order to avoid aliasing in this measurement, the effective bandwidth of the object (or at least its Radon transform) should be $\frac{2\pi}{d}$. This implies that $\hat{f}(r\omega)$ should be essentially supported in the disk of radius of $\frac{\pi}{d}$. Along each radial line, $l_{0,\omega(k\Delta\theta)}$ the Fourier transform, $\hat{f}$ is sampled at N points and therefore the sample spacing in this direction is about $\frac{\pi}{L}$. The widest sample spacing of $\hat{f}$ in the angular direction, at least within the effective bandwidth of f is therefore $\Delta\theta\frac{\pi}{d}$, see figure 8.10(b). A reasonable criterion is to choose parameters so that the worst angular resolution, in the Fourier domain, is equal to the resolution in the radial direction. This means that

$$\Delta\theta\frac{\pi}{d} = \frac{\pi}{L} \Rightarrow M + 1 = \frac{\pi}{2}(2N + 1).$$

We are sampling a function supported in the $2L \times 2L$ square and then using a filtered backprojection algorithm to reconstruct the image. This means that we must pad the measured data. As a periodic function, we can regard our data as defined in the square $[-2L, 2L]\times[-2L, 2L]$ and then extended periodically. We have essentially $(2N+1)\times(2N+1)$ samples of the Fourier transform of this function and therefore we should use an equal number of grid points in the square with side length $4L$. This implies that we should take $K \simeq N$, so that $\tau \approx d$. With this choice we are not sacrificing any resolution which is in our measurements nor are we interpolating by using partial sums of the Fourier series. A slightly different way to view this question is to choose the sample spacing in the reconstruction grid so that there are an equal number of measurements as reconstruction points. There are approximately πK^2 grid points in the circle of radius L. About $2MN$ samples of $\mathrm{R}f$ are collected. Using $M \approx \pi N$, from above, implies that K should be approximately $\sqrt{2}N$, so that $\tau \approx \frac{d}{\sqrt{2}}$.

A careful analysis of this question requires a knowledge of the full "point spread function," which incorporates the measurement process, sampling and the reconstruction al-

gorithm. This is complicated by the observation that the map from the "input" f to the output $\{\tilde{f}(x_m, y_l)\}$ is not, in any reasonable sense, a shift invariant filter and so it does not have a point spread function, *per se*. We return to this question after we have considered various physical limitations of the measurement process. This problem is sufficiently complicated that a final determination for the various parameters must be done empirically.

Exercise 8.3.7. Explain why the effective bandwidth of a function $f(x, y)$ and its Radon transform $\mathrm{R}f(t, \omega)$, in the t variable, are the same.

8.4 Filtered backprojection in the fan-beam case

We now consider reconstruction algorithms a fan beam scanner. A view for a parallel beam scanner consists of measurements of $\mathrm{R}f$ for a family of parallel lines and therefore the central slice theorem applies to give an approximation to the Fourier transform of the attenuation coefficient along lines passing through the origin. A view, for either type of fan beam scanner, consists of samples of $\mathrm{R}f$ for a family of lines passing through a point and so the central slice theorem is not directly applicable. There are two general approaches to reconstructing images from the data collected by a fan beam machine: 1. Re-sort and interpolate to obtain the data needed to apply the parallel beam algorithms discussed earlier, 2: Work with the fan geometry and find algorithms well adapted to it. Herman, Lakshminarayanan and Naparstek first proposed an algorithm of the second type in [21]. Our derivation of this algorithm closely follows the presentation in [39].

8.4.1 Fan beam geometry

It is convenient to use a different parameterization for the lines in the plane from that used earlier. As before (t, θ) denotes parameters for the line with oriented normal $\omega(\theta)$ at distance t from the origin. Consider the geometry shown in figure 8.11. Here S denotes the intersection point of the lines defining a view. It lies a distance D from the origin. The central ray (through S and $(0, 0)$) makes an angle β with the positive y-axis. The other lines through S are parametrized by the angle, γ they make with the central ray. These are the natural fan beam coordinates on the space of oriented lines, they are related to the (t, θ) variables by

$$\theta = \gamma + \beta \text{ and } t = D \sin \gamma. \tag{8.24}$$

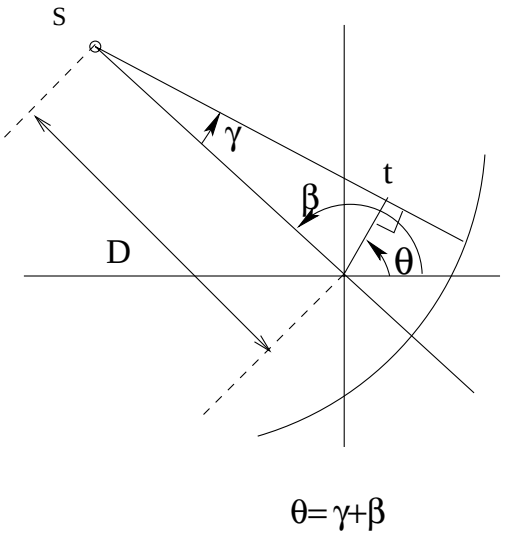


Figure 8.11: Fan beam geometry.

We now derive the continuous form of the approximate reconstruction formula used in fan beam algorithms. This is the analogue of the formula,

$$f_\phi(x,y) = \frac{1}{2\pi}\int_0^\pi \int_{-L}^{L} \mathrm{R}f(t,\theta)\phi(x\cos\theta + y\sin\theta - t)dtd\theta, \tag{8.25}$$

with ϕ the filter function, used for a parallel beam scanner. In (8.25) the weighting of different lines in the filtering operation depends only on their distance from the point (x,y). This general form is retained for the moment, though by the end of the calculation we end up with approximations that do not satisfy this condition. It is convenient to use polar coordinates in the reconstruction plane,

$$(x,y) = (r\cos\varphi, r\sin\varphi).$$

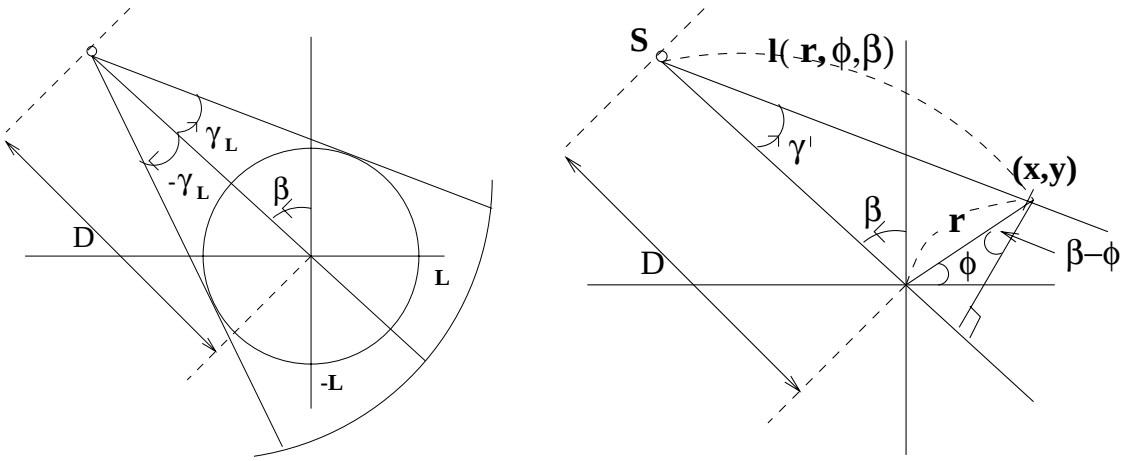
Let κ denote the filter function, so as not to confuse it with the polar angle or the parallel beam filter function. This function is assumed to be smooth and decaying at infinity. In these coordinates (8.25) becomes

$$f_\kappa(r,\varphi) = \frac{1}{2}\int_0^{2\pi} \int_{-L}^{L} \mathrm{R}f(t,\theta)\kappa(r\cos(\theta-\varphi) - t)dtd\theta. \tag{8.26}$$

Our goal is to re-express the reconstruction formula in fan beam coordinates, as a filtered backprojection. Because of the geometry of the fan beam coordinates this is not quite possible. Instead the final formula is a *weighted*, filtered backprojection.

Using the relations in (8.24) to re-express this integral in fan beam coordinates gives

$$f_\kappa(r,\varphi) = \frac{1}{2}\int_0^{2\pi} \int_{-\gamma_L}^{\gamma_L} \mathrm{R}f(D\sin\gamma, \beta+\gamma)\kappa(r\cos(\beta+\gamma-\varphi) - D\sin\gamma)D\cos\gamma d\gamma d\beta.$$



(a) Physical parameters in a fan beam scanner.

(b) Variables for the reconstruction formula.

Figure 8.12: Quantities used in the fan beam, filtered backprojection algorithm.

The function f is supported in the disk of radius L. The limits of integration in the γ-integral, $\pm\gamma_L$ are chosen so that the lines corresponding to the parameters

$$\{(\beta,\gamma)\ :\ \beta\in[0,2\pi),\quad -\gamma_L\le\gamma\le\gamma_L\}$$

include all those intersecting D_L. The data actually collected with a fan beam machine is an approximation to uniformly spaced samples in the (β,γ)-coordinates. To simplify the notation we introduce

$$Pf(\beta,\gamma)\stackrel{d}{=}\mathrm{R}f(D\sin\gamma,\beta+\gamma).$$

In terms of this data the formula reads

$$f_\kappa(r,\varphi)=\frac{1}{2}\int_0^{2\pi}\int_{-\gamma_L}^{\gamma_L}Pf(\beta,\gamma)\kappa(r\cos(\beta+\gamma-\varphi)-D\sin\gamma)D\cos\gamma d\gamma d\beta.$$

From the trigonometric identity

$$\cos(a+b)=\cos(a)\cos(b)-\sin(a)\sin(b)$$

it follows that

$$r\cos(\beta+\gamma-\varphi)-D\sin\gamma=r\cos(\beta-\varphi)\cos\gamma-[r\sin(\beta-\varphi)+D]\sin\gamma. \tag{8.27}$$

Let $l(r,\varphi,\beta)$ be the distance from S to the point (x,y) and let γ' be the angle between the ray $\overline{SO}$ and the ray $\overline{S(x,y)}$. As S is outside the reconstruction grid the function $l(r,\varphi,\beta)$ is strictly positive for points of interest, with

$$l\cos\gamma' = D+r\sin(\beta-\varphi), \tag{8.28}$$
$$l\sin\gamma' = r\cos(\beta-\varphi). \tag{8.29}$$

As functions of (r, φ, β), the distance, l and the angle γ' are given by:

$$\begin{aligned} l(r,\varphi,\beta) &= \sqrt{[D + r\sin(\beta-\varphi)]^2 + [r\cos(\beta-\varphi)]^2}, \\ \gamma'(r,\varphi,\beta) &= \tan^{-1}\frac{r\cos(\beta-\varphi)}{D + r\sin(\beta-\varphi)}. \end{aligned}$$

Using (8.27) and (8.29), the argument of κ can be rewritten

$$r\cos(\beta+\gamma-\varphi) - D\sin\gamma = l(r,\varphi,\beta)\sin(\gamma'(r,\varphi,\beta)-\gamma). \tag{8.30}$$

With these substitutions, the expression for f_κ becomes

$$f_\kappa(r,\varphi) = \frac{1}{2}\int_0^{2\pi}\int_{-\gamma_L}^{\gamma_L} Pf(\beta,\gamma)\kappa(l(r,\varphi,\beta)\sin(\gamma'(r,\varphi,\beta)-\gamma))D\cos\gamma\, d\gamma\, d\beta. \tag{8.31}$$

Using a slightly different functional form for the filter function κ leads to a much simpler formula. In the next section we explicitly incorporate the condition that $\hat{\kappa}(\xi) \approx |\xi|$ for small values of $|\xi|$.

8.4.2 Fan beam filtered backprojection

Let χ_ϵ be a family of functions of bounded support so that

$$\lim_{\epsilon\to 0}\chi_\epsilon = 1 \text{ for all } \xi.$$

Tracing backwards to (8.26) shows that the exact reconstruction is obtained as the limit:

$$f(r,\varphi) = \lim_{\epsilon\to\infty}\frac{1}{4\pi}\int_0^{2\pi}\int_{-\gamma_L}^{\gamma_L} Pf(\beta,\gamma)D\cos\gamma\left[\int_{-\infty}^{\infty} e^{il\sin(\gamma'-\gamma)\xi}|\xi|\chi_\epsilon(\xi)d\xi\right]d\gamma\, d\beta.$$

The β-integral is essentially a weighted backprojection and is innocuous. For the analysis of the γ and ξ integrals we let $h(\gamma)$ be a bounded function with bounded support, and set

$$H(\gamma') = \lim_{\epsilon\to\infty}\frac{1}{4\pi}\int_{-\gamma_L}^{\gamma_L} h(\gamma)D\cos\gamma\left[\int_{-\infty}^{\infty} e^{il\sin(\gamma'-\gamma)\xi}|\xi|\chi_\epsilon(\xi)d\xi\right]d\gamma\, d\beta.$$

Change coordinates in the ξ-integral letting

$$\eta = \left[\frac{l\sin(\gamma'-\gamma)}{\gamma'-\gamma}\right]\xi$$

to obtain

$$H(\gamma') = \lim_{\epsilon\to 0}\frac{1}{4\pi}\int_{-\gamma_L}^{\gamma_L}\int_{-\infty}^{\infty} h(\gamma)\left[\frac{\gamma'-\gamma}{l\sin(\gamma'-\gamma)}\right]^2|\eta|\chi_\epsilon\left(\eta\left[\frac{\gamma'-\gamma}{l\sin(\gamma'-\gamma)}\right]\right)e^{i(\gamma'-\gamma)\eta}d\eta\, d\gamma.$$

The function, $h(\gamma)$ has compact support and therefore the order of integration can be interchanged. As an iterated integral,

$$H(\gamma') = \frac{1}{4\pi} \int_{-\infty}^{\infty} \int_{-\gamma_L}^{\gamma_L} h(\gamma) \left[\frac{\gamma' - \gamma}{l \sin(\gamma' - \gamma)} \right]^2 |\eta| e^{i(\gamma' - \gamma)\eta} d\gamma d\eta.$$

This is an exact formula for the filtering step expressed in fan beam coordinates. A slightly different approximation to this integral is given by

$$H_\epsilon(\gamma') = \frac{1}{2} \int_{-\gamma_L}^{\gamma_L} h(\gamma) \left[\frac{\gamma' - \gamma}{l \sin(\gamma' - \gamma)} \right]^2 \kappa_\epsilon(\gamma' - \gamma) d\gamma,$$

where

$$\kappa_\epsilon(\gamma) = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\eta| \chi_\epsilon(\eta) e^{i\eta\gamma} d\eta.$$

From this calculation it follows that a reasonable approximation to $f(r, \varphi)$ is given by

$$f_\kappa(r, \varphi) = \frac{1}{2} \int_0^{2\pi} \int_{-\gamma_L}^{\gamma_L} Pf(\beta, \gamma) \left[\frac{\gamma' - \gamma}{l \sin(\gamma' - \gamma)} \right]^2 \kappa(\gamma' - \gamma) D \cos\gamma d\gamma d\beta,$$

where, as before κ should be chosen so that $\hat{\kappa}(\xi) \simeq |\xi|$ over the essential bandwidth (in the γ-variable) of $Pf(\beta, \gamma)$. This formula can be rewritten as

$$f_g(r, \varphi) = \int_0^{2\pi} \frac{1}{l^2(r, \varphi, \beta)} \int_{-\gamma_L}^{\gamma_L} Pf(\beta, \gamma) g(\gamma' - \gamma) D \cos\gamma d\gamma d\beta$$

where

$$g(\gamma) = \frac{1}{2} \left(\frac{\gamma}{\sin\gamma} \right)^2 \kappa(\gamma).$$

The weight factor, which accounts for the geometry of the fan beam variables, is included in the definition of the filter function. To interpret this formula as a weighted backprojection set

$$\begin{aligned} Q_g f(\beta, \gamma') &= \int P' f(\beta, \gamma - \gamma') g(\gamma) d\gamma, \text{ and} \\ f_g(r, \varphi) &= \int_0^{2\pi} \frac{1}{l^2(r, \varphi, \beta)} Q_g f(\beta, \gamma') d\beta. \end{aligned} \tag{8.32}$$

Here $P'f(\beta, \gamma) = Pf(\beta, \gamma) D \cos\gamma$ and

$$\gamma'(r, \varphi, \beta) = \tan^{-1} \left[\frac{r \cos(\beta - \varphi)}{D + r \sin(\beta - \varphi)} \right].$$

8.4.3 Implementing the fan beam algorithm

Using (8.32) we can describe a reasonable algorithm for reconstructing an image from fan beam data which is well adapted to the geometry of this type of scanner. The fan beam data is

$$Pf(\beta_j, n\alpha) \text{ where } \beta_j = \frac{2\pi j}{M+1}, \quad j = 0, \ldots, M$$

and n takes integer values. The image is reconstructed in three steps.

Step 1.: Replace the measurements by weighted measurements, that is multiply by the factor $D \cos n\alpha$ to obtain:

$$P'f(\beta_j, n\alpha) = Pf(\beta_j, n\alpha) D \cos n\alpha$$

Step 2.: Discretely convolve the scaled projection data $P'f(\beta_j, n\alpha)$ with $g(n\alpha)$ to generate the filtered projection at the sample points:

$$\begin{aligned} Q_g\tilde{f}(\beta_j, n\alpha) &= \alpha[P'f(\beta_j, \cdot) \star g](n\alpha), \\ \text{where } g(n\alpha) &= \frac{1}{2}\left(\frac{n\alpha}{\sin n\alpha}\right)^2 \kappa(n\alpha). \end{aligned}$$

The filter function κ is selected according to the criteria used in the selection of ϕ for the parallel beam case: it should be real, even and decay at infinity. For ξ in the effective bandwidth of the data $\hat{\kappa}(\xi) \approx |\xi|$.

Step 3.: Perform a weighted backprojection of each filtered projection

$$\tilde{f}_g(x_m, y_l) \approx \Delta\beta \sum_{k=0}^{M} \frac{1}{l^2(x_m, y_l, \beta_k)} Q_g f(\beta_k, \gamma'(x_m, y_l, \beta_k)).$$

As before, the values $\{Q_g\tilde{f}(\beta_k, \gamma'(x_m, y_l, \beta_k))\}$ were **not** computed in the previous step. They are obtained by using interpolation from the values, $\{Q_g f(\beta_k, n\alpha)\}$, which were computed.

The values of the functions

$$\{l(x_m, y_l, \beta_k), \gamma'(x_m, y_l, \beta_k)\}$$

as well as the interpolation coefficients can be pre-computed. In that case the computational load of this algorithm is the same order of magnitude as that of the parallel beam Ram-Lak algorithm. Note that as in that case, steps 1 and 2 can be performed as soon as the data from a given view has been collected and the backprojection step can also be parallelized. For a third generation machine, a view is defined by the source position, so the filter step can begin almost immediately. For a fourth generation machine, a view is defined by a detector position, hence the filtering step must wait until all the data for first view has been collected. Once this threshold is reached, the filtering can again be effectively parallelized.

Example 8.4.1. Using the filter function defined by Shepp and Logan for $g(n\alpha)$ gives

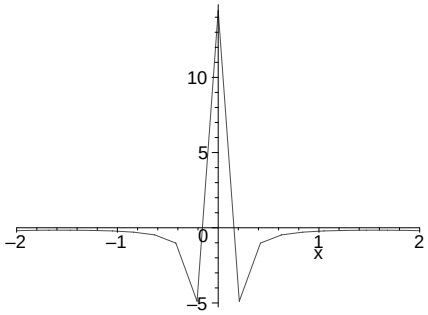
$$g_{SL}(n\alpha) = \begin{cases} \frac{2}{\pi\alpha^2} & \text{if } n = 0, \\ \frac{-2}{\pi \sin^2 n\alpha} \frac{n^2}{4n^2-1} & \text{if } \neq 0. \end{cases} \tag{8.33}$$

The graph of this function with $\alpha = .21$ is shown in figure 8.4.3(a).

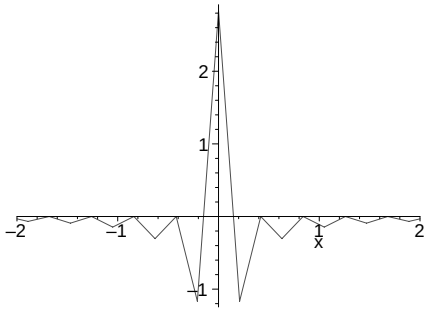
Example 8.4.2. In their original paper, Herman, Lakshminarayanan and Naparstek used a slightly different function defined by

$$g_{HLN}(n\alpha) = \begin{cases} \frac{1}{8\alpha^2} & \text{if } n = 0 \text{ or even}, \\ -\frac{1}{2\pi^2 \sin^2(n\alpha)} & \text{if } n \text{ is odd}. \end{cases} \tag{8.34}$$

The graph of this function for $\alpha = .21$ is shown in figure 8.4.3(b).



(a) The impulse response for a Shepp-Logan, fan beam filter.



(b) The impulse response for the HLN-fan beam filter.

8.4.4 Data collection for a fan beam scanner

In the derivation of the algorithm above it is assumed that data is collected for $\beta \in [0, 2\pi)$. This means that every projection is measured twice, as two pairs of fan beam coordinates (β_1, γ_1) and (β_2, γ_2) define the same line if and only if

$$\begin{gathered} \gamma_1 = -\gamma_2, \\ \beta_1 - \gamma_1 = \beta_2 - \gamma_2 + \pi \Rightarrow \beta_1 = \beta_2 + 2\gamma_1 + \pi \end{gathered} \tag{8.35}$$

then $Pf(\beta_1, \gamma_1) = Pf(\beta_2, \gamma_2)$, see figure8.13(a). In a parallel beam machine it suffices to collect data for $\theta \in [0, \pi)$, for a fan beam machine one does not need to have measurements for all $\beta \in [0, 2\pi)$. However some care is required in collecting and processing smaller data sets. The Radon transform satisfies

$$\mathrm{R}f(t, \theta) = \mathrm{R}f(-t, \pi + \theta).$$

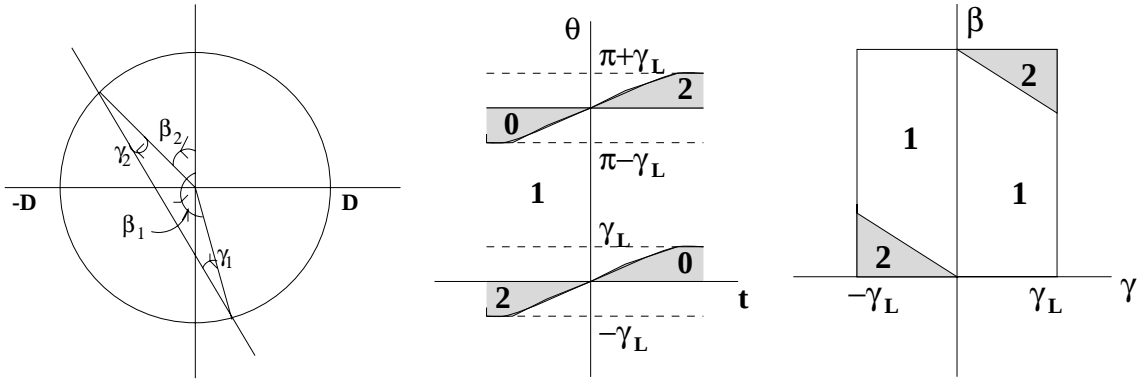
In fan beam coordinates this is equivalent to

$$Pf(\beta,\gamma) = Pf(\beta + 2\gamma + \pi, -\gamma).$$

Sampling Pf on the range

$$\beta \in [0,\pi], \text{ and } -\gamma_L \le \gamma \le \gamma_L$$

and using $t = D\sin\gamma$, and $\theta = \beta + \gamma$, gives the diagram, in (t,θ) space, shown in figure 8.13(b). The numbers show how many times a given projection is measured for (β,γ) in this range. Some points are measured once, some twice and some not at all.



(c) Symmetry in (γ,β)-coordinates.

(d) Data collection in (t,θ)-coordinates.

(e) Data collection in (γ,β)-coordinates.

Figure 8.13: Collecting data for fan beam scanners.

In order to gather a complete data set it is necessary for β go from 0 to $\pi + 2\gamma_L$. Of course even more values are now sampled twice. The algorithm above can be used with measurements gathered over such a range, however care must be taken to count each projection exactly once. This can be done by multiplying the projection data by a windowing function. For example one could use

$$w_\beta(\gamma) = \begin{cases} 0 & 0 \le \beta \le 2\gamma_L + 2\gamma, \\ 1 & \text{otherwise} \end{cases}.$$

As usual a sharp cutoff produces its own artifacts so it is preferable to use a smoother window. The window should be continuous and satisfy the conditions

(a) $w_{\beta_1}(\gamma_1) + w_{\beta_2}(\gamma_2) = 1$, for pairs (β_1,γ_1), (β_2,γ_2) satisfying (8.35),

(b) $w_\beta(\gamma) \ge 0$.

Better results are obtained if w has continuous first derivatives. Given the known regularity of the data, a bounded though not necessarily continuous, first derivative should usually suffice. An example is given by

$$w_\beta(\gamma) = \begin{cases} \sin^2\left[\frac{\pi\beta}{4(\gamma_L-\gamma)}\right] & 0 \le \beta \le 2\gamma_L - 2\gamma, \\ 1 & 2\gamma_L - 2\gamma \le \beta \le \pi - 2\gamma, \\ \sin^2\left[\frac{\pi}{4}\frac{\pi+2\gamma_L-\beta}{\gamma+\gamma_L}\right] & \pi - 2\gamma \le \beta \le \pi + 2\gamma_L \end{cases}$$

8.4.5 Rebinning

It is also possible to re-sort the fan beam data into collections of approximately parallel projections, interpolate and then use a parallel beam algorithm. One such method is called *rebinning.* It requires that $\Delta\beta = \alpha$ from which it follows that

$$Pf(m\alpha, n\alpha) = \mathrm{R}f(D\sin n\alpha, \omega((n+m)\alpha)).$$

If $n + m = q$, then these are samples belonging to a single "parallel beam" view. Since $\sin(n+1)\alpha - \sin n\alpha$ depends on n these samples are not equally spaced in the t-direction. Equally spaced samples of the parallel projections can be obtained by interpolating in this direction. The interpolated data set could then be processed using a standard parallel beam algorithm to reconstruct the image.

Another possibility is to select angles $< \gamma_{-P}, \ldots, \gamma_P >$ so that

$$D(\sin(\gamma_j) - \sin(\gamma_{j-1})) = \Delta t, \text{ for } j = 1-P, \ldots, P.$$

The fan beam data $\{Pf(\beta_j, \gamma_k)\}$ is then equally spaced in the t direction. Interpolating in the β parameter gives data which again approximates the data collected by a parallel beam machine. Beyond the errors introduced by interpolations these algorithms cannot be effectively parallelized. This is because all the data from a slice needs to be collected before the interpolation and rebinning can begin.

8.5 The effect of a finite width X-ray beam

Up to now, we have assumed that an X-ray "beam" is just a line with no width and that the measurements are integrals over such lines. What is really measured is better approximated by averages of such integrals. We now consider how the finite width of the X-ray beam affects the measured data. Our treatment closely follows the discussion in [69].

A simple linear model for this effect is to replace the line integrals of f

$$\mathrm{R}f(t,\omega) = \int f(t\omega + s\hat{\omega})ds,$$

by a weighted average of these line integrals (colloquially, "a strip integral")

$$\mathrm{R}_W f(t,\omega) = \int w(u)\,\mathrm{R}f(t-u,\omega)du.$$

Here $w(u)$ is weighting function. It models the distribution of energy *across* the X-ray beam as well as the detector used to make the measurements. This function is sometimes called the *beam profile*.

The relationship between $\mathrm{R}f$ and R_W is a consequence of the convolution theorem for the Radon transform. In the imaging literature it is due to Shepp and Logan.

Theorem 8.5.1 (Shepp and Logan). *The weighted Radon transform* $\mathrm{R}_W f(t,\theta)$ *is the Radon transform of*

$$f^k(x,y) = \int_0^{2\pi}\int_0^{\infty} f(x-\rho\cos\alpha, y-\rho\sin\alpha)k(\rho)\rho d\rho d\alpha$$

where

$$k(\rho) = -\frac{1}{\pi\rho}\partial_\rho \int_\rho^{\infty} \frac{w(u)u}{\sqrt{u^2-\rho^2}}du$$

and

$$w(u) = \int_{-\infty}^{\infty} k(\sqrt{u^2+v^2})dv.$$

Remark 8.5.1. The function f^k is the convolution of f with $k(\sqrt{x^2+y^2})$ expressed in polar coordinates. If $w(u)$ has a bounded support, the integrand of k is zero for sufficiently large ρ, hence $k(\sqrt{x^2+y^2})$ also has bounded support in the plane. Similarly if k has bounded support, so does $w(u)$.

Proof. The theorem is an immediate consequence of Proposition 4.1.1. The function $\mathrm{R}_W f(t,\omega)$ is the convolution in the t-parameter of $\mathrm{R}f$ with $w(t)$. If k is a function on $\mathbb{R}^2$ such that $\mathrm{R}k = f$ then the proposition states that

$$\mathrm{R}_W f = \mathrm{R}(f * k).$$

Since w is independent of ω it follows that k must also be a radial function. The formula for k is the Radon inversion formula for radial functions derived in section 2.5. □

Some simple examples of w, k pairs are

$$w_1(u) = \begin{cases} \frac{1}{2\delta} & u\in[-\delta,\delta], \\ 0 & |u|>\delta \end{cases}, \text{then} \quad k_1(\rho) = \begin{cases} \frac{1}{2\pi\delta}\frac{1}{\sqrt{\delta^2-\rho^2}} & 0\le\rho<\delta, \\ 0 & \rho\ge\delta, \end{cases}$$

and

$$w_2(u) = \begin{cases} \frac{1}{\pi\delta^2}(\delta^2-u^2)^{1/2} & |u|<\delta, \\ 0 & |u|>\delta \end{cases}, \text{then} \quad k_2(\rho) = \begin{cases} \frac{1}{\pi\delta^2} & \rho<\delta, \\ 0 & \rho\ge\delta. \end{cases}$$

A consequence of finite strip width is that the actual measurements are samples of the Radon transform of $f * k$, which is a somewhat smoothed version of f. Indeed

$$\widetilde{\mathrm{R}_W f}(r,\omega) = \hat{w}(r)\widetilde{\mathrm{R}f}(r,\omega) \tag{8.36}$$

and therefore the finite strip width leads to low pass filtering of $\mathrm{R}f$ in the affine parameter. This has the desirable effect of reducing the aliasing artifacts that result from sampling in the t-parameter. In X-ray tomography this is essentially the only way to low pass filter the data before it is sampled. Of course this averaging process also leads to a loss of resolution; so the properties of the averaging function w need to be matched with the sample spacing. As we saw in section 7.3 the effects of such averaging, can to some extent be removed, nonetheless algorithms are often evaluated in terms of their ability to reconstruct samples of $f * k$ rather than f itself.

As mentioned above the beam profile models both the source and detector. If $I(u)$ describes the intensity of the X-ray beam incident on the detector then the output of the detector is modeled as

$$\int_{-\infty}^{\infty} w_d(u) I(u) du.$$

Suppose that that $w_s(u)$ models the source; if the source and detector are fixed, relative to one another then the combined effect of this source-detector pair would be modeled by the product $w(u) = w_s(u) w_d(u)$. This is the geometry in a third generation, fan beam scanner. In some parallel beam scanners and fourth generation scanners the detectors are fixed and the source moves. In this case the model for the source detector pair is the convolution $w = w_s * w_d$. The detector is often modeled by a simple function like w_1 or w_2 defined above while the source is sometimes described by a Gaussian $ce^{-\frac{u^2}{\sigma}}$. In this case the X-ray source is said to have a *Gaussian focal spot.*

Remark 8.5.2. The problem of modeling X-ray sources and detectors is somewhat beyond the scope of this text. A very thorough treatment is given in [4].

Exercise 8.5.1. Explain why these two models of source-detector pairs are reasonable for the different hardware. In particular, explain how relative motion of the source and detector leads to a convolution.

8.5.1 A non-linear effect

Unfortunately the effect of finite beam width is a bit more complicated than this. If we could produce a 1-dimensional X-ray beam then, what is measured would actually be

$$I_o = I_i \exp[-\mathrm{R}f(t, \omega)],$$

where I_i is the intensity of the X-ray source and I_o is the measured output. The difficulty is that this is a nonlinear relation. For a weighted strip what is actually measured is closer to

$$\log \frac{I_o}{I_i} \approx -\log \int_{-\infty}^{\infty} w(u) \exp[-\mathrm{R}f(t-u, \omega)] du.$$

If $w(u)$ is very concentrated near $u = 0$ and $\int w(u) du = 1$ then, the Taylor expansion gives

$$\log \frac{I_o}{I_i} \approx \int_{-\infty}^{\infty} w(u) \, \mathrm{R}f(t-u, \omega) du.$$

To derive this expression we use the Taylor expansions:

$$\begin{aligned} e^{-x} =& 1 - x + \frac{x^2}{2} - \frac{x^3}{6} + O(x^4), \\ \log(1+x) =& x - \frac{x^2}{2} + \frac{x^3}{3} + O(x^4). \end{aligned} \tag{8.37}$$

This analysis assumes that the oscillation of $\mathrm{R}f(t-u,\omega)$ is small over the support of $w(u)$:

$$\begin{aligned} &\int_{-\infty}^{\infty} w(u) \exp(-\mathrm{R}f(t-u,\omega))du \\ &= \int_{-\infty}^{\infty} w(u) \exp(-\mathrm{R}f(t,\omega)) \exp[\mathrm{R}f(t,\omega) - \mathrm{R}f(t-u,\omega)]du \\ &= \exp(-\mathrm{R}f(t,\omega)) \int_{-\infty}^{\infty} w(u) \exp[\mathrm{R}f(t,\omega) - \mathrm{R}f(t-u,\omega)]du \\ &= \exp(-\mathrm{R}f(t,\omega)) \int_{-\infty}^{\infty} w(u)[1 - (\mathrm{R}f(t,\omega) - \mathrm{R}f(t-u,\omega)) + O((\mathrm{R}f(t-u,\omega) - \mathrm{R}f(t,\omega))^2)]du \\ &= \exp(-\mathrm{R}f(t,\omega))[1 - \int_{-\infty}^{\infty} w(u)(\mathrm{R}f(t-u,\omega)du - \mathrm{R}f(t,\omega))du + \\ &\qquad O\left(\int w(u)(\mathrm{R}f(t-u,\omega) - \mathrm{R}f(t,\omega))^2 du\right)] \end{aligned}$$

In the third line we have used the Taylor expansion for e^x. Taking $-\log$, using the Taylor expansion for $\log(1+x)$ and the assumption that $\int w(u) = 1$ gives

$$\log \frac{I_o}{I_i} \approx \int w(u)\, \mathrm{R}f(t-u,\omega)du + O\left(\int w(u)(\mathrm{R}f(t-u,\omega) - \mathrm{R}f(t,\omega))^2 du\right).$$

The leading order error is proportional to the mean square oscillation of Rf, weighted by w.

8.5.2 The partial volume effect

If the variation of $Rf(t,\omega)$ is large over the width of the strip then the analysis above is not valid. In practice this happens if part of the X-ray beam intercepts bone and the remainder passes through soft tissue. In imaging applications this is called the *partial volume effect*. To explain this we consider a simple special case. Suppose that the intensity of the X-ray is constant across a strip of width 1. Half the strip is blocked by a rectangular object of height 1 with attenuation coefficient 2 and half the strip is empty. If we assume that

μ_0	μ_1	non-linear	linear	relative error
0	.01	.00499	.005	2%
0	.1	.0488	.05	2.4%
0	.5	.2191	.25	12.4%
0	1	.3799	.5	24%
0	2	.5662	1	43%
.3	.4	.34875	.35	3.5%
.3	1.3	.6799	.8	15%
1	2	1.38	1.5	8%

Table 8.1: Errors due to the partial volume effect.

$w(u) = \chi_{[0,1]}(u)$ then

$$-\log\left[\int_{-\infty}^{\infty} w(u)\exp[-\mathrm{R}f(t-u,\omega)]du\right] = -\log\left[\frac{1+e^{-2}}{2}\right] \simeq 0.5662,$$

whereas

$$\int_{-\infty}^{\infty} w(u)\,\mathrm{R}f(t-u,\omega)du = 1.$$

In the table below we give the linear and non-linear computations for an absorbent unit square with two absorption coefficients μ_0, μ_1 each occupying half, see figure 8.14.

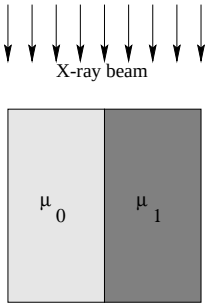


Figure 8.14: Absorbing square.

An even more realistic example is provided by a long rectangle of absorbing material with a small inclusion of more absorbent material, as shown in figure (8.15). The graph shows the relative errors with $\mu_0 = 1$ and $\mu_1 \in \{1.5, 2, 2.5, 3\}$. This is a model for a long stretch of soft abdominal tissue terminating at a piece of a rib.

The partial volume effect is the discrepancy between the non-linear data which is actually collected and the linear model for the data collection, used in the derivation of the

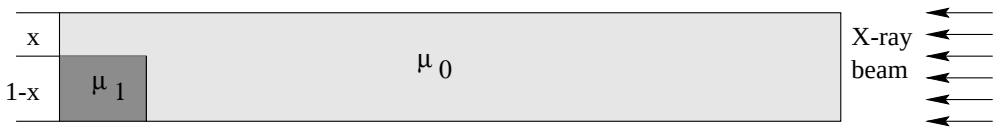


Figure 8.15: Rectangle with small inclusion

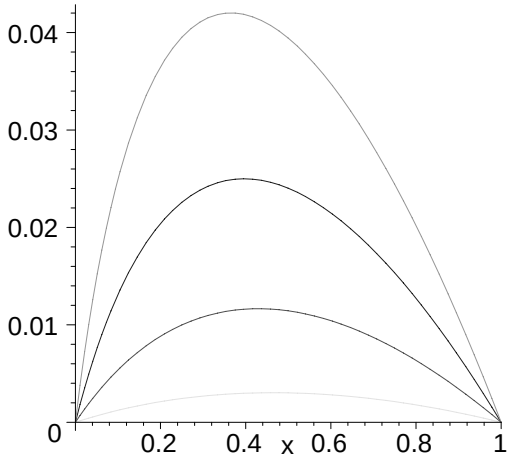


Figure 8.16: Relative errors with small inclusion

reconstruction algorithms. That is, our algorithm assumes that what is collected are samples of $Rf^k(t,\omega)$, because of the finite width of a real strip and the partial volume effect this is not so. Even if we could measure a projection for all relevant pairs (t,ω) our algorithm would not reconstruct f^k exactly but rather some further *non-linear* transformation applied to f. In real images the partial volume effect appears as abnormally bright spots near hard object or streaks.

8.5.3 Some mathematical remarks*

In the foregoing sections we consider algorithms for approximately reconstructing an unknown density function f from finitely many samples of its Radon transform $\{\mathrm{R}f(t_j,\omega_k)\}$. It is reasonable the enquire what is the "best" one can do in approximating f from such data and, if g is the "optimal solution," then what does $f-g$ "look like." A considerable amount of work has been done on these two questions, we briefly describe the results of Logan and Shepp, see [47] and [46].

Assume that f is an L^2-function which is supported in the disk of radius 1. It is assumed that $\mathrm{R}f(t,\omega_j)$ is known for n-distinct directions. Logan and Shepp examine the following problem: Find the function $g \in L^2(D_1)$ such that

$$\mathrm{R}f(t,\omega_j) = \mathrm{R}g(t,\omega_j), \quad j=0,\ldots,n-1, \tag{8.38}$$

which minimizes

$$\begin{aligned} \mathrm{Var}(g) &= \int_{D_1} (g(x,y) - \bar{g})^2 dxdy, \text{ where} \\ \bar{g} &= \frac{1}{\pi} \int_{D_1} g(x,y) dxdy. \end{aligned} \tag{8.39}$$

Briefly: find the function g with the specified projection data and minimum L^2-variation from its mean. In light of the fact that many imaging artifacts are highly oscillatory, the solution to this variational problem is a reasonable candidate for the "optimal reconstruction" from finitely many (complete) projections.

In [47] it is shown that this problem has a unique solution of the form

$$g(x,y) = \sum_{j=0}^{n-1} \alpha_j(\langle (x,y), \omega_j \rangle),$$

where, as indicated, $\{\alpha_0(t), \ldots, \alpha_{n-1}(t)\}$ are functions of one variable. An explicit solution, as Fourier series for the $\{\alpha_j\}$ is obtained in the case that the angles are equally spaced. In the general case, an algorithm is provided to determine these Fourier coefficients. In the process of deriving the formula for the optimal function, necessary and sufficient conditions are found for n-functions $\{r_j(t)\}$ to satisfy

$$r_j(t) = \mathrm{R}f(t, \omega(\frac{j\pi}{n})), \quad j = 0, \ldots, n-1$$

for some function $f \in L^2(D_1)$. These are complicated relations satisfied by the sine-series coefficients of the functions $\{r_j(\cos\tau)\}$.

In [46] the question of how large n must be taken to get a good approximation is considered. The precise answer is rather complicated. Roughly speaking, if the Fourier transform of f is essentially supported in a disk of radius about $n - n^{\frac{1}{3}}$ then, at least for n large, n-projections suffice to find a good approximation to f. Moreover the error $e = f - g$ is a "high frequency" function in the sense that

$$\int_{\|\xi\| < n - n^{\frac{1}{3}}} |\hat{e}(\xi)|^2 d\xi_1 d\xi_2 \leq \lambda_n \int_{D_1} |f(x,y)|^2 dxdy.$$

Here λ_n is a function which tends to zero as n tends to infinity. These results are precise versions of a general, heuristic principle in image reconstruction that a function f for which $\mathrm{R}f(t,\omega) = 0$ for "many" values of (t,ω) is necessarily highly oscillatory. This is indicative of the general experience that, with a good reconstruction algorithm, the error $f - \tilde{f}_\phi$ is a highly oscillatory function.

8.6 The point spread function and linear artifacts

See: 4.4, A.4.7.

The remainder of this chapter is concerned with analyzing artifacts which arise in image reconstruction using realistic data and a reasonable model for the source-detector pair. The explanation for a given artifact is usually found by isolating the features of the image which produce it. At the center of this discussion are the point spread and modulation transfer functions (PSF and MTF), characterizing the measurement and reconstruction process. Once the data is sampled the measurement process is no longer translation invariant and therefore it is not described by a single PSF. Instead, for each point (x, y) there is a (generalized) function $\Phi(x, y; a, b)$ so that the reconstructed image at (x, y) is given by

$$f_\Phi(x, y) = \int_{\mathbb{R}^2} \Phi(x, y; a, b) f(a, b) da db.$$

Our derivation of Φ is done in two steps. First we find a PSF that incorporates a model for the source-detector pair and the filter used in the filtered backprojection step. This part is both translation invariant and isotropic. Afterward we incorporate the effects of sampling the measurements to obtain an expression for $\Phi(x, y; a, b)$. In this section only the parallel beam geometry is considered, our presentation follows [56]. The results for the fan beam geometry are similar but a little more complicated to derive, see [36] and [34].

As a function of (x, y), $\Phi(x, y; a, b)$ is the output of the measurement-reconstruction process applied to a unit point source at (a, b). Mathematically this is modeled by

$$\delta_{(a,b)}(x, y) = \delta((x, y) - (a, b)).$$

To facilitate the computation of Φ is it useful to determine the Radon transform of this generalized function, which should itself be a generalized function on $\mathbb{R} \times S^1$. Since $\delta_{(a,b)}$ is $\delta_{(0,0)}$ translated by (a, b) it suffices to determine $\mathrm{R}\delta_{(0,0)}$. Let φ_ϵ be a family of smooth functions converging to the $\delta_{(0,0)}$ in the sense that $\varphi_\epsilon * f$ converges uniformly to f, for f a continuous function with bounded support. Proposition 4.1.1, the convolution theorem for the Radon transform says that

$$\mathrm{R}(\varphi_\epsilon * f)(t, \omega) = \mathrm{R}\varphi_\epsilon *_t \mathrm{R}f(t, \omega).$$

Since the left hand side converges to $\mathrm{R}f$, as $\epsilon \to 0$ it follows that

$$\mathrm{R}\delta_{(0,0)}(t, \omega) = \lim_{\epsilon \to 0} \mathrm{R}\varphi_\epsilon(t, \omega) = \delta(t).$$

Using Proposition 4.1.2 we obtain the general formula

$$\mathrm{R}\delta_{(a,b)}(t, \omega) = \delta(t - \langle \omega, (a, b) \rangle). \tag{8.40}$$

In the early days of imaging, machines were calibrated by making measurements of composite objects with known absorption coefficients. These objects are called *phantoms*.

The problem with this approach is that it mixes artifacts caused by physical measurement errors with those caused by algorithmic errors. A very important innovation in medical imaging was introduced by Larry Shepp. In order to isolate the algorithmic errors he replaced the (physical) phantom with a *mathematical phantom*.

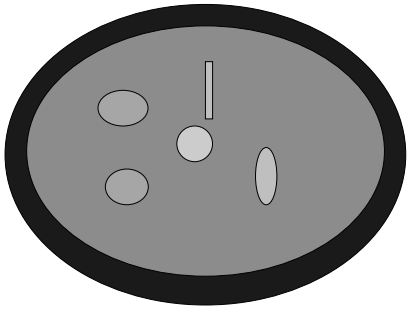


Figure 8.17: A mathematical phantom.

Instead of using real measurements of a known object, he suggested that one give a mathematical description of a phantom to create simulated and *controlled* data. In this way, algorithmic errors could be separated from measurement errors. A mathematical phantom is created as follows: first a simplified model of a slice of the human head (or other object) is described as an arrangement of ellipses and polygons. Each region is then assigned a density or attenuation coefficient, see figure 8.17. Next the continuous model is digitized by superimposing a regular grid and replacing the piecewise continuous densities by their averaged values over the squares that make up the grid. Finally measurements are simulated by integrating the digitized model over a collection of strips, arranged to model some particular measurement apparatus. One can incorporate different sorts of measurement errors, e.g. noise, beam hardening, patient motion, etc. into the simulated measurements to test the robustness of an algorithm to different sorts of measurement errors. The main point is that by using mathematical phantoms you know, *a priori,* exactly what is being measured and can therefore compare the reconstructed image to a known, exact model. Mathematical phantoms are very useful in the study of artifacts caused by sampling errors and noise.

Exercise 8.6.1. Derive (8.40) by using the family of functions

$$\varphi_\epsilon(x,y) = \frac{1}{\epsilon^2}\chi_{[0,\epsilon^2]}(x^2+y^2).$$

8.6.1 The PSF without sampling

See: A.3.6.

For the purposes of this discussion we use the simpler, linear model of a finite width X-ray beam. Let $w(u)$ be a non-negative function with total integral 1. Our model for a

measurement is a sample of

$$\mathrm{R}_W f(t,\omega) = \int_{-\infty}^{\infty} w(u)\,\mathrm{R}f(t-u,\omega).$$

If "all" the data

$$\{\mathrm{R}_W f(t,\omega) \,:\, t \in [-L,L], \omega \in S^1\}$$

were available, then the filtered backprojection reconstruction, with filter function ϕ, would be

$$f_{\phi,w}(x,y) = (\mathrm{R}^* Q_\phi\, \mathrm{R}_W f)(x,y).$$

Recall that R^* is the backprojection operation and

$$\begin{aligned} Q_\phi g(t,\omega) &= \int_{-\infty}^{\infty} g(t-s,\omega)\phi(s)ds \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{g}(r,\omega)\hat{\phi}(r)e^{irt}dr. \end{aligned} \tag{8.41}$$

Because $\mathrm{R}_W f$ is defined by convolving $\mathrm{R}f$ with w in the t-parameter, it is a simple computation to see that

$$f_{\phi,w}(x,y) = \mathrm{R}^* Q_{\phi * w}\,\mathrm{R}f, \tag{8.42}$$

where $\phi * w$ is a 1-dimensional convolution.

Using the central slice theorem in (8.41) gives

$$f_{\phi,w}(x,y) = \frac{1}{[2\pi]^2}\int_0^{\infty}\int_0^{2\pi} \hat{f}(r\omega)\hat{\phi}(r)\hat{w}(r)e^{ir\langle(x,y),\omega\rangle}drd\omega. \tag{8.43}$$

As $\hat{\phi}(r) \approx |r|$ for small r it is reasonable to assume that $\hat{\phi}(0) = 0$ and define $\hat{\psi}(r)$ by the equation

$$\hat{\phi}(r) = |r|\hat{\psi}(r).$$

Substituting this into (8.43) we recognize $rdrd\omega$ as the area element on $\mathbb{R}^2$, to get

$$f_{\phi,w}(x,y) = \frac{1}{[2\pi]^2}\int_{\mathbb{R}^2} \hat{f}(\xi)\hat{\psi}(\|\xi\|)\hat{w}(\|\xi\|)e^{i\langle(x,y),\xi\rangle}d\xi. \tag{8.44}$$

The MTF is therefore

$$\hat{\Psi}_0(\xi) = \hat{\psi}(\|\xi\|)\hat{w}(\|\xi\|). \tag{8.45}$$

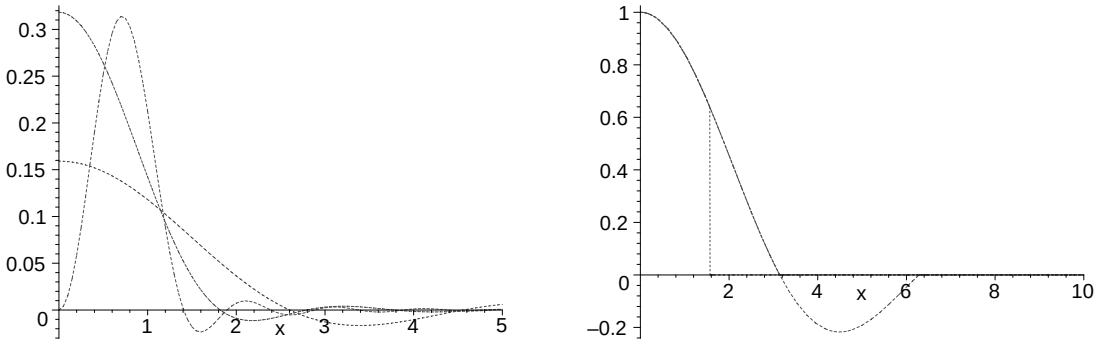
It is important to keep in mind that the Fourier transforms on the right hand side of (8.45) are *one* dimensional, while that on the left is a *two* dimensional transform. The PSF is

therefore

$$\begin{aligned}\Psi_0(x,y) &= \frac{1}{[2\pi]^2}\int_{\mathbb{R}^2}\hat{\psi}(\|\xi\|)\hat{w}(\|\xi\|)e^{i\langle(x,y),\xi\rangle}d\xi \\ &= \frac{1}{2\pi}\int_0^\infty \hat{\phi}(r)\hat{w}(r)J_0(r\|(x,y)\|)dr.\end{aligned} \tag{8.46}$$

Let $\rho = \|(x,y)\|$ denote the radius function in the spatial variables.

Example 8.6.1. For the first example we consider the result of using a sharp cutoff in frequency space. The transfer function of the filter is $\hat{\phi}(r) = \chi_{[-\frac{\pi}{d},\frac{\pi}{d}]}(r)$, with the beam profile function w_1. Figure 8.18(a) shows the PSFs with $d = .5, 1$ and 2, figure 8.18(b) shows the corresponding MTFs.



(a) Several PSFs.

(b) And their MTFs

Figure 8.18: Examples of PSF and MTF with band limited regularization.

Notice the large oscillatory side lobes. Using a filter of this type leads to severe Gibbs artifacts, that is, a sharp edge in the original image produces large oscillations, parallel to the edge in the reconstructed image.

Example 8.6.2. We consider the family of examples with

$$w_\delta(t) = \begin{cases} \frac{1}{2\delta} & \text{if } |t| \le \delta, \\ 0 & \text{if } |t| > \delta, \end{cases}$$

and

$$\hat{\phi}_\epsilon(r) = |r|e^{-\epsilon|r|}$$

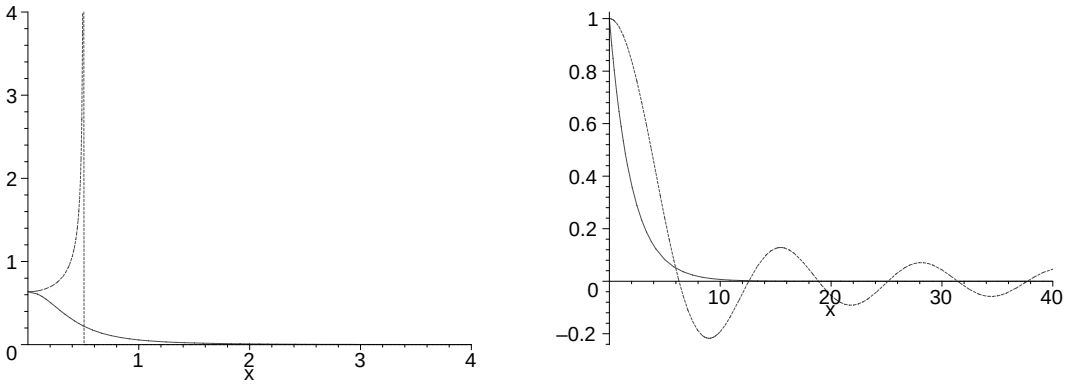
and therefore $\hat{\psi}_\epsilon(r) = e^{-\epsilon|r|}$. As $\hat{w}_\delta(r) = \operatorname{sinc}(\delta r)$ the MTF is given by

$$\hat{\Psi}_{0(\delta,\epsilon)}(\xi) = \operatorname{sinc}(\delta\|\xi\|)e^{-\epsilon\|\xi\|}.$$

If $\epsilon = 0$ (no regularizing function) or $\delta = 0$ (1-dimensional X-ray beam) then the integrals defining Ψ_0 exists as an improper Riemann integrals,

$$\Psi_{0(\delta,0)}(\rho) = \frac{1}{2\pi\delta} \cdot \frac{\chi_{[0,\delta)}(\rho)}{\sqrt{\delta^2 - \rho^2}}, \quad \Psi_{0(0,\epsilon)}(\rho) = \frac{\epsilon}{2\pi} \cdot \frac{1}{[\epsilon^2 + \rho^2]^{\frac{3}{2}}}.$$

The graphs of these functions are shown figure 8.19(a), the dotted curve shows $\Psi_{0(.5,0)}$ and the solid line is $\Psi_{0(0,.5)}$. Figure 8.19(b) shows the corresponding MTFs.

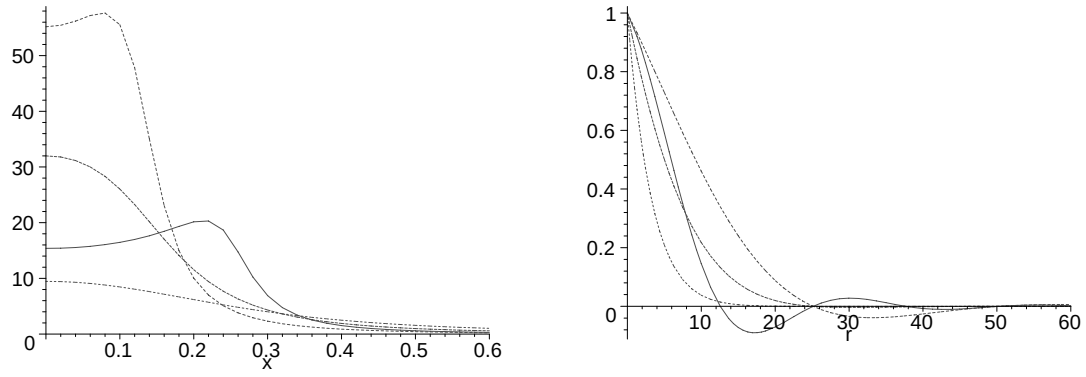


(a) Limiting PSFs.

(b) Limiting MTFs

Figure 8.19: Limits for the PSF and MTF in the filtered backprojection algorithm.

Graphs of $\Phi_{0(\delta,\epsilon)}(\rho)$, for several values of (δ, ϵ) are shown in figure 8.20(a). The MTFs are shown in figure 8.20(b). The values used are $(.25, .05), (.125, .05), (.125, .125), (.125, .3))$, smaller values of ϵ producing a sharper peak in the PSF. When $\epsilon << \delta$ the PSF resembles the singular case $\epsilon = 0$ for small values of ρ. Due to the shape of the PSF this is sometimes called the "volcano effect."



(a) Several PSFs.

(b) And their MTFs

Figure 8.20: Examples of PSF and MTF with exponential regularization.

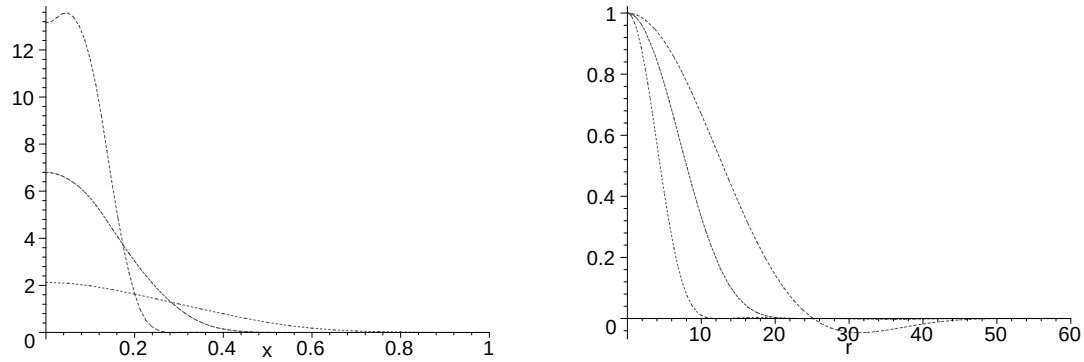
Example 8.6.3. As a final example we consider the Shepp-Logan filter. The regularizing filter has (one dimensional) transfer function

$$\hat{\phi}(r) = |r| \left| \operatorname{sinc}\left(\frac{dr}{2}\right) \right|^3$$

Using the same model for the source-detector pair as before gives the total MTF:

$$\hat{\Psi}_{0(\delta,d)}(\xi) = \operatorname{sinc}(\delta\|\xi\|) \left| \operatorname{sinc}\left(\frac{d\|\xi\|}{2}\right) \right|^3.$$

Recall that the Shepp-Logan filter is linearly interpolated and d represents the sample spacing. Here 2δ is the width of the source-detector pair. Graphs, in the radial variable of the PSFs and corresponding MTFs, for the pairs $(.125, .05), (.125, .125), (.125, .3)$ are shown in figure 8.21. Again smaller values of d produce a more sharply peaked PSF.



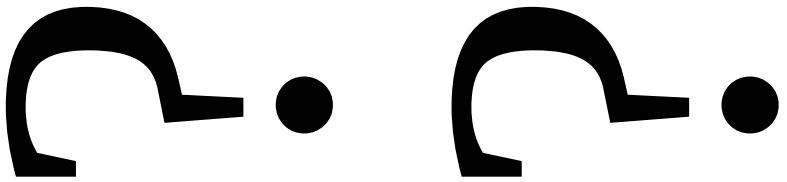
(a) Several PSFs.

(b) And their MTFs

Figure 8.21: Examples of PSF and MTF with Shepp-Logan regularization.

It is apparent in the graphs of the PSFs with exponential and Shepp-Logan regularization that these functions do not have long oscillatory tails and so the effect of convolving a piecewise continuous, bounded function with Ψ_0 should be an overall blurring, *without* oscillatory artifacts. They are absent because the MTF decays *smoothly* and sufficiently rapidly to zero. The PSFs obtained using a sharp cut-off in frequency have long oscillatory tails which, in turn produce serious Gibbs artifacts in the reconstructed images. Oscillatory artifacts can also result from sampling. This is considered in the following section. From both (8.46) and (8.47) it is clear that the roles of the beam width function $w(t)$ and the regularizing function $\psi(t)$ are entirely interchangeable in the total, unsampled PSF. This is no longer the case after sampling is done in the t-parameter.

These examples all indicate that, once the beam width is fixed, the full width half maximum of the PSF is not very sensitive to the sample spacing. However, smaller sample spacing produces a sharper peak which should in turn lead to less blurring in the reconstructed image. From the limiting case shown in figure 8.19(a) it is clear that the resolution is ultimately limited by the beam width. Since the PSF tends to infinity the FWHM definition of resolution is not applicable. Half of the volume under the PSF (as a radial function on $\mathbb{R}^2$) lies in the disk of radius $d/2$, indicating that the maximum available resolution, with the given beam profile is about half the width of the beam. This is in good agreement with experimental results that show that having several samples per beam width leads to a better reconstruction, though little improvement is seen beyond 4 samples per beam width, see [56] or [38]. To measure the resolution of a CT-machine or reconstruction algorithm it is customary to use a "resolution phantom." This is an array of disks of various sizes with various spacings. An example is shown in figure 8.22.



(a) A resolution phantom.

(b) Its reconstruction using a fan beam algorithm.

Figure 8.22: Resolution phantoms are used to gauge the resolution of a CT-machine or reconstruction algorithm.

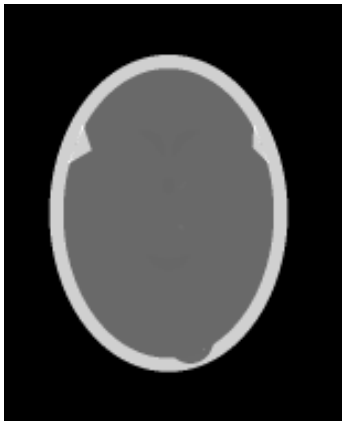
Exercise 8.6.2. Using the formula for $\mathrm{R}\delta_{(a,b)}$ derive the alternate expression for $\Psi_0(x,y)$:

$$\Psi_0(x,y) = \frac{1}{2\pi}\int_0^{\pi}\int_{-\infty}^{\infty} w(\langle(x,y),\omega\rangle) - s)\phi(s)dsd\omega. \tag{8.47}$$

Exercise 8.6.3. By considering the decay properties of the MTFs, in examples 8.6.2 and 8.6.3 explain why one does not expect the PSFs to have slowly decaying, oscillatory tails.

Exercise 8.6.4. Exercises on applying different measures of resolution to these PSFs.

8.6.2 The PSF with sampling



(a) A head phantom.

(b) Reconstruction using parallel beam data.

(c) Reconstruction using fan beam data.

Figure 8.23: Reconstructions of a mathematical phantom using filtered backprojection algorithms.

Real measurements entail both ray and view sampling. For a parallel beam machine, ray sampling refers to sampling in the t parameter and view sampling to the ω (or θ) parameter. For the sake of simplicity these effects are usually considered separately. We follow this procedure, first finding the kernel function incorporating ray sampling and then view sampling. Each produces distinct artifacts in the reconstructed image. As ray sampling is not a shift invariant operation, the measurement and reconstruction process can no longer be described by a single PSF, but instead requires a different integrand for each point in

the reconstruction grid. For the purpose of comparison, the PSFs are often evaluated at the center of reconstruction grid (i.e. $(0,0)$), though it is also interesting to understand how certain artifacts depend on the location of the input. In the previous section we obtained the PSF for unsampled data, with a reasonable filter function it was seen to produce an overall blurring of the image, without oscillatory effects. Both aliasing artifacts and the Gibbs phenomenon are consequences of slow decay in the Fourier transform which is typical of functions that change abruptly. One therefore would expect to see a lot of oscillatory artifacts produced by inputs with sharp edges. To test algorithms one typically uses the characteristic functions of disks or polygons placed at various locations in the image. Figure 8.23 shows a mathematical phantom and its reconstructions made with filtered backprojection algorithms. Note the oscillatory artifacts parallel to sharp boundaries as well as the pattern of oscillations in the exterior region.

Ray sampling

Suppose that d is the sample spacing in the t-parameter and that the image is reconstructed using a Ram-Lak (linearly interpolated) filter. For the purposes of this paragraph we suppose that sampling is only done in the affine parameter so that

$$\{\mathrm{R}_W f(jd,\omega) \,:\, j=-N,\ldots,N\}$$

is collected for all $\omega \in S^1$. With ϕ the Ram-Lak filter, the reconstructed image is

$$\tilde{f}_{\phi,w}(x,y) = \frac{1}{2\pi}\int_0^{2\pi} Q_{\phi,w}\tilde{f}(\langle (x,y),\omega\rangle,\omega)d\omega, \tag{8.48}$$

where

$$Q_{\phi,w}\tilde{f}(t,\omega) = d\sum_{j=-\infty}^{\infty}\phi(t-jd)\,\mathrm{R}_W f(jd,\omega).$$

For the derivation of the PSF it is very useful to let $f=\delta_{(a,b)}$, as noted above this implies that

$$\mathrm{R}_W f(jd,\omega) = w(jd-\langle (a,b),\omega\rangle).$$

The linear interpolation in the Ram-Lak filter is easiest to incorporate in the Fourier representation. If

$$\hat{\Phi}_d(r) = \sum_{j=-\infty}^{\infty}\phi(jd)e^{-ijdr}$$

then, as shown in section 8.3.4,

$$\hat{\phi}(r) = \mathrm{sinc}^2\left(\frac{rd}{2}\right)\hat{\Phi}_d(r).$$

The Fourier transform of $Q_{\phi,w}\tilde{f}$ in the t-variable is given by

$$\widetilde{Q_{\phi,w}\tilde{f}}(r,\omega) = d\,\mathrm{sinc}^2\left(\frac{rd}{2}\right)\hat{\Phi}_d(r)\sum_{j=-\infty}^{\infty} w(jd-\langle (a,b),\omega\rangle)e^{-ijdr}. \tag{8.49}$$

To evaluate the last sum we use the dual Poisson summation formula, (6.12), obtaining

$$d\sum_{j=-\infty}^{\infty} w(jd-\langle(a,b),\omega\rangle)e^{-ijdr}=e^{-i\langle(a,b),r\omega\rangle}\sum_{j=-\infty}^{\infty}\hat{w}(r+\frac{2\pi j}{d})e^{-i\frac{2\pi j}{d}\langle(a,b),\omega\rangle}. \tag{8.50}$$

For this computation to be valid we need to assume that both w and $\hat{w}$ decay sufficiently rapidly for the Poisson summation formula to be applicable. In particular w must be smoother than the functions, w_δ, used in example 8.6.2.

Using the Fourier inversion formula to express $Q_{\phi,w}\tilde{f}$ in (8.48) gives the PSF,

$$\begin{aligned}\Phi(x,y;a,b)=\frac{1}{[2\pi]^2}\int_0^\infty\int_0^{2\pi}\operatorname{sinc}^2\left(\frac{rd}{2}\right)\hat{\Phi}_d(r)e^{i\langle(x-a,y-b),r\omega\rangle}\times\\ \left[\sum_{j=-\infty}^{\infty}\hat{w}(r+\frac{2\pi j}{d})e^{-i\frac{2\pi j}{d}\langle(a,b),\omega\rangle}\right]drd\omega.\end{aligned} \tag{8.51}$$

It is quite apparent that Φ is not a function of $(x-a,y-b)$. Moreover the symmetry in roles played by ϕ and w has also been lost. The infinite sum in (8.51) leads to aliasing errors, resulting from the shape of the beam profile. A sharp beam profile produces larger errors. For a given beam profile, the sample spacing d must be selected so that the infinite sum

$$\sum_{j\neq 0}\hat{w}(r+\frac{2\pi j}{d})e^{-i\frac{2\pi j}{d}\langle(a,b),\omega\rangle}$$

is "small" for values of r where $\operatorname{sinc}^2\left(\frac{rd}{2}\right)$ is "large."

Example 8.6.4. As remarked above the very simple beam profile functions used in the previous paragraph cannot, strictly speaking be used in this discussion. We instead use w_δ convolved with a Gaussian. The window function w_δ models the detector while the Gaussian models the X-ray source. A Gaussian focal spot of "width" $a/\sqrt{2}$ is described by the function:

$$s_a(u)=\frac{1}{\pi a^2}e^{-\left(\frac{u}{a}\right)^2}.$$

For our examples we fix $a=\frac{1}{2}$ and $\delta=1$. It is reasonable to fix these parameters once and for all, as they model hardware which is difficult to adjust. The total beam profile is then

$$w(u)=\frac{1}{2}\int_{-1}^{1}s_{\frac{1}{2}}(u-v)dv,$$

with

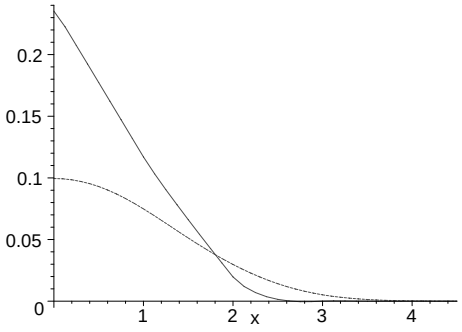
$$\hat{w}(r)=\operatorname{sinc}(r)e^{-\left(\frac{r^2}{16}\right)}.$$

We use the Shepp-Logan regularizing filter, for which

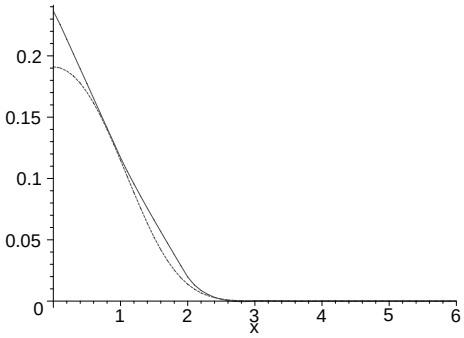
$$\hat{\Phi}_d(r)=|r\cdot\operatorname{sinc}\left(\frac{rd}{2}\right)|.$$

At issue here is the relationship between d, the sample spacing and the "width" of the w. Colloquially one asks for the "number of samples per beam width." With the given parameters, the FWHM(w) is very close to 2.

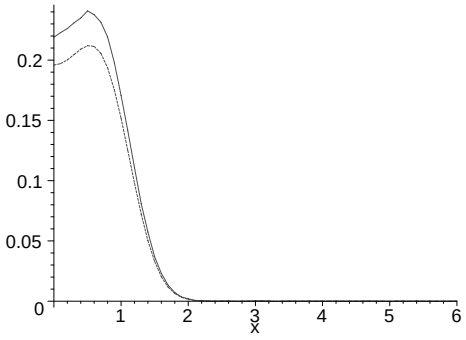
The overall filtering operation is no longer shift invariant. The function $\Phi(x, y; 0, 0)$ is a radial, figure (8.24) shows this function (of ρ) with various choices for d, with and without the effects of ray sampling. The dotted line is the unaliased PSF and the solid line the aliased. As before, smaller values of d give rise to sharper peaks in the PSF. The corresponding MTFs are shown in figure 8.24(d).



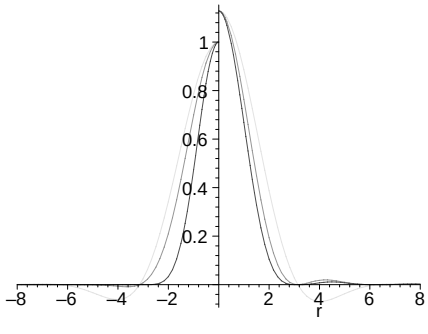
(a) $d = 2$



(b) $d = 1$



(c) $d = .25$



(d) The corresponding MTFs.

Figure 8.24: The effect of ray sampling on the PSF.

The graphs on the right hand side include the effect of aliasing while those on the left are the unaliased MTFs, as d decreases, the "passband" of the MTF broadens. With this choice of beam profile and regularizing filter, once there is at least one sample per beam width, the resolution, measured by FWHM, is not affected very much by aliasing. Though it is not evident in the pictures, these PSFs have long oscillatory tails. The very

small amplitude of these tails is a result of using a smooth, rapidly decaying regularizing function.

Exercise 8.6.5. When doing numerical computations of Φ it can be very useful to use the fact that

$$d \sum_{j=-\infty}^{\infty} w(jd - \langle (a,b), \omega \rangle) e^{-ijdr}$$

is a periodic function. Explain this observation and describe how it might figure in a practical computation. It might be helpful to try to compute this sum using both representations.

Exercise 8.6.6. Continue the computation begun in example 8.6.4 and draw graphs of $\Phi(x, y; a, b)$ for some $(a, b) \neq (0, 0)$. Note that Φ is no longer a radial function of (x, y).

8.6.3 View sampling

We now turn to the artifacts which result from using finitely many views and begin by considering the reconstruction of a mathematical phantom made out of constant density elliptical regions. In figure 8.25 note the pattern of oscillations in the exterior region along lines, tangent to the boundary of ellipse and the absence of such oscillations in the interior. A somewhat subtler observation is that the very pronounced, coherent pattern of oscillations does not begin immediately but rather at a definite distance from the boundary of the ellipse. This phenomenon is a consequence of the sampling in the angular parameter and the filtering operations needed to approximately invert the Radon transform. Our discussion of these examples closely follows that in [70].

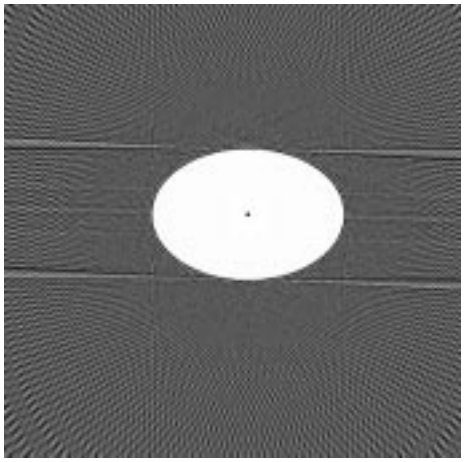


Figure 8.25: Filtered backprojection reconstruction of elliptical phantom

Example 8.6.5. Suppose the object, E is of constant density 1 with boundary the locus of points, $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$. The line integral of $f = \chi_E(x, y)$ along a line (t, θ) in simply the length of the intersection of the line with E. Let $s_\pm(t, \theta)$ denote the s-parameters for the intersection points of the line $l_{t,\theta}$ with the boundary of E. The distance between these two

points $|s_+(t,\theta) - s_-(t,\theta)|$ is the Radon transform of f. Plugging the parametric form of the line into the equation for the ellipse and expanding gives

$$s^2(\frac{\sin^2\theta}{a^2} + \frac{\cos^2\theta}{b^2}) + 2st\sin\theta\cos\theta(\frac{1}{b^2} - \frac{1}{a^2}) + \frac{t^2\cos^2\theta}{a^2} + \frac{t^2\sin^2\theta}{b^2} - 1 = 0.$$

Re-write the equation as

$$p(t,\theta)s^2 + q(t,\theta)s + r(t,\theta) = 0,$$

p, q and r are the corresponding coefficients. The two roots are given by

$$s_\pm = \frac{-q \pm \sqrt{q^2 - 4ps}}{2p},$$

the distance between the two roots is therefore

$$s_+ - s_- = \frac{\sqrt{q^2 - 4ps}}{p}.$$

This gives the formula for $\mathrm{R}f$:

$$\mathrm{R}f(t,\omega(\theta)) = \begin{cases} 2\beta(\theta)\sqrt{\alpha(\theta)^2 - t^2} & |t| \le \alpha(\theta), \\ 0 & |t| > \alpha(\theta), \end{cases}$$

where

$$\alpha(\theta) = \sqrt{\frac{a^4\cos^2(\theta) + b^4\sin^2(\theta)}{a^2\cos^2(\theta) + b^2\sin^2(\theta)}},$$

$$\beta(\theta) = \sqrt{\frac{(a^2\cos^2(\theta) + b^2\sin^2(\theta))(b^2\cos^2(\theta) + a^2\sin^2(\theta))}{a^4\cos^2(\theta) + b^4\sin^2(\theta)}}.$$

Both α and β are smooth, non-vanishing functions of θ.

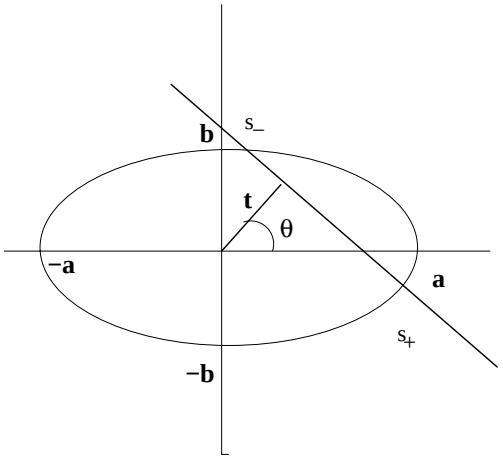


Figure 8.26: Parameters describing the Radon transform of χ_E.

Doing the exact filtration,

$$Qf(t,\omega(\theta)) = \frac{1}{i}\mathcal{H}\partial_t \,\mathrm{R}\, f(t,\omega(\theta))$$

gives

$$Qf(t,\omega(\theta)) = \begin{cases} 2\beta(\theta) & |t| \le \alpha(\theta), \\ 2\beta(\theta)(1 - \frac{|t|}{\sqrt{t^2-\alpha^2(\theta)}}) & |t| > \alpha(\theta). \end{cases}$$

In an actual reconstruction a regularizing factor is used to compute $Q_\phi f$. Approximating the backprojection with a Riemann sum gives

$$\tilde{f}_\phi(x,y) \approx \frac{1}{2(M+1)} \sum_{k=0}^{M} Q_\phi f(\langle (x,y), \omega(k\Delta\theta)\rangle, \omega(k\Delta\theta)).$$

For points (x,y) inside the ellipse

$$\tilde{f}_\phi(x,y) \approx \frac{1}{2(M+1)} \sum_{k=0}^{M} \beta(k\Delta\theta).$$

This is well approximated by the Riemann sum because $\beta(\theta)$ is a smooth bounded function. For the points (x,y) outside the ellipse there are three types of lines:

1 Lines which pass through E,

2 Lines which are distant from E,

3 Lines outside E which pass very close to bE.

The first two types of lines are not problematic. However for any point in the exterior of the ellipse, the backprojection involves lines which are exterior to the ellipse but pass very close to it. This leads to the oscillations apparent in the reconstruction along lines tangent to the boundary of the regions. This is a combination of the Gibbs phenomenon and aliasing. To compute an accurate value for $\tilde{f}_\phi(x,y)$ at an exterior point requires delicate cancelations between the positive and negative values assumed by $Q_\phi f$. For points near enough to the boundary there is a sufficient density of samples to obtain the needed cancelations. For more distant points the cancelation does not occur and the pattern of oscillations appears. In the next paragraph we derive a "far field" approximation to the reconstruction of such a phantom. This gives, among other things, a formula for the radius where the oscillatory artifacts first appear. In [36] such a formula is derived for a fan beam scanner, using their approach we derive a similar formula for the parallel beam case. Also apparent in figure 8.25 is an oscillation very near and *parallel* to the boundary of E; this is the usual combination of the Gibbs phenomenon and aliasing caused by ray sampling.

Example 8.6.6. An very striking example of this phenomenon can be seen in the reconstruction of a rectangular region. If f is the characteristic function of the square with vertices

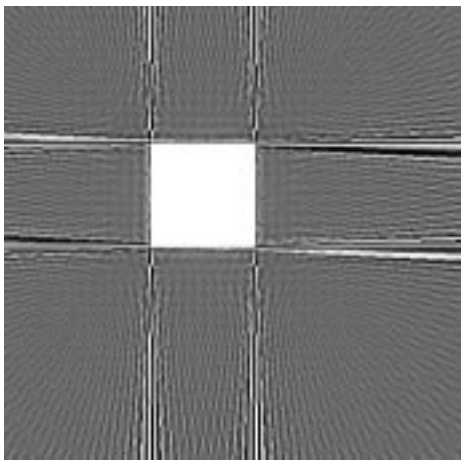


Figure 8.27: Filtered backprojection reconstruction of square phantom

$\{(\pm 1, \pm 1)\}$ then for $|\theta| < \frac{\pi}{4}$

$$\mathrm{R}f(t,\omega(\theta)) = \begin{cases} 0 & \text{if } |t| > \cos\theta + \sin\theta, \\ \frac{\cos\theta+\sin\theta-t}{\cos\theta\sin\theta} & \text{if } \cos\theta - \sin\theta < t < \cos\theta + \sin\theta, \\ \frac{2}{\cos\theta} & \text{if } \sin\theta - \cos\theta < t < \cos\theta - \sin\theta, \\ \frac{\cos\theta+\sin\theta+t}{\cos\theta\sin\theta} & \text{if } -(\cos\theta + \sin\theta) < t < \sin\theta - \cos\theta. \end{cases}$$

The function $\mathrm{R}f(t,\omega(\theta))$ is periodic of period $\frac{\pi}{2}$ in the θ parameter. If $\theta \neq 0$ then $\mathrm{R}f(t,\theta)$ is a continuous, piecewise differentiable function in θ, whereas

$$\mathrm{R}f(t,(1,0)) = \chi_{[-2,2]}(t)$$

has a jump discontinuity. Note the pronounced oscillatory artifact in the exterior of the square along lines tangent to the sides of the square in figure 8.27. As before there is also a Gibbs oscillation in the reconstructed image, parallel to the boundary of the square.

Exercise 8.6.7. Derive the formula for $\mathrm{R}f$ and Qf in example 8.6.5.

Exercise 8.6.8. Compute $-i\mathcal{H}\partial_t \mathrm{R}f$ for example 8.6.6.

A far field approximation for the reconstructed image

In this paragraph we obtain an approximate formula for the Ram-Lak reconstruction of a small radially symmetric object at points far from the object. Let ϕ denote a Ram-Lak filter function and w a function describing the source-detector pair. If f is the data then the approximate, filtered backprojection reconstruction is given by

$$f_{\phi,w}(x,y) = \frac{1}{2\pi}\int_0^\pi Q_{\phi,w}f(\langle(x,y),\omega),\omega)d\omega,$$

where

$$Q_{\phi,w}f(t,\omega) = \int_{-\infty}^{\infty} \mathrm{R}_W f(s,\omega)\phi(t-s)ds.$$

Here $\mathrm{R}_W f$ denotes the "averaged Radon transform" of f. We now consider the effect of sampling in the ω-parameter, leaving t as a continuous parameter. Equation (8.42) shows that, in this case, ϕ and w are interchangeable; the effects of finite beam width and regularizing the filter are both captured by using $\phi * w$ as the filter function. We analyze the difference between $f_{\phi,w}(x,y)$ and the Riemann sum approximation

$$\tilde{f}_{\phi,w}(x,y) = \frac{1}{4\pi}\sum_{j=0}^{M} Q_{\phi * w}f(\langle (x,y), \omega(j\Delta\theta)\rangle, \omega(j\Delta\theta))\Delta\theta.$$

For simplicity we restrict attention to functions of the form

$$f^{(a,b)}(x,y) = f(\|(x,y)-(a,b)\|).$$

The averaged Radon transform of f, is independent of ω and therefore

$$\mathrm{R}_W f^{(a,b)}(t,\omega) = \mathrm{R}_W f(t - \langle (a,b),\omega\rangle),$$

moreover

$$Q_{\phi * w}f^{(a,b)}(t,\omega) = Q_{\phi * w}f(t - \langle (a,b),\omega\rangle). \tag{8.52}$$

From equation (8.52) we obtain

$$f^{(a,b)}_{\phi,w}(x,y) = \frac{1}{2\pi}\int_0^{\pi} Q_{\phi * w}f(\langle (x,y)-(a,b),\omega\rangle)d\omega.$$

Letting $(x-a, y-b) = R(\cos\varphi, \sin\varphi)$ and $\omega(\theta) = (\cos\theta, \sin\theta)$ gives

$$f^{(a,b)}_{\phi,w}(x,y) = \frac{1}{2\pi}\int_0^{\pi} Q_{\phi * w}f(R\cos(\theta-\varphi))d\theta$$

as well as the Riemann sum

$$\tilde{f}^{(a,b)}_{\phi,w}(x,y) = \frac{1}{2\pi}\sum_{j=0}^{M} Q_{\phi * w}f(R\cos(j\Delta\theta-\varphi))\Delta\theta. \tag{8.53}$$

The objects of principal interest in this analysis are small hard objects. From the examples presented in the previous paragraph, our primary interest is the reconstruction at points in the exterior of the object. The function $f^{(a,b)}_{\phi,w}$ is an approximation to $f * k$, where k is the inverse Radon transform of w. If w has bounded support and ϕ provides a good approximation to $-i\partial_t\mathcal{H}$, then $f^{(a,b)}_{\phi,w}(x,y)$ should be very close to zero for points outside the support of $f * k$. Indeed, if w has small support then so does k and therefore

the support of $f * k$ is a small enlargement of the support of f itself. We henceforth assume that, for points of interest,

$$f_{\phi,w}^{(a,b)}(x,y) \approx 0 \tag{8.54}$$

and therefore any deviation of $\tilde{f}_{\phi,w}^{(a,b)}(x,y)$ from zero is an error.

If f is the characteristic function of a disk of radius r then $Q_{\phi,w}f(t)$ falls off rapidly for $|t| >> r$, see example 8.3.1. There is a j_0 so that

$$\begin{aligned} j_0\Delta\theta - \varphi &< \frac{\pi}{2}, \\ (j_0+1)\Delta\theta - \varphi &\geq \frac{\pi}{2}. \end{aligned} \tag{8.55}$$

If we let $\Delta\varphi = \frac{\pi}{2} - j_0\Delta\theta + \varphi$ then $0 < \Delta\varphi \leq \Delta\theta$ and

$$\tilde{f}_{\phi,w}^{(a,b)}(x,y) = \frac{1}{2\pi}\sum_{j=0}^{M} Q_{\phi * w}f(R\sin(j\Delta\theta - \Delta\varphi))\Delta\theta.$$

As the important terms in this sum are those with $|j|$ close to zero, we approximate it by using

$$\sin(j\Delta\theta - \Delta\varphi) \approx j\Delta\theta - \Delta\varphi$$

obtaining,

$$\tilde{f}_{\phi,w}^{(a,b)}(x,y) \approx \frac{1}{2\pi}\sum_{j=-\infty}^{\infty} Q_{\phi * w}f(R(j\Delta\theta - \Delta\varphi))\Delta\theta.$$

The limits of summation have also been extended from $-\infty$ to ∞. The error this introduces is small as $\phi(t) = O(t^{-2})$.

The Poisson summation formula can be used to evaluate the last expression, it gives

$$\tilde{f}_{\phi,w}^{(a,b)}(x,y) \approx \frac{1}{2\pi R}\sum_{j=-\infty}^{\infty} \widetilde{Q_{\phi * w}f}\left(\frac{2\pi j}{R\Delta\theta}\right) e^{-\frac{2\pi i j\Delta\varphi}{\Delta\theta}}.$$

From the central slice theorem

$$\widetilde{Q_{\phi * w}f}(\rho) = \hat{\phi}(\rho)\hat{w}(\rho)\hat{f}(\rho).$$

Assuming the $\hat{\phi}(0) = 0$ and that w is a even function gives the simpler formula

$$\tilde{f}_{\phi,w}^{(a,b)}(x,y) \approx \frac{1}{\pi R}\sum_{j=1}^{\infty} \hat{\phi}\cdot\hat{w}\cdot\hat{f}\left(\frac{2\pi j}{R\Delta\theta}\right)\cos\left(\frac{2\pi j\Delta\varphi}{\Delta\theta}\right). \tag{8.56}$$

In order for this sum to be negligible at a point whose distance to (a,b) is R, the angular sample spacing, $\Delta\theta$ must be chosen so that the effective support of $\hat{\phi}\cdot\hat{w}\cdot\hat{f}$ is contained in

$$\left(-\frac{2\pi}{R\Delta\theta}, \frac{2\pi}{R\Delta\theta}\right).$$

This explains why the oscillatory artifacts only appear at points the are at a definite distance from the object: for small values of R the sum itself is very small however, for sufficiently large R the terms of the sum start to become significant.

Suppose, for example that $R\Delta\theta$ is such that all but the first term in this sum are negligible, then

$$\tilde{f}^{(a,b)}_{\phi,w}(x,y) \approx \frac{1}{\pi R}\hat{\phi}\cdot\hat{w}\cdot\hat{f}\left(\frac{2\pi}{\|(x,y)-(a,b)\|\Delta\theta}\right)\cos\left(\frac{2\pi\Delta\varphi}{\Delta\theta}\right). \tag{8.57}$$

The cosine factor produces an oscillation in the sign of the artifact whose period equals $\Delta\theta$. This quite apparent in figures 8.28 and 8.29. The amplitude of the artifact depends on the distance to the object through the product $\hat{\phi}\cdot\hat{w}\cdot\hat{f}$. This allows us to relate the angular sample spacing, needed to obtain an artifact free reconstruction in a disk of given radius, to the source-detector function w. For simplicity suppose that

$$w(u) = \frac{1}{2\delta}\chi_{[-\delta,\delta]}(u) \text{ so that } \hat{w}(\rho) = \operatorname{sinc}(\rho\delta).$$

The first zero of $\hat{w}$ occurs at $\rho = \pm\frac{\pi}{\delta}$ which suggests that taking

$$\Delta\theta < \frac{2\delta}{R}$$

is a minimal requirement to get "artifact free" reconstructions in the disk of radius R. This ignores the possible additional attenuation of the high frequencies which results from $\hat{\phi}$, which is consistent with our desire to get a result which is independent of the sample spacing in the t-parameter. The estimate for $\Delta\theta$ can be re-written

$$\frac{\pi R}{2\delta} < \frac{\pi}{\Delta\theta}.$$

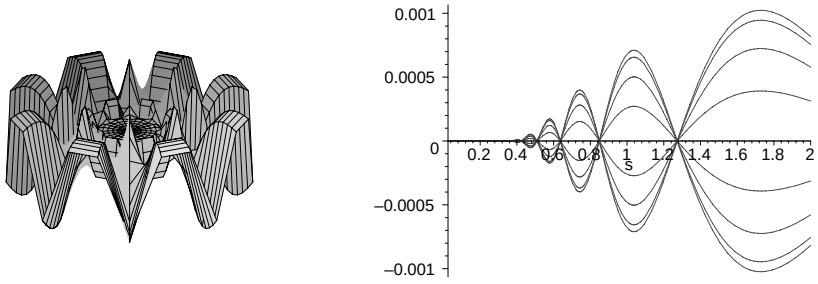
The quantity on the right hand side is the number of samples, $M+1$ in the ω direction. As 2δ is the width of the source, Nyquist's theorem implies that the maximum spatial frequency available in the data is about $(4\delta)^{-1}$. If we denote this by ν then the estimate reads

$$2\pi R\nu < M+1.$$

Essentially the same result was obtained in section 8.3.5, with much less effort! The difference in the analyses is that, in the earlier discussion, it was *assumed* that the data is essentially bandlimited to $[\frac{-\pi}{2\delta},\frac{\pi}{2\delta}]$. Here this bandlimiting is a consequence of the low pass filtering which results from averaging over the width of the X-ray beam.

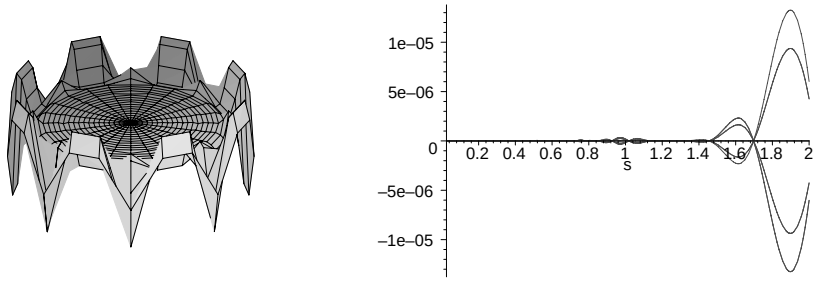
It is very important to note that the artifacts which result from view sampling are present whether or not the data is sampled in the t-parameter. These artifacts can be reduced by making either ϕ or w smoother. This is in marked contrast to the result obtained for ray sampling. In that case the aliasing errors are governed solely by w and cannot be reduced by changing ϕ. If f describes a smooth object, so that $\hat{f}$ decays rapidly, then it is unlikely that view sampling, aliasing artifacts will appear in the reconstruction region.

Example 8.6.7. To understand formula (8.57) we consider $(a, b) = (0, 0)$, and $f(x, y) = \chi_{D_{\frac{1}{10}}}(x, y)$. For simplicity we use w_1 and the Shepp-Logan filter with $d = .25$. We consider the right hand side of 8.57 with $\Delta\theta = \frac{2\pi}{8}$ and $\frac{2\pi}{32}$. The 3-dimensional plots give an idea of how the artifacts appear in a reconstructed image. Notice that the sign of the error reverses along a circle. The ordinary graphs are sections of the 3d-plot along lines of constant φ. These graphs allow for a quantitative appreciation for the size of the errors and their dependence on $\Delta\theta$.



(a) 3d-plot. (b) Sections along radii.

Figure 8.28: View sampling artifacts with $\Delta\theta = \frac{2\pi}{8}$

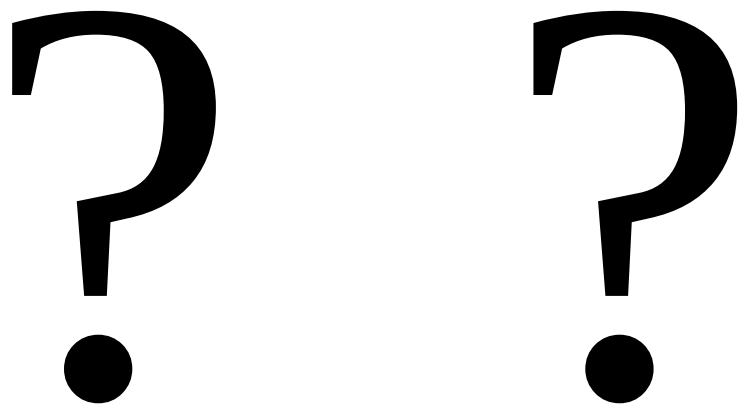


(a) 3d-plot. (b) Sections along radii.

Figure 8.29: View sampling artifacts with $\Delta\theta = \frac{2\pi}{32}$

Remark 8.6.1. In [36] a similar analysis is presented for a fan beam machine. The results are similar though a bit more complicated. The artifacts produced by view sampling in a fan beam machine are different in important ways: In a parallel beam machine the pattern of oscillations is circularly symmetric and depends on the distance from the center of a

radially symmetric object. For a fan beam machine the pattern displays a similar circular symmetry but the center of the circle no longer agrees, in general, with the center of the object, see figure 8.30.



(a) Parallel beam. (b) Fan beam.

Figure 8.30: Examples comparing view aliasing in parallel beam and fan beam scanners.

Exercise 8.6.9. Given that we use the Poisson summation formula, why is it still possible to use $w(u) = (2\delta)^{-1}\chi_{[-\delta,\delta]}(u)$ in this analysis.

Exercise 8.6.10. Show that the PSF for the Ram-Lak reconstruction, incorporating the beam width function w and sampling in the ω-parameter is

$$\Phi(x, y; a, b) = \frac{\Delta\theta}{2\pi} \sum_{j=0}^{M} \phi * w(\langle (x - a, y - b), \omega(j\Delta\theta)\rangle).$$

Note that this operation is shift invariant.

8.6.4 Bad rays versus bad views

The artifacts considered in the previous sections are algorithmic artifacts, resulting from the sampling and approximation used in any practical reconstruction method. The final class of linear artifacts we consider are the effects of *systematic* measurement errors. This should be contrasted to the analysis in Chapter 12 of the effects of *random* measurement errors or noise. Recall that the measurements made by a CT-machine are grouped into views and each view is comprised of a collection of rays. We now consider the consequences of having a bad ray in a single view, a bad ray in every view and a single bad view. These analyses illustrate a meta-principle, called the *smoothness principle*, often invoked in medical imaging,

- The filtered backprojection algorithm is very sensitive to errors that vary abruptly from ray to ray or view to view and is relatively tolerant of errors that very gradually. This feature is a reflection of the fact that the filter function, $\hat{\phi}(\xi) \approx |\xi|$, attenuates low frequencies and amplifies high frequencies.

Our discussion is adapted from [71], [34] and [35].

For this analysis we suppose that the measurements, made with a parallel beam scanner, are the samples

$$\{P(t_j, \omega(k\Delta\theta)), k = 0, \ldots, M, j = 1, \ldots, N\}$$

of $\mathrm{R}_W f(t, \omega)$. The coordinates are normalized so that the object lies in $[-1, 1] \times [-1, 1]$. The angular sample spacing is

$$\Delta\theta = \frac{\pi}{M+1}$$

and the rays are uniformly sampled at

$$t_j = -1 + (j - \frac{1}{2})d.$$

If ϕ is the filter function, which is specified at the sample points and linearly interpolated in between, then the approximate reconstruction is given by

$$\tilde{f}_\phi(x, y) = \frac{1}{N(M+1)} \sum_{k=0}^{M} \sum_{j=1}^{N} P(t_j, \omega(k\Delta\theta))\phi(x\cos(k\Delta\theta) + y\sin(k\Delta\theta) - t_j). \qquad (8.58)$$

For the purposes of comparison we use the Shepp-Logan filter

$$\phi(0) = \frac{4}{\pi d^2}, \quad \phi(jd) = \frac{-4}{\pi d^2(4j^2 - 1)}.$$

A single bad ray

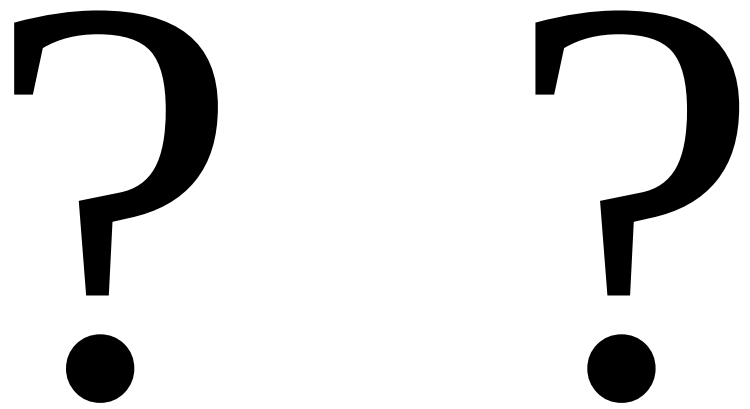
The first effect we consider is an isolated measurement error in a single ray, from a single view. Suppose that $P(t_{j_0}, \omega(k_0\Delta\theta))$ differs from the "true" value by ϵ. As formula (8.58) is linear, this measurement error produces a reconstruction error at (x, y) equal to

$$\Delta\tilde{f}_\phi(x, y) = \frac{\epsilon\phi(x\cos(k_0\Delta\theta) + y\sin(k_0\Delta\theta) - t_{j_0})}{N(M+1)} = \frac{\epsilon}{N(M+1)}\phi(\mathrm{dist}((x, y), l_{t_{j_0}, \omega(k_0\Delta\theta)})).$$

The effect of this error at (x, y) depends on the distance from (x, y) to the "bad ray," $l_{t_{j_0}, \omega(k_0\Delta\theta)}$. In light of the form of the function ϕ the error is worst along the ray itself, where it equals

$$\frac{\epsilon N}{\pi(M+1)}.$$

Figure 8.31(a) shows the Shepp-Logan phantom and figure 8.31(b) shows the Shepp-Logan phantom with errors ?????? in the ????? ray.



(a) Shepp-Logan phantom. (b) A few bad rays.

Figure 8.31: A reconstruction with a few bad rays.

A bad ray in each view

A bad ray might result from a momentary surge in the output of the X-ray tube. If, on the other hand, a single detector in the detector array is malfunctioning then the same ray in each view will be in error. Let ϵ_k denote the error in $P(t_{j_0}, \omega(k\Delta\theta))$. In light of the linearity of (8.58), the error at (x, y) is now

$$\Delta\tilde{f}_\phi(x,y) = \sum_{k=0}^{M} \frac{\epsilon_k \phi(x\cos(k\Delta\theta) + y\sin(k\Delta\theta) - t_{j_0})}{N(M+1)} \tag{8.59}$$

If $(x, y) = r(\cos\varphi, \sin\varphi)$ in polar coordinates and $\epsilon = \epsilon_k$ for all k then

$$\begin{aligned} \Delta\tilde{f}_\phi(x,y) &= \sum_{k=0}^{M} \frac{\epsilon\phi(r\cos(\varphi - k\Delta\theta) - t_{j_0})}{N(M+1)} \\ &\approx \frac{\epsilon}{\pi N} \int_0^\pi \phi(r\cos(\varphi - s) - t_{j_0})ds. \end{aligned} \tag{8.60}$$

Because the function ϕ is sharply peaked at zero and

$$\int_{-\infty}^{\infty} \phi(s)ds = 0,$$

this artifact is worst for points where $r = t_{j_0}$ and $0 < \varphi < \pi$. At other points the integrand is either uniformly small or the integral exhibits a lot of cancelation. Due to the periodicity of the integrand, it is clear that the largest error occurs where $r = t_{j_0}$ and $\varphi = \frac{\pi}{2}$. The

reason the error is only large in half the circle is that samples are only collected for $0 \leq \theta \leq \pi$. If data is collected over the full circle then the result of an error ϵ in the j_0^{th} ray is approximately

$$\Delta \tilde{f}_\phi \approx \frac{\epsilon}{2\pi N} \int_0^{2\pi} \phi(r\cos(s) - t_{j_0}) ds. \tag{8.61}$$

If the data is collected over half the circle then $[4\pi]^{-1}\Delta\tilde{f}_\phi$ is the *average* error for points on the circle of radius r.

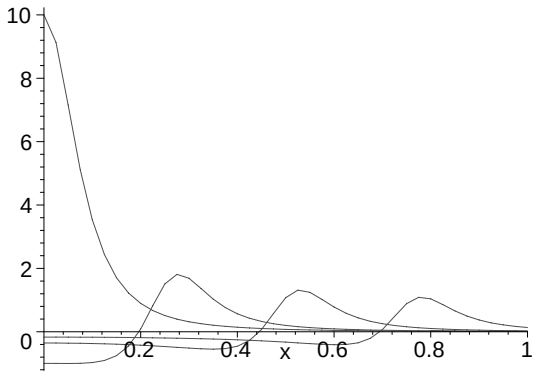
Since all the bad rays have the same affine parameter, t_{j_0}, they are all tangent to the circle, centered at $(0,0)$ of radius t_{j_0}. In [71] it is shown that the average error along this circle is given approximately by

$$\frac{\epsilon\sqrt{N}}{\pi^2\sqrt{t_{j_0}}} \text{ if } t_{j_0} >> 0 \text{ and } \frac{\epsilon N}{\pi} \text{ if } t_{j_0} = 0.$$

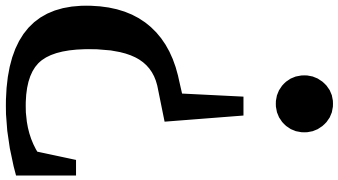
Figure 8.32(a) shows graphs of the average error as a function of r with $t_{j_0} = 0, .25, .5$ and $.75$. The integral in (8.61) is difficult to numerically evaluate if ϕ is a linearly interpolated function. Instead we have used the approximate filter function

$$\phi(t) = \frac{d^2 - t^2}{(d^2 + t^2)^2},$$

which was introduced in section 4.4. These graphs bear out the prediction that the error is worst where $r = t_{j_0}$, however the sharpness of the peak also depends on t_{j_0}. Figure 8.32 shows a reconstruction of the Shepp-Logan phantom with an error of size ???? in the ???? ray.



(a) Radial graphs of the average error.



(b) The Shepp-Logan phantom with a bad ray.

Figure 8.32: A systematic bad ray with $\Delta\theta = \frac{2\pi}{8}$.

Remark 8.6.2. In a third generation machine a single bad detector would result in the situation we have just analyzed: the measurement of the same ray would be erroneous in every view. This is because a view, for a third generation machine, is determined by the *source* position. In a fourth generation scanner a mis-calibrated detector could instead result in every ray from a single *view* being in error. This is because a view is determined, in a fourth generation machine, by a *detector*.

Exercise 8.6.11. Explain why the error is only large in a semi-circle if samples are only collected for $0 \leq \theta \leq \pi$.

Exercise 8.6.12. Provide a more detailed explanation for the smallness of the errors in the half of the circle where $\pi < \varphi < 2\pi$.

A bad view

We now analyze the effect on the reconstruction of having an error in every measurement of a single view. For simplicity we consider the consequence of such an error using the parallel beam algorithm. Suppose that ϵ_j is the error in the measurement $P(t_j, \omega(k_0\Delta\theta))$, then the reconstruction error at (x, y) is

$$\Delta\tilde{f}_\phi(x,y) = \frac{1}{N(M+1)} \sum_{j=1}^{N} \epsilon_j \phi(\langle (x,y), \omega(k_0\Delta\theta)\rangle - t_j).$$

If the error $\epsilon_j = \epsilon$ for all rays in the k_0^{th} view then the error can be approximated by an integral

$$\Delta\tilde{f}_\phi(x,y) \approx \frac{\epsilon}{2(M+1)} \int_{-1}^{1} \phi(\langle (x,y), \omega(k_0\Delta\theta)\rangle - t)dt. \tag{8.62}$$

As before, the properties of ϕ make this artifact most severe at the points $(x, y) = \pm\omega(k_0\Delta\theta)$ and least severe along the line

$$\langle (x,y), \omega(k_0\Delta\theta)\rangle = 0.$$

This is the "central ray" of the k_0^{th} view. Using the facts that $\phi(t) = \phi(-t)$ and that the total integral vanishes,

$$\int_{-\infty}^{\infty} \phi(t)dt = 0,$$

we conclude that the worst error

$$\Delta\tilde{f}_\phi(\pm\omega(k_0\Delta\theta)) \approx \frac{\epsilon}{4(M+1)}\phi(0) = \frac{\epsilon N^2}{8\pi(M+1)}.$$

On the other hand for points near to $(0, 0)$ the error is approximately

$$\frac{\epsilon}{2(M+1)} \int_{-1+\delta}^{1+\delta} \phi(t)dt,$$

where δ is the distance between (x, y) and the central ray of the bad view. The integral from $(-1+\delta)$ to $(1-\delta)$ is approximated by $\frac{2}{\pi(1-\delta)}$ and the integral from $(1-\delta)$ to $(1+\delta)$ is $O(\delta)$, hence

$$\Delta\tilde{f}_\phi(x,y) \approx \frac{\epsilon}{\pi(M+1)}\frac{1}{1-\delta}.$$

Figure 8.33 shows the Shepp-Logan phantom with a systematic error of size $\epsilon =$???? in view ????. This is a slowly varying error which grows to be quite large near to the boundary of the reconstruction region. Nonetheless, comparing this image with 8.32(b) shows that it is visually less disturbing than the result of a systematic bad ray. This example demonstrates the smoothness principle.

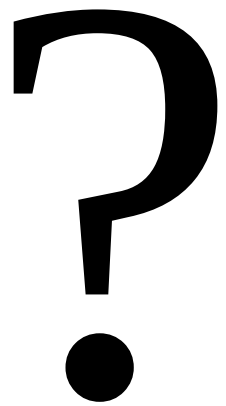


Figure 8.33: A reconstruction with one bad view.

Remark 8.6.3. Many other artifacts have been analyzed in [71] and [35]. We have selected examples that have a fairly simple mathematical structure and illustrate the usage of the tools developed in the earlier chapters. In large part due to their successful analyses, these artifacts are largely absent from modern CT-images.

Exercise 8.6.13. Justify this approximation for the integral in (8.62).

Exercise 8.6.14. Suppose that *every* measurement is off by ϵ. Show that the reconstructed image has a error

$$\Delta\tilde{f}_\phi(x,y) \approx \frac{\epsilon}{\pi\sqrt{1-x^2-y^2}}.$$

8.6.5 Beam hardening

We close our discussion of imaging artifacts with a very short discussion of beam hardening. Because it is non-linear, beam hardening is qualitatively quite different from the foregoing phenomena. It is, instead rather similar to the partial volume effect. Beam hardening is caused by the fact that the X-ray beam is not monochromatic and the absorption coefficient, depends, in a non-trivial way, on the energy of the incident beam. Recall that an actual measurement is the ratio I_i/I_o, where I_i is the *total* energy in the incident beam and I_o is

the total energy in the output. The energy content of the X-ray beam is described by its *spectral function*, $S(\mathcal{E})$; it satisfies

$$I_i = \int_0^\infty S(\mathcal{E}) d\mathcal{E}.$$

If a (thin) X-ray beam is directed along the line $l_{t,\omega}$, then the measured output is

$$I_{o,(t,\omega)} = \int_0^\infty S(\mathcal{E}) \exp\left[- \int_{-\infty}^{\infty} f(s\hat{\omega} + t\omega; \mathcal{E}) ds \right] d\mathcal{E}.$$

Here $f(x, y; \mathcal{E})$ is the attenuation coefficient, with its dependence on the energy explicitly noted. A typical spectral function is shown in figure 2.7. Due to this non-linear distortion, the raw measurements are *not* the Radon transform of f; in the imaging literature it is often said that such measurements are *inconsistent*. Applying the Radon inversion formula to such data leads to streaking in the reconstructed image, see figure 8.34.

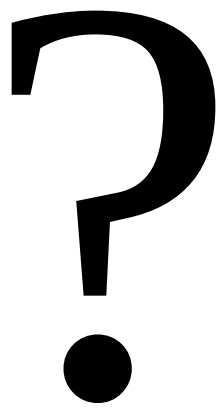


Figure 8.34: Streaks caused by beam hardening.

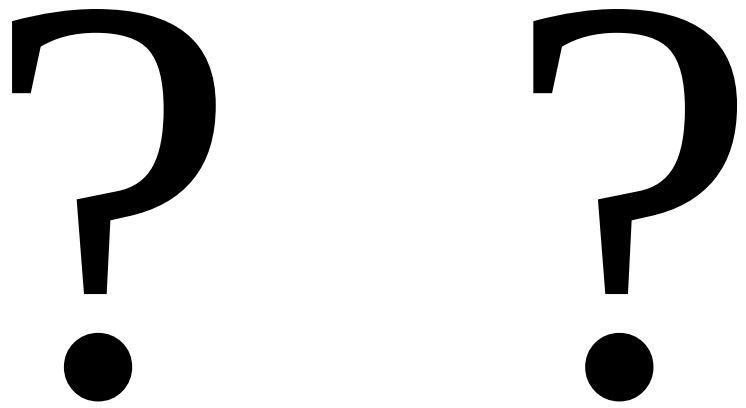
Suppose that D is a bounded object whose absorption coefficient, $f(\mathcal{E})$ only depends on the energy. Even in this case, the function,

$$\log\left[\frac{I_i}{I_{o,(t,\omega)}} \right]$$

is *not* a linear function of length of the intersection of the line $l_{t,\omega}$ with D. If T denotes the length of this line segment then

$$\log\left[\frac{I_i}{I_{o,(t,\omega)}} \right] = H_f(T) \stackrel{d}{=} \log\left[\frac{\int S(\mathcal{E}) e^{-Tf(\mathcal{E})} d\mathcal{E}}{\int S(\mathcal{E}) d\mathcal{E}} \right]. \tag{8.63}$$

Because $S(\mathcal{E})$ and $f(\mathcal{E})$ are non-negative functions it is immediate that $H_f(T)$ is a strictly, monotonely decreasing function. This implies that the inverse function, H_f^{-1} is well defined. Thus by measuring or computing $H_f(T)$, for T in the relevant range, its inverse function can be tabulated. The absorption coefficient of water, $f_w(\mathcal{E})$ as well as $H_w(T)$, for a typical spectral function, are shown in figure 8.35(a-b).



(a) The absorption coefficient of water.

(b) The non-linear absorption function for water.

Figure 8.35: Beam hardening through water.

Using H_f^{-1} the Radon transform of $\chi_D(x, y)$ can be determined from X-ray absorption measurements

$$\mathrm{R}\chi_D(t, \omega) = H_f^{-1}\left(\log\left[\frac{I_i}{I_{o,(t,\omega)}}\right]\right). \tag{8.64}$$

The region D can now be determined using the methods described above for approximately inverting the Radon transform.

The absorption coefficients of the soft tissues in the human body are very close to that of water and their dependence on the energy is almost identical. If $f_w(\mathcal{E})$ is the absorption coefficient of water then this hypothesis amounts to the statement that the ratio

$$\rho(x, y) = \frac{f(x, y; \mathcal{E})}{f_w(\mathcal{E})}$$

is essentially independent of the energy. Let H_w denote the function defined in (8.63) with $f = f_w$. For slices which contain little or no bone the function H_w^{-1} can be used as in (8.64) to correct for beam hardening. This substantially reduces the inconsistencies in the measurements and allows the usage of the Radon formalism to reconstruct ρ.

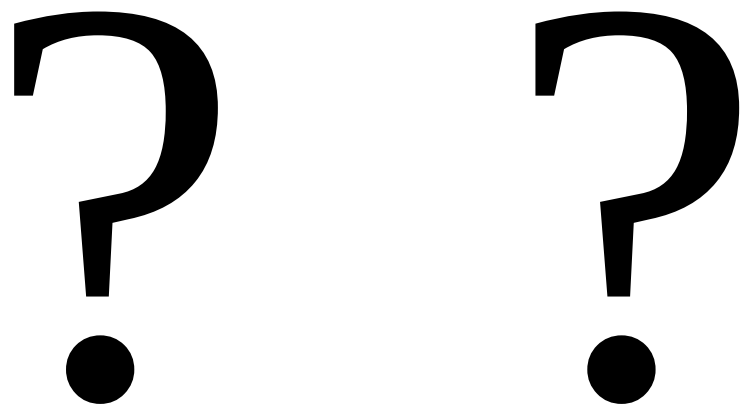
The measurement is re-expressed in terms of ρ as

$$I_{o,(t,\omega)} = \int_0^\infty S(\mathcal{E}) \exp\left[-f_w(\mathcal{E}) \int_{-\infty}^{\infty} \rho(s\hat{\omega} + t\omega) ds\right] d\mathcal{E}.$$

Applying H_w^{-1} to these measurements gives the Radon transform of ρ,

$$H_w^{-1}\left(\log\left[\frac{I_i}{I_{o,(t,\omega)}}\right]\right) = \int_{-\infty}^{\infty} \rho(s\hat{\omega} + t\omega)ds.$$

The function ρ is a density function which reflects the internal structure of the slice in much the same way as the mono-energetic, absorption coefficient.



(a) Dark streaks in a phantom with dense objects.

(b) Dark streaks in an image of the head produced by bone.

Figure 8.36: Due to beam hardening, dense objects produce dark streaks.

Having materials of very different densities in a slice leads to a much more difficult beam hardening problem; one which is, as of this writing, not completely solved. In human CT this is principally the result of bones intersecting the slice. It causes dark streak artifacts as seen in figure 8.36. The analysis of this problem is beyond the scope of this text. In [37] an effective algorithm is presented to substantially remove these artifacts. Another method, requiring two sets of measurements with X-ray beams having different spectral functions, is described in [2]. The discussion in this section is adapted from [37].

Exercise 8.6.15. Prove that $H_f(T)$ is a strictly monotone decreasing function.

Exercise 8.6.16. Find a Taylor series expansion for H_f^{-1}.

8.7 The gridding method

The algorithms described above are generally known as filtered backprojections. Let f denote the density function we would like to reconstruct and $\hat{f}$ its Fourier transform. If the numbers of samples collected in the radial and angular directions are $O(N)$ and the reconstruction is performed on an $N \times N$ grid then these methods require $O(N^3)$ arithmetic

operations. Samples of the $\hat{f}(\xi)$ on a uniform rectangular grid allows the use a *fast* direct Fourier inversion to obtain an approximation to f in $O(N^2 \log_2 N)$ operations. Neither the parallel beam nor fan beam machine collects data which is easily converted to uniformly spaced samples of $\hat{f}$. In this section we discuss *the gridding method* which is a family of efficient methods for passing from realistic data sets to uniformly spaced samples of $\hat{f}$. These methods work in any number of dimensions, so we describe this method in $\mathbb{R}^n$ for any $n \in \mathbb{N}$. Our discussion follows the presentation in [68].

Let f denote a function defined on $\mathbb{R}^n$ which is supported in a cube

$$D = [-L, L] \underset{n-\text{times}}{\times \cdots \times} [-L, L]$$

and $\hat{f}$ its Fourier transform. Suppose that $\hat{w}(\xi)$ is a function with small support, such that $w(x)$ does not vanish in D. The basis of the gridding method is the following observation: Suppose that we can efficiently compute an approximation to the convolution $\hat{g}(\xi_k) = \hat{w} * \hat{f}(\xi_k)$ for ξ_k on a uniformly spaced grid. The exact inverse Fourier transform of $\hat{g}$ equals $w(x)f(x)$. If this too can be efficiently computed on a uniformly spaced grid, then at grid points $x_j \in D$

$$f(x_j) \approx \frac{g(x_j)}{w(x_j)}.$$

The operation of approximating $\hat{g}(\xi_k)$ is essentially that of interpolating $\hat{f}$ to the points $\{\xi_k\}$ using a weight defined by $\hat{w}$. Choosing the weight carefully this step becomes a smoothing operation and so is less sensitive to noise. In the last step we effectively remove the choice of the interpolating weight function $\hat{w}$. We now describe the details of how this method is implemented.

Let $S_{\text{data}} = \{y_j : j = 1, \ldots, M\}$ denote a set of points in $\mathbb{R}^n$; the data which is available are approximations to the samples $\{\hat{f}_j \approx \hat{f}(y_j) : j = 1, \ldots, M\}$. Suppose that these points lie in a cube

$$E = [-B, B] \underset{n-\text{times}}{\times \cdots \times} [-B, B]$$

in Fourier space. In order for this (or any other) method to provide good results, it is necessary that the actual Fourier transform of f be small outside of E, i.e. $2B$ should equal or exceed the essential bandwidth of f. The goal is to approximate the values of $\hat{f}$ on a uniformly spaced grid in E and use the FFT to approximately reconstruct f. Let $d = B/N$ denote the sample spacing, (equal in all directions) in E. Use the bold letters $\mathbf{j}, \mathbf{k}, \mathbf{l}$ to denote points in $\mathbb{Z}^n$. The subset $S_{\text{unif}} \subset \mathbb{Z}^n$ indexes the sample points in E with $(k_1, \ldots, k_n) = \mathbf{k} \in S_{\text{unif}}$ if

$$-N \leq k_i \leq N, \quad i = 1, \ldots, n;$$

and then

$$S_{\text{unif}} \ni \mathbf{k} \leftrightarrow (dk_1, \ldots, dk_n) \in E.$$

From the discussion in section 6.2.1 it follows that the effective *field of view* of our sampled Fourier data is $2\pi/d$. That is, in order to avoid spatial aliasing, the function f must be supported in

$$G = [-\frac{\pi}{d}, \frac{\pi}{d}] \underset{n-\text{times}}{\times \cdots \times} [-\frac{\pi}{d}, \frac{\pi}{d}].$$

We suppose that $L \leq \pi/d$ so that spatial aliasing does not occur if our sample spacing in the Fourier domain is d. The sample points in G are also indexed by S_{unif} with

$$S_{\text{unif}} \ni \mathbf{k} \leftrightarrow \frac{\pi}{B}\mathbf{k} \in G.$$

Here we use the fact that $dN = B$.

Let $\hat{w}(\xi)$ be a function defined in $\mathbb{R}^n$ supported in the cube

$$W = [-Kd, Kd] \underset{n-\text{times}}{\times \cdots \times} [-Kd, Kd]$$

whose Fourier transform satisfies

$$w(x) \neq 0 \text{ for } x \in G.$$

As we eventually need to divide by w it is important that $w(x) > c > 0$ for x in the field of view. We need to approximately compute $\hat{w} * \hat{f}(d\mathbf{k})$ but cannot use a Fourier method to approximate this convolution because we do not have uniformly spaced samples of $\hat{f}$. The integral is therefore directly approximated as a Riemann sum. One imagines that E is divided into disjoint polyhedra $\{P_j : j = 1, \ldots, M\}$ such that each P_j contains exactly one of the sample points y_j. Let c_j denote the n-dimensional volume of P_j then

$$\begin{aligned} \hat{w} * \hat{f}(d\mathbf{k}) &\approx \sum_{j=1}^{M} \int_{P_j} \hat{w}(d\mathbf{k} - y)\hat{f}(y)dy \\ &\approx \sum_{j=1}^{M} \hat{w}(d\mathbf{k} - y_j)\hat{f}_j c_j. \end{aligned} \tag{8.65}$$

The approximation in the first line is due to the fact that $\hat{f}$ is not assumed to be B-bandlimited, only that it is *essentially* B-bandlimited. In the second line the integral is replaced with a Riemann sum. The accuracy of these approximations depends on many things: the smoothness of $\hat{f}$ and $\hat{w}$ as well the geometry of the polyhedra. If $\hat{f}$ and $\hat{w}$ are smooth then a more uniform division should lead to a more accurate result. Because $\hat{w}$ is supported in W this calculation should require $O(K^n)$ operations for each point $\mathbf{k} \in S_{\text{unif}}$. The efficient evaluation of these integrals requires the *a priori* determination of which values of $\hat{w}$ are needed for each index $\mathbf{k}$.

The values computed in (8.65) are approximations to uniformly spaced samples of $\hat{w} * \hat{f}$. After zero padding to obtain a power of 2, the FFT can be used to compute approximate values of $g_{\mathbf{j}} = w \cdot f(\mathbf{j}\frac{\pi}{B})$. The gridding method is completed by setting

$$f_{\mathbf{j}} = \frac{g_{\mathbf{j}}}{w(\mathbf{j}\frac{\pi}{B})}.$$

If N is a power of 2 then this step requires $O((N \log_2 N)^n)$ operations.

One might ask what the advantage of the gridding method is over using a simpler interpolation method. Using interpolation to assign a value to $\hat{f}$ at a grid point, $d\mathbf{k}$ one would typically average only the samples $\{\hat{f}(y_j)\}$ for the small number of y_j closest $d\mathbf{k}$.

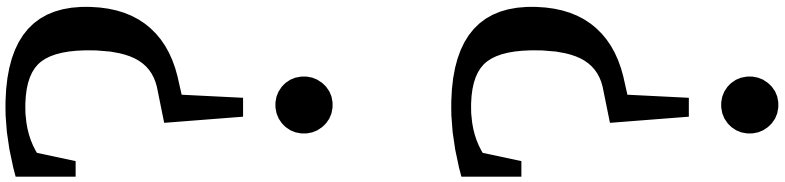
Noise in the measurements is expected to have "mean zero," however the larger the region one averages over, the more the noise from different samples has the opportunity to cancel. For simple interpolation one uses a weight function $\hat{w}$ with $K = 1$. For a larger value of K the computation of $\hat{g}$ involves averaging $\hat{f}$ over a larger region and therefore leads to greater suppression of the noise in the measurements.

Two considerations limit how large K can be taken. The amount of computation needed to do the gridding step is $O((KN)^n)$. For example if $n = 2$ this is $O(K^2N^2)$. The amount of computation needed for the filtered backprojection algorithms is $O(N^3)$. For the gridding method to remain competitive we see that $K << \sqrt{N}$. The second constraint comes from the requirement that $w(x) \neq 0$ for points, x in the cube D. From our discussion of the Fourier transform we know that the smaller the support of $\hat{w}(\xi)$, the larger the set around zero in which $w(x)$ is non-vanishing. Thus K needs to be chooser small enough so that $w(x)$ is non-vanishing throughout D.

A careful analysis of the errors entailed in this method is presented in [68] as well as examples using different choices of $\hat{w}$. Again this algorithm is not well suited to parallelization so may not be able to compete with filtered backprojection running on the specialized hardware used in commercial CT-machines.

8.8 Concluding remarks

X-ray CT is now a highly developed field in medical imaging. Most of our discussion dates from the early days of computed tomography, 1970-1980. At the end of the 1960s it was quite a remarkable idea that slices through a human body could be reconstructed using, what amounted to a large collection of carefully measured, tiny X-rays. The secret of course was mathematics. The original approach of Hounsfield was a variant of the ART method. It did not give very good images. Next a method using (unfiltered) backprojection was tried, which also gives rather poor images. Finally the Radon inversion formula was *rediscovered* and with it the possibility of accurately reconstructing images from projection data. These images also had many artifacts caused by aliasing, the Gibbs phenomenon, mis-calibration of the detectors, unstable X-ray sources, beam hardening, geometric mis-calibration of the gantry, etc. The introduction of the mathematical phantom and the mathematical analysis of errors in algorithms were essential steps in the removal of imaging artifacts. Figure 8.37(a) shows a CT-image of a slice of the head from a 1970s era machine. Note the enormous improvement in the 1990s era image of the same section, shown in figure 8.37(b).



(a) 1970s era brain section. (b) 1990s era brain section.

Figure 8.37: Mathematical analysis has led to enormous improvements in CT-images.

Our discussion of X-ray CT applies to large parts of the reconstruction problems for any "non-diffracting" imaging technique, e.g. positron emission tomography and magnetic resonance imaging. Much of the mathematical technology is also used in the study of "diffracting" modalities such as ultrasound, impedence tomography and infrared imaging. The "inverse problems" for these latter modalities are essentially non-linear and to date they lack complete mathematical solutions. In consequence of this fact these modalities have not yet come close to attaining their full potential. It stands to reason that the first step in finding a good approximation is having a usable formula for the exact result. The challenge for tomorrow is to "solve" the mathematical reconstruction form, for these modalities.

Chapter 9

Algebraic reconstruction techniques

Algebraic reconstruction techniques (ART) are techniques for reconstructing images which have no direct connection to the Radon inversion formula. Instead these methods make direct use of the fact that the measurement process is linear and therefore the reconstruction problem can be posed a system of linear equations. Indeed the underlying mathematical concepts of ART can be applied to approximately solve many types of large systems of linear equations. This chapter contains a very brief introduction to these ideas. An extensive discussion can be found in [24].

In this chapter boldface letters are used to denote vector or matrix quantities, while the entries of a vector or matrix are denoted in regular type with subscripts. For example $\mathbf{r}$ is a matrix, $\{\mathbf{r}_i\}$ are its rows (which are vectors) and r_{ij} its entries (which are scalars).

9.1 Algebraic reconstruction

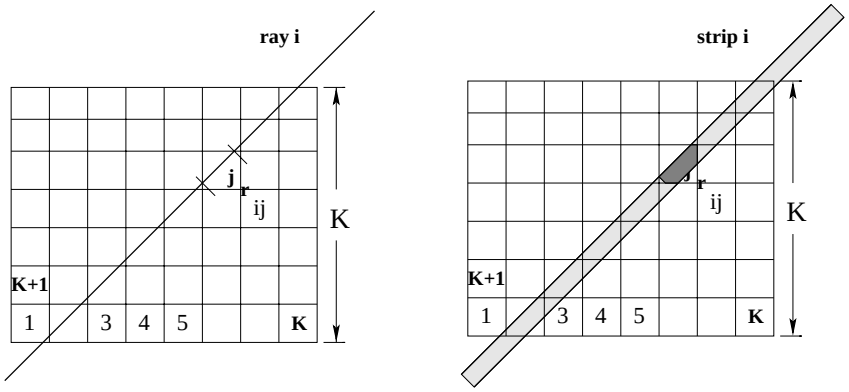
The main features of the Radon transform of interest in ART are (1) the map $f \mapsto \mathrm{R}f$ is linear, (2) for a function defining a simple object with bounded support, $\mathrm{R}f$ has a geometric interpretation. The first step in an algebraic reconstruction technique is the choice of a finite collection of basis functions,

$$\{b_1(x, y), \ldots, b_J(x, y)\}.$$

Certain types of *a priori* knowledge about the expected data and the measurement process itself can be "encoded" in the choice of the basis functions. At a minimum, it is assumed that the absorption coefficients one is likely to encounter are "well approximated" by functions in the linear span of the basis functions. This means that for some choice of constants $\{x_j\}$, the difference

$$f(x, y) - \sum_{j=1}^{J} x_j b_j(x, y),$$

is small, in an appropriate sense. For medical imaging, it is reasonable to require that the finite sum approximate f in that the grey scale images that they define look similar, see



(a) Using 1-dimensional rays. (b) Using 2-dimensional strips.

Figure 9.1: Pixel basis

section 7.6. It is also important that

$$\mathrm{R}f(x,y) \approx \sum_{j=1}^{J} x_j \, \mathrm{R}b_j.$$

A second criterion is to choose basis function for which $\mathrm{R}b_j$ can be efficiently approximated.

The *pixel* basis is a piecewise constant family of functions often used in ART. Suppose that the support of f lies in the square, $[-1,1]\times[-1,1]$. The square is uniformly subdivided into a $K \times K$ grid. When using ART methods it is convenient to label the sub-squares sequentially one after another as in figure 9.1(a). The elements of the $K \times K$ *pixel* basis are defined by

$$b_j^K(x,y) = \begin{cases} 1 & \text{if } (x,y) \in j^{\text{th}} - \text{square}, \\ 0 & \text{otherwise.} \end{cases}$$

If x_j is the average of f in the j^{th}-square then

$$\bar{f}^K(x,y) = \sum_{j=1}^{J} x_j b_j^K$$

provides an approximation to f in terms of the pixel basis. It is easy to see that the $\{b_j^K\}$ are orthogonal with respect to the usual inner product on $L^2(\mathbb{R}^2)$ and that $\bar{f}^K$ is the orthogonal projection of f into the span of the $\{b_j^K\}$.

For f a continuous function with bounded support the sequence $< \bar{f}^K >$ converges uniformly to f as $K \to \infty$. If f represents an image, in the usual sense of the word, then as $K \to \infty$ the image defined by $\bar{f}^K$ also converges to that defined by f. Because the Radon transform is linear

$$\mathrm{R}\bar{f}^K(x,y) = \sum_{j=1}^{J} x_j \, \mathrm{R}b_j^K.$$

Another advantage of the pixel basis is that $\mathrm{R}b_j^K(t,\omega)$ is, in principle, very easy to compute, being simply the length of the intersection of $l_{t,\omega}$ with the j^{th} square.

This is a good basis to keep in mind, it has been used in many research papers on ART as well as in commercial applications. Expressing a function as a linear combination of basis functions is, in fact, the same process used in our earlier analysis of the Radon inversion formula. The only difference lies in the choice of basis functions. In ART methods one typically uses a localized basis like the $\{b_j^K\}$, where each function has support in a small set. For our analysis of the Radon transform we adopted the basis provided by the exponential functions, $\{e^{i\boldsymbol{\xi}\cdot\mathbf{x}}\}$. These basis functions are well localized in frequency space but are not localized in physical space. The exponential basis is very useful because it diagonalizes the linear transformations used to invert the Radon transform. However it suffers from artifacts like the Gibbs phenomenon which are a consequence of its non-localized, oscillatory nature. Wavelets bases are an attempt to strike a balance between these two extremes, they are localized in space but also have a fairly well defined frequency. A good treatment of wavelets can be found in [27].

Returning now to our description of ART, assume that $\{b_j\}$ is a localized basis, though not necessarily the pixel basis. As before the measurements are modeled as samples of $\mathrm{R}f$. The samples are labeled sequentially by $i \in \{1,\ldots,I\}$, with $\mathrm{R}f$ sampled at

$$\{(t_1,\omega_1),(t_2,\omega_2),\ldots,(t_I,\omega_I)\}.$$

Unlike the filtered backprojection algorithm, ART methods are insensitive to the precise nature of the data set. Define the *measurement matrix* by setting

$$r_{ij} = \mathrm{R}b_j(t_i,\omega_i), \quad i = 1,\ldots,I.$$

The measurement matrix models the result of applying the measurement process to the basis functions. Define the entries in the vector of measurements, $\mathbf{p}$ as

$$p_i = \mathrm{R}f(t_i,\omega_i), \quad i = 1,\ldots,I.$$

The reconstruction problem is now phrased as a system of I equations in J unknowns:

$$\boxed{\sum_{j=1}^{J} r_{ij}x_j = p_i \text{ for } i = 1,\ldots,I} \tag{9.1}$$

or more succinctly, $\mathbf{rx} = \mathbf{p}$. A further flexibility of ART methods lies in the definition of the measurement matrix. As we shall see, simplified models are often introduced to compute its components.

The easiest type of linear system to solve is one defined by a diagonal matrix. In the Fourier approach to image reconstruction, the F.F.T. is used to reduce the reconstruction problem to a diagonal system of equations. This explains why it was not necessary to explicitly address the problem of solving linear equations. The difficulty of using ART comes from the size of the linear system (9.1). If the square is divided into $J = 128\times128 \simeq 16,000$ sub-squares then, using the pixel basis there are $16,000$ unknowns. A reasonable number of measurements is 150 samples of the Radon transform at each of 128 equally spaced angles,

so that $I \simeq 19,000$. That gives a $19,000 \times 16,000$ system of equations. Even today, it is not practical, to solve a system of this size directly. Indeed, as is typical in ART, this is an overdetermined system, so it is unlikely to have an exact solution.

Consulting figure (9.1) it is apparent that for each i there are about K values of j such that $r_{ij} \neq 0$. A matrix with "most" of its entries equal to zero is called a *sparse* matrix. Since $K \simeq \sqrt{I}$, r_{ij} is a sparse matrix. With the pixel basis and a one dimensional X-ray beam, the "exact" measurement matrix would be

$$r_{ij} = \text{ length of the intersection of the } i^{\text{th}} \text{ ray with the } j^{\text{th}} \text{ pixel.}$$

This is a reason that localized bases are used in ART methods: it is essential for the measurement matrix to be sparse.

If the X-ray beam has a 2-dimensional cross section then the lines above are replaced by strips. The value of r_{ij} could then be the area of intersection of the i^{th}-strip with the j^{th}-pixel. A more complicated beam profile could also be included by weighting different parts of the strips differently. In either case, the calculation of r_{ij} requires lot of work. Much of the literature on ART discusses the effects of using various schemes to approximate the measurement matrix. A very crude method, which was actually used in the earliest commercial machines, is to set $r_{ij} = 1$ if the center of the j^{th}-pixel is contained in the i^{th}-strip and 0 otherwise. For such a simple scheme, the values of r_{ij} can be computed at run time and do not have to be stored. An undesirable consequence of approximating r_{ij} is that it leads to inconsistent systems of equations. If the measurement matrix is not an accurate model for the measurement process then, given the overdetermined character of (9.1), one should neither expect an actual vector of measurements to satisfy

$$\mathbf{rx} = \mathbf{p}, \tag{9.2}$$

for any choice of $\mathbf{x}$, nor should one expect that a solution of this equation gives a good approximation to the actual attenuation coefficient. Here $\mathbf{x}$ is the $J \times 1$ column matrix of unknown coefficients and $\mathbf{r}$ is the $I \times J$ measurement matrix. A practical ART method needs to strike a balance between the computational load of accurately computing $\mathbf{r}$ and the inconsistencies which result from crude approximations.

A useful approach for handling inconsistent or over-determined problems is to look for a vector $\tilde{\mathbf{x}}$ which minimizes an error function,

$$e(\mathbf{r}\tilde{\mathbf{x}} - \mathbf{p}).$$

The most common choice of error function is Euclidean (or l_2) norm of the difference,

$$e_2(\mathbf{rx} - \mathbf{p}) = \|e(\mathbf{rx} - \mathbf{p})\|^2.$$

In this chapter $\|\cdot\|$ refers to the Euclidean norm. Minimizing $e_2(\mathbf{rx} - \mathbf{p})$ leads to the *least squares method*. This method is often reasonable on physical grounds and from the mathematical standpoint it is a very simple matter to derive the *linear* equations that an optimal vector, $\tilde{\mathbf{x}}$ satisfies. Using elementary calculus and the bilinearity of the inner product defining the norm gives the variational equation:

$$\frac{d}{dt}\langle \mathbf{r}(\tilde{\mathbf{x}} + t\mathbf{v}) - \mathbf{p}, \mathbf{r}(\tilde{\mathbf{x}} + t\mathbf{v}) - \mathbf{p}\rangle\Big|_{t=0} = 0 \quad \text{for all vectors } \mathbf{v}.$$

Expanding the inner product gives

$$t^2\langle \mathbf{rv}, \mathbf{rv}\rangle + 2t\langle \mathbf{rv}, \mathbf{r\tilde{x}} - \mathbf{p}\rangle + \langle \mathbf{r\tilde{x}} - \mathbf{p}, \mathbf{r\tilde{x}} - \mathbf{p}\rangle.$$

Hence, the derivative at $t = 0$ vanishes if and only if

$$2\langle \mathbf{rv}, \mathbf{r\tilde{x}} - \mathbf{p}\rangle = 2\langle \mathbf{v}, \mathbf{r}^t(\mathbf{r\tilde{x}} - \mathbf{p})\rangle = 0.$$

Since $\mathbf{v}$ is an arbitrary vector it follows that

$$\mathbf{r}^t\mathbf{r\tilde{x}} = \mathbf{r}^t\mathbf{p}. \tag{9.3}$$

These are sometimes called the *normal equations*. If $\mathbf{r}$ has maximal rank then $\mathbf{r}^t\mathbf{r}$ is invertible, which implies that the minimizer is unique. One might consider solving this system of equations. However for realistic imaging data, it is about a $10^4 \times 10^4$ system, which again is too large to solve directly. Moreover, the matrix $\mathbf{r}^t\mathbf{r}$ may fail to be sparse even though $\mathbf{r}$ is.

Exercise 9.1.1. If f is a continuous function with bounded support show that $\bar{f}^K$ converges uniformly to f as $K \to \infty$.

Exercise 9.1.2. Let $f = \chi_{[-a,a]}(x)\chi_{[-b,b]}(y)$. By examining $\bar{f}^K$ show that there is no "Gibbs phenomenon" for the pixel basis. In what sense does $\bar{f}^K$ converge to f?

Exercise 9.1.3. Suppose that f is a piecewise continuous function, find norms $|\cdot|_1, |\cdot|_2$ so that $|\bar{f}^K - f|_1$ and $|\,\mathrm{R}\bar{f}^K - \mathrm{R}f|_2$ tend to zero as $K \to \infty$.

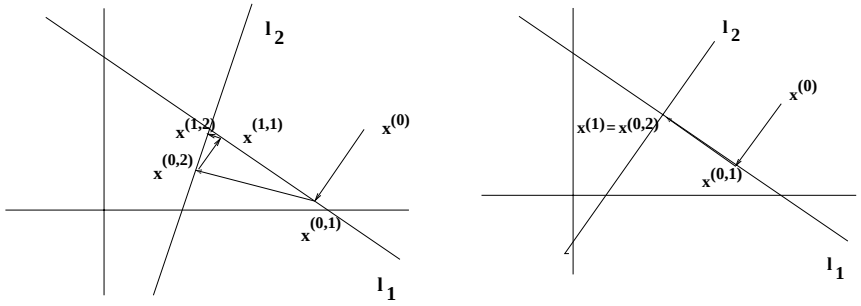
Exercise 9.1.4. Prove directly that if $\mathbf{r}$ has maximal rank then the normal equations have a unique solution.

9.2 Kaczmarz' method

Most of the techniques used in medical imaging are iterative. Instead of attempting to solve an equation like (9.3), we use an algorithm which defines a sequence $< \mathbf{x}^{(k)} >$, of vectors that get closer and closer to a solution (or approximate solution). The principal method used in medical imaging derives from the *Kaczmarz method* or *method of projections*. The idea can be explained using a very simple 2×2-example

$$\begin{aligned} r_{11}x_1 + r_{12}x_2 &= p_1, \\ r_{21}x_1 + r_{22}x_2 &= p_2. \end{aligned}$$

For $i = 1, 2$ $r_{i1}x_1 + r_{i2}x_2 = p_i$ defines a line l_i in the plane. The solution for the system of equation is the point of intersection of these two lines. The method of projections is very simple to describe geometrically:



(a) Convergence in infinitely many steps.

(b) Convergence in one iteration.

Figure 9.2: Method of projections

(1) Choose an arbitrary point and call it $\mathbf{x}^{(0)}$.

(2) Orthogonally project $\mathbf{x}^{(0)}$ onto l_1, denote the projected point by $\mathbf{x}^{(0,1)}$. Orthogonally Project $\mathbf{x}^{(0,1)}$ onto l_2, denote the projected point by $\mathbf{x}^{(0,2)}$. This completes one iteration, set $\mathbf{x}^{(1)} \stackrel{d}{=} \mathbf{x}^{(0,2)}$.

(3) Go back to (2), replacing $\mathbf{x}^{(0)}$ with $\mathbf{x}^{(1)}$, etc.

This gives a sequence $< \mathbf{x}^{(j)} >$ which, in case the lines intersect, converges, as $j \to \infty$ to the solution of the system of equations. If the two lines are orthogonal, a single iteration is enough. However, the situation is not always so simple. Figure 9.3(a) shows that it does not converge for two parallel lines - this corresponds to an inconsistent system which has no solution. Figure 9.3(b) depicts an over-determined, inconsistent 3×2 system, the projections are trapped inside the triangle but do not converge as $j \to \infty$.

The equations which arise imaging applications can be rewritten in the form

$$\mathbf{r}_i \cdot \mathbf{x} = p_i, \quad i = 1, \ldots, I,$$

where $\mathbf{r}_i$ is the i^{th}-row of the measurement matrix, $\mathbf{r}$. Each pair $(\mathbf{r}_i, p_i)$ defines a hyperplane in $\mathbb{R}^J$

$$\{\mathbf{x} : \mathbf{r}_i \cdot \mathbf{x} = p_i\}.$$

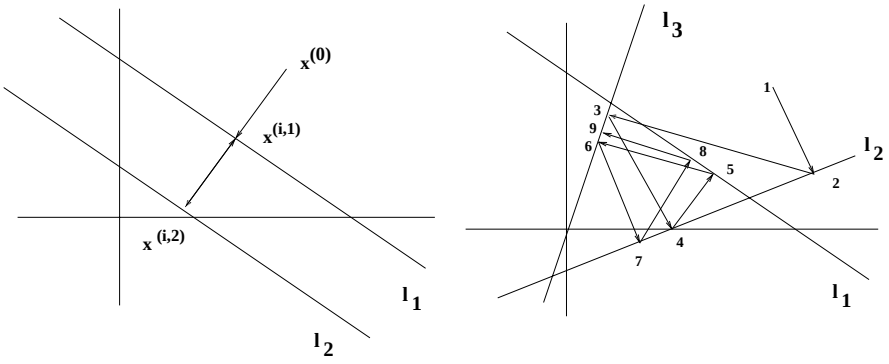
Following exactly the same process used above gives the basic Kaczmarz iteration:

(1) Choose an initial vector $\mathbf{x}^{(0)}$.

(2) Orthogonally project $\mathbf{x}^{(0)}$ into $\mathbf{r}_1 \cdot \mathbf{x} = p_1 \to \mathbf{x}^{(0,1)}$,
Orthogonally project $\mathbf{x}^{(0,1)}$ into $\mathbf{r}_2 \cdot \mathbf{x} = p_2 \to \mathbf{x}^{(0,2)}$,

$$\vdots$$

Orthogonally project $\mathbf{x}^{(0,I-1)}$ into $\mathbf{r}_I \cdot \mathbf{x} = p_I \to \mathbf{x}^{(0,I)} \stackrel{d}{=} \mathbf{x}^{(1)}$.



(a) Two parallel lines.

(b) Overdetermined system.

Figure 9.3: Examples where the projection algorithm does not converge.

(3) Go back to (2) replacing $\mathbf{x}^{(0)}$ with $\mathbf{x}^{(1)}$, etc.

To do these computations requires a formula for the orthogonal projection of a vector into a hyperplane. The vector $\mathbf{r}_i$ is orthogonal to the hyperplane $\mathbf{r}_i \cdot \mathbf{x} = p_i$. The orthogonal projection of a vector $\mathbf{y}$ onto $\mathbf{r}_i \cdot \mathbf{x} = p_i$ is found by subtracting a multiple of $\mathbf{r}_i$ from $\mathbf{y}$. Let $\mathbf{y}^{(1)} = \mathbf{y} - \alpha \mathbf{r}_i$, then α must satisfy

$$p_i = \mathbf{y}^{(1)} \cdot \mathbf{r}_i = \mathbf{y} \cdot \mathbf{r}_i - \alpha \mathbf{r}_i \cdot \mathbf{r}_i.$$

Solving this equation gives

$$\alpha = \frac{\mathbf{y} \cdot \mathbf{r}_i - p_i}{\mathbf{r}_i \cdot \mathbf{r}_i}.$$

The explicit algorithm is therefore:

$$\begin{aligned}
\mathbf{x}^{(k)} &\mapsto \mathbf{x}^{(k)} - \frac{\mathbf{x}^{(k)} \cdot \mathbf{r}_1 - p_1}{\mathbf{r}_1 \cdot \mathbf{r}_1} \mathbf{r}_1 = \mathbf{x}^{(k,1)}, \\
\mathbf{x}^{(k,1)} &\mapsto \mathbf{x}^{(k,1)} - \frac{\mathbf{x}^{(k,1)} \cdot \mathbf{r}_2 - p_2}{\mathbf{r}_2 \cdot \mathbf{r}_2} \mathbf{r}_2 = \mathbf{x}^{(k,2)}, \\
&\vdots \\
\mathbf{x}^{(k,I-1)} &\mapsto \mathbf{x}^{(k,I-1)} - \frac{\mathbf{x}^{(k,I-1)} \cdot \mathbf{r}_I - p_I}{\mathbf{r}_I \cdot \mathbf{r}_I} \mathbf{r}_I = \mathbf{x}^{(k+1)}.
\end{aligned}$$

Does the sequence $< \mathbf{x}^{(k)} >$ converge and if so, to what does it converge? As was already apparent in the trivial cases considered above, the answer depends on the situation. The fundamental case to consider is when the system $\mathbf{rx} = \mathbf{p}$ has a solution. In this case $< \mathbf{x}^{(k)} >$ *does* converge to a solution. This fact is very important, even though this case is unusual in imaging. For under-determined systems Tanabe has shown that this sequence converges to the solution $\mathbf{x}_s$, closest to the initial vector, $\mathbf{x}^{(0)}$, see [74].

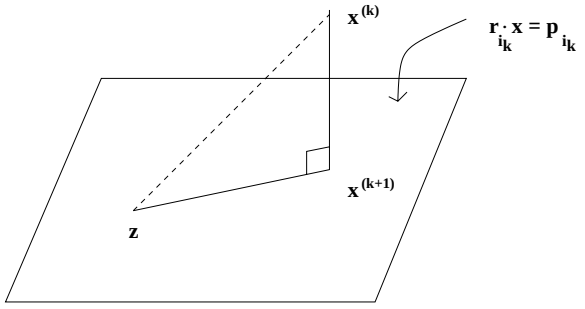


Figure 9.4: One step in the Kaczmarz algorithm.

Theorem 9.2.1. *Let $< \mathbf{r}_i >$ be a sequence of vectors in $\mathbb{R}^J$. If the system of equations*

$$\mathbf{r}_i \cdot \mathbf{x} = p_i, \quad i = 1, \ldots, I.$$

has a solution, then the Kaczmarz iteration converges to a solution.

Proof. For the proof of this theorem it is more convenient to label the iterates sequentially

$$\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \ldots, \mathbf{x}^{(k)}, \ldots$$

instead of

$$\mathbf{x}^{(0,1)}, \ldots, \mathbf{x}^{(0,I)}, \mathbf{x}^{(1,0)}, \ldots, \mathbf{x}^{(1,I)}, \ldots.$$

Thus $\mathbf{x}^{(j,k)} \leftrightarrow \mathbf{x}^{(jI+k)}$.

Let $\mathbf{z}$ denote any solution of the system of equations. To go from $\mathbf{x}^{(k)}$ to $\mathbf{x}^{(k+1)}$ entails projecting into a hyperplane $\mathbf{r}_{i_k} \cdot \mathbf{x} = p_{i_k}$, see figure 9.4. The difference $\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}$ is orthogonal to this hyperplane, as both $\mathbf{x}^{(k+1)}$ and $\mathbf{z}$ lie in this hyperplane it follows that

$$\langle \mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}, \mathbf{x}^{(k+1)} - \mathbf{z} \rangle = 0.$$

The Pythagorean theorem implies that

$$\|\mathbf{x}^{(k+1)} - \mathbf{z}\|^2 + \|\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}\|^2 = \|\mathbf{x}^{(k)} - \mathbf{z}\|^2 \Rightarrow \|\mathbf{x}^{(k+1)} - \mathbf{z}\|^2 \leq \|\mathbf{x}^{(k)} - \mathbf{z}\|^2. \tag{9.4}$$

The sequence $< \|\mathbf{x}^{(k)} - \mathbf{z}\|^2 >$ is a non-negative and decreasing hence, it converges to a limit. This shows that $< \mathbf{x}^{(k)} >$ lies in a ball of finite radius and so the Bolzano-Weierstrass theorem implies that it has a convergent subsequence $\mathbf{x}^{(k_j)} \to \mathbf{x}^*$.

Observe that each index k_j is of the form $l_j + nI$ where $l_j \in \{0, \ldots, I-1\}$. This means that, for some l, there must be an infinite sub-sequence, $\{k_{j_i}\}$ so that $k_{j_i} = l + n_i I$. All the vectors $\{\mathbf{x}^{(k_{j_i})} : i = 1, 2, \ldots\}$ lie in the hyperplane $\{\mathbf{r}_l \cdot \mathbf{x} = p_l\}$. As a hyperplane is a closed set this implies that the limit,

$$\mathbf{x}^* = \lim_m \mathbf{x}^{(k_{j_m})}$$

also belongs to the hyperplane $\mathbf{r}_l \cdot \mathbf{x} = p_l$.

On the other hand, it follows from (9.4) and the fact that $\|\mathbf{x}^{(k)} - \mathbf{z}\|$ converges that

$$\lim_{j \to \infty} \|\mathbf{x}^{(k_j+1)} - \mathbf{x}^{(k_j)}\| = 0.$$

Thus $\mathbf{x}^{(k_j+1)}$ also converges to $\mathbf{x}^*$. The definition of the Kaczmarz algorithm implies that $\mathbf{x}^{(k_j+1)} \in \{\mathbf{x} : \mathbf{r}_{l+1} \cdot \mathbf{x} = p_{l+1}\}$. As above, this shows that $\mathbf{x}^*$ is in this hyperplane as well. Repeating this argument I times we conclude that

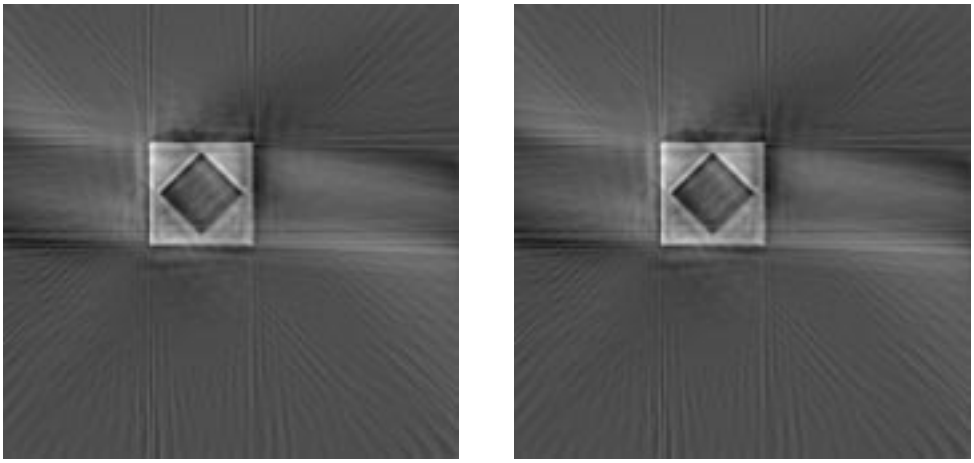
$$\mathbf{x}^* \in \{\mathbf{r}_i \cdot \mathbf{x} = p_i\}, \quad \text{for all } i = 1, \ldots I.$$

That is, $\mathbf{x}^*$ is a solution of the original system of equations. To complete the proof we need to show that the original sequence, $< \mathbf{x}^{(k)} >$ converges to $\mathbf{x}^*$. Recall that $\|\mathbf{x}^{(k)} - \mathbf{z}\|$ tends to a limit as $k \to \infty$ for *any* solution $\mathbf{z}$. Let $\mathbf{z} = \mathbf{x}^*$ then $\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \to \lambda$. For the subsequence $\{k_j\}$, it follows that

$$\lim_{j\to\infty} \|\mathbf{x}^{(k_j)} - \mathbf{x}^*\| = 0$$

Thus $\lambda = 0$ and $\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}^*$. □

As it generally requires an infinite number of iterates to find the solution, the result is largely of theoretical interest. In fact, in medical imaging applications only a few complete iterations are actually used. One reason is that the size of the system prevents using more. More importantly, it is an empirical fact that the quality of the reconstructed image improves for a few iterates but then begins to rapidly deteriorate. This is thought to be a consequence of noise in the data and inconsistencies introduced by approximating the measurement matrix. The image in figure 9.5(a)is obtained using one iteration while three iterations are used for figure 9.5(b). Note the absence of Gibbs artifacts parallel to the sides of the squares, though view sampling artifacts are still apparent.



(a) One iteration. (b) Three iterations.

Figure 9.5: Reconstructions using ART

With an algorithm of this type it is easy to take advantage of the sparseness of $\mathbf{r}$. For each i let $(j_1^i, \ldots, j_{k_i}^i)$ be a list of the indices of non-zero entries in row i. Knowing the locations of the non-zero entries greatly reduces the amount of computation needed to find $\mathbf{r}_i \cdot \mathbf{x}^{(k,i)}$ as well as $\mathbf{r}_i \cdot \mathbf{r}_i$. Note also, that in passing from $\mathbf{x}^{(k,i)}$ to $\mathbf{x}^{(k,i+1)}$ only entries at locations where $r_{(i+1)j} \neq 0$ are changed. This makes these methods practical even for the very large, sparse systems which arise in imaging applications.

If the equation has more than one solution, then using the Kaczmarz iteration with initial vector 0 gives the least squares solution.

Lemma 9.2.1. *If* $\mathbf{x}^{(0)} = 0$*, then* $\mathbf{x}^*$ *is the solution of* (9.1) *with minimal* l^2*-norm.*

Proof. Suppose that $A : \mathbb{R}^J \to \mathbb{R}^I$ with $I \leq J$, a matrix of maximal rank. The solution to $A\mathbf{x} = \mathbf{y}$ with minimal l^2-norm is given by $A^t\mathbf{u}$, where $\mathbf{u}$ is the unique solution to $AA^t\mathbf{u} = \mathbf{y}$. To see this let $\mathbf{x}_0$ be any solution to $A\mathbf{x} = \mathbf{y}$, and let $\mathbf{v} \in \ker A$ be such that $\|\mathbf{x}_0 + \mathbf{v}\|^2$ is minimal. The minimal norm solution is the one perpendicular to $\ker A$ so that

$$0 = \frac{d}{dt}\langle \mathbf{x}_0 + \mathbf{v} + t\mathbf{w}, \mathbf{x}_0 + \mathbf{v} + t\mathbf{w}\rangle\Big|_{t=0} = 2\langle \mathbf{x}_0 + \mathbf{v}, \mathbf{w}\rangle, \text{ for all } \mathbf{w} \in \ker A.$$

The assertion follows from the fact that the range of A^t is the orthogonal complement of $\ker A$, see Theorem 1.3.2.

Suppose in our application of the Kaczmarz method we use an initial vector $\mathbf{x}^{(0)} = \mathbf{r}^t\mathbf{u}$ for some $\mathbf{u}$. From the formula for the algorithm, it follows that all subsequent iterates are also of this form. Hence $\mathbf{x}^* = \mathbf{r}^t\mathbf{u}$ for some $\mathbf{u}$ and it satisfies $\mathbf{r}\mathbf{r}^t\mathbf{u} = \mathbf{p}$. By the claim above $\mathbf{x}^*$ is the least squares solution. Taking $\mathbf{u} = 0$ gives $\mathbf{x}^{(0)} = 0$ and this completes the proof of the lemma. □

This lemma and its proof are taken from [18].

In Chapter 8 it is shown that many reconstruction artifacts appear as rapid oscillations, hence it is of interest to find a solution with the smallest possible variation. The minimal norm solution is often the minimal variance solution as well. Set

$$\mathbf{e} = (1, \ldots, 1)$$

then

$$\mu_{\mathbf{x}} = \frac{\langle \mathbf{e}, \mathbf{x}\rangle}{J}$$

is the *average value* of the coordinates of $\mathbf{x}$. The variance is then defined to be

$$\sigma_{\mathbf{x}}^2 = \|\mathbf{x} - \mu_{\mathbf{x}}\mathbf{e}\|^2.$$

Proposition 9.2.1. *If $\mathbf{e} = \sum \alpha_i \mathbf{r}_i$ for some α_i then the minimum variance solution is also the minimum norm solution.*

Proof. If $\mathbf{x}$ is a solution of $\mathbf{r}\mathbf{x} = \mathbf{p}$, then

$$\begin{aligned} \|\mathbf{x} - \frac{1}{J}\langle \mathbf{e}, \mathbf{x}\rangle \mathbf{e}\|^2 &= \|\mathbf{x}\|^2 - 2\frac{\langle \mathbf{e}, \mathbf{x}\rangle^2}{J} + \frac{\langle \mathbf{e}, \mathbf{x}\rangle^2}{J} \\ &= \|\mathbf{x}\|^2 - \frac{1}{J}\sum \alpha_i \langle \mathbf{r}_i, \mathbf{x}\rangle \\ &= \|\mathbf{x}\|^2 - \frac{1}{J}\sum \alpha_i p_i. \end{aligned}$$

The second line follows from the fact that $\mathbf{x}$ is assumed to satisfy $\langle \mathbf{r}_i, \mathbf{x}\rangle = p_i$ for all i. Hence the $\|\mathbf{x}\|^2$ and its variance differ by a constant. This shows that minimizing the variance is equivalent to the minimizing the Euclidean norm. If $\mathbf{x}^{(0)} = 0$ then the lemma shows that the Kaczmarz solution has minimal Euclidean norm and therefore also minimal variance. □

Exercise 9.2.1. Prove the assertion, made in the proof, that if $\mathbf{x}^{(0)} = \mathbf{r}^t\mathbf{u}$, for some vector $\mathbf{u}$ then this is true of all subsequent iterates as well.

9.3 A Bayesian estimate

A small modification of the ART algorithm leads to an algorithm which produces a "Bayesian estimate" for an optimal solution to (9.1). Without going into the details, in this approach one has prior information that the solution should be close to a known vector $\mathbf{v}_0$. Instead of looking for a least squares solution to the original equation we try to find the vector which minimizes the combined error function:

$$\mathcal{B}_\rho(\mathbf{x}) \stackrel{d}{=} \rho\|\mathbf{rx} - \mathbf{p}\|^2 + \|\mathbf{x} - \mathbf{v}_0\|.$$

Here ρ is a fixed, positive number. It calibrates the relative weight given to the measurements versus the prior information. If $\rho = 0$ then the measurements are entirely ignored, as $\rho \to \infty$ less and less weight is given to the prior information. In many different measurement schemes it is possible to use the measurements alone to compute the average value, $\mu_{\mathbf{x}}$ of the entries $\mathbf{x}$. If we set

$$\mathbf{v}_0 = \mu_{\mathbf{x}}\mathbf{e},$$

then the "prior information" is the belief that the variance of the solution should be as small as possible.

The vector $\mathbf{x}_\rho$ which minimizes $\mathcal{B}_\rho(\mathbf{x})$ can be found as the minimal norm solution of a **consistent** system of linear equations. In light of Theorem 9.2.1 and Lemma 9.2.1 this vector can then be found using the Kaczmarz algorithm. The trick is to think of the error

$$\mathbf{u} = \mathbf{rx} - \mathbf{p}$$

as an independent variable. Let $\begin{pmatrix}\mathbf{u}\\ \mathbf{z}\end{pmatrix}$ denote an $I + J$-column vector and E the $I \times I$ identity matrix. The system of equations we use is

$$[E\ \rho\mathbf{r}]\begin{pmatrix}\mathbf{u}\\ \mathbf{z}\end{pmatrix} = \rho\left[\mathbf{p} - \mathbf{rv}_0\right]. \tag{9.5}$$

Theorem 9.3.1. *The system of equations* (9.5) *has a solution. If* $\begin{pmatrix}\mathbf{u}_\rho\\ \mathbf{z}_\rho\end{pmatrix}$ *is its minimal norm solution then* $\mathbf{x}_\rho = \mathbf{z}_\rho + \mathbf{v}_0$ *minimizes the function* $\mathcal{B}_\rho(\mathbf{x})$.

Proof. That (9.5) has solutions is easy to see. For any choice of $\mathbf{x}$ setting

$$\mathbf{u} = \rho\left[\mathbf{p} - \mathbf{rz} - \mathbf{rv}_0\right]$$

gives a solution to this system of equations. A minimal norm solution to (9.5) is orthogonal to the null space of $[E \quad \rho\mathbf{r}]$. This implies that it belongs to the range of the transpose, that is

$$\mathbf{z}_\rho = \rho\mathbf{r}^t\mathbf{u}_\rho. \tag{9.6}$$

On the other hand a vector $\mathbf{x}_\rho$ minimizes $\mathcal{B}_\rho$ if and only if it satisfies the variational equation:

$$\rho\mathbf{r}^t(\mathbf{rx}_\rho - \mathbf{p}) = \mathbf{v}_0 - \mathbf{x}\rho. \tag{9.7}$$

The relation, (9.6) between $\mathbf{u}_\rho$ and $\mathbf{z}_\rho$ implies that

$$\rho(\mathbf{rz}_\rho - \mathbf{p}) = -(\mathbf{u}_\rho + \rho\mathbf{rv}_0)$$

and therefore

$$\rho \mathbf{r}^t(\mathbf{r}\mathbf{z}_\rho - \mathbf{p}) = -\mathbf{z}_\rho - \rho \mathbf{r}^t \mathbf{r}\mathbf{v}_0.$$

This in turn shows that

$$\rho \mathbf{r}^t(\mathbf{r}\mathbf{x}_\rho - \mathbf{p}) = -\mathbf{z}_\rho = \mathbf{v}_0 - \mathbf{x}_\rho.$$

Thus $\mathbf{x}_\rho$ satisfies the variational equation (9.7) and therefore minimizes $\mathcal{B}_\rho$. □

Because (9.5) is consistent, the Kaczmarz method applied to this system, starting with the zero vector converges to the minimum norm solution of (9.7) which therefore also minimizes $\mathcal{B}_\rho$. This algorithm is easy to describe explicitly in terms of $\mathbf{u}$ and $\mathbf{x}$. The initial vector is

$$\begin{pmatrix} \mathbf{u}^{(0)} \\ \mathbf{x}^{(0)} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{v}_0 \end{pmatrix},$$

suppose we have found $\mathbf{u}^{(k,i)}$ and $\mathbf{x}^{(k,i)}$ then

$$\begin{aligned} \mathbf{u}^{(k,i+1)} &= \mathbf{u}^{(k,i)} + c^{(k,i)}\mathbf{e}_i, \\ \mathbf{x}^{(k,i+1)} &= \mathbf{x}^{(k,i)} + \rho c^{(k,i)}\mathbf{r}_i, \\ \text{where } c^{(k,i)} &= \frac{\rho(p_i - \langle \mathbf{r}_i, \mathbf{x}^{(k,i)} \rangle) - u_i^{(k,i)}}{1 + \rho^2 \|\mathbf{r}_i\|^2}. \end{aligned} \tag{9.8}$$

Here $\{\mathbf{e}_i, i = 1, \ldots, I\}$ is the standard basis for $\mathbb{R}^I$. This theorem and its proof are taken from [24].

Exercise 9.3.1. Suppose that the measurements are obtained using a parallel beam scanner, explain how to compute an approximation to the average value $\mu_{\mathbf{x}}$. How would you try to minimize the effects of measurement error?

Exercise 9.3.2. Explain (9.6) and derive the variational equation (9.7). Show that any vector which satisfies (9.7) also minimizes $\mathcal{B}_\rho$.

Exercise 9.3.3. If I and J are comparable and the pixel basis is used, how does the amount of computation required in (9.8) compare to that required in the normal Kaczmarz algorithm?

9.4 Variants of the Kaczmarz method

There are many ways to modify the basic Kaczmarz algorithm to obtain algorithms which give better results in a few iterations or reduce the effects of noise and modeling error. We give a very small taste of this vast subject.

9.4.1 Relaxation parameters

The systems of equations encountered in medical imaging are often over-determined and inconsistent because the data itself is noisy or the measurement matrix is only computed approximately. All these problems call for some kind of smoothing to be included in the algorithm. A common way to diminish noise and speed up convergence is to use *relaxation*

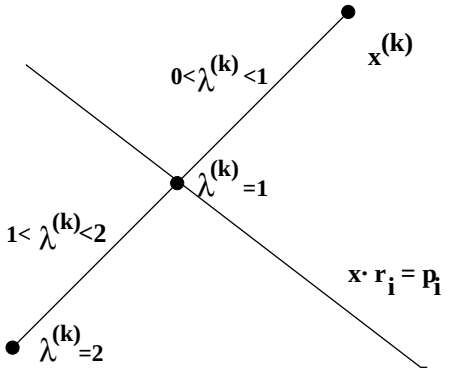


Figure 9.6: Ranges relaxation parameters.

parameters. Instead of applying the full correction, a scalar multiple is used instead. To that end, the ART algorithm is modified by putting in factors, $\{\lambda_k\}$ to obtain

$$\mathbf{x}^{(k,i)} \rightarrow \mathbf{x}^{(k,i)} - \lambda_k \frac{\mathbf{x}^{(k,i)} \cdot \mathbf{r}_i - p_i}{\mathbf{r}_i \cdot \mathbf{r}_i} \mathbf{r}_i.$$

The $\{\lambda_k\}$ are called *relaxation parameters*.

If $\lambda_k = 0$ then $\mathbf{x}^{(k,i)} = \mathbf{x}^{(k,i+1)}$, while $\lambda_k = 1$ gives the original algorithm. If $0 < \lambda_k < 1$ then $\mathbf{x}^{(k,i+1)}$ is on the same side of the hyperplane as $\mathbf{x}^{(k,i)}$ but closer. If $1 < \lambda_k < 2$ then $a^{(k,i+1)}$ is on the other side of the hyperplane and again closer than $\mathbf{x}^{(k,i)}$. If $\lambda_k = 2$ then $\mathbf{x}^{(k,i+1)}$ is the reflection of $\mathbf{x}^{(k,i)}$ in the hyperplane. So long as $0 < a \leq \lambda_k \leq b < 2$, for all k, and the system of equations has a solution, the modified algorithm also converges to the solution. If $\mathbf{x}^{(0)} = \mathbf{r}^t \mathbf{u}$ then the limit is again the minimum norm solution. Proofs of these facts are given in [24]. By making the sequence $< \lambda_k >$ tend to zero, a limit can be obtained even though the system of equations has no solution. Using algorithms of this type, one can find approximate solutions which are optimal for several different criteria, see [9] and [24]. As explained in these papers the choice of relaxation parameters is largely empirical, with some choices suppressing noise and other choices improving the contrast in the final image.

Another trick used in actual reconstruction algorithm stems from the following observation. Two adjacent rays produce measurement vectors $\mathbf{r}_i, \mathbf{r}_{i+1}$ which are very nearly parallel and therefore

$$\frac{\langle \mathbf{r}_i, \mathbf{r}_{i+1} \rangle}{\|\mathbf{r}_i\| \cdot \|\mathbf{r}_{i+1}\|} \approx 1.$$

Going from i to $i+1$ will, in general, lead to a very small change in the approximate solution; small corrections often get lost in the noise and round-off error. To speed up the convergence the successive hyperplanes are ordered to be as close to orthogonal as possible. The quality of the image produced by a few iterations is therefore likely to be improved by ordering the hyperplanes so that successive terms of the iteration come from hyperplanes which are not close to parallel. This is sometimes accomplished by "randomly" ordering

the measurements, so that the expected correlation between successive measurements is small.

9.4.2 Other related algorithms

There are many variants of the sort of iteration used in the Kaczmarz method. For example one can think of

$$\Delta x_j^{(k,i)} = x_j^{(k,i)} - x_j^{(k,i-1)} = \frac{p_i - \mathbf{r}_i \cdot \mathbf{x}^{(k,i-1)}}{\mathbf{r}_i \cdot \mathbf{r}_i} r_{ij}$$

as a correction that is applied to the j^{th}-entry of our vector but defer applying the corrections until we have cycled once through all the equations. Define

$$\delta x_j^{(k,i)} = \frac{p_i - \mathbf{r}_i \cdot \mathbf{x}^{(k)}}{\mathbf{r}_i \cdot \mathbf{r}_i} r_{ij}.$$

After this quantity is computed for all pairs $1 \leq i \leq I$, and $1 \leq j \leq J$, the approximate solution is updated

$$\begin{aligned} x_j^{(k+1)} &= x_j^{(k)} + \frac{1}{N_j} \sum_i \delta x_j^{(k,i)} \\ &= x_j^{(k)} + \frac{1}{N_j} \sum_i \frac{p_i - \mathbf{r}_i \cdot \mathbf{x}^{(k)}}{\mathbf{r}_i \cdot \mathbf{r}_i} r_{ij}. \end{aligned}$$

Here N_j is the number of i for which $r_{ij} \neq 0$. This is the number of iterates in which the value of x_j actually changes. In this algorithm we use the average of the corrections. This type of an algorithm is sometimes called a "simultaneous iteration reconstruction technique" or *SIRT*. A slight variant of the last algorithm which is used in real applications, is to set

$$x_j^{(k+1)} = x_j^{(k)} + \sum_i \frac{\left[r_{ij} \frac{p_i - \mathbf{r}_i \cdot \mathbf{x}^{(k)}}{\sum_{j=1}^N r_{ij}} \right]}{\sum_{j=1}^N r_{ij}}.$$

The denominator, $\sum_j r_{ij}$ equals the length of the intersection of the i^{th} ray with the image region. Using $\sum_{j=1}^N r_{ij}$ instead of $\sum_{j=1}^N r_{ij}^2$ is done for dimensional reasons and because it appears to give superior results.

Note finally, that a density function is normally assumed to be non-negative. Bases used in ART methods usually consist of non-negative functions and therefore the coefficients of a density function should also be non-negative. This observation can be incorporated into ART algorithms in various ways. The simplest approach is to replace the final coefficients with the maximum of the computed value and 0. It is also possible to do this at each step of the algorithm, replacing the entries of $\mathbf{x}^{(k,i)}$ with the maximum of $x_l^{(k,i)}$ and 0 for each $1 \leq l \leq J$. Beyond this, ART methods can be used to find vectors which satisfy a collection of *inequalities*, rather than equations. Such algorithms are described in [24].

ART algorithms provide a very flexible alternative to filtered backprojection algorithms. Unlike FBP, they are insensitive to the details of the data set. Through the usage of

relaxation parameters, noisy or inconsistent data can also be effectively handled. Though most present day machines use some form of filtered backprojection, the first algorithm, in the first commercial CT-scanner was of this general type. A very complete discussion of these methods, along with references to the extensive literature in given in [24]. A more recent description of the usage of these techniques in the context of positron emission tomography in given in [25].

Chapter 10

Probability theory and random variables

Up to this point we have considered only deterministic systems. These are systems where a known input produces a known output. We now begin to discuss probability theory, which is the language of noise analysis. But what is noise? There are two essentially different sources of noise in the mathematical description of a physical system or measurement process. The first step in building a mathematical model is to isolate a physical system from the world in which it sits. Once such a separation is fixed, the effects of the outside world on the state of the system are often modeled, *a posteriori* as noise. In our model for CT-imaging it is assumed that every X-ray which is detected is produced by our X-ray source. In reality, there are many other sources of X-rays which might also impinge on our detectors. Practically speaking it is not possible to give a complete description of all such external sources, sometimes it is possible to describe them probabilistically.

Many physical processes are inherently probabilistic and that is the second source of noise. In CT-imaging the interaction of an X-ray "beam" with an object is a probabilistic phenomenon. The beam is in fact a collection of discrete photons. Whether or not a given photon, entering an object on one side re-emerges on the other side, traveling in the same direction, depends on the very complicated interactions this photon has with the microscopic components of the object it passes through. Practically speaking one cannot model the details of these interactions in a useful way. If μ is the absorption coefficient at $\mathbf{x}$ and I is the incident flux of photons then Beer's law says that the change in the flux, $I(t+\Delta t) - I(t)$ over a small distance Δt is

$$I(t+\Delta t) - I(t) \approx -\mu \Delta t I(t)$$

or

$$I(t+\Delta t) = (1-\mu\Delta t) I(t).$$

This can be interpreted as the statement that an incident photon has probability $(1-\mu\Delta t)$ of being emitted. Beer's law describes the average behavior of "a photon" and is only useful if the X-ray beam is composed of a very large number of photons.

For a particular measurement it is not possible to predict the exact discrepancy between Beer's law and the actual life history of the given X-ray beam. Instead one has a probabilistic description which describes the *statistics* of this discrepancy. A certain number of

transmitted photons, $N(t)$ are measured, one can think of the measurement as having two parts:

$$N(t) = N_d(t) + v(t)$$

where N_d is a deterministic quantity, i.e. the part predicted by Beer's law and v is a random process which models the noise in the system. A good probabilistic model for the noise component aids in interpretation of the measurements. Note that even the model of an X-ray beam as a constant flux of energy is an approximation to the truth; the actual output of an X-ray source also has a useful probabilistic description.

In this chapter we review some of the basic concepts of measure theory and probability theory. It is not intended as a development of either subject *ab ovo*, but merely a presentation of the main ideas so that we can later discuss random processes and noise in image reconstruction. These concepts are presented using the mathematical framework provided by measure theory. It is not necessary to have a background in measure theory, we use it as a language to give precise definitions of the concepts used in probability theory and later to pass to random processes. Historically, probability theory preceded measure theory by many decades, but was found, in the early twentieth century to be somewhat lacking in rigorous foundations. Kolmogorov discovered that measure theory provided the missing foundations. A basic introduction to probability can be found in [11], an introduction to random processes in [13] or [54] and an introduction to measure theory in [16] or [65].

10.1 Measure theory

Mathematical probability theory is a subset of measure theory. This is the branch of mathematics which gives a framework in which to study integration. For concreteness we discuss probability theory from the point of view of doing an "experiment." We are then interested in quantifying the likelihood that the experiment has this or that outcome. In measure theory one works with a measure space. This is a pair $(X, \mathcal{M})$ where X is the underlying space and $\mathcal{M}$ is a collection of subsets of X called a σ-algebra.

10.1.1 Allowable events

From the point view of probability theory, X is called the *sample space*, it is the set of all possible outcomes of the experiment. Subsets of X are collections of possible outcomes, which in probability theory are called *events*. The subsets of X in $\mathcal{M}$ are "allowable" events. These are events which can be assigned a well defined probability of occurring. A simple physical example serves to explain, in part why it may not be possible to assign a probability of occurrence to every event, i.e. to every subset of X.

Example 10.1.1. Suppose that the experiment involves determining where a particle strikes a line. The sample space $X = \mathbb{R}$. For each $n \in \mathbb{Z}$, the measuring device can determine whether on not the particle fell in $[n, n+1)$ but not where in this interval the particle fell. For each n, the event

$$A_n = \text{ the particle fell in } [n, n+1)$$

is therefore an allowable event. On the other hand the event: "the particle fell in $[.3, .7)$" is not an allowable event, as the measurements cannot determine whether or not this occurs. For the set of allowable events we take arbitrary unions of the subsets $\{A_n : n \in \mathbb{Z}\}$.

Example 10.1.2. Perhaps the simplest, interesting experiment is that of "flipping a coin." This is an experiment that has two possible outcomes which we label H and T. The sample space

$$X_1 = \{H, T\}.$$

In ordinary language the possible events are

- We get a head.
- We get a tail.
- We get a head or a tail.

These events correspond to the following subsets of X_1 :

$$\{H\}, \{T\}, \{H, T\}.$$

Example 10.1.3. Suppose that instead of flipping a coin once, the experiment involves flipping a coin two times. The possible outcomes are the sequences of length two in the symbols H and T :

$$X_2 = \{(H, H), (H, T), (T, H), (T, T)\}.$$

If the experiment involves flipping a coin N-times then the sample space is all sequences of length N using the symbols H and T. We denote this space by X_N. In each of these examples the sample space is a finite set. In such cases all events are usually allowable and therefore we take $\mathcal{M}_N$ to be **all** the subsets of X_N.

The collection, $\mathcal{M}$ of allowable events provides a mathematical framework for describing the operation of a measuring apparatus. It has the following axiomatic properties:

(1) $X \in \mathcal{M}$, this is the statement that X is the collection of all possible outcomes.

(2) If $A \in \mathcal{M}$ and $B \in \mathcal{M}$, then $A \cup B \in \mathcal{M}$, in other words if the events A and B are each allowable then the event "A or B" is also allowable.

(3) If $A \in \mathcal{M}$, then $X \setminus A \in \mathcal{M}$, in other words if the event "A occurs" is allowable then the event "A does not occur" is also allowable.

(4) If we have a countable collection of allowable events $A_j \in \mathcal{M}$ then

$$\bigcup_{j=1}^{\infty} A_j \in \mathcal{M}$$

as well.

Condition (4) is a technical condition which is essential to have a good mathematical theory of integration. It is very important when taking limits of sequences, but does not have a simple intuitive explanation. Subsets belonging to $\mathcal{M}$ are called allowable events or *measurable sets*. A collection of subsets of a space which satisfy these axioms is called a *σ-algebra*.

As a consequence of these axioms we need to introduce the notion of the *empty set*. It is the subset of X which contains **no** elements and is denoted by $\emptyset$. This arises because (2) and (3) imply that if A and B are allowable events then so is $A \cap B$. However, if A and B have no points in common then $A \cap B = \emptyset$. This is mostly a linguistic device, encapsulating the idea that the experiment has no outcome. As we shall see this "event" always has probability zero of occurring. The list given in example 10.1.2 is not quite all of $\mathcal{M}_1$, but rather

$$\mathcal{M}_1 = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}.$$

Example 10.1.4. Some examples of allowable events for the case X_N defined above are

- All N flips produce heads.
- Half the flips are tails.
- If $k \leq N$ then the k^{th} flip is a head.
- At most $N/3$ flips are heads.

Example 10.1.5. Suppose that the result of our experiment is a real number, then $X = \mathbb{R}$. Allowable events might include

- $\{x\}$ for an $x \in \mathbb{R}$, the outcome of the experiment is the number x.
- $[0, \infty)$, the outcome of the experiment is a non-negative number.
- $(1, 2)$, the outcome of the experiment is a number between 1 and 2.

The smallest collection of subsets of $\mathbb{R}$ which includes all intervals and is a σ-algebra is called the *Borel sets*. It is discussed in [16].

Exercise 10.1.1. Show that the collection of allowable events, $\mathcal{M}$ in example 10.1.1 has the following description. To each $A \in \mathcal{M}$ there are, possibly bi-infinite, sequences $\{a_i\}$ and $\{b_i\}$ of integers so that

$$\cdots < a_i < b_i < a_{i+1} < b_{i+1} < \cdots$$

and

$$A = \bigcup_{j=-\infty}^{\infty} [a_j, b_j).$$

Exercise 10.1.2. Show that the collection of sets defined in example 10.1.1 is a σ-algebra.

Exercise 10.1.3. Show that the space in example 10.1.3 X_N has 2^N elements. How many different allowable events are there, that is how large is $\mathcal{M}_N$?

Exercise 10.1.4. Let $A \subset X$ then *the complement of A in X* is the subset of X defined by

$$A^c = X \setminus A = \{x \in X \, : \, x \notin A\}.$$

Show that if $A, B \subset X$ then

$$(A \cup B)^c = A^c \cap B^c.$$

Conclude that if $A, B \in \mathcal{M}$ then $A \cap B \in \mathcal{M}$ as well.

10.1.2 Measures and probability

See: B.8, B.9.

So far we have no notion of the "probability of an event" occurring. Since events are subsets of X what is required is a way to measure the size of a subset. Mathematically this is described by a function from allowable events to the interval $[0,1]$:

$$\nu : \mathcal{M} \longrightarrow [0,1]$$

This function is called a *probability measure* provided it has the following properties:

(1) $\nu(A) \geq 0$ for all $A \in \mathcal{M}$, that is an allowable even occurs with non-negative probability.

(2) $\nu(X) = 1$, so that X is the list of *all* possible outcomes.

(3) If $A, B \in \mathcal{M}$ and $A \cap B = \emptyset$, i.e., A and B are *mutually exclusive events* then

$$\nu(A \cup B) = \nu(A) + \nu(B).$$

This is called additivity.

(3′) If we have a countable collection of subsets $A_i \in \mathcal{M}$ such that $A_i \cap A_j = \emptyset$ for $i \neq j$, then

$$\nu\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \nu(A_j).$$

Conditions (2) and (3) imply that $\nu(A) \leq 1$ for any $A \in \mathcal{M}$. A triple $(X, \mathcal{M}, \nu)$ consisting of a space X, with a σ-algebra of subsets $\mathcal{M}$ and a probability measure $\nu : \mathcal{M} \to [0,1]$ is called a *probability space*. In the sequel we often use

$$\text{Prob}(A) \stackrel{d}{=} \nu(A)$$

to denote the "probability of the event A." Note that (2) implies that $\text{Prob}(\emptyset) = 0$.

The last, rather technical condition (3′) is called *countable additivity.* This condition becomes important when the underlying measure space is uncountably infinite (for example $\mathbb{R}$) and is needed in order to have a good mathematical theory of integration. It is the reason why one needs to have a notion of allowable event. It turns out that many interesting functions which satisfy these conditions cannot be extended to all the subsets of a given set in a reasonable way. For example, there is *no* way, satisfying these axioms to extend the naive notion of length on the real line, to arbitrary subsets. Subsets which do not belong to $\mathcal{M}$ are called *non-measurable*. An simple example of this is given in example 10.1.1. In most cases non-measurable sets are very complicated and do not arise naturally in applications.

Example 10.1.6. For the case of a single coin toss, the allowable events are

$$\mathcal{M}_1 = \{\{H\}, \{T\}, \{H, T\}, \emptyset\}.$$

The function $\nu_1 : \mathcal{M}_1 \to [0, 1]$ is fixed once we know the probability of H. For say that $\nu_1(H) = p$. Because $\nu_1(H \cup T) = 1$ and H and T are mutually exclusive events, it follows that $\nu_1(T) = 1 - p$.

Example 10.1.7. For the case of general N the allowable events are all collections of sequences $\mathbf{a} = (a_1, \ldots, a_N)$ where $a_j \in \{H, T\}$ for $j = 1, \ldots, N$. The most general probability function on $\mathcal{M}_N$ is defined by choosing numbers $\{p_{\mathbf{a}} : \mathbf{a} \in X_N\}$ so that

$$0 \leq p_{\mathbf{a}} \leq 1$$

and

$$\sum_{\mathbf{a} \in X_N} p_{\mathbf{a}} = 1.$$

If $A \in \mathcal{M}_N$ is an event then

$$\nu(A) = \sum_{\mathbf{a} \in A} p_{\mathbf{a}}.$$

In most instances one uses a much simpler measure to define the probability of events in $\mathcal{M}_N$. Instead of directly assigning probabilities to each sequence of length N we use the assumption that an H occurs (at any position in the sequence) with probability p and a T with probability $1 - p$. We also assume that outcomes of the various flips are independent of one another. With these assumptions one can show that

$$\nu_{p,N}(\{\mathbf{a}\}) = p^{m_{\mathbf{a}}}(1 - p)^{N - m_{\mathbf{a}}}$$

where $m_{\mathbf{a}}$ is the number of Hs in the sequence $\mathbf{a}$.

Example 10.1.8. Suppose that $X = \mathbb{R}$, with $\mathcal{M}$ the Borel sets. For any $a < b$ the half open interval $[a, b) \in \mathcal{M}$. Let $f(t)$ be a non-negative, continuous function defined on X with the property that

$$\int_{-\infty}^{\infty} f(t)dt = 1.$$

Define the probability of the event $[a, b)$ to be

$$\text{Prob}([a, b)) = \int_a^b f(t)dt.$$

It is not difficult to show that this defines a function on $\mathcal{M}$ satisfying the properties enumerated above, see [16].

Example 10.1.9. In the situation described in example 10.1.1 $X = \mathbb{R}$. Choose a bi-infinite sequence $\{a_n : n \in \mathbb{Z}\}$ of non-negative numbers such that

$$\sum_{n=-\infty}^{\infty} a_n = 1.$$

and define

$$\nu([n, n+1)) = a_n \text{ for } n \in \mathbb{Z}.$$

This means that a_n is the probability that the particle fell in $[n, n+1)$. Using the properties above, ν is easily extended to define a measure on the allowable sets defined in example 10.1.1. Again note that it is not possible to assign a probability to the event "the particle fell in $[.3, .7)$" as our measurements are unable to decide whether or not this happens.

Example 10.1.10. Suppose that X is the unit disk in the plane and let $\mathcal{M}$ be the "Lebesgue measurable" subsets of X. These are the subsets whose surface area is well defined. The outcome of our experiment is a point in X. For a measurable set A, define the probability of the event "the point lies in A" to be

$$\text{Prob}(A) = \frac{\text{area}(A)}{\pi}.$$

In a reasonable sense this means that each point in X is "equally likely" to be the outcome of the experiment. Note that if the set A consists of a single point $\{(x, y)\}$ then $\text{Prob}(A) = 0$.

Exercise 10.1.5. In example 10.1.10, explain why $\text{Prob}(\{(x, y)\}) = 0$.

Exercise 10.1.6. In example 10.1.10, let A_r denote the circle of radius r. What is $\text{Prob}(A_r)$.

Exercise 10.1.7. In section 2.4 we defined a "set of measure zero." What is the probabilistic interpretation of such a set?

10.1.3 Integration

If $(X, \mathcal{M}, \nu)$ is a probability space then one can define the notion of a measurable function as well as an integral.

Definition 10.1.1. Let $(X, \mathcal{M}, \nu)$ be a probability space. A real valued function, f defined on X is *measurable* if, for every $t \in \mathbb{R}$, the set $\{x \in X : f(x) \leq t\}$ belongs to $\mathcal{M}$.

The basic examples of measurable functions are the "indicator" functions of sets in $\mathcal{M}$. For an arbitrary subset define the *indicator function*

$$\chi_A(x) = \begin{cases} 1 \text{ if } x \in A, \\ 0 \text{ if } x \notin A. \end{cases}$$

Remark 10.1.1. Earlier in the book these sort of function would have been called the "characteristic function" of the set A. In probability theory 'characteristic function' has a different meaning, so we use (the also standard terminology), indicator function.

For an indicator function it is clear that the only reasonable way to define the integral is to set

$$\int_X \chi_A(x) d\nu(x) = \nu(A).$$

As $\nu(X) = 1$ this defines $\nu(A)$ as the fraction of X occupied by A.

Definition 10.1.2. A function f is called a simple function if there are sets $A_j \in \mathcal{M}, \quad j = 1, \ldots, m$ and real constants $\{a_j\}$ so that

$$f(x) = \sum_{j=1}^{N} a_j \chi_{A_j}(x).$$

It is not difficult to show that a simple function is measurable. Since the integral should be linear it is clear that we must define

$$\int_X \sum_{j=1}^{N} a_j \chi_{A_j}(x) d\nu(x) = \sum_{j=1}^{N} a_j \nu(A_j).$$

While this formula is intuitively obvious, it requires proof that the integral is well defined. This is because a simple function can be expressed as a sum, in different ways

$$\sum a_j \chi_{A_j} = \sum_{k=1}^{M} b_k \chi_{B_k}.$$

It is necessary to show that

$$\sum_{j=1}^{N} a_j \nu(A_j) = \sum_{k=1}^{M} b_k \nu(B_k).$$

for any other representation. This is left as an exercise for the interested reader.

Suppose that f is a bounded, measurable function. Fix a positive integer N, for each $j \in \mathbb{Z}$ set

$$A_{N,j} = f^{-1}\left([\frac{j}{N}, \frac{j+1}{N}) \right). \tag{10.1}$$

Since f is a bounded function, the function

$$F_N(x) = \sum_{j=-\infty}^{\infty} \frac{j}{N} \chi_{A_{N,j}}(x)$$

is a simple function with the following properties

(1).

$$0 \leq f(x) - F_N(x) \leq N^{-1},$$

(2).

$$\int_X F_N(x)d\nu(x) = \sum_{j=-\infty}^{\infty} \frac{j}{N}\nu(A_{N,j}).$$

(3).

$$F_N(x) \leq F_{N+1}(x).$$

In other words a bounded measurable function can be approximated by simple functions for which the integral is defined. This explains, in part, why we introduce the class of measurable functions. For a bounded, non-negative, measurable function the ν-integral is defined by:

$$\int_X f(x)d\nu(x) = \lim_{N\to\infty} \int_X F_N(x)d\nu(x).$$

Condition (3) implies that $\int F_N d\nu$ is an increasing function of N so this limit exists. By approximating non-negative, measurable functions f in a similar way the definition of the integral can be extended to this class. If f is a non-negative, measurable function then its integral over X is denoted by

$$\int_X f(x)d\nu(x).$$

The integral may equal $+\infty$. If f is measurable then the functions

$$f_+(x) = \max\{0, f(x)\}, \quad f_-(x) = \max\{0, -f(x)\}$$

are also measurable. If either $\int_X f_\pm d\nu$ is finite then the integral of f is defined to be

$$\int_X f(x)d\nu(x) \stackrel{d}{=} \int_X f_+(x)d\nu(x) - \int_X f_-(x)d\nu(x).$$

Definition 10.1.3. Let $(X, \mathcal{M}, \nu)$ be a probability space. A measurable function f is *integrable* if both of the integrals

$$\int_X f_\pm d\nu$$

are both finite. In this case

$$\int_X f(x)d\nu(x) = \int_X f_+(x)d\nu(x) - \int_X f_-(x)d\nu(x).$$

A more complete discussion of integration can be found in [16] or [65]. For our purposes, certain formal properties of the integral are important. Let f, g be measurable functions and $a \in \mathbb{R}$ then the integral defined by ν is *linear* in the sense that

$$\int_X (f+g)d\nu = \int_X f d\nu + \int_X g d\nu \text{ and } \int_X af d\nu = a\int_X f d\nu.$$

These conditions imply that if $\{A_j\}$ is a collection of pairwise disjoint subsets belonging to $\mathcal{M}$ and $\{a_j\}$ is a bounded sequence of numbers then the function

$$f(x) = \sum_{j=1}^{\infty} a_j \chi_{A_j}(x)$$

is an integrable function with

$$\int_X f d\nu = \sum_{j=1}^{\infty} a_j \nu(A_j).$$

Note that here we consider infinite sums whereas previously we only considered finite sums.

Example 10.1.11. Let $(X_N, \mathcal{M}_N, \nu_{p,N})$ be the probability space introduced in the example 10.1.7. Since $\mathcal{M}_N$ contains all subsets of X_N, any function on X_N is measurable. Using the properties of the integral listed above it is not difficult to show that if f is a function on X_N then

$$\int_{X_N} f(x) d\nu_{p,N}(x) = \sum_{\mathbf{a} \in X_N} f(\mathbf{a}) \nu_{p,N}(\mathbf{a}).$$

For a finite probability space the integral reduces to an ordinary sum.

Example 10.1.12. If $f(x)$ is a non-negative, continuous function on $\mathbb{R}$ with

$$\int_{\mathbb{R}} f(x) dx = 1$$

then we can define a probability measure by setting

$$\nu_f(A) = \int_A f(x) dx = \int_{\mathbb{R}} \chi_A(x) f(x) dx.$$

Here A is assumed to be a Borel set. With this definition it is not difficult to show that for any bounded, measurable function g we have

$$\int_{\mathbb{R}} g(x) d\nu(x) = \int_{\mathbb{R}} g(x) f(x) dx.$$

Example 10.1.13. Let $X = \mathbb{R}$ and let $\{a_n\}$ be a sequence of non-negative numbers such that

$$\sum_{j=-\infty}^{\infty} a_j = 1.$$

We let $\mathcal{M}$ be the collection of all subsets of $\mathbb{R}$ and define the measure ν by letting

$$\nu(A) = \sum_{\{n \,:\, n \in A\}} a_n.$$

Any function $g(x)$ is measurable and

$$\int_{\mathbb{R}} g(x)d\nu(x) = \sum_{n=-\infty}^{\infty} a_n g(n).$$

Example 10.1.14. Suppose that $F(x)$ is a non-negative function defined on $\mathbb{R}$ which satisfies the conditions

(1). F is a monotone non-decreasing function: $x < y$ implies that $F(x) \leq F(y)$,

(2). F is continuous from the right: for all x, $F(x) = \lim_{y \to x^+} F(y)$,

(3). $\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to \infty} F(x) = 1.$

Such a function defines a measure on the Borel subsets of $\mathbb{R}$. The measure of a half ray is defined to be

$$\nu_F((-\infty, a]) \stackrel{d}{=} F(a),$$

and the measure of an interval $(a, b]$ is defined to be

$$\nu_F((a, b]) \stackrel{d}{=} F(b) - F(a).$$

Note that if $(a, b]$ is written as a disjoint union

$$(a, b] = \bigcup_{j=1}^{\infty} (a_j, b_j]$$

then

$$F(b) - F(a) = \sum_{j=1}^{\infty} F(b_j) - F(a_j). \tag{10.2}$$

This condition shows that the ν_F-measure of an interval $(a, b]$ is well defined.

If $a_1 < b_1 \leq a_2 < b_2 \leq \cdots$ then the measure of the union of intervals is defined by linearity (of the integral)

$$\nu_F \left(\bigcup_{j=1}^{\infty} (a_j, b_j] \right) = \sum_{j=1}^{\infty} F(b_j) - F(a_j).$$

The measure can be extended to an arbitrary Borel set by approximating it by intersections of unions of intervals. The measure ν_F is called the *Lebesgue-Stieltjes measure* defined by F. The ν_F-integral of a function g is usually denoted

$$\int_{-\infty}^{\infty} g(x)dF(x).$$

It is called the *Lebesgue-Stieltjes integral* defined by F.

A similar construction can be used to define measures on $\mathbb{R}^n$. Here $F(x_1, \ldots, x_n)$ is assumed to be a non-negative function which is monotone, non-decreasing and continuous from the right in each variable separately, satisfying

$$\lim_{x_j \to -\infty} F(x_1, \ldots, x_j, \ldots, x_n) = 0 \text{ for } j = 1, \ldots, n.$$

The measure of $(-\infty, a_1] \times \cdots \times (-\infty, a_n]$ is defined to be

$$\nu_F((-\infty, a_1] \times \cdots \times (-\infty, a_n]) = F(a_1, \ldots, a_n).$$

By appropriately adding and subtracting one defines the measure of a finite rectangle. For example, if $n = 2$ then

$$\nu_F((a_1, b_1] \times (a_2, b_2]) = F(b_1, b_2) + F(a_1, a_2) - F(a_2, b_1) - F(a_1, b_2). \tag{10.3}$$

A discussion of Lebesgue-Stieltjes integrals can be found in [12].

How is the notion of probability introduced above related to the outcomes of experiments? We consider the case of flipping a coin, suppose that heads occurs with probability p and tails occurs with probability $1-p$. What is meant by that? Suppose we flip a coin N times, and let $H(N)$ and $T(N)$ be the number of heads and tails respectively. Intuitively the statement $\text{Prob}(\{H\}) = p$ is interpreted to mean

$$\lim_{N \to \infty} \frac{H(N)}{N} = p.$$

This is sometimes called a *time average*: we perform a sequence of identical experiments and take the average of the results. We expect that the average over time will approach the theoretical probability. Another way to model the connection between the outcomes of experiments and probability theory is to consider *ensemble averages.* For the example of coin tossing, we imagine that we have N *identical* coins. We flip each coin once; again let $H(N)$ be the number of heads. If

$$\lim_{N \to \infty} \frac{H(N)}{N} = p$$

then we say that heads occurs, for 'this coin', with probability p. The existence of either of these limits means that the process of flipping coins has some sort of statistical regularity. It is a simple matter to concoct sequences $< H(N) >$ for which the ratio $H(N)/N$ does not converge to a limit. It is, in essence an experimental fact that such sequences are very unlikely to occur. Barlow, [3] presents a very nice of discussion of the philosophical issues underlying the application of probability to experimental science.

Exercise 10.1.8. Suppose that $(X, \mathcal{M}, \nu)$ is a probability space and $A, B \in \mathcal{M}$ are two allowable events. Show that

$$\nu(A \cup B) = \nu(A) + \nu(B) - \nu(A \cap B).$$

Explain why this is reasonable from the point of view of probability.

Exercise 10.1.9. Show that if $A \in \mathcal{M}$ then χ_A is a measurable function.

Exercise 10.1.10. For the σ-algebra $\mathcal{M}$ defined in example 10.1.1 what are the measurable functions?

Exercise 10.1.11. Prove that if f is a measurable function then the sets $A_{N,j}$ defined in (10.1) belongs to $\mathcal{M}$ for every j and N.

Exercise 10.1.12. Show that

$$\int_X F_N d\nu(x) \leq \int_X F_{N+1} d\nu(x).$$

Exercise 10.1.13. Show that if f is a measurable function such that $|f(x)| \leq M$ for all $x \in X$ then

$$\int_X f(x) d\nu(x) \leq M.$$

What can be concluded from the equality condition in this estimate?

Exercise 10.1.14. If f is a bounded function then, for each N there exists an M_N so that

$$A_{N,j} = \emptyset$$

if $|j| \geq M_N$.

Exercise 10.1.15. With $\mathcal{M}$ defined in example 10.1.1 and ν defined in example 10.1.9, which functions are integrable and what is the integral of an integrable function?

Exercise 10.1.16. Give a method for extending the definition of the integral to non-negative functions, which may not be bounded.

Exercise 10.1.17. If $(X, \mathcal{M}, \nu)$ is a probability space, $\{A_j\} \subset \mathcal{M}$ are pairwise disjoint subsets and $\{a_j\}$ is a bounded sequence then show that

$$\sum_{j=1}^{\infty} |a_j| \nu(A_j) < \infty.$$

Exercise 10.1.18. Suppose that $F(x)$ is a differentiable function with derivative f, show that

$$\int_{-\infty}^{\infty} g(x) \nu_F(x) = \int_{-\infty}^{\infty} g(x) f(x) dx. \tag{10.4}$$

Exercise 10.1.19. Suppose that F has a weak derivative f. Does (10.4) still hold?

Exercise 10.1.20. Let $F(x) = 0$ for $x < 0$ and $F(x) = 1$ for $x \geq 0$. For a continuous function, g what is

$$\int_{-\infty}^{\infty} g(x) \nu_F(x)?$$

10.1.4 Independent events

Suppose that a coin is flipped several times in succession. The outcome of one flip should not affect the outcome of a successive flip, nor is it affected by an earlier flip. They are *independent events.* This is a general concept in probability theory.

Definition 10.1.4. Let $(X, \mathcal{M}, \nu)$ be a probability space. Two allowable events, A and B are called *independent* if

$$\text{Prob}(A \cap B) = \text{Prob}(A)\,\text{Prob}(B). \tag{10.5}$$

Earlier we said that two events A and B were mutually exclusive if $A \cap B = \emptyset$, in this case

$$\text{Prob}(A \cup B) = \text{Prob}(A) + \text{Prob}(B).$$

Note the difference between these concepts.

Example 10.1.15. Let $X_N = \{(a_1, a_2, \ldots, a_N) | a_i \in \{H, T\}\}$ be the sample space for flipping a coin N times. Suppose that $\text{Prob}(H) = p$ and $\text{Prob}(T) = 1 - p$. If successive flips are independent then formula (10.5) implies that

$$\text{Prob}((a_1, a_2, \ldots, a_N)) = \prod_{i=1}^{N} \text{Prob}(a_i) = p^{m_{\mathbf{a}}}(1-p)^{N - m_{\mathbf{a}}}.$$

To see this, observe that the event $a_1 = H$ is the set $\{(H, a_2, \ldots, a_N) : a_j \in \{H, T\}\}$; it has probability p because it evidently only depends on the outcome of the first flip. Similarly the event $a_1 = T$ has probability $1 - p$. Indeed for any fixed j,

$$\text{Prob}(a_j = H) = p \text{ and } \text{Prob}(a_j = T) = 1 - p.$$

The event $A_k = \{a_i = H, \quad i = 1, \ldots k \text{ and } a_j = T, \quad j = k+1, \ldots, N\}$ can be expressed

$$\left[\bigcap_{i=1}^{k} \{a_i = H\}\right] \bigcap \left[\bigcap_{j=k+1}^{N} \{a_j = T\}\right].$$

Since this is an intersection of independent events it follows that

$$\text{Prob}(A_k) = p^k (1-p)^{N-k}.$$

A similar argument applies if we permute the order in which the heads and tails arise.

For each integer $0 \leq k \leq N$ define the event

$$H_k = \{\mathbf{a} : k \text{ of the } a_i \text{ are } H\}.$$

An element of H_k is a sequence of length N consisting of k H's and $(N - k)$ T's. Let us choose one of them:

$$(\underbrace{H, \ldots, H}_{k}, \underbrace{T, \ldots, T}_{N-k}).$$

The probability of this event is

$$\text{Prob}((\underbrace{H,\ldots,H}_{k},\underbrace{T,\ldots,T}_{N-k})) = p^k(1-p)^{N-k}.$$

To obtain $\text{Prob}(H_k)$, it is enough to calculate the number of different sequences which contain k H's and $(N-k)$ T's. A moments consideration shows that this equals the number of ways to choose k numbers out of N, which is

$$\binom{N}{k} = \frac{N!}{k!(N-k)!};$$

hence,

$$\text{Prob}(H_k) = \binom{N}{k} p^k(1-p)^{N-k}.$$

Exercise 10.1.21. Suppose that we perform the experiment of flipping a coin N-times. The outcome is the number of heads. The sample space for this experiment is $X = \{0,1,\ldots,N\}$. Show that for each $0 \leq p \leq 1$ the function defined on X

$$\text{Prob}(\{k\}) = \binom{N}{k} p^k(1-p)^{N-k}$$

defines a probability on X.

Exercise 10.1.22. Suppose that we perform the experiment of flipping a coin N-times. Suppose that the probability that the i^{th} flip is a head equals p_i and that all the flips are independent. What is the probability of getting exactly k heads in N-flips? Find a plausible physical explanation for having different probabilities for different flips while maintaining the independence of the successive flips.

Exercise 10.1.23. Suppose that the probability that a coin lands on heads is p. Describe an experiment to decide whether or not successive flips are independent.

10.1.5 Conditional probability

Another important notion in probability theory is called *conditional probability*. Here we suppose that we have a probability space $(X,\mathcal{M},\nu)$ and that $B \in \mathcal{M}$ is an event such that $\text{Prob}(B) > 0$. Assume that we know, *a priori* that the event B occurs. The conditional probability of an event "$A \in \mathcal{M}$ given B" is defined by

$$\text{Prob}(A|B) = \frac{\text{Prob}(A \cap B)}{\text{Prob}(B)}.$$

Notice $\text{Prob}(B|B) = 1$, as it must since we *know* that B occurs. We can also define the probability of B given A:

$$\text{Prob}(B|A) = \frac{\text{Prob}(A \cap B)}{\text{Prob}(A)}.$$

They are related by *Bayes' law*

$$\text{Prob}(B|A) = \frac{\text{Prob}(A|B)\,\text{Prob}(B)}{\text{Prob}(A)}.$$

If A and B are independent events than

$$\text{Prob}(A|B) = \text{Prob}(A).$$

This shows that the terminology is consistent, for if A is independent of B then the fact that B occurs can have no bearing on the probability of A occurring.

Example 10.1.16. In our coin tossing experiment we could consider the following question: Suppose that we know that k-heads turn up in N flips, what is the probability that $a_1 = T$? That is what is $\text{Prob}(a_1 = T|H_k)$? It is clear that

$$\begin{aligned}\text{Prob}(\{a_1 = T\} \cap H_k) &= \text{Prob}(k \text{ of } (a_2, \dots a_N) \text{ are } H \text{ and } a_1 = T) \\ &= (1-p)\binom{N-1}{k} p^k (1-p)^{N-1-k}.\end{aligned} \tag{10.6}$$

On the other hand

$$\text{Prob}(H_k) = \binom{N}{k} p^k (1-p)^{N-k}$$

hence

$$\text{Prob}(a_1 = T|H_k) = 1 - \frac{k}{N}.$$

Example 10.1.17. Let $X = \mathbb{R}$ and $\mathcal{M}$, the Borel sets. A non-negative, measurable function f with

$$\int_{-\infty}^{\infty} f(t)dt = 1$$

defines a probability measure

$$\nu_f(A) = \int_{-\infty}^{\infty} f(t)\chi_A(t)dt.$$

Let $B \in \mathcal{M}$ be a set for which $\nu_f(B) > 0$, then the conditional probability of A given B is defined by

$$\text{Prob}(A|B) = \frac{\nu_f(A \cap B)}{\nu_f(B)} = \frac{\int_{A\cap B} f(t)dt}{\int_B f(t)dt}.$$

10.2 Random variables

The sample space for the coin tossing experiment is X_N. For many questions X_N contains more information than is needed. For example, if we are only interested in the number of heads then we could use the simpler sample space $\{0, 1, \ldots, N\}$. The probability of the event $\{k\} \subset \{0, 1, \ldots, N\}$, which is the same as the event $H_k \subset X_N$, is

$$\text{Prob}(\{k\}) = \binom{N}{k} p^k (1-p)^{N-k}$$

The sample space $\{0, \ldots, N\}$ contains strictly less information than X_N, but for the purpose of counting the number of heads, it is sufficient. It is often useful to employ the simplest possible sample space.

Another way of thinking of the sample space $\{0, \ldots, N\}$ is as the range of a function on the "full" sample space X_N. In our example, define the function χ^H on $X = \{H, T\}$ by

$$\chi^H(H) = 1, \chi^H(T) = 0.$$

Similarly, on X_N define

$$\chi_N^H(\mathbf{a}) = \sum_{i=1}^{N} \chi^H(a_i),$$

where $\mathbf{a} = (a_1, \ldots, a_N)$. The set $\{0, \ldots, N\}$ is the range of this function. The event that k heads arise is the event $\{\mathbf{a} : \chi_N^H(\mathbf{a}) = k\}$; its probability is

$$\text{Prob}(H_k) = \nu(\{\mathbf{a} : \chi_N^H(\mathbf{a}) = k\}) = \binom{N}{k} p^k (1-p)^{N-k}.$$

The expression on the right hand side can be thought of as defining a probability measure on the space $\{0, \ldots, N\}$ with

$$\text{Prob}(\{k\}) = \binom{N}{k} p^k (1-p)^{N-k}.$$

Definition 10.2.1. Let $(X, \mathcal{M}, \text{Prob})$ be a probability space. Recall that a real valued function f is measurable if for every $t \in \mathbb{R}$ the set $f^{-1}((-\infty, t])$ belongs to $\mathcal{M}$. A real valued, measurable function on the sample space is called a *random variable.* A complex valued function is measurable if its real and imaginary parts are measurable. A complex valued, measurable function is a *complex random variable.*

Example 10.2.1. The function χ_N^H is a random variable on X_N.

Example 10.2.2. Let X be the unit interval $[0, 1]$, $\mathcal{M}$ the Borel sets in $[0, 1]$ and $\nu([a, b]) = b - a$. The function $f(x) = x$ is a measurable function. It is therefore a random variable on X. For each $k \in \mathbb{Z}$ the functions $e^{2\pi i k x}$ are measurable and are therefore complex random variables.

Thinking of the sample space X as all possible outcomes of an experiment, its points give a complete description of the possible states of the system under study. With this interpretation, one would not expect to be able to determine, completely which $x \in X$ is the outcome of an experiment. Instead one expects to be able to measure some function of x and this is the way that random variables enter in practical applications.

Example 10.2.3. Consider a system composed of a very large number, N of gas particles contained in a fixed volume. The sample space, $X = \mathbb{R}^{6N}$ describes the position and momentum of every particle in the box. To each configuration in X we associate a temperature T and a pressure P. These are real valued, random variables defined on the sample space. In a realistic experiment one can measure T and P, though not the actual configuration of the system at any given moment.

Given the probabilistic description of an experiment $(X, \mathcal{M}, \nu)$ and a random variable $\chi(x)$, which can be measured, it is reasonable to enquire what value of χ we should *expect* to measure. Because $\nu(X) = 1$, the integral of a function over X is a weighted average. In probability theory this is called the *expected value*.

Definition 10.2.2. Let $(X, \mathcal{M}, \nu)$ be a probability space and χ a random variable. Define the expected value or mean of the random variable χ by setting

$$\mu_\chi = E[\chi] = \int_X \chi(x) d\nu(x).$$

If either $\chi_\pm$ has integral $+\infty$ then χ does **not** have an expected value.

In the literature $< \chi >$ is often used to denote the expected value of χ. We avoid this notation as we have already used $< \cdot >$ to denote sequences. In this text μ is also used to denote the absorption coefficient. The meaning should be clear from the context.

Because the expectation is an integral and an integral depends linearly on the integrand, the expectation does as well.

Proposition 10.2.1. *Suppose that $(X, \mathcal{M}, \nu)$ is a probability space and the random variables χ and ψ have finite expected valued, then so does their sum and*

$$E[\chi + \psi] = E[\chi] + E[\psi].$$

Example 10.2.4. One may ask how many heads will occur, on average among N tosses. This is the expected value of the function χ_N^H :

$$E[\chi_N^H] = \sum_{k=0}^{N} k \operatorname{Prob}(\{\chi_N^H(x) = k\}) = \sum_{k=0}^{N} k \binom{N}{k} p^k (1-p)^{N-k} = pN.$$

The expected value can be expressed as the integral over X_N:

$$E[\chi_N^H] = \int_{X_N} \chi_N^H(\mathbf{a}) d\nu_{p,N}(\mathbf{a}).$$

Example 10.2.5. Suppose we play a game: we get one dollar for each head, and lose one dollar for each tail. What is the expected outcome of this game? Note that the number of tails in a sequence $\mathbf{a}$ is $\chi_N^T(\mathbf{a}) = N - \chi_N^H(\mathbf{a})$. The expected outcome of this game is the expected value of $\chi_N^H - \chi_N^T$. It is given by

$$E[\chi_N^H - (N - \chi_N^H)] = E[2\chi_N^H - N] = 2E[\chi_N^H] - E[N] = 2pN - N = (2p-1)N.$$

If $p = \frac{1}{2}$, then this is a fair game: the expected outcome is 0. If $p > \frac{1}{2}$, we expect to make money from this game.

Example 10.2.6. Suppose that X is the unit disk and the probability measure is dA/π. The radius $r = \sqrt{x^2 + y^2}$ is a measurable function. Its expected value is

$$E[r] = \int_X r \frac{\mathrm{dA}}{\pi} = \frac{1}{\pi} \int_0^{2\pi} \int_0^1 r^2 dr d\theta = \frac{2}{3}.$$

In other words, if a point is picked "randomly" in the unit disk its expected radius is 2/3.

Exercise 10.2.1. Derive the result found in example 10.2.4.

Exercise 10.2.2. Instead of making or losing one dollar for each toss, we could adjust the amount of money for each outcome to make the game in example 10.2.5 into a fair game. For a given p find the amount we should receive for each head and pay for each tail to make this a fair game.

Exercise 10.2.3. In example 10.2.6 what are $E[x]$ and $E[y]$?

Exercise 10.2.4. Let X be the unit circle in $\mathbb{R}^2$ with

$$\nu(A) = \int_A \frac{d\theta}{2\pi}.$$

The exponential functions $\{e^{2\pi i k\theta}\}$ are random variables. What is $E[e^{2\pi i k\theta}]$ for $k \in \mathbb{Z}$.

10.2.1 Cumulative distribution function

Associated to a real-valued, random variable is a probability measure on the *real line.* The *cumulative distribution function* for χ is defined to be

$$P_\chi(t) \stackrel{d}{=} \mathrm{Prob}(\{x \,:\, \chi(x) \leq t\}).$$

A cumulative distribution function has several basic properties

(1). It is monotone: $P_\chi(s) \leq P_\chi(t)$ if $s < t$ and continuous from the right.

(2). $\lim_{t \to -\infty} P_\chi(t) = 0$.

(3). $\lim_{t \to \infty} P_\chi(t) = 1$.

From example 10.1.14 it follows that a function satisfying these conditions defines a Lebesgue-Stieltjes, probability measure on $\mathbb{R}$ with

$$\nu_\chi((a, b]) = P_\chi(b) - P_\chi(a).$$

This measure is defined on a σ-algebra which contains the Borel subsets.

Often the cumulative distribution function can be expressed as the integral of a non-negative function

$$P_\chi(t) = \int_{-\infty}^t p_\chi(s) ds.$$

The function $p_\chi(t)$ is called the *density* or *distribution function* for χ. In terms of the distribution function

$$\text{Prob}(a \leq \chi \leq b) = \int_a^b p_\chi(t)dt.$$

Heuristically $p_\chi(t)$ is the "infinitesimal" probability that the value of χ lies between t and $t + dt$. Since probabilities are non-negative this implies that

$$p_\chi(t) \geq 0 \text{ for all } t.$$

The third property of the cumulative distribution implies that

$$\int_{-\infty}^{\infty} p_\chi(t)dt = 1.$$

The expected value of χ can be computed from the distribution function:

$$E[\chi] = \int_X \chi(x)d\nu(x) = \int_{-\infty}^{\infty} tp_\chi(t)dt.$$

Notice that we have replaced an integration over the probability space X by an integration over the *range* of the random variable χ. Often the sample space X and the probability measure ν on X are not explicitly defined. Instead one just speaks of a random variable with a given distribution function. The "random variable" can then be thought of as the coordinate on the real line and the cumulative distribution defines a Lebesgue-Stieltjes measure on $\mathbb{R}$.

Example 10.2.7. A random variable f is said to be *Gaussian* with mean zero if

$$P_f(t) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{t} \exp\left[-\frac{x^2}{2\sigma^2}\right] dx.$$

If f describes the outcome of an experiment, then the probability that the outcome lies in the interval $[a, b]$ is

$$\frac{1}{\sqrt{2\pi}\sigma} \int_a^b \exp\left[-\frac{x^2}{2\sigma^2}\right] dx.$$

Notice that we have described the properties of this random variable *without* defining the space X on which it is defined.

Let χ be a random variable, the k^{th} *moment* of χ exists if

$$\int_X |\chi|^k d\nu < \infty.$$

If χ has a distribution function p_χ then this is equivalent to the condition that

$$\int_{-\infty}^{\infty} |t^k| p_\chi(t) dt < \infty.$$

The k^{th} moment of χ. is then defined to be

$$E[\chi^k] = \int_X \chi^k(x) d\nu(x).$$

In terms of a distribution function

$$E[\chi^k] = \int_{-\infty}^{\infty} t^k p_\chi(t) dt.$$

A more useful quantity is the k^{th}-*centered moment* with is $E[(\chi - \mu_\chi)^k]$. The centered moments measure the deviation of a random variable from its mean value.

The moments of a random variable may not be defined. For example suppose that a real valued, random variable, χ has cumulative distribution:

$$\text{Prob}(\chi \leq t) = \frac{1}{\pi} \int_{-\infty}^{t} \frac{dx}{1+x^2}.$$

Neither the expected value of χ nor of $|\chi|$ exists because

$$\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{|x|}{1+x^2} dx = \infty.$$

Exercise 10.2.5. Suppose that $(X, \mathcal{M}, \nu)$ is a probability space and χ is a non-negative random variable. Show that $E[\chi] \geq 0$.

Exercise 10.2.6. Suppose that $(X, \mathcal{M}, \nu)$ is a probability space and χ is a random variable for which there exists a number c so that

$$\text{Prob}(\chi = c) > 0.$$

Show that $p_\chi(t)$ does not exist.

Exercise 10.2.7. Suppose that $(X, \mathcal{M}, \nu)$ is a probability space and χ is a non-negative random variable with $E[\chi] = \alpha$. Show that

$$\text{Prob}(\chi \geq t) \leq \frac{\alpha}{t}. \tag{10.7}$$

The estimate in (10.7) is called the Chebyshev inequality.

10.2.2 The variance

Of particular interest in applications is the second centered moment, or *variance* of a random variable. It is defined by

$$\sigma_\chi^2 \stackrel{d}{=} E[(\chi - \mu_\chi)^2].$$

The variance is a measure of how frequently a random variable differs from its mean. In experimental science it is a measure of the uncertainty in a measured quantity. It can be expressed in terms of the expectation of χ^2 and $E[\chi]$,

$$\begin{aligned} \sigma_\chi^2 &= E[(\chi - E[\chi])^2] \\ &= E[\chi^2] - 2E[\chi]^2 + E[\chi]^2 \\ &= E[\chi^2] - E[\chi]^2. \end{aligned} \tag{10.8}$$

As the expected value of non-negative random variable, it is always non-negative. The positive square root of the variance, σ_χ is called the *standard deviation.* Zero standard deviation implies that, with probability one, χ is equal to its mean. One could also use $E[|\chi - \mu_\chi|]$ as a measure of the deviation of a random variable from its mean. In applications the variance occurs much more frequently because it is customary and because computations involving the variance are much simpler than those for $E[|\chi - \mu_\chi|^k]$ if $k \neq 2$.

Example 10.2.8. In the coin tossing example,

$$E[(\chi_N^H)^2] = \sum_{k=0}^{N} k^2 \binom{N}{k} p^k (1-p)^{N-k} = pN[p(N-1)+1]$$

Using (10.8), the variance is

$$E[(\chi_N^H - E[\chi_N^H])^2] = pN[p(N-1)+1] - p^2N^2 = p(1-p)N.$$

If $p = 0$, the standard deviation is zero and the coin always falls on tails. A fair coin, i.e., $p = \frac{1}{2}$ has the largest standard deviation $\frac{1}{4}N$.

Example 10.2.9. Suppose that χ is a random variable with mean μ and variance σ^2, then

$$\text{Prob}(|\chi - \mu| \geq t) \leq \frac{\sigma^2}{t^2}. \tag{10.9}$$

This is also called Chebyshev's inequality. The proof uses the observation

$$\{x \,:\, |\chi(x) - \mu| \geq t\} = \{x \,:\, |\chi(x) - \mu^2| \geq t^2\}$$

and therefore

$$\begin{aligned} \text{Prob}(|\chi - \mu|) \geq t) &= \int\limits_{\{x\,:\,|\chi(x)-\mu|^2 \geq t^2\}} d\nu \\ &\leq \int\limits_{\{x\,:\,|\chi(x)-\mu|^2 \geq t^2\}} \frac{|\chi - \mu|^2}{t^2} d\nu \\ &\leq \frac{\sigma^2}{t^2}. \end{aligned} \tag{10.10}$$

This indicates why the variance is regarded as a measure of the uncertainty in the value of a random variable.

Exercise 10.2.8. Why does $E[\chi - \mu_\chi]$ not provide a good measure of the deviation of χ from its mean?

Exercise 10.2.9. Let χ be a Gaussian random variable with distribution function

$$p_\chi(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(t-a)^2}{2\sigma^2}\right].$$

What are $E[\chi]$ and σ_χ?

Exercise 10.2.10. In (10.10) justify the transition from the second to the third line.

Exercise 10.2.11. Deduce from (10.10) that

$$\text{Prob}(|\chi - \mu| < t) \geq 1 - \frac{\sigma^2}{t^2}.$$

10.2.3 The characteristic function

Another important function of a random variable, χ is the expected value of $e^{-2\pi i\lambda\chi}$,

$$M_\chi(\lambda) = E[e^{-2\pi i\lambda\chi}] = \int_X e^{-2\pi i\lambda\chi(x)} d\nu(x).$$

This is called the *characteristic function* of χ.. If the cumulative distribution for χ has a density function, p_χ then, up to the factor of 2π,

$$E[e^{-2\pi i\lambda\chi}] = \int_{-\infty}^{\infty} e^{-2\pi i\lambda t} p_\chi(t) dt$$

is the Fourier transform of the density function. As the density function is a non-negative integrable function, its Fourier transform is continuous and, by the Riemann-Lebesgue lemma tends to zero as $\lambda \to \infty$.

As we saw in Chapter 3, decay of the density function at infinity makes the characteristic function differentiable. Its derivatives at $\lambda = 0$ determine the moments of χ. They are given by

$$\left[\frac{-\partial_\lambda}{2\pi i}\right]^k M_\chi(\lambda)\big|_{\lambda=0} = E[\chi^k].$$

Using the Taylor series for the exponential and computing *formally* gives

$$\begin{aligned} E[e^{-2\pi i\lambda\chi}] &= \sum_{j=0}^{\infty} \frac{[-2\pi i\lambda]^j}{j!} t^j \int_{-\infty}^{\infty} p_\chi(t) dt \\ &= \sum_{j=0}^{\infty} \frac{[-2\pi i\lambda]^j}{j!} E[\chi^j]. \end{aligned} \tag{10.11}$$

For this reason the $E[e^{-2\pi i\lambda\chi}]$ is sometimes called the *generating function* for the moments of χ. Note however that the expected value of $e^{-2\pi i\lambda\chi}$ always exists, while the moment themselves may not.

Example 10.2.10. In the coin tossing example, the characteristic function is

$$\begin{aligned} E[e^{-2\pi i\lambda\chi_N^H}] &= \sum_{k=0}^{N} e^{-2\pi i\lambda k}\binom{N}{k}p^k(1-p)^{N-k} \\ &= (1-p)^N\sum_{k=0}^{N}\binom{N}{k}\big[e^{-2\pi i\lambda}\frac{p}{1-p}\big]^k \\ &= (1-p)^N\big(1+\frac{e^{-2\pi i\lambda}p}{1-p}\big)^N = (1-p(1-e^{-2\pi i\lambda}))^N. \end{aligned}$$

Notice again that we do not integrate over the space X_N, where the random variable χ_N^H is defined, but rather over the range of χ_N^H. This shows the utility of replacing a complicated sample space by a simpler one when doing calculations with random variables.

What happens to the distribution function if a random variable is shifted or rescaled? Suppose χ is a random variable and a new random variable is defined by

$$\psi = \frac{\chi-\mu}{\sigma}.$$

This is often done for convenience, for example if $\mu = E[\chi]$ then ψ is random variable with mean zero. The cumulative distribution function is

$$\text{Prob}(\psi \le t) = \text{Prob}(\frac{\chi-\mu}{\sigma} \le t) = \text{Prob}(\chi \le \sigma t + \mu).$$

If $p_\chi(t)$ is the distribution function for χ then the distribution function for ψ is

$$p_\psi = \sigma p_\chi(\sigma t+\mu), \text{ and } M_\psi(\lambda) = e^{\frac{2\pi i\lambda\mu}{\sigma}}M_\chi(\frac{\lambda}{\sigma}). \tag{10.12}$$

Exercise 10.2.12. What is the characteristic function of the Gaussian random variable defined in exercise 10.2.9?

Exercise 10.2.13. Derive the formulæ in (10.12).

10.2.4 A pair of random variables

Often times we have more than one random variable. It is then important to understand how they are related. Suppose that χ_1 and χ_2 are random variables on the same space X. By analogy to the cumulative distribution for a single random variable, we define the *joint cumulative distribution function* of χ_1 and χ_2 by

$$\text{Prob}(\chi_1 \le s \text{ and } \chi_2 \le t) = \nu(\{\chi_1^{-1}(-\infty, s]\} \cap \{\chi_2^{-1}(-\infty, t]\}).$$

This function is monotone non-decreasing and continuous from the right in each variable and therefore defines a Lebesgue-Stieltjes measure ν_{χ_1,χ_2} on $\mathbb{R}^2$, see example 10.1.14. The measure of a rectangle is given by the formula

$$\begin{aligned}\nu_{\chi_1,\chi_2}((a,b]\times(c,d]) = &\operatorname{Prob}(\chi_1 \le b, \text{ and } \chi_2 \le d) + \operatorname{Prob}(\chi_1 \le a, \text{ and } \chi_2 \le c) - \\ &\operatorname{Prob}(\chi_1 \le a, \text{ and } \chi_2 \le d) - \operatorname{Prob}(\chi_1 \le b, \text{ and } \chi_2 \le c)\end{aligned} \tag{10.13}$$

If there is a function $p_{\chi_1,\chi_2}(x,y)$ defined on $\mathbb{R}^2$ such that

$$\operatorname{Prob}(\chi_1 \le s, \text{ and } \chi_2 \le t) = \int_{-\infty}^{s}\int_{-\infty}^{t} p_{\chi_1,\chi_2}(x,y)dydx.$$

then we say that p_{χ_1,χ_2} is the *joint distribution function* for the pair of random variables (χ_1,χ_2).

It is clear that

$$\begin{aligned}&\operatorname{Prob}(\chi_1 \le s, \text{ and } \chi_2 \le \infty) = \operatorname{Prob}(\chi_1 \le s) \text{ and} \\ &\operatorname{Prob}(\chi_2 \le s, \text{ and } \chi_1 \le \infty) = \operatorname{Prob}(\chi_2 \le s).\end{aligned} \tag{10.14}$$

This is reasonable because the condition $\chi_i \le \infty$ places no restriction on χ_i. This is expressed in terms of the distribution functions by the relations

$$\begin{aligned}\int_{-\infty}^{s}\int_{-\infty}^{\infty} p_{\chi_1,\chi_2}(x,y)dydx &= \int_{-\infty}^{s} p_{\chi_1}(x)dx, \\ \int_{-\infty}^{s}\int_{-\infty}^{\infty} p_{\chi_1,\chi_2}(x,y)dxdy &= \int_{-\infty}^{s} p_{\chi_2}(y)dy.\end{aligned} \tag{10.15}$$

The joint distribution function therefore, is not independent of the distribution functions for individual random variables. It must satisfy the consistency conditions:

$$p_{\chi_2}(y) = \int_{-\infty}^{\infty} p_{\chi_1,\chi_2}(x,y)dx, \text{ and } p_{\chi_1}(x) = \int_{-\infty}^{\infty} p_{\chi_1,\chi_2}(x,y)dy.$$

Recall that two events A and B are independent if:

$$\operatorname{Prob}(A\cap B) = \operatorname{Prob}(A)\operatorname{Prob}(B),$$

Similarly two random variables, χ_1 and χ_2 are *independent* if

$$\operatorname{Prob}(\chi_1 \le s \text{ and } \chi_2 \le t) = \operatorname{Prob}(\chi_1 \le s)\operatorname{Prob}(\chi_2 \le t).$$

In terms of their distribution functions this is equivalent to

$$p_{\chi_1,\chi_2}(x,y) = p_{\chi_1}(x)p_{\chi_2}(y).$$

The expected value of a product of random variables, having a joint distribution function is given by

$$E[\chi_1\chi_2] = \iint xy \cdot p_{\chi_1,\chi_2}(x,y)dxdy \tag{10.16}$$

Whether or not χ_1 and χ_2 have a joint distribution function, this expectation is an integral over the sample space and therefore $E[\chi_1\chi_2]$ satisfies the Cauchy-Schwarz inequality.

Proposition 10.2.2. *Let χ_1 and χ_2 be a pair of random variables defined on the same sample space with finite mean and variance then*

$$E[\chi_1\chi_2] \leq \sqrt{E[|\chi_1|^2]E[|\chi_2|^2]}. \tag{10.17}$$

It is useful to have a simple way to quantify of the interdependence of a pair of random variables.

Definition 10.2.3. The *covariance* of χ_1 and χ_2, defined by

$$\mathrm{Cov}(\chi_1,\chi_2) = E[(\chi_1 - \mu_{\chi_1})(\chi_2 - \mu_{\chi_2})] = E[\chi_1\chi_2] - E[\chi_1]E[\chi_2],$$

and *correlation coefficient*

$$\rho_{\chi_1\chi_2} = \frac{\mathrm{Cov}(\chi_1,\chi_2)}{\sigma_{\chi_1}\sigma_{\chi_2}}$$

are the fundamental measures of independence. If χ_1 is measured in units u_1 and χ_2 is measured in units u_2 then the covariance has units $u_1 \cdot u_2$. The correlation coefficient is a more useful measure of the interdependence of χ_1 and χ_2 because it is dimensionally independent; it is a pure number assuming values between ± 1, see exercise 10.1.3.

If χ_1, χ_2 are independent then

$$\begin{aligned} E[\chi_1\chi_2] &= \int_{-\infty}^{\infty} yp_{\chi_2}(y)dy \int_{-\infty}^{\infty} xp_{\chi_1}(x)dx \\ &= E[\chi_1]E[\chi_2]. \end{aligned} \tag{10.18}$$

In this case the covariance, $\mathrm{Cov}(\chi_1,\chi_2)$ is equal to zero. This is a necessary, but not sufficient condition for two random variables to be independent.

Example 10.2.11. Zero covariance does **not** imply the independence of two random variables. We illustrate this point with a simple example. Let $X = [0,1]$ and $d\nu = dx$. Two random variables are defined by:

$$\chi_1 = \cos 2\pi x, \quad \chi_2 = \sin 2\pi x.$$

Their means are clearly zero, $E[\chi_1] = E[\chi_2] = 0$. They are also uncorrelated:

$$\mathrm{Cov}(\chi_1,\chi_2) = \int_0^1 \cos 2\pi x \sin 2\pi x dx = 0.$$

On the other hand we compute the probability:

$$\text{Prob}(0 \leq |\sin 2\pi x| \leq \frac{1}{\sqrt{2}}, \text{ and } \frac{1}{\sqrt{2}} \leq |\cos 2\pi x| \leq 1).$$

From the identity $\cos\theta^2 + \sin\theta^2 = 1$, the first condition is equivalent to the second one. Using the graph of $\sin 2\pi x$, one can easily check that

$$\text{Prob}(0 \leq |\sin 2\pi x| \leq \frac{1}{\sqrt{2}}, \text{ and } \frac{1}{\sqrt{2}} \leq |\cos 2\pi x| \leq 1) = \text{Prob}(0 \leq |\sin 2\pi x| \leq \frac{1}{\sqrt{2}}) = \frac{1}{2}.$$

But, the product is

$$\text{Prob}(0 \leq |\sin 2\pi x| \leq \frac{1}{\sqrt{2}})\,\text{Prob}(\frac{1}{\sqrt{2}} \leq |\cos 2\pi x| \leq 1) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}.$$

Hence these are not independent variables.

Random variables are, in many ways, like ordinary variables. The properties of a single real valued, random variable are entirely specified by its cumulative distribution, which in turn defines a measure on the real line. Indeed nothing is lost by thinking of the random variable χ as *being* the coordinate on $\mathbb{R}$. If f is a function, then $f(\chi)$ is a new random variable and it bears the same relation to χ as $f(x)$ bears to x, for example

$$E[f(\chi)] = \int_{-\infty}^{\infty} f(x)p_\chi(x)dx$$

and

$$\text{Prob}(a \leq f(\chi) \leq b) = \int_{f^{-1}([a,b])} p_\chi(x)dx$$

So long as we are only interested in random variables which are functions of χ we can think of our sample space as being the real line.

Two functions defined on the plane are thought of as being independent if they behave like x, y coordinates. That is, one is in no way a function of the other. This is the essence of the meaning of independence for random variables: One is not a function of the other, with probability 1. Independence is a very strong condition. If χ_1, χ_2 are independent, then

$$E[f(\chi_1)g(\chi_2)] = E[f(\chi_1)]E[g(\chi_2)] \tag{10.19}$$

for any functions f and g such that this makes sense. To work with a pair of random variables we use $\mathbb{R}^2$ as the underlying sample space. If the variables are independent then ν_{χ_1,χ_2} is the induced product measure, $\nu_{\chi_1} \times \nu_{\chi_2}$. As with ordinary functions, there are degrees of dependence between two random variables. If χ_1 and χ_2 are random variables which are **not** independent then it does not mean that one is a function of the other or that there is a third random variable, χ_3 so that $\chi_1 = f(\chi_3)$ and $\chi_2 = g(\chi_3)$. When working with random variables it is often very useful to replace them by coordinates or functions of coordinates on a Euclidean space.

Example 10.2.12. Let X_N be the sample space for N-coin tosses. As noted above we usually assume that the results of the different tosses in the sequence are independent of one another. In example 10.1.15 we showed that the probability that a sequence $\mathbf{a}$ has k heads and $N-k$ tails is $p^k(1-p)^{N-k}$. The corresponding measure on X_N is denoted $\nu_{p,N}$. To translate this example into the language of random variables we define the functions $\chi_j : X_N \to \{0,1\}$ by letting

$$\chi_j(\mathbf{a}) = \begin{cases} 1 & \text{if } a_j = H, \\ 0 & \text{if } a_j = T. \end{cases}$$

With the probability defined by $\nu_{p,N}$ these random variables are pairwise independent. Observe that χ_j is only a function of a_j so the various χ_j are also *functionally* independent of one another. Using a different probability measure we could arrange to have $\sigma_{\chi_i\chi_j} \neq 0$, so that the $\{\chi_j\}$ are no longer independent as random variables.

If χ is a random variable recall that the characteristic function is defined by

$$M_\chi(\lambda) \stackrel{d}{=} E[e^{-2\pi i\lambda\chi}] = \int e^{-2\pi i\lambda\chi(\xi)} d\nu(\xi).$$

If χ has a distribution function, then

$$M_\chi(\lambda) = \int e^{-2\pi i\lambda t} p_\chi(t)dt = \hat{p}_\chi(2\pi\lambda).$$

As a nice application of the characteristic function formalism we compute the distribution function for the sum of a pair of independent random variables. Suppose χ_1, χ_2 are independent random variables with distribution functions p_{χ_1}, p_{χ_2} respectively. What is the distribution function for $\chi_1 + \chi_2$? It is calculated as follows:

$$\begin{aligned} M_{\chi_1+\chi_2}(\lambda) &= E[e^{-2\pi i\lambda(\chi_1+\chi_2)}] = E[e^{-2\pi i\lambda\chi_1} e^{-2\pi i\lambda\chi_2}] \\ &= E[e^{-2\pi i\lambda\chi_1}]E[e^{-2\pi i\lambda\chi_2}] \\ &= M_{\chi_1}(\lambda)M_{\chi_2}(\lambda). \end{aligned}$$

The second line comes from the fact that χ_1 and χ_2 are independent. On the other hand $M_\chi(\lambda) = \hat{p}_\chi(2\pi\lambda)$, hence $\hat{p}_{\chi_1+\chi_2} = \hat{p}_{\chi_1}\hat{p}_{\chi_2}$. This implies

$$p_{\chi_1+\chi_2} = p_{\chi_1} * p_{\chi_2}$$

and therefore

$$\begin{aligned} \text{Prob}(\chi_1 + \chi_2 \leq t) &= \int_{-\infty}^{t} p_{\chi_1+\chi_2}(s)ds \\ &= \int_{-\infty}^{t}\int_{-\infty}^{\infty} p_{\chi_1}(s-y)p_{\chi_2}(y)dyds \qquad (10.20) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{t-y} p_{\chi_1}(x)dxp_{\chi_2}(y)dy. \end{aligned}$$

We can understand the last expression intuitively: Heuristically the given probability can be written as

$$\text{Prob}(\chi_1 + \chi_2 \leq t) = \cup_{y\in(-\infty,\infty)} \text{Prob}(\chi_1 \leq t - y, \chi_2 = y).$$

Note that $p_{\chi_2}(y)dy$ is the "infinitesimal "probability" that $\chi_2 = y$. The argument is not rigorous, in part because the right hand side is an *uncountable* union of events and the axioms of a σ-algebra only assure good behavior for countable unions!

Exercise 10.2.14. Suppose that $(X, \mathcal{M}, \nu)$ is a probability space and χ_1 and χ_2 are random variables. Express $E[\chi_1\chi_2]$ as an integral over X.

Exercise 10.2.15. Give a geometric explanation for formula (10.13). When a joint distribution function exists show that $\nu_{\chi_1,\chi_2}((a, b] \times (c, d])$ reduces to the expected integral.

Exercise 10.2.16. Suppose that χ_1 and χ_2 are random variables with finite mean and variance show that

$$-1 \leq \rho_{\chi_1\chi_2} \leq 1.$$

Exercise 10.2.17. In the situation of the previous exercise show that

$$|\rho_{\chi_1\chi_2}| = 1$$

if and only if $\chi_2 = a\chi_1 + b$ for some constants a, b. More precisely $\text{Prob}(\chi_2 = a\chi_1 + b) = 1$.

Exercise 10.2.18. Prove the expectation version of the Cauchy inequality, (10.2.2).

Exercise 10.2.19. Prove the statement in example 10.2.12, that χ_j and χ_k are independent random variables if $j \neq k$.

Exercise 10.2.20. Suppose that χ is a random variable with distribution function $p_\chi(x)$. Let f and g be functions, show that $\text{Prob}(f(\chi) \leq a, g(\chi) \leq b)$ can be expressed in the form

$$\text{Prob}(f(\chi) \leq a, g(\chi) \leq b) = \int\limits_{E_{f,g}} p_\chi(x)dx,$$

where $E_{f,g}$ is a subset of $\mathbb{R}$.

Exercise 10.2.21. Suppose that χ_1 and χ_2 are independent random variables and f, g are functions. Show that $f(\chi_1)$ and $g(\chi_2)$ are also independent random variables.

Exercise 10.2.22. Suppose that χ_1 and χ_2 are random variables and that f and g are functions. Does

$$E[\chi_1\chi_2] = E[\chi_1] \cdot E[\chi_2]$$

imply that

$$E[f(\chi_1)g(\chi_2)] = E[f(\chi_1)] \cdot E[g(\chi_2)]?$$

Give a proof or counterexample.

Exercise 10.2.23. A probability measure space is defined on $\mathbb{R}^2$ by

$$\text{Prob}(A) = \frac{1}{\pi}\int\limits_A \exp[-(x^2+y^2)]dxdy.$$

Are the functions $x+y$ and $x-y$ independent random variables? How about x and $x+y$?

Exercise 10.2.24. Suppose that (χ_1, χ_2) is a pair of independent random variables with means (μ_1, μ_2) and variances (σ_1^2, σ_2^2). Show that

$$\mu_{\chi_1+\chi_2} = \mu_1 + \mu_2 \text{ and } \sigma^2_{\chi_1+\chi_2} = \sigma_1^2 + \sigma_2^2. \tag{10.21}$$

Exercise 10.2.25. Suppose that (χ_1, χ_2) is a pair of random variables with means (μ_1, μ_2), variances (σ_1^2, σ_2^2) and covariance σ_{12}. Find formulæ for $\mu_{\chi_1+\chi_1}$ and $\sigma^2_{\chi_1+\chi_2}$.

Exercise 10.2.26. In example 10.2.12 find a probability measure on X_N so that $\text{Cov}(\chi_i, \chi_j) \neq 0$.

10.2.5 Several random variables

These concepts can be generalized to more than two random variables. Suppose that $\{\chi_1, \ldots, \chi_m\}$ is a collection of m real valued, random variables. Their *joint cumulative distribution* is a function on $\mathbb{R}^m$ defined by

$$\begin{aligned} P_{\chi_1,\ldots,\chi_m}(t_1,\ldots,t_m) &= \text{Prob}(\chi_1 \leq t_1 \text{ and } \ldots \text{ and } \chi_m \leq t_m) \\ &= \nu\left(\bigcap_{j=1}^m \{\chi_j^{-1}((-\infty, t_j])\}\right). \end{aligned} \tag{10.22}$$

This function is monotone non-decreasing and continuous from the right in each variable separately and therefore defines a Lebesgue-Stieltjes measure on $\mathbb{R}^m$. If there is a function $p_{\chi_1,\ldots,\chi_m}(t_1,\ldots,t_m)$ so that

$$P_{\chi_1,\ldots,\chi_m}(t_1,\ldots,t_m) = \int\limits_{-\infty}^{t_1}\cdots\int\limits_{-\infty}^{t_m} p_{\chi_1,\ldots,\chi_m}(s_1,\ldots,s_m)ds_1\cdots ds_m$$

then $p_{\chi_1,\ldots,\chi_m}$ is called the *joint distribution function* for this collection of random variables.

Definition 10.2.4. If $\{\chi_1, \ldots, \chi_m\}$ is a collection of random variables with joint distribution function $p_{\chi_1,\ldots,\chi_m}$ then we say that the variables are independent if

$$\text{Prob}(\chi_1 \leq t_1 \text{ and } \ldots \text{ and } \chi_m \leq t_m) = \text{Prob}(\chi_1 \leq t_1)\cdots\text{Prob}(\chi_m \leq t_m).$$

If there is a joint distribution function this is equivalent to

$$p_{\chi_1,\ldots,\chi_m}(t_1,\ldots,t_m) = p_{\chi_1}(t_1)\cdots p_{\chi_m}(t_m).$$

Once again it is useful to have a statistical measure of independence. The expected values of the products $E[\chi_j\chi_k]$ is an $m \times m$-matrix called the *correlation matrix* and the difference

$$\mathrm{Cov}(\chi_j, \chi_k) = E[\chi_j\chi_k] - E[\chi_j]E[\chi_k]$$

is called the *covariance matrix*. The dimensionless version is the normalized correlation matrix defined by

$$\rho_{\chi_j,\chi_k} = \frac{\mathrm{Cov}(\chi_j, \chi_k)}{\sigma_{\chi_j}\sigma_{\chi_k}}.$$

If $\{\chi_1, \ldots, \chi_m\}$ are random variables then there is a nice formalism for computing certain conditional probabilities. For example, suppose we would like to compute the probability of the event

$$\chi_1 \leq t_1, \ldots, \chi_k \leq t_k \text{ given that } \chi_{k+1} = s_{k+1}, \ldots, \chi_m = s_m. \tag{10.23}$$

To find the distribution function, the event $\{\chi_{k+1} = s_{k+1}, \ldots, \chi_m = s_m\}$ is thought of as the limit of the events $\{|\chi_{k+1} - s_{k+1}| \leq \epsilon, \ldots, |\chi_m - s_m| \leq \epsilon\}$ as $\epsilon \to 0$.

The limiting distribution function does not exist unless the event $\chi_j = s_j$ for $j = k+1, \ldots, m$ has "non-zero infinitesimal probability," i.e.

$$p_{\chi_{k+1},\ldots,\chi_m}(s_{k+1}, \ldots, s_m) \neq 0$$

To obtain a simple formula we also require that $p_{\chi_{k+1},\ldots,\chi_m}$ be continuous at $(s_{k+1}, \ldots, s_m)$ and $p_{\chi_1,\ldots,\chi_m}$ is as well. The probability in the limiting case is then given by

$$P(\chi_1 \leq t_1, \ldots, \chi_k \leq t_k | \chi_{k+1} = s_{k+1}, \ldots, \chi_m = s_m) = \frac{\int_{-\infty}^{t_k} \cdots \int_{-\infty}^{t_1} p_{\chi_1,\ldots,\chi_m}(x_1, \ldots, x_k, s_{k+1}, \ldots, s_m) dx_1 \cdots dx_k}{p_{\chi_{k+1},\ldots\chi_m}(s_{k+1}, \ldots, s_m)}. \tag{10.24}$$

The joint density for the random variables $\{\chi_1, \ldots, \chi_k\}$ given that

$$\chi_j = s_j \text{ for } j = k+1, \ldots, m$$

is therefore

$$\frac{p_{\chi_1,\ldots,\chi_m}(x_1, \ldots, x_k, s_{k+1}, \ldots, s_m)}{p_{\chi_{k+1},\ldots,\chi_m}(s_{k+1}, \ldots, s_m)}. \tag{10.25}$$

Measure theory and probability theory are two different languages for describing to same set of issues. The following chart summarizes the correspondence between the basic vocabulary of the two subjects:

	Measure Theory	**Probability Theory**
X	space	sample space
$\mathcal{M}$	$\sigma-$ algebra	allowable events
ν	measure	probability
χ	measurable function	random variable.

The properties of a finite collection of random variables are developed in the following exercises.

Exercise 10.2.27. Show that if $\{\chi_1, \ldots, \chi_m\}$ are independent random variables then for each pair $i \neq j$ the variables χ_i and χ_j are independent. Is the converse statement true, i.e. does pairwise independence imply that a collection of random variables are independent?

Exercise 10.2.28. Suppose that the random variables $\{\chi_1, \ldots, \chi_m\}$ are pairwise independent. Show that $\mathrm{Cov}(\chi_j, \chi_k) = 0$.

Exercise 10.2.29. Let $\{\chi_1, \ldots, \chi_m\}$ be random variables with joint distribution function $p_{\chi_1,\ldots,\chi_m}$. Show that

$$p_{\chi_1,\ldots,\chi_{m-1}}(t_1, \ldots, t_{m-1}) = \int_{-\infty}^{\infty} p_{\chi_1,\ldots,\chi_m}(t_1, \ldots, t_{m-1}, s)ds.$$

Exercise 10.2.30. Show that if $\{\chi_1, \ldots, \chi_m\}$ have a joint distribution function and $1 \leq i_1 < \cdots < i_k \leq m$ then $\{\chi_{i_1}, \ldots, \chi_{i_k}\}$ also have a joint distribution function. Give a formula for it.

Exercise 10.2.31. Suppose that $\{\chi_1, \ldots, \chi_n\}$ are independent random variables with means $\{\mu_1, \ldots, \mu_n\}$ and variances $\{\sigma_1^2, \ldots, \sigma_n^2\}$. Let

$$\bar{\chi} = \frac{\chi_1 + \cdots + \chi_n}{n},$$

show that

$$\mu_{\bar{\chi}} = \frac{\mu_1 + \cdots + \mu_n}{n} \text{ and } \sigma_{\bar{\chi}}^2 = \frac{\sigma_1^2 + \cdots + \sigma_n^2}{n^2}. \tag{10.26}$$

Does this formula remain valid if we only assume that $\mathrm{Cov}(\chi_i, \chi_j) = 0$?

Exercise 10.2.32. Suppose that $\{\chi_1, \ldots, \chi_m\}$ are independent random variables with distribution functions $\{p_{\chi_1}, \ldots, p_{\chi_m}\}$. What is the distribution function of $\chi_1 + \cdots + \chi_m$? Hint: Show that the characteristic function of the sum is $M_{\chi_1} \cdots M_{\chi_m}$.

Exercise 10.2.33. Suppose that $\{\chi_1, \ldots, \chi_m\}$ are random variables on a probability space $(X, \mathcal{M}, \nu)$ and let $c_{ij} = E[\chi_i \chi_j]$ be their correlation matrix. Show that this matrix is non-negative definite, that is, if $(x_1, \ldots, x_m) \in \mathbb{R}^m$ then

$$\sum_{i=1}^{m} \sum_{j=1}^{m} c_{ij} x_i x_j \geq 0.$$

Hint: Express this as the expectation of a non-negative random variable.

Exercise 10.2.34. Fill in the details in the derivation of the formula (10.25), for the density function of the conditional probability,

$$P(\chi_1 \leq t_1, \ldots, \chi_k \leq t_k | \chi_{k+1} = s_{k+1}, \ldots, \chi_m = s_m).$$

10.3 Some important random variables

In medical imaging and in physics there are three fundamental probability distributions. We now introduce them and discuss some of their properties.

10.3.1 Bernoulli Random Variables

A Bernoulli random variable is specified by two parameters, $p \in [0,1]$ and $N \in \mathbb{N}$. The variable χ assumes the values $\{0, 1, \ldots, N\}$ with probabilities given by

$$\text{Prob}(\chi = k) = \binom{N}{k} p^k (1-p)^{N-k}. \tag{10.27}$$

The number of heads in N-independent coin tosses is an example of a Bernoulli random variable. Sometimes these are called *binomial random variables*.

Recall that in the coin tossing experiment we defined a function χ_N^H such that

$$\chi_N^H((a_1, \ldots, a_N)) = \text{ number of heads in } \mathbf{a}.$$

There is a similar model for a γ-ray detector. The model is summarized by the following axioms

- Each photon incident on the detector is detected with probability p.
- Independence axiom: The detection of one photon is independent of the detection of any other.

Let χ denote the number of photons detected out of N arriving at the detector. The probability that k out of N incident photons are detected is given by (10.27). We see that

$$\begin{aligned} \text{expected value} &: E[\chi] = pN, \\ \text{variance} &: \sigma^2 = E[(\chi - Np)^2] = p(1-p)N. \end{aligned}$$

If $p = 1$, that means we have a perfect detector, hence there is no variance. There is also no variance if $p = 0$. In the latter case the detector is turned off.

Suppose we know the detector, i.e., p is known from many experiments.. The number N characterizes the intensity of the source. We would like to know how many photons were emitted by the source. If we measure M photons a reasonable guess for N is given by $pN = M$. Of course we do not really expect that this is true as the variance is in general not zero. What this means is that, if all our assumptions are satisfied, **and** we repeat the measurement many times then the average value of the measurements should approach pN.

10.3.2 Poisson Random Variables

A Poisson random variable χ assumes the values $\{0, 1, 2, \ldots\}$ and it is characterized by the following probability distribution

$$\text{Prob}(\chi = k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

where λ is a positive number. This defines a probability measure on the nonnegative integers since

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{\lambda} e^{-\lambda} = 1.$$

The expected value is given by

$$E[\chi] = \sum k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda.$$

The constant λ is called the intensity. Poisson random variables are used to model many different situations. Some examples are

- The arrival of patients at a doctor's office.
- The number of telephone calls passing through a switch.
- The number of radioactive decays occurring in a large quantity of a radioactive element, in a fixed amount of time.
- The generation of X-rays.

The standard deviation is given by

$$\sigma_\chi^2 = E[(\chi - E[\chi])^2] = \lambda.$$

Notice that the variance is equal to the expected value. This has an interesting consequence. The *signal-to-noise ratio*, (SNR) is defined as the expected value divided by the standard deviation. For a Poisson random variable it is given by

$$\frac{\text{expected value}}{\text{standard deviation}} = \frac{\lambda}{\sqrt{\lambda}} = \sqrt{\lambda}.$$

Hence, the intensity of the Poisson random variable measures the relative noise in the system.

Exercise 10.3.1. Derive the formulæ for the mean and standard deviation of a Poisson random variable.

10.3.3 Gaussian Random Variables

The final class of distributions we discuss are Gaussian random variables. These have already been briefly discussed. Gaussian random variables are determined by their first and second moments and have many special properties as a consequence of this fact. They are very important in the context of measurement as the average of a large collection of independent random variables is approximately Gaussian, almost no matter how the individual variables are distributed. This fact, known as the "Central Limit Theorem" is treated in the next section.

A random variable, χ is Gaussian if and only if its cumulative distribution is given by

$$\text{Prob}(\chi \leq t) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{t} \exp\big[-\frac{(x-\mu)^2}{2\sigma^2}\big] dx.$$

Integrating gives formulæ for the mean and variance

$$E[\chi] = \mu, \quad E[(\chi - \mu)^2] = \sigma^2.$$

The standard deviation has the interpretation that the probability that χ lies between $\mu - \sigma$ and $\mu + \sigma$ is about 2/3. From the definition it is clear that the converse statement is also true: The distribution function of a Gaussian random variable is determined by its mean and variance.

The characteristic function of a Gaussian random variable is

$$M_\chi(\lambda) = e^{2\pi i \mu \lambda} e^{-\frac{\sigma^2 (2\pi\lambda)^2}{2}}.$$

Higher moments are easily computed using the fact that if

$$f(t) = \frac{1}{\sqrt{t}} = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\big[-t\frac{(x-\mu)^2}{2\sigma^2}\big] dx$$

then

$$E[(\chi - \mu)^{2k}] = (-1)^k 2^k \sigma^{2k} f^{[k]}(1).$$

Thus

$$E[(\chi - \mu)^k] = \begin{cases} 0 & \text{if } k \text{ is odd,} \\ 1 \cdot 3 \cdots (k-1)\sigma^k & \text{if } k \text{ is even.} \end{cases}$$

Two random variables, χ_1, χ_2 are jointly Gaussian if their joint density function is given by

$$p_{\chi_1,\chi_2}(x,y) = \frac{1}{2\pi\sigma_{\chi_1}\sigma_{\chi_2}} \exp\left[\frac{1}{1-\rho^2}\big[\big(\frac{x-\mu_{\chi_1}}{\sigma_{\chi_1}}\big)^2 - 2\rho\big(\frac{x-\mu_{\chi_1}}{\sigma_{\chi_1}}\big)\big(\frac{y-\mu_{\chi_2}}{\sigma_{\chi_2}}\big) + \big(\frac{y-\mu_{\chi_2}}{\sigma_{\chi_2}}\big)^2\big]\right] \tag{10.28}$$

where

$$\begin{aligned} \mu_{\chi_i} &= E[\chi_i], \\ \sigma_{\chi_i} &= \text{ standard deviation of } \chi_i, \\ \rho &= \frac{E[(\chi_1 - \mu_{\chi_1})(\chi_2 - \mu_{\chi_2})]}{\sigma_{\chi_1}\sigma_{\chi_2}}, \quad \text{normalized correlation coefficient.} \end{aligned}$$

Once again the joint distribution function of a pair of Gaussian random variables is determined by the second order statistics of the pair of variables, $\{E[\chi_i], E[\chi_i\chi_j] \,:\, i,j = 1,2\}$.

Proposition 10.3.1. *Let χ_1 and χ_2 be Gaussian random variables then they are independent if and only if they are uncorrelated, i.e.*

$$E[(\chi_1 - \mu_{\chi_1})(\chi_2 - \mu_{\chi_2})] = 0.$$

Proof. If a pair of random variables has a joint density function then they are independent if and only if

$$p_{\chi_1,\chi_2}(x,y) = p_{\chi_1}(x) p_{\chi_2}(y). \tag{10.29}$$

From the form of the joint density of a pair of Gaussian variables it is clear that (10.29) holds if and only if $\rho = 0$. □

More generally, a collection of m random variables, $\{\chi_1, \cdots, \chi_m\}$ is Gaussian if and only if the density of the joint distribution function has the form

$$p_{\chi_1,\ldots,\chi_m}(t_1,\ldots,t_m) = \sqrt{\frac{\det a_{ij}}{[2\pi]^m}} \exp\left[-\frac{1}{2}\sum_{i,j=1}^{m} a_{ij}(t_i-\mu_i)(t_j-\mu_j)\right]. \tag{10.30}$$

Here a_{ij} is assumed to be a symmetric, positive definite matrix. That is, for some $c > 0$

$$a_{ij} = a_{ji} \text{ and } \sum_{i,j=1}^{m} a_{ij}x_i x_j \geq c\|(x_1,\ldots,x_m)\|^2.$$

Proposition 10.3.2. *If $\{\chi_1,\ldots,\chi_m\}$ are jointly Gaussian random variables with the joint density function given in* (10.30) *then*

$$E[\chi_i] = \mu_i$$

and a_{ij} is the inverse of $\mathrm{Cov}(\chi_i, \chi_j)$.

Evidently the joint distribution of a collection of Gaussian random variables is again determined by the second order statistics. This proposition can also be viewed as an "existence theorem." Given a collection of numbers $\{\mu_1,\ldots,\mu_m\}$ and a positive definite matrix r_{ij} there is a collection of Gaussian random variables $\{\chi_1,\ldots,\chi_m\}$ with

$$E[\chi_i] = \mu_i \text{ and } \mathrm{Cov}(\chi_i,\chi_j) = r_{ij}. \tag{10.31}$$

In practical situations, if all that is known about a collection of random variables is their means and covariance then one is free to assume that they are Gaussian.

Suppose that $\{\chi_1, \cdots, \chi_m\}$ is a collection of independent Gaussian random variables. If $\{a_j\}$ are constants then the linear combination

$$\chi_{\mathbf{a}} = \sum_{j=1}^{m} a_j\chi_j$$

is also a Gaussian random variable. The easiest way to see this is to compute the characteristic function of $\chi_{\mathbf{a}}$. From exercise 10.2.32 it follows that

$$M_{\mathbf{a}}(\lambda) = (a_1\cdots a_m)M_{\chi_1}(\lambda)\cdots M_{\chi_m}(\lambda). \tag{10.32}$$

Because a sum of quadratic functions is a quadratic this implies that

$$p_{\chi_{\mathbf{a}}} = \frac{1}{\sqrt{2\pi}\sigma_{\mathbf{a}}}\exp\big[-\frac{(x-\mu_{\mathbf{a}})^2}{2\sigma_{\mathbf{a}}^2}\big]$$

for some constants $\mu_{\mathbf{a}}, \sigma_{\mathbf{a}}$.

Exercise 10.3.2. Using formula (10.32) find an expression for the mean and variance of $\chi_{\mathbf{a}}$ in terms of the means and variances of the $\{\chi_j\}$.

Exercise 10.3.3. Suppose that χ_1 and χ_2 are Gaussian random variables, show that for any constants a, b the linear combination $a\chi_1 + b\chi_2$ is also a Gaussian random variable. Compute its mean and standard deviation.

Exercise 10.3.4. Suppose that χ_1 and χ_2 are Gaussian random variables, show that there is a invertible matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

so that $a\chi_1 + b\chi_2$ and $c\chi_1 + d\chi_2$ are independent, Gaussian random variables.

Exercise 10.3.5. Suppose that real numbers, $\{\mu_1, \ldots, \mu_m\}$ and a positive definite $m \times m$ matrix, r_{ij} are given. Find a sample space X, a probability measure on X and random variables $\{\chi_1, \ldots, \chi_m\}$ defined on X which are jointly Gaussian, satisfying (10.31).

Exercise 10.3.6. Suppose that $\{\chi_1, \ldots, \chi_m\}$ are jointly Gaussian random variables with means $\{\mu_1, \ldots, \mu_m\}$. Show that they are pairwise independent if and only if

$$\mathrm{Cov}(\chi_i, \chi_j) = \delta_{ij}\sigma^2_{\chi_i}.$$

Exercise 10.3.7. Suppose that $\{\chi_1, \ldots, \chi_m\}$ are jointly Gaussian random variables, show that they are independent if and only if they are pairwise independent.

10.3.4 The Central Limit Theorem

Random variables are often assumed to be Gaussian. This is, of course not always true, but the following theorem explains why it is often a reasonable approximation.

Theorem 10.3.1 (Central Limit Theorem). *Let $\{\chi_1, \chi_2, \ldots, \}$ be a sequence of identically distributed, independent random variables with mean μ and variance σ^2. Let*

$$Z_n = \frac{\chi_1 + \cdots + \chi_n - n\mu}{\sigma\sqrt{n}} = \sqrt{n}\frac{\bar{\chi}_n - \mu}{\sigma}$$

where $\bar{\chi}_n = (\chi_1 + \cdots + \chi_n)/n$. The distribution function for the variable Z_n tends to a normalized Gaussian as $n \to \infty$. That is

$$\lim_{n\to\infty} \mathrm{Prob}(Z_n \leq t) = \int_{-\infty}^{t} e^{-\frac{x^2}{2}} \frac{dx}{\sqrt{2\pi}}.$$

The hypothesis that the variables are independent is quite important. To get a sensible limit we must subtract the mean of $\chi_1 + \cdots + \chi_n$ and divide by $\sqrt{n}$. Also notice that there is an implicit hypothesis: the second moments of the $\{\chi_i\}$ are assumed to exist. Nonetheless, it is a remarkable result. It says that the distribution function of the average of a large collection of independent random variables approaches a Gaussian, *no matter* how the individual random variables are distributed.

Before we prove the theorem, we derive an interesting consequence. Is there any reason to expect that the average of a collection of measurements will converge to the theoretical mean value? Assume that the individual measurements are independent random variables,

$\{\chi_i\}$. Since they are the result of performing the same experiment, over and over, they can also be assumed to be identically distributed. As above set

$$\bar{\chi}_n = \frac{\chi_1 + \cdots + \chi_n}{n}$$

and observe that

$$\begin{aligned}\text{Prob}(|\bar{\chi}_n - \mu| \leq \epsilon) &= \text{Prob}(\Big|\frac{\bar{\chi}_n - \mu}{\sigma/\sqrt{n}}\Big| \leq \frac{\epsilon\sqrt{n}}{\sigma}) \\ &\geq \text{Prob}(|Z_n| < N)\end{aligned}$$

for any N such that $\epsilon\sqrt{n}/\sigma > N$. Hence

$$\lim_{n\to\infty} \text{Prob}(|\bar{\chi}_n - \mu| \leq \epsilon) \geq \int_{-N}^{N} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}.$$

As this holds for any N we conclude that

$$\lim_{n\to\infty} \text{Prob}(|\bar{\chi}_n - \mu| \leq \epsilon) = 1.$$

This is called the *weak law of large numbers*. If we recall that each χ_j is a function on the sample space $(X, \mathcal{M}, \nu)$ the weak law of large numbers says that for any $\epsilon > 0$, the average,

$$\frac{\chi_1(x) + \cdots + \chi_n(x)}{n}$$

will eventually be within ϵ of μ for almost every $x \in X$. That is, if we do many independent trials of the same experiment and average the results there is good reason to expect that this average will approach the theoretical mean value. This explains, in part the importance of the assumption that the individual trials of an experiment are independent of one another.

The weak law of large numbers says that, in a weak sense, the sequence of random variables

$$\bar{\chi}_n = \frac{\chi_1 + \cdots + \chi_n}{n}$$

converges to the random variable which is a constant equal to the common mean value. There are many theorems of this type where one considers different ways of measuring convergence. The central limit theorem itself is such a statement, it asserts that the sequence of random variables $\{Z_n\}$ converges to a normalized Gaussian *in distribution*. Meaning, that the cumulative distribution for Z_n converges to the cumulative distribution of the normalized Gaussian. Other results of this type can be found in [13] and [12].

We now turn to the proof of the Central Limit Theorem.

Proof. Let $p(x)$ be the (common) density function for $\{\chi_i - \mu\}$. The hypotheses imply that

$$\int_{-\infty}^{\infty} p(x)dx = 1, \quad \int_{-\infty}^{\infty} xp(x)dx = 0, \quad \int_{-\infty}^{\infty} x^2 p(x)dx = \sigma^2. \tag{10.33}$$

For simplicity, we assume that the characteristic function of $\chi_i - \mu$ has two derivatives at the origin. Using the Taylor expansion and the relations (10.33) gives

$$\hat{p}(\xi) = 1 - \frac{\sigma^2\xi^2}{2} + o(\xi^2).$$

Let q_n be the density function for the shifted and scaled random variables, $\{(\chi_i - \mu)/\sigma\sqrt{n}\}$. It is given by $q_n(x) = \sigma\sqrt{n} \cdot p(\sigma\sqrt{n}x)$ and therefore

$$\text{Prob}(\frac{\chi_j - \mu}{\sigma\sqrt{n}} \leq t) = \int_{-\infty}^{t} q_n(x)dx.$$

The Fourier transform of q_n and its Taylor expansion are:

$$\begin{aligned} \hat{q}_n(\xi) &= \hat{p}(\frac{\xi}{\sigma\sqrt{n}}), \\ \hat{q}_n(\xi) &= 1 - \frac{\xi^2}{2n} + o(\frac{\xi^2}{n}). \end{aligned}$$

Since $\{(\chi_i - \mu)/(\sigma\sqrt{n})\}$ are independent random variables, the characteristic function of their sum,

$$Z_n = \frac{\chi_1 - \mu}{\sigma\sqrt{n}} + \frac{\chi_2 - \mu}{\sigma\sqrt{n}} + \cdots + \frac{\chi_n - \mu}{\sigma\sqrt{n}}$$

is just the product of the characteristic functions of each:

$$\begin{aligned} \hat{p}_{Z_n}(\xi) = E[e^{-i\xi Z_n}] &= E[e^{-i\xi\frac{\chi_1-\mu}{\sigma\sqrt{n}}}] \cdots E[e^{-i\xi\frac{\chi_n-\mu}{\sigma\sqrt{n}}}] \\ &= [\hat{q}_n(\xi)]^n = \big[1 - \frac{\xi^2}{2n} + o(\frac{\xi^2}{n})\big]^n. \end{aligned}$$

As $n \to \infty$, the last term is negligible, therefore

$$\hat{p}_{Z_n}(\xi) = \big[1 - \frac{\xi^2}{2n} + o(\frac{\xi^2}{n})\big]^n \to e^{-\frac{\xi^2}{2}}.$$

Thus, by the Fourier inversion formula, the density function of Z_n converges to the Gaussian:

$$\mathcal{F}^{-1}[\hat{p}_{Z_n}] \to \int_{-\infty}^{\infty} e^{i\xi x} e^{-\frac{\xi^2}{2}} \frac{dx}{\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

see section 3.2.2. □

10.3.5 Limits of random variables

See: A.3.5.

Often one takes limits of random variables as one parameter or the other tends to infinity. In this section we consider several basic examples. We do not treat the problem of convergence of the random variables themselves but only the behavior of their distributions under limits. The former problem is treated in [13] or [15]. Our first example is an application of the central limit theorem.

Example 10.3.1. Let us denote by X_∞ the set of all infinite sequences of heads and tails. Let

$$\text{Prob}(H) = p, \quad \text{Prob}(T) = 1 - p.$$

We assume that this probability holds for all flips, and each flip is independent of the others. Let χ_i be the random variable defined by

$$\chi_i((a_1, \ldots, a_n, \ldots)) = \begin{cases} 1 & \text{if } a_i = H, \\ 0 & \text{if } a_i = T. \end{cases}$$

These are identically distributed, independent random variables with expected value and variance given by

$$E[\chi_i] = p, \quad \sigma^2_{\chi_i} = p(1-p).$$

The central limit theorem implies that

$$\text{Prob}(Z_n \leq t) \to \int_{-\infty}^{t} e^{-\frac{x^2}{2}} \frac{dx}{\sqrt{2\pi}} \quad \text{as} \quad n \to \infty$$

where

$$Z_n = \frac{\chi_1 + \chi_2 + \cdots + \chi_n - np}{\sqrt{np(1-p)}}.$$

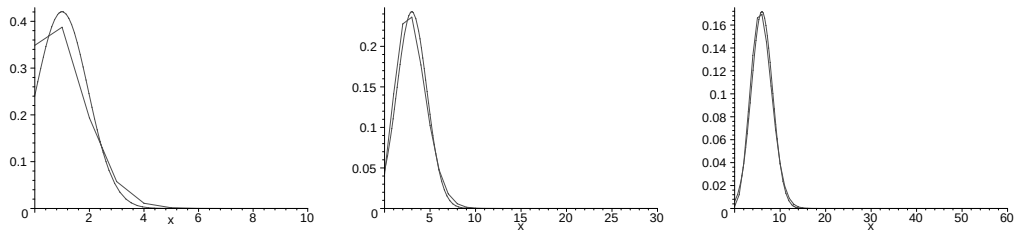
We use this fact to approximate the Bernoulli distribution. The probability for the Bernoulli distribution is given by

$$\text{Prob}(\chi_1 + \cdots + \chi_n \leq k) = \sum_{j \leq k} \binom{n}{j} p^j (1-p)^{n-j}.$$

Let $k = [\sqrt{np(1-p)}t + np]$. The central limit theorem implies that

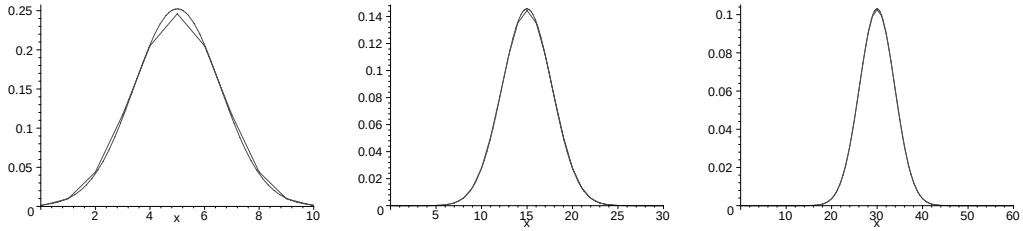
$$\text{Prob}(\chi_1 + \cdots + \chi_n \leq k) \approx \int_{-\infty}^{t} e^{-\frac{x^2}{2}} \frac{dx}{\sqrt{2\pi}}.$$

The combinatorial quantity, on the left is rather complicated to compute, whereas the right side is the integral of a very smooth, rapidly decaying function. It is often more useful to have a rapidly convergent integral approximation rather than an exact combinatorial formula. Note that this approximation is useful even for moderately sized n. The graphs in figures 10.1 and 10.2 show the distribution functions for Bernoulli distributions with $p = .1, .5$ and $n = 10, 30, 60$ along with the Gaussians having the same mean and variance. The Bernoulli distribution is only defined for integral values, for purposes of comparison it has been linearly interpolated in the graphs. Note the much more rapid convergence for $p = .5$.



(a) $p = .1, n = 10$ (b) $p = .1, n = 30$ (c) $p = .1, n = 60$

Figure 10.1: Comparisons of Bernoulli and Gaussian distribution functions with $p = .1$.



(a) $p = .5, n = 10$ (b) $p = .5, n = 30$ (c) $p = .5, n = 60$

Figure 10.2: Comparisons of Bernoulli and Gaussian distribution functions with $p = .5$.

Example 10.3.2. We now consider a different limit of the Bernoulli distribution. The Bernoulli distribution is used to model the number of radioactive decays occurring in a fixed time interval. Suppose there are N particles and each has a probability p of decaying in a fixed time interval, $[0, T]$. If we suppose that the decay of one atom is independent of the decay of any other and let χ denote the number of decays occurring in $[0, T]$ then

$$\text{Prob}(\chi = k) = \binom{N}{k} p^k (1-p)^{N-k}.$$

The number of decays is therefore a Bernoulli random variable. An actual sample of any substance contains $O(10^{23})$ atoms. In other words N is a huge number which means that p must be a very small number. Suppose that we let $N \to \infty$ and $p \to 0$ in such a way that $Np \to \bar{\lambda}$ for some constant $\bar{\lambda} > 0$. It is not difficult to find the limit of the Bernoulli distribution under these hypotheses. Assuming that $Np = \bar{\lambda}$ we get that

$$\begin{aligned}\binom{N}{k} \big(\frac{\bar{\lambda}}{N}\big)^k \big(1 - \frac{\bar{\lambda}}{N}\big)^{N-k} &= \frac{N(N-1)\cdots(N-(k-1))}{k!} \big(1 - \frac{\bar{\lambda}}{N}\big)^N / \big(\frac{N}{\bar{\lambda}} - 1\big)^k \\ &= \frac{1}{k!} \frac{N^k (1 - 1/N)\cdots(1-(k-1)/N)}{N^k (1/\bar{\lambda} - 1/N)^k} \big(1 - \frac{\bar{\lambda}}{N}\big)^N.\end{aligned}$$

Since $(1-\frac{\alpha}{N})^N \to e^{-\alpha}$, as $N \to \infty$, we have that

$$\binom{N}{k}(\frac{\bar{\lambda}}{N})^k(1-\frac{\bar{\lambda}}{N})^{N-k} \to \frac{\bar{\lambda}^k e^{-\bar{\lambda}}}{k!}$$

This explains why the Poisson process provides a good model for radioactive decay. The parameter $\bar{\lambda}$ is a measure of the intensity of the radioactive source.

Example 10.3.3. As a final example consider the behavior of a Poisson random variable as the intensity gets to be very large. The probability density for a Poisson random variable χ with intensity λ is the generalized function

$$p_\chi(x) = \sum_{k=0}^{\infty} \delta(x-k)\frac{\lambda^k e^{-\lambda}}{k!}.$$

This density function has a variety of defects from the point of view of applications: 1. It is combinatorially complex, 2. It is not a function, but rather a generalized function. The second observation is a consequence of the fact that the cumulative distribution has jumps at the integers but is otherwise constant. We are interested in the behavior of $\text{Prob}(\chi = k)$ as λ becomes very large with $|k-\lambda|$ reasonably small compared to λ. Recall that the mean of χ is λ and the standard deviation is $\sqrt{\lambda}$. This means that the

$$\text{Prob}(\lambda - m\sqrt{\lambda} \leq \chi \leq \lambda + m\sqrt{\lambda})$$

is very close to 1 for reasonably small values of m, e.g. $m = 3, 4, 5$, etc. Let

$$p_\lambda(k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

To obtain an asymptotic form for $p_\lambda(k)$ we use Stirling's formula. It implies that for large values of k

$$k! = \sqrt{2\pi k}k^k e^{-k}(1+O(\frac{1}{k})), \tag{10.34}$$

see section A.3.5 or [82]. Using this approximation for large k gives

$$p_\lambda(k) \approx \frac{1}{\sqrt{2\pi\lambda}}\left[\frac{\lambda}{k}\right]^{k+\frac{1}{2}} e^{k-\lambda}. \tag{10.35}$$

To find a useful asymptotic formula we set $k = \lambda + x$ with the understanding that $x\lambda^{-1} << 1$. In terms of x

$$\begin{aligned}\left[\frac{\lambda}{k}\right]^{k+\frac{1}{2}} &= e^{(\lambda+x+\frac{1}{2})\log(1+\frac{x}{\lambda})} \\ &\approx e^{-[x+\frac{x^2}{2\lambda}]}.\end{aligned} \tag{10.36}$$

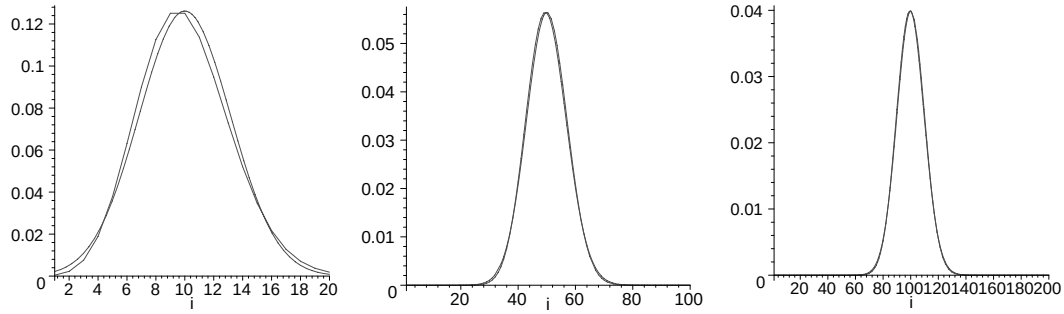
In the second line we use the Taylor polynomial for $\log(1+y)$. Putting this into the formula for p_λ shows that

$$p_\lambda(k) \approx \frac{e^{-\frac{(k-\lambda)^2}{2\lambda}}}{\sqrt{2\pi\lambda}}, \tag{10.37}$$

provided that $|\lambda - k|/\sqrt{\lambda}$ remains bounded and λ is large. Once again we see that the Gaussian distribution provides a limiting form for a random variable. Using formula (10.37) we can approximately compute expected values for functions of χ. If λ is large and f is a reasonably well behaved function then

$$E[f(\chi)] \approx \frac{1}{\sqrt{2\pi\lambda}} \int_{-m\sqrt{\lambda}}^{m\sqrt{\lambda}} f(t) e^{-\frac{(t-\lambda)^2}{2\lambda}} dt.$$

Here m is chosen to make the error as small as needed provided only that it remains small compared to $\lambda^{\frac{1}{4}}$. The figure shows the graphs of the Poisson and Gaussian distributions for $\lambda = 10, 50$ and 100.



(a) $\lambda = 10$ (b) $\lambda = 50$ (c) $\lambda = 100$

Figure 10.3: Comparisons of Poisson and Gaussian distribution functions.

10.3.6 Modeling a source-detector pair

Now suppose that we have a Poisson source with intensity $\bar{\lambda}$, and a *Bernoulli detector*. This means that each incident photon has probability p, $0 \leq p \leq 1$ of being detected and each detection event is independent of any other. What is the statistical description of the output of such a source-detector pair? The probability of observing k photons, given that N photons arrive is a conditional probability:

$$P_d(k|N) = \begin{cases} \binom{N}{k} p^k (1-p)^{N-k} & k = 0, \cdots, N, \\ 0 & k > N. \end{cases}$$

On the other hand, the source is described by

$$P_s(\chi = N) = \frac{\bar{\lambda}^N e^{-\bar{\lambda}}}{N!}.$$

The probability that the detector observes k photons, $P_o(d=k)$ is therefore

$$\begin{aligned} P_o(d=k) &= \sum_{N=k}^{\infty} P_s(N) P_d(k|N) \\ &= \sum_{N=k}^{\infty} \binom{N}{k} p^k (1-p)^{N-k} \frac{\bar{\lambda}^N e^{-\bar{\lambda}}}{N!} \\ &= e^{-\bar{\lambda}} \sum_{N=k}^{\infty} \frac{1}{k!(N-k)!} (\bar{\lambda}p)^k (\bar{\lambda}(1-p))^{N-k} \\ &= \frac{(\bar{\lambda}p)^k}{k!} e^{-\bar{\lambda}} e^{-\bar{\lambda}(1-p)} = \frac{(\bar{\lambda}p)^k}{k!} e^{-\bar{\lambda}p}. \end{aligned} \tag{10.38}$$

Hence the source-detector pair is again a Poisson random variable with the intensity scaled by the probability of detection. This is a general feature of Poisson and Bernoulli random variables: if a Poisson random variable is the "input" to a Bernoulli random variable then the output is again a Poisson random variable.

10.3.7 Beer's Law

It has been asserted several times that Beer's law is essentially a prescription for the behavior of the *mean value* of a random variable. In this section we consider Beer's Law from this perspective. In our analysis we consider a "beam" of N photons traveling, through a material, along an interval $[a,b]$ contained in a line l. Let χ_N be a random variable which equals the numbers of photons that are emitted. The absorption coefficient $\mu(s)$ is a non-negative function defined along this line. Suppose that Δs is a very small number (really an infinitesimal). We assume that the photons are independent and the absorption or transmission of a particle through a thin slab of material is a Bernoulli random variable:

(1) A single particle which is incident upon the material at point s has a probability $(1-\mu(s)\Delta s)$ of being emitted and therefore probability $\mu(s)\Delta s$ of being absorbed.

(2) Each particle is independent of each other particle.

(3) Disjoint subintervals of $[a,b]$ are independent.

To derive Beer's law we subdivide $[a,b]$ into m subintervals

$$J_k = [a + \frac{(k-1)(b-a)}{m}, a + \frac{k(b-a)}{m}), \quad k = 1, \ldots, m.$$

In order for a particle incident at a to emerge at b it must evidently pass through every subinterval. The probability that a particle passes through J_k is approximately

$$p_{k,m} \approx (1 - \mu(a + \frac{k(b-a)}{m}) \frac{b-a}{m}).$$

By hypothesis (3) it follows that the probability that a particle incident at a emerges at b is the product of these probabilities

$$p_{ab,m} \approx \prod_{k=1}^{m} p_{k,m}. \tag{10.39}$$

This is an approximate result because we still need to let m tend to infinity.

If $\mu(s) = \mu_0$ is a constant, then it is an elementary result that the limit of this product, as $m \to \infty$ is $e^{-\mu_0(b-a)}$. Hence a single particle incident at a has a probability $e^{-\mu_0(b-a)}$ of emerging at b.. The independence of the individual photons implies that the probability that k out of N photons emerge is

$$P(k,N) = \binom{N}{k} e^{-k\mu_0(b-a)}(1 - e^{-\mu_0(b-a)})^{N-k}.$$

If N photons are incident then number of photons expected to emerge is therefore

$$E[\chi_N] = e^{-\mu_0(b-a)}N,$$

see example 10.2.4. The variance, computed in example 10.2.8 is

$$\sigma^2_{\chi_N} = Ne^{-\mu_0(b-a)}(1 - e^{-\mu_0(b-a)}).$$

For an experiment where the outcome is a random variable, the *signal-to-noise* ratio is a number which reflects the expected quality of the data. If χ is the random variable then

$$\mathrm{SNR}(\chi) \stackrel{d}{=} \frac{\text{expected value}}{\text{standard deviation}} = \frac{E[\chi]}{\sigma_\chi}.$$

For our experiment

$$\mathrm{SNR}(\chi_N) = \sqrt{N\left(\frac{e^{-\mu_0(b-a)}}{1 - e^{-\mu_0(b-a)}}\right)}.$$

This is a very important result: the quality of the measurements can be expected to increase as the $\sqrt{N}$. Moreover, the greater the fraction of photons absorbed, the less reliable the measurements. This has an important consequence in imaging: measurements corresponding to rays which pass through more (or harder) material have a lower SNR than those passing through less (or softer) material.

In general the absorption coefficient is not a constant and we therefore need to compute the limit in (10.39). After taking the logarithm, this is easily done,

$$\log p_{ab,m} = \sum_{k=1}^{m} \log\left((1 - \mu(a + \frac{k(b-a)}{m})\frac{b-a}{m}\right). \tag{10.40}$$

The Taylor expansion for the logarithm implies that

$$\log p_{ab,m} = -\sum_{k=1}^{m}\left[\mu(a + \frac{k(b-a)}{m})\frac{b-a}{m}\right] + O(m^{-1}). \tag{10.41}$$

As m tends to infinity, the right hand side of (10.41) converges to

$$-\int_a^b \mu(s)ds.$$

Hence the probability that a particle incident at a emerges at b is

$$p_\mu = \exp\left[-\int_a^b \mu(s)ds\right].$$

Arguing exactly as before we conclude that, if N photons are incident then the the probability that $k \leq N$ emerge is

$$P(k,N) = \binom{N}{k} p_\mu^k (1-p_\mu)^{N-k} \tag{10.42}$$

and therefore the expected number to emerge is

$$E[\chi_N] = N \exp\left[-\int_a^b \mu(s)ds\right]. \tag{10.43}$$

This is exactly Beer's law! The variance is

$$\mathrm{Var}(\chi_N) = p_\mu(1-p_\mu)N \tag{10.44}$$

so the signal-to-noise ratio is

$$\mathrm{SNR}(\chi_N) = \sqrt{N\left(\frac{p_\mu}{1-p_\mu}\right)}.$$

In medical imaging N, the number of incident photons is also a random variable. It is usually assumed to satisfy a Poisson distribution. In the previous section it is shown that having a Poisson random variable as the "input" to a Bernoulli process leads to Poisson random variable. This is considered in exercise 10.3.10.

Exercise 10.3.8. In example 10.3.2 we considered a different limit for a Bernoulli distribution from that considered in this section and got a different result. What is the underlying physical difference in the two situations?

Exercise 10.3.9. Suppose that the probability that k out of N photons are emitted is given by (10.42) and that each emitted photon is detected with probability q. Assuming that there are N (independent) incident photons, show that the probability that k are detected is

$$P_{\mathrm{det}}(k,N) = \binom{N}{k} (p_\mu q)^k (1-p_\mu q)^{N-k}. \tag{10.45}$$

Another way of putting this is to say that a cascade of Bernoulli processes is a Bernoulli process.

Exercise 10.3.10. Suppose that the number of X-ray photons emitted is a Poisson random variable with intensity N and the Bernoulli detector has a probability q of detecting each photon. Show that the overall system of X-ray production, absorption and detection is a Poisson random variable with intensity $p_\mu qN$.

Exercise 10.3.11. Suppose that the process of absorption of X-ray photons through a slab is modeled as a Poisson random variable. If the expected number of emitted photons is given by Beer's law, what is the variance in the number of emitted photons? Is this a reasonable model?

10.4 Statistics and measurements

We close our discussion of probability theory by considering how these ideas apply in a simple practical situation. Suppose that χ is a real valued, random variable which describes the outcome of a experiment. By describing the outcome of the experiment in these terms we are acknowledging that the measurements involved in the experiment contain errors. While,at the same time, asserting that the experimental errors have a statistical regularity in that they are distributed according to some definite (but *a priori* unknown) law. Let $p_\chi(x)$ denote the density function for χ so that, for any $a < b$

$$\text{Prob}(a \leq \chi \leq b) = \int_a^b p_\chi(x)dx. \tag{10.46}$$

Often times one knows that p_χ belongs to a family of distributions. For example, if χ is the number of radioactive decays which occur in a fixed time interval then, we know χ is a Poisson random variable and is therefore determined by its intensity, $\lambda = E[\chi]$. On the other hand, the general type of distribution may not be known in advance. For most practical applications one would like estimates for the mean and variance of χ :

$$\mu_\chi = E[\chi] \text{ and } \sigma_\chi^2 = E[(\chi - \mu_\chi)^2].$$

The mean represents the ideal outcome of the experiment while the variance measures the uncertainty in the measurements themselves. Let $\{\chi_i\}$ denote a sequence of independent random variables which are all distributed according to (10.46). This is a model for independent trials of the experiment in question. If the experiment is performed N-times then the probability that the results lie in a rectangle

$$[a_1, b_1] \times \cdots [a_N, b_N]$$

is

$$\text{Prob}(a_1 < \chi_1 b \leq_1, \ldots, a_N < \chi_N \leq b_N) = \int_{a_1}^{b_1} \cdots \int_{a_N}^{b_N} p_\chi(x_1) \cdots p_\chi(x_N)dx_1 \cdots dx_N. \tag{10.47}$$

Our discussion of experimental errors is strongly predicated on the assumption that the various trials of the experiment are independent. Let

$$\bar{\chi}_N = \frac{\chi_1 + \cdots + \chi_N}{N}$$

denote the random variables defined as the average of the results of the first N-trials. Because the trials are independent, formula (10.26) gives the mean and variance of $\bar{\chi}_N$:

$$\mu_{\bar{\chi}_N} = \mu \text{ and } \sigma^2_{\bar{\chi}_N} = \frac{\sigma^2}{N}. \tag{10.48}$$

The various laws of large numbers imply that $\bar{\chi}_N$ converges to the constant function, μ in a variety of different senses. Since the variables in question have finite variance, the Chebyshev inequality, (10.9) gives the estimate

$$\text{Prob}(|\bar{\chi}_N - \mu| < \epsilon) = 1 - \text{Prob}(|\bar{\chi}_N - \mu| \geq \epsilon) \leq 1 - \frac{\sigma^2}{N\epsilon^2}. \tag{10.49}$$

This explains a sense in which the variance is a measure of experimental error. Indeed, the central limit theorem implies that, for large N

$$\text{Prob}(-\epsilon \leq \bar{\chi}_N - \mu \leq \epsilon) \approx \frac{1}{\sqrt{2\pi}} \int_{-\frac{\sqrt{N}\epsilon}{\sigma}}^{\frac{\sqrt{N}\epsilon}{\sigma}} e^{-\frac{x^2}{2}} dx.$$

This is all very interesting but it does not really address the question of how to estimate μ and σ^2 using the actual outcomes $\{x_1, \ldots, x_N\}$ of N trials of our experiment. An estimate for μ is an estimate for the "exact outcome" of the experiment whereas an estimate for σ^2 provides an estimate for the uncertainty in our results. These questions are properly questions in estimation theory, which is a part of statistics, a field distinct from probability *per se.* We consider only very simple answers to these questions. Our answers are motivated by (10.48) and (10.49). The results of N trials defines a point in $\mathbf{x} \in \mathbb{R}^N$, so our answer is phrased in terms of functions on $\mathbb{R}^N$. As an estimate for μ we use the *sample mean*,

$$m(\mathbf{x}) = \frac{x_1 + \cdots + x_N}{N}$$

and for σ^2, the *sample variance*,

$$S^2(\mathbf{x}) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - m(\mathbf{x}))^2.$$

The sample mean is exactly what one would expect; the sample variance requires some explanation. One might expect that we should define S^2 by subtracting μ from the measurements, but of course, we do not know μ and that is why we subtract m instead. One also might have expected a factor $1/N$ instead of $1/(N-1)$, however, due to the non-linearity of S^2 this would lead to an estimate of σ^2 whose expected value is **not** σ^2. An estimate with the wrong expected value is called *biased*.

Since $\mathbb{R}^N$ parametrizes the outcomes of N-independent trials, we think of it as a probability space, with the distribution defined in (10.47). With this interpretation the expected value and variance of m are given by.

$$E[m] = \mu \text{ and } E[(m-\mu)^2] = \frac{\sigma^2}{N}. \tag{10.50}$$

Hence an estimate for σ^2 leads to an estimate for the error in asserting that $m(x_1, \ldots, x_N) = \mu$. For example Chebyshev's inequality gives

$$\text{Prob}(|m(\mathbf{x}) - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{N\epsilon^2}.$$

It is important to recognize that in a situation of this sort, the best that one can hope for is a statement to the effect that

$$|m(\mathbf{x}) - \mu| < \epsilon$$

with a given probability.

Expanding the square to compute $E[S^2]$ gives

$$E[S^2] = \frac{1}{N-1} E\left[\sum_{i=1}^{N} x_i^2 - 2m\sum_{i=1}^{N} x_i + m^2\right] = \frac{1}{N-1}\sum_{i=1}^{N} E[x_i^2 - m^2]. \tag{10.51}$$

In the exercises it is shown that

$$E[m^2] = \frac{\sigma^2}{N} + \frac{(N-1)\mu^2}{N}, \tag{10.52}$$

from which it follows that

$$E[S^2] = \frac{1}{N-1}[N(\sigma^2 + \mu^2) - (\sigma^2 + (N-1)\mu^2)] = \sigma^2.$$

This explains the factor $N - 1$: with this factor the expected valued of S^2 is the true variance. Finally we would like to compute the variance in S^2,

$$E[(S^2 - \sigma^2)^2] = \frac{(N-1)^2}{N^3} E[(\chi - \mu)^4] - \frac{(N-1)(N-3)\sigma^4}{N^3}. \tag{10.53}$$

It depends on the fourth moment of the original random variable χ. If χ is a Gaussian random variable then $E[(\chi - \mu)^4] = 3\sigma^4$ and therefore

$$E[(S^2 - \sigma^2)^2] = \frac{2\sigma^4(N-1)}{N^2}.$$

If the variance is large then m provides a much better estimate for μ then S^2 for σ^2. In the final analysis, all these formulæ involve σ^2, which is *a priori* unknown. To use these formulæ with any confidence therefore requires an *a priori* estimate for σ^2.

We have only made a tiny scratch in the surface of estimation theory, a subject with a very interesting interplay of probability and empirical experience. Underlying its application to real experiments are deep, philosophical assumptions about the nature of the universe we live in. Our treatment of this subject is adapted from [3] and [13]. The former reference is a very good and complete introduction to the application of probability and statistics to experimental science.

Exercise 10.4.1. Show that

$$E[m^2] = \frac{\sigma^2}{N} + \frac{(N-1)\mu^2}{N}.$$

Exercise 10.4.2. Derive the formula for $E[(S^2 - \sigma^2)^2]$.

Exercise 10.4.3. What are $E[(m-\mu)^2]$ and $E[(S^2-\sigma^2)^2]$ if χ is a Poisson random variable with intensity λ?

Chapter 11

Random processes

To model noise in measurement and filtering requires concepts more general than that of a random variable. This is because we need to discuss the results of passing a noisy signal through a linear filter. As was the case in the chapter on filtering, it is easier to present this material in terms of functions of a single variable, though the theory is easily adapted to functions of several variables. Our discussion of random processes is very brief, aimed squarely at the goal of analyzing the effects of noise in the filtered backprojection algorithm.

11.1 Random processes in measurements

To motivate this discussion we think of the familiar example of a radio signal. A radio station broadcasts a signal $s_b(t)$ as an electro-magnetic wave. A receiver detects a signal $s_r(t)$ which we model as a sum

$$s_r(t) = F(s_b)(t) + n(t).$$

Here F is a filtering operation which describes the propagation of the broadcast signal. For the purposes of this discussion we model F as attenuation, that is $F(s_b) = \lambda s_b$ for a $0 < \lambda < 1$. The other term is "noise." The noise is composed of several parts. On the one hand it records "random" aspects of the life history of the broadcast signal which are not modeled by F. On the other hand it is an accumulation of other signals which happen to be present at about the same carrier frequency as the signal of interest. The existence of the second part is easily verified by tuning an (old) radio to a frequency for which there is no local station. Practically speaking it is not possible to give a formula for the noise. Because we cannot give an exact description of the noise we instead describe it in terms of its properties, beginning with the assumption that the noise is a bounded function of t.

What else can be done to specify the noise? Recall the ensemble average definition of probability: it is the average of the results of many "identical experiments." In the radio example, imagine having many different radios, labeled by a set $\mathcal{A}$. For each $\alpha \in \mathcal{A}$ we let $s_{r,\alpha}(t)$ be the signal received by radio α at time t. The collection of radios $\mathcal{A}$ *is* the sample space. The value of $s_{r,\alpha}$ at a time t is a function on the sample space, in other words, a random variable. From the form given above we see that

$$s_{r,\alpha}(t) = \lambda s_b(t) + n_\alpha(t).$$

The noise can then be described in terms of the statistical properties of the random variables $\{n_\alpha(t)\}$ for different values of t. We emphasize that the sample space is $\mathcal{A}$, the collection of different receivers, the time parameter simply labels different random variables defined on the sample space. A family of random variables, defined on the same sample space, parametrized by a real variable is called a *random process* or more precisely a continuous parameter random process.

Once the sample space, $\mathcal{A}$ is equipped with a σ-algebra $\mathcal{M}$ and a probability measure, ν we can discuss the statistics of the noise. For example at each time t the random variable $n_\alpha(t)$ has an expected value

$$E[n_\alpha(t)] = \int_{\mathcal{A}} n_\alpha(t) d\nu(\alpha).$$

In many applications one assumes that the noise has mean zero, i.e. $E[n_\alpha(t)] = 0$ for all t. This means that if we make many different independent measurements $\{s_{r,\alpha}(t)\}$ and average them we should get a good approximation to $\lambda s_r(t)$. The correlation between the noise at one moment in time and another is given by

$$E[n_\alpha(t_1)n_\alpha(t_2)] = \int_{\mathcal{A}} n_\alpha(t_1)n_\alpha(t_2) d\mu(\alpha).$$

How should the sample space be described mathematically? In an example like this, the usual thing is to use the space of all bounded functions as the index set. That is, *any* bounded function is a candidate for the noise in our received signal. In principle, the probabilistic component of the theory should then be encoded in the choice of a σ-algebra and probability measure on the space of bounded continuous functions. These probability measures are rarely made explicit. Instead one specifies the cumulative joint distributions of the noise process at finite collections of times. This means that for any $k \in \mathbb{N}$, any k-times $(t_1, \ldots, t_k)$ and any k values $(s_1, \ldots, s_k)$ the joint probability that

$$n_\alpha(t_j) \leq s_j \text{ for } j = 1, \ldots, k$$

is specified. If the joint distributions satisfy the usual consistency conditions then a result of Kolmogorov states that there is a probability measure on $\mathcal{A}$, inducing the joint distributions, with σ-algebra chosen so that all the sets

$$\{n_\alpha \in \mathcal{A} \,:\, n_\alpha(t_j) \leq s_j\}$$

are measurable. In this chapter we give a brief introduction to the basic concepts of random processes. Our treatment, though adequate for applications to imaging, is neither complete not rigorous. In particular we do not establish the existence of random processes as outlined above. Complete treatments can be found in [12] or [79].

11.2 Basic definitions

Let $(X, \mathcal{M}, \nu)$ be a probability space, as noted above a random process is an indexed family of random variables defined on a fixed probability space. There are two main types

of random processes. If the index set is a subset of integers, e.g. natural numbers $\mathbb{N}$ then the process is a *discrete parameter random process*. The random process is then a sequence of $\{\chi_1, \chi_2, \dots, \}$ of random variables defined on X. A *continuous parameter random process* is a collection of random variables, $\chi(t)$ indexed by a continuous parameter. Often the whole real line is used as the index set, though one can also use a finite interval or a half ray. For each t, $\chi(t)$ is a random variable, that is a measurable function on X. We can think of χ as a function of the pair $(w;t)$ where $w \in X$. For a fixed $w \in X$ the map $t \mapsto \chi(w;t)$ is called a *sample path* for this random process. Depending on the context we use either the standard functional notation, $\chi(t)$ or subscript notation χ_t, for sample paths. The dependence on the point in X is suppressed unless it is required for clarity. In the continuous case, a rigorous treatment of this subject requires hypotheses about the continuity properties of the random variables as functions of t, see [12] or [79].

It would appear that the first step in the discussion of a random process should be the definition of the measure space and a probability measure defined on it. As noted above, this is rarely done. Instead the random process is defined in terms of the properties of the random variables themselves. In the continuous time case, for each k and every k-tuple of times $t_1 \leq \cdots \leq t_k$ the cumulative distributions

$$P_{t_1,\dots,t_k}(s_1, \dots, s_k) = \text{Prob}(\chi(t_1) \leq s_1, \chi(t_2) \leq s_2, \dots, \chi(t_k) \leq s_k)$$

are specified. In favorable circumstances, these distributions are given by integrals of density functions.

$$P_{t_1,\dots,t_k}(s_1, \dots, s_k) = \int_{-\infty}^{s_1} \cdots \int_{-\infty}^{s_k} p_{t_1,\dots,t_k}(x_1, \cdots, x_k) dx_1 \cdots dx_k,$$

They must satisfy the usual consistency conditions:

$$\begin{aligned} p_{t_1,\dots,t_k}(x_1, \dots, x_k) &= \int_{-\infty}^{\infty} p_{t_1,\dots,t_{k+1}}(x_1, \dots, x_k, x_{k+1}) dx_{k+1}, \\ p_{t_{\tau(1)},\dots,t_{\tau(k)}}(x_{\tau(1)}, \dots, x_{\tau(k)}) &= p_{t_1,\dots,t_k}(x_1, \dots, x_k), \end{aligned} \tag{11.1}$$

here τ is any permutation of $\{1, \dots, k\}$.

In the discrete case the joint distribution functions are specified for any finite subset of the random variables. That is for each $k \in \mathbb{N}$ and each k-multi-index, $\mathbf{i} = (i_1, \dots, i_k)$ the cumulative distribution

$$P_{\mathbf{i}}(s_1, \dots, s_k) = \text{Prob}(\chi_{i_1} \leq s_1, \chi_{i_2} \leq s_2, \dots, \chi_{i_k} \leq s_k)$$

is specified. They also need to satisfy the consistency conditions for joint cumulative distributions. If $\{\chi_i\}$ is a discrete parameter random process, we say that the terms of the sequence are *independent* if for any choice of distinct indices $\{i_1, \dots, i_k\}$ the random variables $\{\chi_{1_i}, \dots, \chi_{i_k}\}$ are independent.

The cumulative distribution functions $P_{t_1,\dots,t_k}(s_1, \dots, s_k)$ (or $P_{\mathbf{i}}(s_1, \dots, s_k)$) are called the *finite dimensional distributions* of the random process. A basic result of Kolmogorov states that if finite dimensional distributions are specified which satisfy the compatibility conditions then there is a probability space $(X, \mathcal{M}, \nu)$ and a random process χ_t defined on

it which induces the given finite dimensional distributions. We take this result for granted, for a proof see [12] or [79].

Some examples of random processes will serve to clarify these ideas.

Example 11.2.1. Let X be the set of *all* bounded sequences of real numbers, i.e.

$$X = \{\mathbf{a} = (a_1, a_2, \ldots), \quad a_i \in \mathbb{R} \quad \text{with} \limsup_{i\to\infty} |a_i| < \infty\}.$$

Define a discrete parameter, random process $\{\chi_1, \chi_2, \ldots\}$ by setting

$$\chi_i(\mathbf{a}) = a_i.$$

To describe the measure theoretic aspects of this process we choose a probability measure, ν on $\mathbb{R}$. For all i define

$$\text{Prob}(\chi_i \leq t) = \int_{-\infty}^{t} d\nu.$$

Supposing further that the $\{\chi_i\}$ are independent random variables, we can compute the joint distributions: for each $k \in \mathbb{N}$, multi-index $\{i_1, \ldots, i_k\}$ and $(s_1, \ldots, s_k) \in \mathbb{R}^k$ we have

$$\text{Prob}(\chi_{i_1} \leq s_1, \ldots, \chi_{i_k} \leq s_k) = \int_{-\infty}^{t_1} d\nu \cdots \int_{-\infty}^{t_k} d\nu.$$

These properties serve to characterize a random process, though the proof that a σ-algebra and measure are defined on X inducing these joint distribution functions is by no means trivial.

Example 11.2.2. Another example is the set of infinite sequences of coin flips. The sample space X is a set of all possible infinite sequences of heads and tails. As above, define

$$\chi_i(\mathbf{a}) = \begin{cases} 1 & \text{if } a_i = H, \\ 0 & \text{if } a_i = T. \end{cases}$$

The $\{\chi_i\}$ are then taken to be independent random variables with

$$\text{Prob}(\chi_i = 0) = 1 - p, \quad \text{Prob}(\chi_i = 1) = p.$$

Such a process is called a *Bernoulli random process* because each χ_i is Bernoulli random variable.

Example 11.2.3. An example of a continuous time random process is provided by setting $X = \mathcal{C}_0(\mathbb{R}_+)$ the set of continuous functions on $\mathbb{R}_+$ which vanish at 0. For each $t \in \mathbb{R}_+$ we have the random variable $\chi(t)$ defined at $w \in X$ by

$$\chi(t; w) = w(t)$$

χ is the evaluation of the function w at time t. As before, it is difficult to give a direct description of the σ-algebra and measure on X. Instead the process is described in terms of its joint distribution functions. That is, we need to specify

$$\text{Prob}(\chi(t_1) \leq s_1, \chi(t_2) \leq s_2, \ldots, \chi(t_k) \leq s_k),$$

for all $k \in \mathbb{N}$ and all pairs of real k-tuples, $((t_1, \ldots, t_k), (s_1, \ldots, s_k))$.

An important special case of this construction is given by

$$\text{Prob}(\chi(t_1) \leq s_1, \chi(t_2) \leq s_2, \ldots, \chi(t_k) \leq s_k) = \int_{-\infty}^{s_k} \cdots \int_{-\infty}^{s_1} \frac{e^{-\frac{x_1^2}{2t_1}}}{\sqrt{2\pi t_1}} \prod_{j=2}^{k} \frac{e^{-\frac{(x_j - x_{j-1})^2}{2(t_j - t_{j-1})}}}{\sqrt{2\pi(t_j - t_{j-1})}} dx_1 \cdots dx_k, \tag{11.2}$$

if $t_1 < t_2 < \cdots < t_k$. This process is called *Brownian motion*. For each t the random variable $\chi(t)$ is Gaussian and for any finite set of times, $(t_1, \ldots, t_k)$ the random variables $\{\chi(t_1), \ldots, \chi(t_k)\}$ are jointly Gaussian. We return to this example latter on.

Exercise 11.2.1. Show that the cumulative distributions defined in example 11.2.1 satisfy the consistency conditions.

Exercise 11.2.2. In many applications we need to approximate a random process by a finite dimensional sample space. Suppose that in example 11.2.1 we consider finite, real sequences of length N. The sample space is then $\mathbb{R}^N$. The random variables $\{\chi_1, \ldots, \chi_N\}$ are defined on this space. Find a probability measure on $\mathbb{R}^N$ which gives the correct joint distributions functions for these variables. Are there others which would also work?

11.2.1 Statistical properties of random processes

In practical applications a random process is described by its statistical properties. The simplest are the mean

$$\mu_\chi(t) = E[\chi(t)]$$

and variance

$$\sigma_\chi^2(t) = E[(\chi(t) - \mu_\chi(t))^2].$$

A measure of the relationship of χ at two different times is the *autocorrelation function* :

$$R_\chi(t_1, t_2) = < \chi(t_1)\chi(t_2) > .$$

As before, the *covariance* is defined by

$$\begin{aligned} \text{Cov}(\chi(t_1), \chi(t_2)) &= R_\chi(t_1, t_2) - < \chi(t_1) >< \chi(t_2) > \\ &= E[(\chi(t_1) - \mu_\chi(t_1))(\chi(t_2) - \mu_\chi(t_2))]. \end{aligned} \tag{11.3}$$

The normalized correlation coefficient is

$$\rho(t_1, t_2) = \frac{\text{Cov}(\chi(t_1), \chi(t_2))}{\sigma_{\chi(t_1)}\sigma_{\chi(t_2)}}.$$

If the cumulative joint distribution for $\chi(t_1)$ and $\chi(t_2)$ has a distribution function then

$$R_\chi(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy p_{t_1,t_2}(x, y) dx dy.$$

Using the Cauchy-Schwartz inequality we deduce that,

$$\begin{aligned} R_\chi(t_1,t_2) &= \iint xyp_{t_1,t_2}(x,y)dxdy \\ &\leq \big[\iint x^2 p_{t_1,t_2}(x,y)dxdy\big]^{1/2}\big[\iint y^2 p_{t_1,t_2}(x,y)dxdy\big]^{1/2} \\ &= \big[\iint x^2 p_{t_1}(x)dx\big]^{1/2}\big[\iint y^2 p_{t_2}(y)dy\big]^{1/2}. \end{aligned}$$

Hence, we have the estimate

$$|R_\chi(t_1,t_2)| \leq \sqrt{E[\chi(t_1)^2]E[\chi(t_2)^2]}. \tag{11.4}$$

Exercise 11.2.3. Show how to derive (11.4) using the formalism of a probability space, i.e. by integrating over X with respect to $d\nu$.

11.2.2 Stationary random processes

A important notion is that of a *stationary random process*. Heuristically a noise process is stationary if it does not matter when you start looking, the noise is always "the same."

Definition 11.2.1. Let $\chi(t)$ be a continuous parameter random process. It is a stationary process if

(1) $\text{Prob}(\chi(t) \leq \lambda)$ is independent of t.

(2) For any $\tau \in \mathbb{R}$,

$$\text{Prob}(\chi(t_1) \leq r, \chi(t_2) \leq s) = \text{Prob}(\chi(t_1+\tau) \leq r, \chi(t_2+\tau) \leq s).$$

(2)′ Similarly, for any collection of $(t_1,\ldots,t_k)$, $(s_1,\ldots,s_k)$ and τ we have

$$\text{Prob}(\chi(t_1) \leq s_1,\ldots,\chi(t_k) \leq s_k) = \text{Prob}(\chi(t_1+\tau) \leq s_1,\ldots,\chi(t_k+\tau) \leq s_k).$$

If $\chi(t)$ is a stationary random process, then

$$R_\chi(t_1,t_2) = E[\chi(t_1)\chi(t_2)] = E[\chi(0)\chi(t_2-t_1)]. \tag{11.5}$$

Setting $r_\chi(\tau) = E[\chi(0)\chi(\tau)]$ it follows that

$$R_\chi(t_1,t_2) = r_\chi(t_2-t_1).$$

On the other hand, the fact that $R_\chi(t_1,t_2)$ is a function of t_2-t_1 does **not** imply that $\chi(t)$ is a stationary process. A process satisfying this weaker condition is called a *weak sense stationary random process*. For a weak sense stationary process,

$$r_\chi(\tau) \leq\ E[\chi^2(0)] = r_\chi(0). \tag{11.6}$$

This coincides well with intuition. If something is varying in a "random" but stationary way then it is unlikely to be better correlated at two different times than at a given time.

The reason for introducing the autocorrelation function is that it allows the use of Fourier theory to study weak sense stationary processes.

Definition 11.2.2. If χ_t is a weak sense stationary random process and r_χ is integrable then its Fourier transform,

$$S_\chi(\xi) = \int_{-\infty}^{\infty} r_\chi(\tau) e^{-i\tau\xi} d\tau$$

is called the *spectral density function* for the process χ.

The autocorrelation function is not always integrable but, as shown below, it is a "non-negative, definite function." It then follows from a theorem of Herglotz that its Fourier transform is well defined as a measure on $\mathbb{R}$. This means that, while $S_\chi(\xi)$ may not be well defined at points, for any $[a, b]$ the integral

$$\frac{1}{2\pi} \int_a^b S_\chi(\xi) d\xi$$

is meaningful. This measure is also called the spectral density function. The integral *defines* the power contained in the process in the interval $[a, b]$.

The proposition enumerates the important properties of r_χ and S_χ.

Proposition 11.2.1. *If $r_\chi(\tau)$ is the autocorrelation function of a real, weak sense stationary random process and $S_\chi(\xi)$ is its Fourier transform then the following statements hold:*

(1) $r_\chi(\tau)$ is real valued.

(2) The autocorrelation is an even function: $r_\chi(\tau) = r_\chi(-\tau)$.

(3) S_χ is a real valued even function.

(4) S_χ is non negative.

(5) The total power of the process is the variance.

$$r_\chi(0) = E[\chi(t)^2] = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_\chi(\xi) d\xi.$$

Proof. By definition, we have

$$r_\chi(\tau) = E[\chi(t)\chi(t+\tau)] = E[\chi(t-\tau)\chi(t)] = r_\chi(-\tau).$$

This implies that S_χ is even. To show that the spectral density function is real valued, take the conjugate of S_χ:

$$\bar{S}_\chi(\xi) = \int_{-\infty}^{\infty} r_\chi(\tau) e^{i\xi\tau} d\tau = \int_{-\infty}^{\infty} r_\chi(-\tau) e^{i\xi\tau} d\tau = S_\chi(\xi).$$

The fourth fact is not obvious from the definition. This follows because the autocorrelation function $r_\chi(\tau)$ is a *non-negative definite function*. This means that for any vectors $(x_1, \ldots, x_N)$ and $(\tau_1, \ldots, \tau_N)$, we have that

$$\sum_{i,j=1}^{N} r_\chi(\tau_i - \tau_j) x_i x_j \geq 0.$$

This is a consequence of the fact that the expected value of a non-negative, random variable is non-negative. If $\chi(t)$ is any continuous time random process with finite mean and covariance then

$$0 \leq < |\sum_{1}^{N} x_i \chi(\tau_i))|^2 >= \sum_{i,j} < x_i x_j \chi_i(\tau_i) \chi_j(\tau_j) >= \sum_{i,j} x_i x_j R_\chi(\tau_i, \tau_j).$$

Hence, $\sum_{i,j=1}^{N} R_\chi(\tau_i, \tau_j) x_i x_j \geq 0$. For a weak sense stationary process, $R_\chi(\tau_1, \tau_2) = r_\chi(\tau_1 - \tau_2)$ and thus

$$\sum_{i,j=1}^{N} r_\chi(\tau_i - \tau_j) x_i x_j \geq 0. \qquad (*)$$

The Herglotz theorem states that a function is the Fourier transform of a positive measure if and only if it is non-negative definite. Hence, the fact that r_χ is non-negative definite implies that $S_\chi d\xi$ is a non-negative measure, that is

$$\int_a^b S_\chi(\xi) d\xi = \int_a^b \int_{-\infty}^{\infty} r_\chi(\tau) e^{-i\tau\xi} d\tau \geq 0,$$

see [40]. Fact (5) follows from the Fourier inversion formula. □

If $f(t)$ is a real valued, bounded, integrable function then we see that its "autocorrelation" function

$$r_f(t) = \int_{-\infty}^{\infty} f(t) f(t+\tau) dt$$

is well defined. Computing the Fourier transform of r_f gives

$$\hat{r}_f(\xi) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(t) f(t+\tau) e^{-i\xi\tau} dt d\tau = |\hat{f}(\xi)|^2.$$

These computations are not immediately applicable to a random process but hold in a probabilistic sense.

Let $\chi(t)$ be a real valued, weak sense stationary random process. For each $T > 0$ define

$$\widehat{\chi}_T(\xi) = \int_{-T}^{T} \chi(t) e^{-it\xi} dt.$$

We compute the expected value of $|\widehat{\chi}_T(\xi)|^2$,

$$\begin{aligned} E[|\widehat{\chi}_T(\xi)|^2] &= E\left[\int_{-T}^{T} \chi(t)e^{-it\xi}dt \int_{-T}^{T} \chi(s)e^{is\xi}ds\right] \\ &= \int_{-T}^{T}\int_{-T}^{T} r_\chi(t-s)e^{-i(t-s)\xi}dtds. \end{aligned} \tag{11.7}$$

Letting $\tau = t - s$ we obtain

$$\begin{aligned} E[|\widehat{\chi}_T(\xi)|^2] &= \int_{-2T}^{2T} (2T - |\tau|)r_\chi(\tau)e^{-i\tau\xi}d\tau \\ &= (2T)\int_{-2T}^{2T} \left(1 - \frac{|\tau|}{2T}\right) r_\chi(\tau)e^{-i\tau\xi}d\tau. \end{aligned} \tag{11.8}$$

Proposition 11.2.2. *If $\chi(t)$ is a weak sense stationary random process and $r_\chi(t)$ is integrable then*

$$S_\chi(\xi) = \lim_{T\to\infty} \frac{1}{2T} E[|\widehat{\chi}_T(\xi)|^2]. \tag{11.9}$$

Proof. Under the hypotheses, the Lebesgue dominated convergence theorem applies to shows that the limit, as $T \to \infty$ can be taken inside the integral in the second line of (11.8), giving the result. □

This proposition justifies the description of $S_\chi(\xi)$ as the "power spectral density" of the process. For a given point w in X, the sample space $t \mapsto \chi(t; w)$ is a sample path. A reasonable definition of the autocorrelation function for a single sample path is

$$r_w(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \chi(t; w)\chi(t + \tau; w)dt.$$

This is sometimes called the *time autocorrelation function*. For a given choice of w this limit may not exist. It turns out that for many interesting classes of random processes this time average, exists, defines a random variable on X and with probability one, does not depend on the choice of path. In fact

$$\text{Prob}(r_w = r_\chi) = 1.$$

Exercise 11.2.4. Derive (11.5).

Exercise 11.2.5. Prove (11.6).

Exercise 11.2.6. Derive the right hand side of (11.8).

Exercise 11.2.7. Suppose that $\chi(t)$ is a random process so that $E[|\chi(t)|]$ is independent of t. Show either $\chi(t) \equiv 0$, with probability one or

$$\int_{-\infty}^{\infty} |\chi(t)|dt = \infty,$$

with probability one.

11.2.3 Independent and stationary increments

Many processes encountered in imaging applications are not themselves stationary but satisfy the weaker hypothesis of having *stationary increments*.

Definition 11.2.3. Let χ_t be a continuous parameter random process such that for any finite sequence of times $t_1 < t_2 < \cdots < t_n$, $n \geq 3$ the random variables

$$\chi_{t_2} - \chi_{t_1}, \chi_{t_3} - \chi_{t_2}, \ldots, \chi_{t_n} - \chi_{t_{n-1}}$$

are independent. The process is said to have *independent increments*. If moreover $\text{Prob}(\chi_t - \chi_s \leq \lambda)$ depends only on $t - s$ then the process has *stationary increments*.

A weaker condition is that a process have uncorrelated increments, that is

$$E[(\chi_{t_2} - \chi_{s_2})(\chi_{t_1} - \chi_{s_1}) = E[(\chi_{t_2} - \chi_{s_2})]E[(\chi_{t_1} - \chi_{s_1})],$$

provided that $[t_2, s_2] \cap [t_1, s_1] = \emptyset$. If $E[|\chi_t - \chi_s|^2]$ only depends on $t - s$ then the process is said to have *wide sense stationary increments*.

Example 11.2.4. Brownian motion is a random process, parametrized by $[0, \infty)$ which describes, among other things the motion of tiny particles in a fluid. It is defined as a process χ_t with independent increments, such that for every s, t the increment $\chi_t - \chi_s$ is a Gaussian random variable with

$$E[\chi_t - \chi_s] = 0 \text{ and } E[(\chi_t - \chi_s)^2] = \sigma^2|t - s|^2$$

This process is often normalized by fixing $\chi_0 = a \in \mathbb{R}$, with probability 1. A very important fact about Brownian motion is that it is essentially the only process with independent increments whose sample paths are continuous, with probability one. Brownian motion is frequently called the *Wiener process*.

Exercise 11.2.8. Show that if a Gaussian process has $E[\chi_t] = 0$ and uncorrelated increments then it has independent increments.

11.3 Examples of random processes

For many applications a small collection of special random processes suffice. Several have already been defined, a few more are described in this section.

11.3.1 Gaussian random process

A Gaussian random process is a family, $\{\chi(t)\}$ or sequence of random variables $\{\chi_i\}$ which, for each t (or i) is a Gaussian random variable. The finite dimensional distributions are also assumed to be Gaussian. As we saw in section 10.3.3, the joint distributions for Gaussian random variables are determined by their means and covariance matrix. This remains true of Gaussian processes and again the converse statement is also true. Suppose that T is the parameter space for a random process and that there are real valued functions $\mu(t)$ defined on T and $r(s,t)$ defined on $T \times T$. The function r is assumed to satisfy the conditions

(1). For any pair $s, t \in T$ $r(s,t) = r(t,s)$ and,

(2). If $\{t_1, \ldots, t_m\} \subset T$ then the matrix $[r(t_i, t_j)]$ is non-negative definite.

There exists a Gaussian random process $\{\chi_t \,:\, t \in T\}$ such that

$$E[\chi_t] = \mu(t) \text{ and } E[\chi_s \chi_t] = r(s,t).$$

If one is only concerned with the second order statistics of a random process then one is free to assume that there process is Gaussian.

Brownian motion, defined in example 11.2.4, is an important example of a Gaussian process. As remarked there, we can fix $\chi_0 = 0$ with probability one. Since $E[\chi_t - \chi_s] = 0$ for all t, s it follows that

$$E[\chi_t] = 0 \text{ for all } t \in [0, \infty).$$

The autocorrelation function can now be computed using the hypothesis

$$E[(\chi_t - \chi_s)^2] = \sigma^2 |t - s|. \tag{11.10}$$

Let $0 < s < t$ then as $\chi_0 = 0$ with probability one,

$$E[\chi_s \chi_t] = E[(\chi_s - \chi_0)(\chi_t - \chi_0)].$$

This can be rewritten as

$$\begin{aligned} E[\chi_s \chi_t] &= E[(\chi_s - \chi_0)^2] + E[(\chi_s - \chi_0)(\chi_t - \chi_s)] \\ &= \sigma^2 s^2 = \sigma^2 \min\{|s|, |t|\}. \end{aligned} \tag{11.11}$$

In passing from the first line to the second we use the independence of the increments and (11.10). Thus Brownian motion is not a weak sense stationary process.

11.3.2 The Poisson counting process

The Poisson counting process is another example of a process with independent, stationary increments. This process is a family of random variables $\{\chi(t)\}$ defined for $t \geq 0$, which take values in $\{0, 1, 2, \ldots\}$. The Poisson counting process has a nice axiomatic characterization. For convenience let:

$$P(k,t) = \text{Prob}(\chi(t) = k).$$

Here are the axioms phrased in terms of counting "emitted particles:"

Independent increments:
The number of particles emitted in $[t_1, t_2]$ is independent of the number in $[t_3, t_4]$ if $[t_1, t_2] \cap [t_3, t_4] = \emptyset$.

Short time behavior:
The probability that one particle is emitted in a very short time interval is given by

$$P(1, \Delta t) = \lambda \Delta t + o(\Delta t) \quad \text{for some constant } \lambda$$

where $o(\Delta t)$ denotes a term such that $\lim_{\Delta t \to 0} o(\Delta t)/\Delta t = 0$. As a consequence $P(1, 0) = 0$.

Stationary increments:
The process has stationary increments.

$$\text{Prob}(\chi(t) - \chi(s) = k) = \text{Prob}(\chi(t + \tau) - \chi(s + \tau) = k), \quad \forall \tau \geq 0, 0 \leq s \leq t.$$

We can now estimate the probability that two particles are emitted in a short interval $[0, \Delta t]$: In order for this to happen there must be a $0 < \tau < \Delta t$ such that one particle is emitted in $[0, \tau]$ and one is emitted in $(\tau, \Delta]$. The hypothesis of independent, stationary increments implies that

$$P(2, \Delta t) \leq \max_{\tau \in (0, \Delta t)} P(1, \tau) P(1, \Delta t - \tau) = O((\Delta t)^2).$$

From the independent, stationary increments axiom, we have that

$$P(0, t + \Delta t) = P(0, t) P(0, \Delta t).$$

For any time Δt it is clear that $P(k, \Delta t) \leq P(k + 1, \Delta t) \leq \dots$ and that

$$\sum_{k=0}^{\infty} P(k, \Delta t) = 1$$

In fact arguing as above one can show that $P(k, \Delta t) \leq [P(1, \Delta t)]^k$, combining these observation leads to

$$P(0, \Delta t) + P(1, \Delta t) = 1 + o(\Delta t). \tag{11.12}$$

Hence

$$P(0, \Delta t) = 1 - \lambda \Delta t + o(\Delta t), \tag{11.13}$$

$$P(0, t + \Delta t) = P(0, t) P(0, \Delta t) = P(0, t)[1 - \lambda \Delta t + o(\Delta t)]. \tag{11.14}$$

Letting $\Delta t \to 0$, we have

$$\lim_{\Delta t \to 0} \frac{P(0, t + \Delta t) - P(0, t)}{\Delta t} = \frac{P(0, t)(-\lambda \Delta t + o(\Delta t))}{\Delta t} = -\lambda P(0, t).$$

This provides a differential equation for $P(0, t)$:

$$\frac{dP}{dt}(0, t) = -\lambda P(0, t), \text{ with } P(0, 0) = 1.$$

The solution of this equation is

$$P(0,t) = e^{-\lambda t}.$$

The probabilities $\{P(k,t)\}$ for $k > 1$ are obtained recursively. For each $t \geq 0$ and $j \leq k$ suppose that

$$P(j,t) = \frac{(\lambda t)^j}{j!} e^{-\lambda t}.$$

The hypothesis of independent, stationary increments implies that

$$\begin{aligned} P(k+1, t+\Delta t) &= P(k+1,t)P(0,\Delta t) + P(k,t)P(1,\Delta t) + \sum_{j=0}^{k-1} P(j,t)P(k+1-j,\Delta t) \\ &= P(k,t)P(1,\Delta t) + P(k+1,t)P(0,\Delta t) + o(\Delta t). \end{aligned} \tag{11.15}$$

Using the induction hypothesis, the last term behaves as $o(\Delta t)$. From equations (11.13) and (11.15) we obtain

$$P(k+1,t+\Delta t) - P(k+1,t) = P(k,t)(\lambda \Delta t + o(\Delta t)) + P(k+1,t)[P(0,\Delta t) - 1] + o(\Delta t),$$

which leads to

$$\frac{dP}{dt}(k+1,t) = \lambda P(k,t) - \lambda P(k+1,t) \tag{11.16}$$

Hence, we obtain that

$$\text{Prob}(\chi(t) = k+1) = P(k+1,t) = \frac{(\lambda t)^{k+1}}{(k+1)!} e^{-\lambda t}.$$

For each t, $\{\chi(t)\}$ is a Poisson random variable with intensity λt. As the intensity changes with time it follows that this cannot be a stationary random process. The expected value and the variance are

$$E[\chi(t)] = \lambda t, \tag{11.17}$$

$$E[(\chi(t) - \lambda t)^2] = \lambda t, \tag{11.18}$$

as follows from the formulæ in section 10.3.2.

Suppose we know that one particle is emitted in an interval $[0,T]$, what is the probability distribution for the time the particle is emitted? This is a question about conditional probability. We formulate it as follows, for $0 \leq t \leq T$:

$$\begin{aligned} \text{Prob}(\chi(t) = 1 | \chi(T) = 1) &= \frac{\text{Prob}(\chi(t) = 1, \text{ and } \chi(T) = 1)}{\text{Prob}(\chi(T) = 1)} \\ &= \frac{\text{Prob}(\chi(t) = 1, \chi(T) - \chi(t) = 0)}{\text{Prob}(\chi(T) = 1)}. \end{aligned}$$

Using the distributions obtained above we see that this equals

$$\text{Prob}(\chi(t) = 1 | \chi(T) = 1) = \frac{P(1,t)P(0,T-t)}{P(1,T)} = \frac{\lambda t e^{-\lambda t} e^{-\lambda(T-t)}}{\lambda T e^{-\lambda T}} = \frac{t}{T}.$$

This says that each time in $[0,T]$ is equally probable. The Poisson counting process is used to describe radioactive decay. If it is known that one decay was observed in certain interval then, the time of decay is uniformly distributed over the interval. This is why it is said that the time of decay is "completely random."

Next we compute the autocorrelation function, $E[\chi(t)\chi(s)]$. It follows from the identity identity:

$$\begin{aligned} E[(\chi(t)-\chi(s))^2] &= E[\chi(t)^2 - 2\chi(t)\chi(s) + \chi(s)^2] \\ &= E[\chi(t)^2] - 2E[\chi(t)\chi(s)] + E[\chi(s)^2]. \end{aligned}$$

From the stationary increments property, and $\chi(0)=0$, it follows that

$$E[(\chi(t)-\chi(s))^2] = E[\chi(t-s)-\chi(0))^2] = E[\chi(t-s)^2].$$

Assume that $t \geq s$, then

$$\begin{aligned} E[\chi(t-s)^2] &= \sum_{k=0}^{\infty} k^2 \,\mathrm{Prob}(\chi(t-s)=k) \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} k^2 [\lambda(t-s)^k e^{-\lambda(t-s)}] = \lambda(t-s) + \lambda^2(t-s)^2. \end{aligned}$$

The autocorrelation is now easily obtained

$$E[\chi(t)\chi(s)] = \frac{1}{2}(E[\chi(t)^2] + E[\chi(s)^2] - E[(\chi(t)-\chi(s))^2]) = \lambda \min(t,s) + \lambda^2 ts.$$

The Poisson counting process is also not a weak sense stationary process. Substituting $E[\chi(t)] = \lambda t$, gives

$$\mathrm{Cov}(\chi(t),\chi(s)) = \lambda \min(t,s).$$

Exercise 11.3.1. Prove that $P(k,t)$ satisfies (11.16).

11.3.3 Poisson arrival process

Let $\chi(t)$ be a continuous parameter, Poisson counting process. Strictly speaking, a continuous parameter, Poisson process is a function $\chi(t;w)$ of two variables (t,w), where w is a point in the sample space. The second variable is usually suppressed. Several interesting processes, with the same underlying sample space, can be built out of the counting process. We now describe the Poisson arrival process. Let $T_1(w)$ be the time the first particle arrives, and $T_2(w)$ the second arrival time and recursively, $T_n(w)$ the n^{th} arrival time. This is called the Poisson arrival process. Taking differences gives a further process

$$\begin{aligned} Z_1 &= T_1, \\ Z_2 &= T_2 - T_1, \\ &\vdots \\ Z_i &= T_i - T_{i-1}. \end{aligned}$$

The hypothesis that the original process $\{\chi(t)\}$ has independent increments implies that $\{Z_i\}$ are independent random variables. They are identically distributed because the counting process has stationary increments. The original process is a function of a continuous parameter which takes integer values. The arrival process and its increments are sequences of random variables indexed by positive integers taking continuous values.

We now work out the distribution function for these two processes. The probability that the first particle arrives after time t equals the probability that $\chi(t)$ is zero:

$$\text{Prob}(Z_1 > t) = \text{Prob}(\chi(t) = 0) = e^{-\lambda t}.$$

Hence,

$$\text{Prob}(Z_1 \leq t) = 1 - e^{-\lambda t} = \int_{-\infty}^{t} \lambda e^{-\lambda t} \chi_{[0,\infty]}(t) dt. \tag{11.19}$$

The density function of Z_1, hence that of Z_i for each i, is $\lambda e^{-\lambda t}\chi_{[0,\infty]}(t)$. The expected value of Z_1 is

$$E[Z_1] = \int_{0}^{\infty} t\lambda e^{-\lambda t} dt = \frac{1}{\lambda}.$$

For radioactive decay this says that the expected length of time before the first decay is $1/\lambda$ which agrees well with intuition. The more intense a process is, the less time one expects to wait for the first decay. The variance and standard deviation are

$$E[Z_1^2] = \int_{0}^{\infty} t^2 \lambda e^{-\lambda t} dt = \frac{2}{\lambda^2} \Rightarrow \sigma_{Z_1}^2 = E[Z_1^2] - E[Z_1]^2 = \frac{1}{\lambda^2}.$$

The arrival time, T_n is the sum of the differences of arrival times, i.e., $T_n = Z_1 + \cdots + Z_n$. Thus each T_n is a sum of independent, identically distributed random variables. The $E[e^{-i\xi Z_n}]$ is

$$\begin{aligned} \hat{p}(\xi) &= \int_{-\infty}^{\infty} e^{-i\xi t} \lambda e^{-\lambda t} \chi_{[0,\infty]}(t) dt \\ &= \lambda \int_{0}^{\infty} e^{-t(i\xi+\lambda)} dt \\ &= \frac{\lambda}{\lambda + i\xi}. \end{aligned}$$

From exercise 10.2.32 it follows that $E[e^{-i\xi T_n}]$ is

$$E[e^{-i\xi T_n}] = E[e^{-i\xi Z_1}] \cdots E[e^{-i\xi Z_n}] = [\hat{p}(\xi)]^n.$$

Using a complex contour integral, the Fourier inversion of $[\hat{p}(\xi)]^n$ is obtained:

$$\int_{-\infty}^{\infty} \frac{\lambda^n}{(\lambda + i\xi)^n} e^{i\xi t} d\xi = \begin{cases} 0 & t < 0, \\ \frac{1}{(n-1)!} \lambda e^{-\lambda t} (\lambda t)^{n-1} & t \geq 0. \end{cases}$$

The probability distribution for T_n is therefore:

$$\text{Prob}(T_n \leq t) = \int_0^t \frac{\lambda e^{-\lambda t}(\lambda t)^{n-1}}{(n-1)!} = \frac{1}{(n-1)!}\int_0^{\lambda t} e^{-\tau}\tau^{n-1}d\tau.$$

Recall that the Gamma function is defined by

$$\Gamma(x) = \int_0^\infty e^{-s}s^{x-1}ds.$$

The Γ-function satisfies $\Gamma(n+1) = n!;$. For any $n > 0$ this implies that

$$\lim_{t\to\infty} \text{Prob}(T_n \leq t) = 1$$

Since the $\{Z_i\}$ are identically distributed, independent random variables, the central limit theorem applies to show

$$\lim_{n\to\infty} \text{Prob}(\frac{T_n - n/\lambda}{\sqrt{n}/\lambda} \leq t) \to \int_{-\infty}^t e^{-\frac{x^2}{2}}\frac{dx}{\sqrt{2\pi}}.$$

Exercise 11.3.2. Prove that the $\{Z_i\}$ are independent , identically distributed random variables.

11.3.4 Fourier coefficients for periodic processes

Suppose that χ_t is a weak sense stationary random process with finite variance and correlation function $r_\chi(\tau)$. A process is said to be *mean square T-periodic* if

$$E[|\chi_{t+T} - \chi_t|^2] = 0 \text{ for all } t.$$

Since χ_t has finite variance, with probability one, the sample paths have finite square integral over the interval $[0, T]$

$$E\left[\int_0^T |\chi_t|^2 dt\right] = \left[\int_0^T E[|\chi_t|^2]dt\right] = Tr_\chi(0).$$

Because the process is weak sense stationary $E[|\chi_t|^2] = r_\chi(0)$ is constant. In light of these estimates the Fourier coefficients

$$\zeta_k = \frac{1}{T}\int_0^\infty \chi_t e^{-\frac{2\pi i k t}{T}}dt, \quad k \in \mathbb{Z}$$

are defined, with probability one. These are *complex valued* functions on the same sample space as χ_t. The integral can be understood as a limit of finite sums; using standard measure theory one can show that the $\{\zeta_k\}$ are measurable functions on the sample space. Hence $\{\zeta_k\}$ defines a complex, discrete parameter random process.

We first consider the autocorrelation function of the original process.

Proposition 11.3.1. *If χ_t is a mean square T-periodic, weak sense stationary random process then $r_\chi(\tau)$ is T-periodic.*

Proof. We need to show that $r_\chi(\tau + T) = r_\chi(\tau)$ for any τ. The proof is a simple computation:

$$\begin{aligned} r_\chi(\tau + T) &= E[\chi_0 \chi_{\tau+T}] \\ &= E[\chi_T \chi_{\tau+T}] + E[(\chi_0 - \chi_T)\chi_{\tau+T}] \\ &= r_\chi(\tau) + E[(\chi_0 - \chi_T)\chi_{\tau+T}]. \end{aligned} \tag{11.20}$$

The Cauchy-Schwarz inequality, (10.2.2) gives the estimate

$$E[(\chi_0 - \chi_T)\chi_{\tau+T}] \leq \sqrt{E[(\chi_0 - \chi_T)^2]E[(\chi_{\tau+T})^2]} = 0.$$

The right hand side is zero because χ_t is mean square T-periodic. This completes the proof of the proposition. □

Since $r_\chi(\tau)$ is bounded by $r_\chi(0)$ and T-periodic it has a Fourier series expansion

$$r_\chi(\tau) = \sum_{k=-\infty}^{\infty} r_k e^{\frac{2\pi i k \tau}{T}}.$$

Here equality is in the "limit in the mean" sense, see Proposition 5.3.1 and

$$r_k = \frac{1}{T} \int_0^T r_\chi(\tau) e^{-\frac{2\pi i k \tau}{T}} d\tau.$$

According to the Parseval formula the coefficients satisfy

$$\sum_{k=-\infty}^{\infty} |r_k|^2 = \frac{1}{T} \int_0^T |r_\chi(\tau)|^2 d\tau.$$

Using the definition of the autocorrelation function gives the formula

$$r_k = E[\chi_0 \zeta_k] \tag{11.21}$$

This brings us to the main result of this section

Proposition 11.3.2. *The random variables $\{\zeta_k\}$ are pairwise uncorrelated, i.e. $E[\zeta_k \overline{\zeta_l}] = 0$ if $k \neq l$ and*

$$E[|\zeta_k|^2] = r_k. \tag{11.22}$$

Remark 11.3.1. It is natural to use $E[f\bar{g}]$ when working with complex valued random variables so that this pairing defines an hermitian inner product.

Proof. Once again the proof is a simple computation interchanging an expectation with integrals over $[0,T]$:

$$\begin{aligned} E[\zeta_k\overline{\zeta_l}] &= \frac{1}{T^2}E\left[\int_0^T \chi_t e^{-\frac{2\pi ikt}{T}}dt \int_0^T \chi_s e^{\frac{2\pi iks}{T}}ds\right] \\ &= \frac{1}{T^2}\int_0^T\int_0^T \left[e^{-\frac{2\pi ikt}{T}}e^{\frac{2\pi iks}{T}}r_\chi(t-s)\right]dtds \\ &= \frac{r_k}{T}\int_0^T e^{\frac{2\pi i(l-k)t}{T}}dt. \end{aligned} \tag{11.23}$$

The passage from the second to the third lines is a consequence of the T-periodicity of r_χ. The proposition follows as the integral in last line of (11.23) is $\delta_{kl}T$. □

Thus a mean square T-periodic, weak sense stationary process leads to a discrete parameter process of uncorrelated variables. The final question we need to address is the convergence of the Fourier series

$$\sum_{k=-\infty}^{\infty} \zeta_k e^{\frac{2\pi ikt}{T}} \tag{11.24}$$

to χ_t. Given that the χ_t is *mean square* periodic, it is reasonable to examine the convergence in $L^2([0,T])$.

Proposition 11.3.3. *The Fourier series in* (11.24) *converges to* χ_t *in* $L^2([0,T])$ *with probability one if and only if*

$$\lim_{N\to\infty}\sum_{k=-N}^{N} r_k = r_\chi(0).$$

Proof. We need to compute the expected value of

$$\int_0^T \left|\chi_t - \sum_{k=-\infty}^{\infty}\zeta_k e^{\frac{2\pi ikt}{T}}\right|^2 dt.$$

Once again, this reduces to interchanging an expectation with an ordinary integral,

$$\begin{aligned} E\left[\int_0^T \left|\chi_t - \sum_{k=-\infty}^{\infty}\zeta_k e^{\frac{2\pi ikt}{T}}\right|^2 dt\right] & \\ &= E\left[\int_0^T |\chi_t|^2 dt - T\sum_{k=-\infty}^{\infty}|\zeta_k|^2\right] \\ &= T(r_\chi(0) - \sum_{k=-\infty}^{\infty} r_k). \end{aligned} \tag{11.25}$$

The definition of r_χ and equation (11.22) are used to go from the second to the third line. The statement of the proposition follows easily from the last line of (11.25). □

Another way to state the conclusion of the proposition is that the series in (11.24) represents χ_t, in the mean, with probability one if and only if r_χ is represented pointwise by its Fourier series at 0. This in turn depends on the regularity of $r_\chi(\tau)$ for τ near to zero.

Remark 11.3.2. This material is adapted from [13] where a thorough treatment of eigenfunction expansions for random processes can be found.

Exercise 11.3.3. Prove (11.21).

11.3.5 White noise

See: A.4.6, A.4.7.

In applications it is very often assumed that the noise component is modeled by *white noise*. This is a mean zero random process which is uncorrelated from time to time. The continuous parameter version of this process turns out to be rather complicated. We begin with a discussion of the discrete parameter case.

A random process $\{\chi_n\}$, indexed by $\mathbb{Z}$ is called a *white noise* process if

$$\begin{gathered} E[\chi_n] = 0 \text{ and } E[|\chi_n|^2] < \infty \text{ for all n and} \\ E[\chi_m \chi_n] = 0 \text{ if } m \neq n. \end{gathered} \tag{11.26}$$

A white noise process is simply an orthogonal collection of random variables on the sample space where the inner product is defined by $E[fg]$. The Fourier coefficients of a mean square periodic process therefore define a (complex valued) discrete, white noise process.

In the continuous parameter case we would like to do the same thing. White noise should be defined to be a random process, W_t which is weak sense stationary and satisfies

$$E[W_t] = 0, \quad r_W(\tau) = E[W_t W_{t+\tau}] = \sigma^2 \delta(\tau).$$

These properties are intuitively appealing because they imply that the noise is completely uncorrelated from one time to another. On the other hand its variance, $\sigma_W = \sqrt{r_W(0)}$, is infinite. The power spectral density is given by

$$S_W(\xi) = \int_{-\infty}^{\infty} \sigma^2 \delta(\tau) e^{-i\xi\tau} d\tau = \sigma^2.$$

Thus white noise has the same amount of power at every frequency. The problem is that there is no real valued random process with these properties. For example it makes no sense to ask for the value of $\text{Prob}(W_t \leq \lambda)$.

However, this concept is constantly used, so what does it mean? White noise is, in a sense, a 'generalized function' random process. One cannot speak of the value of the process at any given time, in much the same way that the δ-function does not have a well defined value at 0. On the other hand, if f is continuous then

$$\int_{-\infty}^{\infty} f(t)\delta(t)dt = f(0)$$

makes perfect sense. Similarly, one can give as precise meaning to time averages of white noise. If $f(t)$ is a continuously differentiable function, then

$$W_f = \int_a^b f(t) W_t dt$$

makes sense as a random variable; it has the "expected" mean and variance:

$$E[W_f] = \int_a^b f(t) E[W_t] dt = 0, \text{ and } E[W_f^2] = \sigma^2 \int_a^b f^2(t) dt.$$

In a similar way it makes sense to pass white noise through an sufficiently smoothing, linear filter. It is much more complicated to make sense of non-linear operations involving white noise.

The sample paths for a white noise process are usually described as the derivatives of the sample paths of an ordinary continuous time random process. Of course the sample paths of a random process are essentially never classically differentiable, so these derivatives must be interpreted as weak derivatives. We close this section by explaining, formally why white noise can be thought of in this way.

Let $\{\chi_t : t \geq 0\}$ denote Brownian motion and recall that

$$E[\chi_t] = 0 \text{ and } E[\chi_s \chi_t] = \sigma^2 \min\{s, t\}.$$

Formally we set $W_t = \partial_t \chi_t$. Commuting the derivative and the integral defining the expectation gives

$$E[W_t] = E[\partial_t \chi_t] = \partial_t E[\chi_t] = 0,$$

hence W_t has mean zero. To compute the variance we again commute the derivatives and the expectation to obtain

$$E[W_t W_s] = \partial_t \partial_s E[\chi_t \chi_s] = \partial_t \partial_s \sigma^2 \min\{s, t\}.$$

The right most expression is well defined as the weak derivative of a function of two variables:

$$\partial_t \partial_s \sigma^2 \min\{s, t\} = 2\sigma^2 \delta(t - s). \tag{11.27}$$

Let $\varphi(t, s)$ be a smooth function with bounded support in $[0, \infty) \times [0, \infty)$, the weak derivative in (11.27) is defined by the condition

$$\int_0^\infty \int_0^\infty \partial_t \partial_s \min\{s, t\} \varphi(s, t) ds dt = \int_0^\infty \int_0^\infty \min\{s, t\} \partial_t \partial_s \varphi(s, t) ds dt,$$

for every test function φ. Writing out the integral on the right hand side gives

$$\begin{aligned}\int_0^\infty \int_0^\infty \min\{s,t\}\partial_t\partial_s\varphi(s,t)dsdt &= \int_0^\infty \int_s^\infty s\partial_t\partial_s\varphi(s,t)dtds + \int_0^\infty \int_t^\infty t\partial_s\partial_t\varphi(s,t)dsdt \\ &= -\int_0^\infty s\partial_s\varphi(s,s)ds - \int_0^\infty t\partial_t\varphi(t,t)dt \\ &= 2\int_0^\infty \varphi(s,s)ds.\end{aligned} \tag{11.28}$$

The last line follows by integration by parts, using the bounded support of φ to eliminate the boundary terms. At least formally, this shows that the first (weak) derivative of Brownian motion is white noise.

Exercise 11.3.4. Give a detailed justification for the computations in (11.28).

Exercise 11.3.5. Show that the first derivative of the Poisson arrival process is also, formally a white noise process. What do the sample paths for this process look like?

11.4 Random inputs to linear systems

In the analysis of linear filters, we often interpret the input and output as a deterministic part plus noise. The noise is modeled as a random process and therefore the output is also a random process. One often wishes to understand the statistical properties of the output in terms of those of the input. In this connection is is useful to think of a continuous parameter, random process as a function of two variables $\chi(t;w)$, with w a point in the sample space X and t a time.

Recall that for w a point in the sample space X, $t \mapsto \chi(t;w)$ is called a sample path. For a shift invariant filter H with the impulse response function h, the output for such a random input is again a random process on the same sample space given *formally* by

$$\Upsilon(t;w) = H(\chi(t;w)) = \int_{-\infty}^{\infty} h(t-s)\chi(s;w)ds.$$

When writing such an expression we are asserting that it makes sense with probability one.

The statistics of the output process Υ are determined by the impulse response and the statistics of the input process. The expected value of the output of a linear system is an iterated integral:

$$\begin{aligned}E[\Upsilon(t)] &= \int_X \Upsilon(t;w)d\nu(w) \\ &= \int_X \int_{-\infty}^{\infty} h(t-s)\chi(s;w)dsd\nu(w).\end{aligned}$$

Under reasonable hypotheses, e.g. χ is bounded and h is integrable, the order of the two integrations can be interchanged. Though this exchange of order may not be trivial because X is usually an infinite dimensional space. Interchanging the order of the integrations gives

$$\begin{aligned} E[\Upsilon(t)] &= \int_{-\infty}^{\infty} \int_X h(t-s)\chi(s;w)d\nu(w)ds \\ &= \int_{-\infty}^{\infty} h(t-s)E[\chi(s)]ds. \end{aligned}$$

The expected output of a linear filter applied to a random process is the result of applying the linear filter to the expected value of the random process. Some care is necessary, even at this stage. If $E[\chi(s)]$ is a constant, μ_χ, then the expected value of Υ is

$$E[\Upsilon(t)] = \mu_\chi \int_{-\infty}^{\infty} h(t-s)ds = \mu_\chi \int_{-\infty}^{\infty} h(s)ds = \mu_\chi \hat{h}(0).$$

For this to make sense, we should assume that $\int |h(s)| < \infty$. If the input random process χ is stationary then the output process $H\chi$ is as well.

Exercise 11.4.1. Suppose that χ_t is a stationary random process and H is a linear shift invariant filter for which $H\chi$ makes sense (with probability one). Show that $(H\chi)_t$ is also a stationary process.

11.4.1 The autocorrelation of the output

To analyze shift invariant, linear systems we used the Fourier transform. In this case, it cannot be used directly since noise does not usually have a Fourier transform in the ordinary sense. Observe that

$$E[\int_{-\infty}^{\infty} |\chi(s)|ds] = \int_{-\infty}^{\infty} E[|\chi(s)|]ds.$$

For a stationary process the integral diverges unless $E[|\chi(s)|] = 0$, which would imply that the process equals zero, with probability one! For a non-trivial, stationary process the Fourier transform does not exist as an absolutely convergent integral. To get around this difficulty we consider the autocorrelation function. It turns out that the autocorrelation function for a stationary process is frequently square integrable.

For a non-stationary process, the autocorrelation function $R_\chi$is defined by

$$R_\chi(t_1, t_2) = E[\chi(t_1)\chi(t_2)].$$

The process is weak sense stationary if there is a function r_χ so that the autocorrelation function is

$$R_\chi(t_1, t_2) = r_\chi(t_1 - t_2).$$

Given two random processes, $\chi(t), \Upsilon(t)$ on the same underlying probability space the *cross-correlation function* is defined to be

$$R_{\chi,\Upsilon}(t_1, t_2) = E[\chi(t_1)\Upsilon(t_2)].$$

For two stationary processes $R_{\chi,\Upsilon}$ is only a function of $t_2 - t_1$, we define

$$r_{\chi,\Upsilon}(\tau) = E[\chi(t)\Upsilon(t+\tau)].$$

Now suppose that H is a linear shift filter, with impulse response h and that χ_t is a random process for which $H\chi_t$ makes sense. The autocorrelation of the output process is

$$\begin{aligned} R_{H\chi}(t_1, t_2) &= E[H\chi(t_1)H\chi(t_2)] \\ &= E[\int_{-\infty}^{\infty} h(t_1 - s_1)\chi(s_1)ds_1 \int_{-\infty}^{\infty} h(t_2 - s_2)\chi(s_2)ds_2]. \end{aligned} \tag{11.29}$$

The expected value is itself an integral, interchanging the order of the integrations leads to

$$\begin{aligned} R_{H\chi}(t_1, t_2) &= E[\int_{-\infty}^{\infty} h(t_1 - s_1)\chi(s_1; w)ds_1 \int_{-\infty}^{\infty} h(t_2 - s_2)\chi(s_2; w)ds_2] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t_1 - s_1)h(t_2 - s_2)E[\chi(s_1; w)\chi(s_2; w)]ds_1 ds_2 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t_1 - s_1)h(t_2 - s_2)R_\chi(s_1, s_2)ds_1 ds_2 \\ &= [h^{(2)} * R_\chi](t_1, t_2). \end{aligned}$$

where $h^{(2)}(x, y) := h(x)h(y)$. Hence, $R_{H\chi} = h^{(2)} * R_\chi$ is expressible as a two dimensional convolution with R_χ.

For the case of a weak sense stationary process the result is simpler, recall that

$$R_\chi(t_1, t_2) = r_\chi(t_1 - t_2).$$

Letting $\tau_i = t_i - s_i$, $i = 1, 2$ we obtain

$$R_{H\chi}(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\tau_1)h(\tau_2)r_\chi(\tau_1 - \tau_2 + t_2 - t_1).d\tau_1 d\tau_2.$$

Thus the output is also weak sense stationary with

$$r_{H\chi}(\tau) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(s+t)h(t)r_\chi(s - \tau)dtds.$$

In Proposition 11.2.1 the properties of the power spectral density of a stationary random process are enumerated. Using the formula for $r_{H\chi}$ we compute the spectral power density of the output in terms of the spectral power density of the input obtaining

$$S_{H\chi}(\xi) = |\hat{h}(\xi)|^2 S_\chi(\xi). \tag{11.30}$$

This is consistent with the "determinate" case for if x is a finite energy signal, with $y = Hx$ and $\hat{y} = \hat{h}\hat{x}$ we have

$$|\hat{y}(\xi)|^2 = |\hat{h}(\xi)|^2 |\hat{x}(\xi)|^2. \tag{11.31}$$

If $h(\xi)$ is large for large values of ξ, then the linear filter amplifies the noise. Note that the total power of the output is given by

$$E[(H\chi)^2] = r_{H\chi}(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\hat{h}(\xi)|^2 S_\chi(\xi) d\xi.$$

which we compare this with the power in the input

$$E[\chi^2] = R_\chi(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_\chi(\xi) d\xi.$$

The variance in the input and output are given by

$$\begin{aligned} \sigma_\chi^2 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} S_\chi(\xi) d\xi - \mu_\chi^2 \\ \frac{1}{2\pi} \sigma_{H\chi}^2 &= \int_{-\infty}^{\infty} |\hat{h}(\xi)|^2 S_\chi(\xi) d\xi - |\hat{h}(0)|^2 \mu_\chi^2. \end{aligned} \tag{11.32}$$

To compute the power or variance of the output requires a knowledge of both the spectral density function $S_\chi(\xi)$ of the process as well as the transfer function of the filter.

Exercise 11.4.2. If $\chi(t)$ and $\Upsilon(t)$ are stationary processes show that $R_{\chi,\Upsilon}(t_1, t_2)$ only depends on $t_2 - t_1$.

Exercise 11.4.3. Derive (11.30).

11.4.2 Thermal or Johnson noise

Current is the flow of electrons through a conductor. The electrons can be thought of as discrete particles which, at normal room temperature, move in a random way through the conductor. Even with no applied voltage, the randomness of this motion produces fluctuations in the voltage measured across the conductor. The thermal motion of electrons produce noise in essentially any electrical circuit which is known as *Johnson noise*. While not an important source of noise in CT-imaging, Johnson noise is the main source of noise in MRI. The intensity of this noise is related to the impedence of the electrical circuit. To

understand this dependence we examine the result of using a white noise voltage source as the input to the simple electrical circuit shown in figure 11.1.

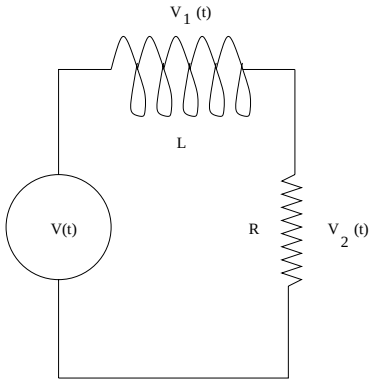


Figure 11.1: An RL-circuit.

Thermodynamic considerations show that the expected power through the circuit, due to the thermal fluctuations of the electrons is

$$E[\frac{LI^2}{2}] = \frac{kT}{2}, \tag{11.33}$$

where T is the absolute temperature and k is Boltzmann's constant. The voltage source $V(t)$ is a white noise process with intensity σ. This means that the spectral density of the noise is constant with

$$S_V(\xi) = \sigma^2.$$

In example 7.1.21 it is shown that the transfer function for the current through this circuit is

$$\hat{h}(\xi) = \frac{1}{R + iL\xi}.$$

Since the input is a random process the output, $I(t)$ is also a random process. According to (11.30) its spectral density function is

$$S_I(\xi) = \sigma^2 \frac{1}{R^2 + (L\xi)^2}.$$

This allows the computation of $E[I^2]$,

$$E[I^2] = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\sigma^2 d\xi}{R^2 + (L\xi)^2} = \frac{\sigma^2}{2RL}.$$

Comparing this result with the (11.33) gives the intensity of the white noise process:

$$\sigma^2 = 2RkT.$$

This result and its generalizations are also known as Nyquist's theorem. At room temperature (about 300°K) with a resistance $R = 10^6$ Ohms, the intensity of the Johnson noise process is

$$\sigma^2 \approx 4 \times 10^{-15} (\text{volt})^2 \text{sec}.$$

Of course, in a real physical system the spectrum of the thermal noise cannot be flat, for this would imply that the noise process contains an infinite amount of energy. It is an empirical fact that the spectrum is essentially flat up to a fairly high frequency. Indeed Johnson noise is sometimes describes as a random process, χ with

$$S_\chi(\xi) = \sigma^2 \chi_{[0,B]}(|\xi|),$$

or briefly, as *bandlimited white noise*. The integral above from $-\infty$ to ∞ is then replaced by an integral from $-B$ to B. If B is reasonably large then the result is nearly the same. The total power of the (bandlimited) Johnson noise is therefore

$$S_{\text{tot}} = \frac{RkTB}{\pi}.$$

In many practical applications, the spectrum of the noise is bandlimited because the data itself is bandlimited. The formula for S_{tot} shows that any attempt to increase the bandwidth of the data increases the total power of the Johnson noise commensurately.

Remark 11.4.1. Our treatment of Johnson noise is adapted from [84].

11.4.3 Optimal filters

As a final application of these ideas, we consider the design of a noise reducing filter which is "optimal" in some sense. Let the signal x be modeled as

$$x(t) = s(t) + n(t)$$

where s stands for the signal and n, the noise. Both s and n are assumed to be weak sense stationary, finite variance, random processes; which are not correlated:

$$E[s(t_1)n(t_2)] = 0 \text{ for all } t_1 \text{ and } t_2. \tag{11.34}$$

We would like to design a filter H which minimizes the error in the detected signal in the sense that expected mean square error, $E[|s - Hx|^2(t)]$ is minimized. We look for the optimal filter among linear, shift invariant filters. In this case both s and Hx are weak sense stationary processes and therefore the value of the error is independent of t.

The solution of the minimization problem is characterized by an orthogonality condition,

$$E[(Hx - s)(t_1)x(t_2)] = 0, \text{ for all } t_1, t_2. \tag{11.35}$$

We give a formal derivation of this condition. Suppose that h is the impulse response of an optimal filter and that k is an "arbitrary" impulse response. The optimality condition is that

$$\frac{d}{d\lambda} E[|(s - (h + \lambda k) * x)(t_1)|^2]\big|_{\lambda=0} = 0.$$

Expanding the square and differentiating in t gives

$$E[((h * x - s)k * x)(t_1)] = 0 \text{ for any } k \text{ and } t_1. \tag{11.36}$$

Given that the various convolutions make sense, the derivation up to this point has been fairly rigorous. Choose a smooth, non-negative function $\varphi(t)$ with bounded support and total integral 1. For any t_2, taking

$$k_\epsilon(t) = \frac{1}{\epsilon}\varphi\left(\frac{t-(t_1-t_2)}{\epsilon}\right)$$

gives a sequence of very smooth test functions which "converge" to $\delta(t-(t_1-t_2))$. Assuming that the limit makes sense, (11.36) implies that

$$0 = \lim_{\epsilon \downarrow 0} E[((h * x - s)k_\epsilon * x)(t_1)] = E[(h * x - s)(t_1)x(t_2)], \tag{11.37}$$

which is the desired orthogonality condition. By using finite sums to approximate $k * x$ the condition in (11.36) is easily deduced from (11.35).

Using (11.34), the orthogonality condition can be rewritten in terms of s and n as

$$\begin{aligned} 0 &= E[(h * s(t_1) + h * n(t_1) - s(t_1))(s(t_2) + n(t_2))] \\ &= E[h * s(t_1)s(t_2) + h * n(t_1)n(t_2) - s(t_1)s(t_2)] \\ &= \int_{-\infty}^{\infty} E[s(\tau)s(t_2)]h(t_1-\tau)d\tau + \int_{-\infty}^{\infty} E[n(\tau)n(t_2)]h(t_1-\tau)d\tau - E[s(t_1)s(t_2)]. \end{aligned}$$

Let r_s and r_n be the autocorrelation functions for the signal and noise, respectively. Letting $t = \tau - t_2$ and $\sigma = t_1 - t_2$ gives

$$\int_{-\infty}^{\infty} r_s(t)h(\sigma - t)dt + \int_{-\infty}^{\infty} r_n(t)h(\sigma - t)dt = r_s(\sigma).$$

This is a convolution equation, so taking the Fourier transform gives the relation:

$$S_s(\xi)\hat{h}(\xi) + S_n(\xi)\hat{h}(\xi) = S_s(\xi).$$

Recalling that the spectral density function is non-negative we divide, obtaining the transfer function for the optimal filter

$$(\xi) = \frac{S_s(\xi)}{S_s(\xi) + S_n(\xi)} = \frac{1}{1 + S_n(\xi)/S_s(\xi)} \approx \begin{cases} 1 & S_n(\xi)/S_s(\xi) << 1, \\ 0 & S_n(\xi)/S_s(\xi) >> 1. \end{cases}$$

This shows how one can use the power spectrum of the noise and a probabilistic description of the signal to design an optimal filter. This example is called the *Wiener filter*, it is a very simple example of an optimal filter. There are many variants on this approach using different classes of filters and different kinds of random processes. Kalman and Bucy found a different approach to the problem of optimal filtering. More complete treatments of this subject can be found in [13] or [8].

Exercise 11.4.4. Prove that if H defines the optimal filter then

$$E[|Hx|^2] = E[sHx]. \tag{11.38}$$

Exercise 11.4.5. Using (11.38), compute the expected mean squared error for the optimal filter, H

$$E[|Hx - s|^2] = r_s(0) - \int_{-\infty}^{\infty} h(t) r_s(t) dt. \tag{11.39}$$

Exercise 11.4.6. Using the Parseval formula and (11.39) prove that

$$E[|Hx - s|^2] = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{S_s(\xi) S_n(\xi) d\xi}{S_s(\xi) + S_n(\xi)}.$$

Exercise 11.4.7. If the signal and noise have non-zero cross-correlation of the form

$$r_{sn}(\tau) = E[s(t)n(t+\tau)]$$

show that the orthogonality condition for the optimal filter becomes

$$\int_{-\infty}^{\infty} r_s(t) h(\sigma - t) dt + \int_{-\infty}^{\infty} r_n(t) h(\sigma - t) dt = r_s(\sigma) + r_{sn}(\sigma).$$

Find the transfer function for the optimal filter in this case.

Chapter 12

Resolution and noise in the filtered backprojection algorithm

In Chapter 8 we determined the point spread function of the measurement and reconstruction process for a parallel beam scanner. It is shown in examples that if there is at least one sample per beam width (as defined by the beam profile function) then the resolution in the reconstructed image is determined by the beam width. Using a rectangular profile, the resolution is about half the beam width. A second conclusion of that analysis is that the effects of aliasing, resulting from ray sampling are well controlled by using the Shepp-Logan filter and a Gaussian focal spot. Decreased sample spacing sharpens the peak of the PSF and does not produce oscillatory side lobes. Finally, the effect of view sampling is an oscillatory artifact, appearing at a definite distance from a hard object. The distance is proportional to $\Delta\theta^{-1}$. This analysis, and considerable empirical evidence shows that, by decreasing $\Delta\theta$ one can obtain an artifact free region of any desired size.

In this chapter we consider how noise in the measurements obtained in a CT scanner propagates through a reconstruction algorithm. The principal source of noise in CT-imaging is *quantum noise*. This is a consequence of the fact that X-rays "beams" are really composed of discrete photons, whose number fluctuates in a random way. An X-ray source is usually modeled as Poisson random process, with the intensity equal to the expected number of photons per unit time. Recall that the signal-to-noise ratio in a Poisson random variable, χ equals $\sqrt{E[\chi]}$. Because each photon carries a lot of energy, safety considerations severely limit the intensity of the X-ray source. This is why quantum noise is an essentially unavoidable problem in X-ray imaging. Matters are further complicated by the fact that the X-ray photons have different energies, but we do not consider this effect, see [4].

In most of this section we assume that the absorption and detection processes are Bernoulli. As shown in section 10.3.7 (especially exercise 10.3.10) the complete system of X-ray production, absorption and detection is modeled as a Poisson process. In X-ray CT, the actual measurement is the number of photons emerging from an object along a finite collection of rays, $N_{\text{out}}(t_j, \omega(k\Delta\theta))$. This number is compared to the number of photons which entered, $N_{\text{in}}(t_j, \omega(k\Delta\theta))$. Both of these numbers are Poisson random variables and Beer's law is the statement that

$$\mathcal{E}[N_{\text{out}}(t_j, \omega(k\Delta\theta))] = E[N_{\text{in}}] \exp\left[-\mathrm{R}_W f(t_j, \omega(k\Delta\theta))\right].$$

Here f is the absorption coefficient of the object and $\mathrm{R}_W f$ is the Radon transform averaged with the beam profile. The *signal-to-noise ratio* provides a measure of the useful information in the reconstructed image. If χ is a random variable then recall that the signal-to-noise ratio is defined by

$$\mathrm{SNR} = \frac{E[\chi]}{\sqrt{E[(\chi - E[\chi])^2]}}. \tag{12.1}$$

Either the totality of the reconstructed image or its value at each point can be regarded as a random variable. With either model we examine how uncertainty in the measurement produces uncertainty in the reconstructed image, or briefly the effect of the reconstruction algorithm on measurement noise.

12.1 The continuous case

Let $f(x, y)$ denote a function supported in the disk of radius L, representing the function we would like to reconstruct. We begin our analysis of noise in the filtered backprojection algorithm assuming that $\mathrm{R}f(t, \omega)$ can be measured for all $(t, \omega) \in [-L, L] \times S^1$ and that f is approximately reconstructed using filtered backprojection,

$$f_\phi(x, y) = \frac{1}{2\pi} \int_0^\pi \int_{-L}^{L} \mathrm{R}f(\langle (x, y), \omega \rangle - s, \omega)\phi(s) ds d\omega. \tag{12.2}$$

To simplify the notation in this section we omit explicit reference to the beam width function. The results in section 8.6.1 show that this does not reduce the generality of our results.

The uncertainty in the measurements can be modeled in two different ways. On the one had we can imagine that f itself is corrupted by noise, so that the measurements are of the form $\mathrm{R}(f + n_i)$. Here n_i is a random process, represented by functions on $\mathbb{R}^2$, which models the uncertainty in the input, f. On the other hand f can be considered to be determinate but the measurements themselves are corrupted by noise. In this case the measurements are modeled as $\mathrm{R}f + n_m$. Here n_m is a random process, represented by functions on $\mathbb{R} \times S^1$, which models uncertainty in the measurements. Of course, the real situation involves a combination of these effects. We analyze these sources of error assuming that f itself is zero.

The first case is very easy to analyze as we can simply the results in section 11.4. The map from f to f_ϕ is a shift invariant linear filter with MTF given by

$$\hat{\Psi}(\xi) = \hat{\psi}(\|\xi\|),$$

where

$$\hat{\phi}(r) = |r|\hat{\psi}(r).$$

Assume that n_i is a stationary random process with mean zero for each (x, y) is $\mathbb{R}^2$. Denote the autocorrelation function by

$$r_i(x, y) = E[n_i(0, 0)n_i(x, y)].$$

Its Fourier transform $S_i(\xi)$ is the power spectral density in the input noise process. The power spectral density of the output is given by (11.30),

$$S_o(\xi) = S_i(\xi)|\hat{\Psi}(\xi)|^2.$$

The total noise power in the output is therefore

$$S_{\text{tot}} = \frac{1}{[2\pi]^2} \int_0^\infty S_i(r\omega)|\hat{\psi}(r)|^2 r dr d\omega. \tag{12.3}$$

A useful, though not too realistic example, is to assume that n_i is a white noise process with $S_i(\xi) = \sigma^2$. If ϕ is the Shepp-Logan filter with

$$\hat{\phi}(r) = |r| \left| \text{sinc}\left(\frac{dr}{2}\right)\right|^3$$

then the total noise in the output, which equals the variance $r_o(0)$ is

$$S_{\text{tot}} = C\frac{\sigma^2}{d^2}, \tag{12.4}$$

here C is a positive constant. The reconstruction algorithm amplifies the uncertainty in f by a factor proportional to d^{-2}.

The other possibility is that the noise is measurement noise. In this case n_m is a function on $\mathbb{R} \times S^1$. Using an angular coordinate we can think of n_m as a function of (t, θ) which is 2π-periodic. This noise process is then weak sense stationary in that

$$E[n_m(t_1, \theta_1)n_m(t_2, \theta_2)] = r_m(t_1 - t_2, \theta_1 - \theta_2),$$

where $r_m(\tau, \theta)$ is also 2π-periodic in θ. The backprojection algorithm applied to the noise gives

$$n_{m\phi}(x, y) = \frac{1}{4\pi} \int_0^\pi \int_{-\infty}^\infty n_m(\langle (x, y), \omega\rangle - s, \omega)\phi(s) ds d\omega.$$

For convenience, we have replaced the finite limits of integration with infinite limits. Because the noise is bounded and the Shepp-Logan filter is absolutely integrable this does not significantly affect the outcome.

The auto correlation of the noise in the output is

$$E[n_{m\phi}(x, y)n_{m\phi}(0, 0)] = \frac{1}{[2\pi]^2} \int_0^\pi \int_0^\pi \int_{-\infty}^\infty \int_{-\infty}^\infty r_m(\langle (x, y), \omega(\theta_1)\rangle + s_2 - s_1, \theta_1 - \theta_2)\phi(s_1)\phi(s_2) ds_1 ds_2 d\theta_1 d\theta_2. \tag{12.5}$$

Without further information this expression is very difficult to evaluate. We make the hypothesis that the measurement noise is white, i.e. the errors in one ray are uncorrelated with the errors in another. This means that

$$r_m(\tau, \theta) = \sigma^2 \delta(\tau)\delta(\theta),$$

where, strictly speaking θ should be understood in this formula as $\theta \mod 2\pi$. That the errors from ray to ray are weakly correlated is not an unreasonable hypothesis, however the analysis in section 10.3.7, particularly equation (10.44), shows that the variance is unlikely to be constant. These assumptions give

$$E[n_{m\phi}(x,y)n_{m\phi}(0,0)] = \frac{\sigma^2}{[2\pi]^2}\int_0^{\pi}\int_{-\infty}^{\infty}\phi(s_1)\phi(s_1 - \langle(x,y),\omega(\theta_1)\rangle)ds_1 d\theta_1 \tag{12.6}$$

Because ϕ is an even function and $\hat{\phi}(r) = 0$ this can be re-expressed as a 2-dimensional inverse Fourier transform

$$E[n_{m\phi}(x,y)n_{m\phi}(0,0)] = \frac{\sigma^2}{[2\pi][4\pi]^2}\int_0^{\pi}\int_{-\infty}^{\infty}\frac{|\hat{\phi}(r)|^2}{|r|}e^{ir\langle(x,y),\omega\rangle}|r|drd\omega. \tag{12.7}$$

The power spectral density in the output is therefore

$$S_o(\xi) = \frac{\sigma^2}{8\pi}\frac{|\hat{\phi}(r)|^2}{|r|}.$$

Using the same filter as above, the total noise power in the output is

$$S_{\text{tot}} = C'\frac{\sigma^2}{d^3}, \tag{12.8}$$

where again C' is a positive constant. The total noise power in the measurements is amplified by a factor proportional to d^{-3}. Recalling that the resolution is proportional to d, it follows that, as the resolution increases, errors in measurement have a much greater affect on the reconstructed image than uncertainty in f itself. In either case the noise is assumed to have mean zero so a non-zero f would only change the variance computations by a bounded function of d. Note finally that with either sort of noise, the variance tends to infinity as d goes to zero. This substantiates our claim that noise necessitates the use of regularization in the reconstruction process. This discussion is adapted, in part from [33].

Exercise 12.1.1. Repeat the computations in this section with a non-zero input f.

12.2 A simple model with sampled data

In this section we consider sampled data with uncertainty in the measurements. The variance in the value of the reconstruction of each pixel is estimated in terms of the variance of the noise. Let $\{P(t_k,\theta_j)\}$ denote approximate samples of the Radon transform of an absorption coefficient $f(x,y)$. In this section the spatial coordinates are normalized so that f is supported in $[-1,1]\times[-1,1]$. The Radon transform is sampled at $M+1$ equally spaced angles,

$$\{j\Delta\theta \,:\, j = 0,\ldots,M\} \text{ with } \Delta\theta = \frac{\pi}{M+1}.$$

The sample spacing in the affine parameter is denoted by d. Given the normalization of the spatial coordinates

$$N = \frac{2}{d}$$

is the number of samples in the t-direction. With ϕ, a choice of filter function, the filtered backprojection formula gives

$$\tilde{f}_\phi(x,y) = \frac{d}{2(M+1)} \sum_{j=0}^{M} \sum_{k=-\infty}^{\infty} P(t_k, \theta_j) \phi(\langle (x,y), \omega(j\Delta\theta) - t_k). \tag{12.9}$$

This is a Riemann sum approximation for the formula, (12.2) used in the previous section. The basic constraint on the filter function ϕ is that

$$\hat{\phi}(\xi) \approx |\xi| \text{ for } |\xi| < \Omega \tag{12.10}$$

where Ω represents the effective bandwidth of the measured data.

The measurement is modeled as the "true value" plus an additive noise process, $\{\eta_{kj}\}$,

$$Q_{kj} = P(t_k, \omega(j\Delta\theta)) + \sigma\eta_{kj}.$$

The statistical assumptions made on the noise are

$$E[\eta_{kj}] = 0, \quad E[\eta_{kj}\eta_{lm}] = \delta_{kl}\delta jm. \tag{12.11}$$

The condition on the mean implies that there are no systematic errors in the measurements. The second assumption asserts that the errors made in each measurement are uncorrelated. Again the variance is assumed to be constant, which, as remarked above, is not a realistic assumption. The mean and variance of the individual measurements are

$$E[Q_{kj}] = P(t_k, \omega(j\Delta\theta)), \text{ and } E[(Q_{kj} - < Q_{kj} >)^2] = \sigma^2.$$

Given these measurements, the reconstructed image is

$$\breve{f}_\phi(x,y) = \frac{d}{2(M+1)} \sum_{j=0}^{M} \sum_{k=-\infty}^{\infty} Q_{kj} \phi(\langle (x,y), \omega(j\Delta\theta)\rangle - t_k). \tag{12.12}$$

Since $E[Q_{kj}] = P(t_k, \omega(j\Delta\theta))$, the expected value of the output is

$$E[\breve{f}_\phi(x,y)] = \frac{d}{2(M+1)} \sum_{j=0}^{M} \sum_{k=-\infty}^{\infty} E[Q_{kj}] \phi(\langle (x,y), \omega(j\Delta\theta)\rangle - t_k) = \tilde{f}_\phi(x,y).$$

The variance at each point (x,y) is

$$E[(\breve{f}_\phi(x,y) - \tilde{f}_\phi(x,y))^2] = E[\breve{f}_\phi^2(x,y)] - \tilde{f}_\phi^2(x,y).$$

Remark 12.2.1 *(***Important notational remark)**. In the remainder of this chapter, the notation $\breve{f}_\phi$ refers to a reconstruction using noisy data, as in (12.12). This allows us to distinguish, such approximate reconstructions from the approximate reconstruction, $\tilde{f}_\phi$ made with "exact" data.

Expanding the square in the reconstruction formula gives

$$\begin{aligned}
\breve{f}_\phi^2 &= (\frac{d}{2(M+1)})^2\left[\sum_{j=0}^{M}\sum_{k=-\infty}^{\infty}(P(t_k,\omega(j\Delta\theta))+\sigma\eta_{kj})\phi(\langle(x,y),\omega(j\Delta\theta)\rangle - t_k)\right]^2 \\
&= (\frac{d}{2(M+1)})^2\left[\sum_{j=0}^{M}\sum_{k=-\infty}^{\infty}P(t_k,\omega(j\Delta\theta))\phi(\langle(x,y),\omega(j\Delta\theta)\rangle - t_k)\right]^2 \\
&+ 2\sigma(\frac{d}{2(M+1)})^2\sum_{j,k}\sum_{l,m}P(t_k,\omega(j\Delta\theta))\phi(\langle(x,y),\omega(j\Delta\theta)\rangle - t_k)\phi(\langle(x,y,\omega(m\Delta\theta)\rangle - t_l)\eta_{l,m} \\
&+ \sigma^2(\frac{d}{2(M+1)})^2\sum_{j,k}\sum_{l,m}\phi(\langle(x,y),\omega(j\Delta\theta)\rangle - t_k)\phi(\langle(x,y),\omega(m\Delta\theta)\rangle - t_l)\eta_{j,k}\eta_{l,m}
\end{aligned}$$

Using the hypothesis that the noise for different measurements is uncorrelated leads to

$$E[\breve{f}_\phi^2] = \tilde{f}_\phi^2 + \sigma^2\left[\frac{d}{2(M+1)}\right]^2\sum_{j=0}^{M}\sum_{k=-\infty}^{\infty}\phi^2(\langle(x,y),\omega(j\Delta\theta)\rangle - t_k), \tag{12.13}$$

and therefore

$$\sigma_\phi^2 = \sigma^2\left[\frac{d}{2(M+1)}\right]^2\sum_{j=0}^{M}\sum_{k=-\infty}^{\infty}\phi^2(\langle(x,y),\omega(j\Delta\theta)\rangle - t_k). \tag{12.14}$$

For each fixed j, the sum on k is an approximation to the integral of ϕ^2,

$$d\sum_{k=-\infty}^{\infty}\phi^2(\langle(x,y),\omega(j\Delta\theta)\rangle - t_k) \approx \int_{-\infty}^{\infty}\phi^2(t)dt.$$

Therefore,

$$\sigma_\phi^2 \approx \sigma^2\frac{d}{4(M+1)}\int_{-\infty}^{\infty}\phi^2(t)dt = \frac{\sigma^2}{2N(M+1)}\int_{-\infty}^{\infty}\phi^2(t)dt. \tag{12.15}$$

The noise variance *per pixel*, is therefore proportional to the integral of the square of the filter function. Note that this variance is independent of the point in image. This is a consequence of assuming that the variance in the number of measured photons is constant.

Note that Parseval's theorem says that

$$\int_{-\infty}^{\infty}\phi^2(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty}|\hat{\phi}(\xi)|^2d\xi.$$

If ϕ satisfies (12.10), then

$$\int_{-\infty}^{\infty} \phi^2(t)dt \approx \frac{1}{\pi}\int_0^{\Omega} \xi^2 d\xi = \frac{\Omega^3}{3\pi}.$$

The ratio σ_ϕ/σ is called the *noise amplification factor*. Its square is given approximately by:

$$\frac{\sigma_\phi^2}{\sigma^2} \approx \frac{\Omega^3}{6\pi N(M+1)}.$$

From Nyquist's theorem, $\Omega \approx N/2$, so we see that

$$\frac{\sigma_\phi^2}{\sigma^2} \approx \frac{N^2}{48\pi(M+1)}.$$

In order to have the resolution in the angular direction equal that in the radial direction we need to take $(M+1) \approx 2\pi d^{-1}$ and therefore:

$$\frac{\sigma_\phi^2}{\sigma^2} \approx \frac{N}{48\pi} = \frac{1}{24\pi^2 d}.$$

This is an estimate of the noise variance for a single pixel. As the number of pixels in the reconstruction grids is $O(d^{-2})$ this result agrees with equation (12.8). This discussion is adapted from [70].

12.3 A computation of the variance

In the previous section we considered the effect of an additive noise process, where

$$Q_{kj} = P(t_i, \omega(j\Delta\theta)) + \sigma\eta_{jk},$$

assuming that the variance in all the measurements are same. This is not a reasonable assumption because the variance in the number of photons counted is proportional to the number of measured photons. This is true whether the number of detected photons is modeled as a Bernoulli process (deterministic source) or a Poisson process (Poisson source). These numbers can vary quite a lot due to difference thicknesses and absorbencies encountered along different rays. Using the same geometry as in the previous calculation - a parallel beam scanner with sample spacing d for the affine parameter, we derive an estimate for $\mathrm{Var}(Q_{kj})$ from the assumption that the number of photons counted is a Poisson random variable. The computation of the variance is complicated by the fact that the input to the reconstruction algorithm is not the *number* of measured photons, but rather

$$\log\left(\frac{N_{\text{in}}}{N_{\text{out}}}\right).$$

The non-linearity of the log renders the estimation of the variance in Q_{kj} a non-trivial calculation.

Let θ denote the direction, $\omega(\theta)$ and $N_\theta(kd)$ the number of photons measured for the ray $l_{kd,\omega(\theta)}$. Let ideal result would give

$$P_\theta(kd) = \int\limits_{l_{kd,\omega(\theta)}} f ds.$$

For each ray the number of measured photons is a Poisson random variable. Let $\bar{N}_\theta(kd)$ denote the expected value $E[N_\theta(kd)]$. For simplicity we assume that, N_{in} the number of incident photons in each beam is a deterministic fixed, large number. Beer's law is the statement that

$$\bar{N}_\theta(kd) = N_{\text{in}} e^{-P_\theta(kd)}.$$

Because $N_\theta(kd)$ is a Poisson random variable its probability distribution is determined by its expected value,

$$\text{Prob}(N_\theta(kd) = l) = \frac{[\bar{N}_\theta(kd)]^l e^{-\bar{N}_\theta(kd)}}{l!}, \tag{12.16}$$

its variance is

$$\text{Var}(N_\theta(kd)) = \bar{N}_\theta(kd). \tag{12.17}$$

The SNR of the individual measurements is therefore

$$\frac{E[N_\theta(kd)]}{\sigma_{N_\theta(kd)}} = \frac{1}{\sqrt{\bar{N}_\theta(kd)}}.$$

This is characteristic of Poisson random variables: the signal-to-noise ratio is inversely proportional the square root of the expected value.

12.3.1 The variance of the Radon transform

Let $P^m_\theta(kd)$ denote the measured value of $P_\theta(kd)$

$$P^m_\theta(kd) = \log \frac{N_{\text{in}}}{N_\theta(kd)}.$$

The expected value of the measurement is given by

$$E[P^m_\theta(kd)] = E[\log N_{\text{in}} - \log N_\theta(kd)] = E[N_{\text{in}}] - E[\log N_\theta(kd)] \tag{12.18}$$

where

$$E[\log N_\theta(kd)] = \sum_{l=0}^{\infty} \frac{\text{Ln}(l)[\bar{N}_\theta(kd)]^l e^{-\bar{N}_\theta(kd)}}{l!}.$$

Because $\log 0$ is infinity, we define $\text{Ln}(0) = 0$ in this summation. Unfortunately there is no simple closed form for this expression. Since the logarithm is not a linear function,

$$E[\log N_\theta(kd)] \neq \log E[N_\theta(kd)].$$

Using Taylor's formula we derive an expression for the difference,

$$E[\log N_\theta(kd)] - \log E[N_\theta(kd)].$$

Let y be a non-negative random variable with density function $p(y)$, for which

$$\bar{y} = \int_{-\infty}^{\infty} yp(y)dy \text{ and } \sigma^2 = \int_{-\infty}^{\infty} (y-\bar{y})^2 p(y)dy.$$

Assuming that $\bar{y}$ is a large number and that p is sharply peaked around its mean,

$$\begin{aligned} E[\log y] &= \int_0^\infty (\log y)p(y)dy \\ &= \int_{-\bar{y}}^\infty \log(x+\bar{y})p(x+\bar{y})dx \\ &= \int_{-\bar{y}}^\infty \left[\log \bar{y} + \log(1+\frac{x}{\bar{y}})\right] p(x+\bar{y})dx \\ &\approx \log \bar{y} + \int_{-\bar{y}}^{\bar{y}} \left[\frac{x}{\bar{y}} - \frac{1}{2}(\frac{x}{\bar{y}})^2 + \cdots\right] p(x+\bar{y})dx \\ &\approx \log \bar{y} - \frac{1}{2\bar{y}^2}\sigma^2 \end{aligned} \tag{12.19}$$

To apply this computation we approximate the distribution function for a Poisson random variable, which is in fact a sum of δ-functions by a smooth Gaussian distribution. As shown in section 10.3.3 the distribution of a Poisson random variable with intensity $\lambda >> 1$ is well approximated by

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

For large λ, the standard deviation, $\sqrt{\lambda}$ is much smaller than λ. So long as λ is large the approximation of $E[\log y]$ in (12.19) is applicable.

Using our assumption that $\bar{N}_\theta(kd)$ is a large number and that

$$\sigma^2 = \bar{N}_\theta(kd),$$

the foregoing computation gives the estimate

$$E[\log N_\theta(kd)] \approx \log E[N_\theta(kd)] - \frac{\sigma^2}{2E[N_\theta(kd)]} = \log \bar{N}_\theta(kd) - \frac{1}{2\bar{N}_\theta(kd)}.$$

Equation (12.18) with thus above approximation gives

$$E[P_\theta^m(kd)] \approx \log\left(\frac{N_{\text{in}}}{\bar{N}_\theta(kd)}\right) = P_\theta(kd). \tag{12.20}$$

The variance is

$$\begin{aligned}\mathrm{Var}(P_\theta^m(kd)) &= E[(P_\theta^m(kd) - P_\theta(kd))^2] \\ &= E[(\log\left(\frac{N_\theta(kd)}{N_{\text{in}}}\right) - \log\left(\frac{\bar{N}_\theta(kd)}{N_{\text{in}}}\right))^2] \\ &= E[(\log\left(\frac{N_\theta(kd)}{\bar{N}_\theta(kd)}\right))^2].\end{aligned} \tag{12.21}$$

Assuming that the X-ray source is deterministic, the variance in the measurements is independent of the source intensity The variance is therefore given by

$$\mathrm{Var}(P_\theta^m(kd)) \approx \int_{-\bar{y}}^{\bar{y}} \left[\log\left(1 + \frac{x}{\bar{y}}\right)\right]^2 dx,$$

which is easily seen to give

$$\mathrm{Var}(P_\theta^m(kd)) \approx \frac{1}{\bar{N}_\theta(kd)}.$$

This verifies the claim that the variance in a measurement of $\mathrm{R}f$ is inversely proportional to the number of photons measured. This computation assumes that the number of incident photons N_{in} is a fixed number.

Exercise 12.3.1. Compute the variance in $P_\theta^m(kd)$ assuming that the source is also a Poisson random variable with intensity N_{in}.

12.3.2 The variance in the reconstructed image

Assuming that the measurement errors in different rays are uncorrelated, we now find a more accurate computation for the variance in the reconstructed image. The reconstructed image is given by (12.9) with the actual measurements in place of the ideal values

$$\breve{f}_\phi = \frac{\pi d}{(M+1)} \sum_{j=0}^{M} \sum_k P_{\theta_i}^m(kd)\phi(\langle(x,y),\omega(j\Delta\theta)\rangle - kd).$$

As the errors have mean zero, the linearity of the reconstruction formula implies that

$$E[\breve{f}_\phi(x,y)] = \tilde{f}_\phi(x,y).$$

The variance of the reconstructed image is given by

$$\mathrm{Var}(\breve{f}_\phi(x,y)) = \left(\frac{\pi d}{(M+1)}\right)^2 \sum_{j=0}^{M} \sum_k \frac{1}{\bar{N}_{\theta_i}(kd)} \phi^2(\langle(x,y),\omega(j\Delta\theta)\rangle - kd).$$

This is quite similar to what was obtained before with the small modification that the contribution to the variance of each projection is weighted according to the expected number of measured photons, $1/\bar{N}_{\theta_i}$. The thicker parts of the object contribute more to the variance.

Using the formula

$$\bar{N}_{\theta_i}(kd) = N_{in}e^{-P_\theta(kd)},$$

the variance can be rewritten

$$\mathrm{Var}(\breve{f}_\phi(x,y)) = \big(\frac{\pi d}{(M+1)}\big)^2 \frac{1}{N_{in}} \sum_{j=0}^{M} \sum_{k} e^{P_{\theta_j}(kd)} \phi^2(\langle (x,y), \omega(j\Delta\theta)\rangle - kd).$$

At the center of the image the variance is

$$\mathrm{Var}(\breve{f}_\phi(0,0)) \approx \left(\frac{\pi d}{(M+1)}\right)^2 \sum_{k} \sum_{j=0}^{M} \frac{\phi^2(-kd)}{\bar{N}_{\theta_i}(kd)}.$$

Assuming that the object is radially symmetric and of constant absorption coefficient m implies that

$$\bar{N}_{\theta_j}(kd) = N_{\mathrm{in}} e^{-2m\sqrt{1-(kd)^2}}$$

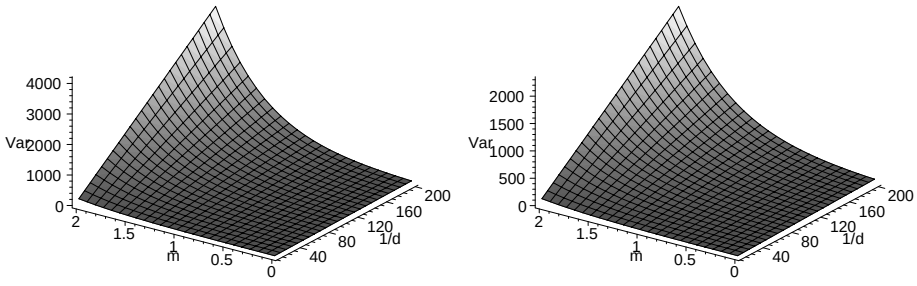
for all j. This gives

$$\mathrm{Var}(\breve{f}_\phi(0,0)) \approx \frac{(\pi d)^2}{(M+1)N_{\mathrm{in}}} \sum_{k} \phi^2(kd) e^{2m\sqrt{1-(kd)^2}} \approx \frac{\pi^2 d}{(M+1)N_{\mathrm{in}}} \int_{-1}^{1} \phi^2(t) e^{2m\sqrt{1-t^2}} dt. \tag{12.22}$$

The graphs in figure 12.1 show $\mathrm{Var}(\breve{f}_\phi(0,0))$ as functions of m and $1/d$. In figure 12.1(a) formula 12.22 is used, while in figure 12.1(b) formula 12.15 is used with

$$\sigma^2 = \int_0^1 e^{2m\sqrt{1-t^2}} dt,$$

the average. Note that for larger m the constant variance estimate is much smaller.



(a) Image variance from 12.22 as a function of m and $1/d$.

(b) Image variance from 12.15 as a function of m and $1/d$.

Figure 12.1: Comparison of the image variance using different models for the variance in the measurements.

Exercise 12.3.2. Find a formula for $\mathrm{Var}(\breve{f}_\phi(x,y))$ for other points in the disk. Graph the result, for fixed values of (d,m).

Exercise 12.3.3. Find a formula for $\mathrm{Cov}(\breve{f}_\phi(0,0),\breve{f}_\phi(x,y))$. Graph the result, for fixed values of (d,m).

12.3.3 Signal-to-noise ratio, dosage and contrast

We now compute the signal-to-noise ratio at the center of a homogeneous disk of radius R and absorption coefficient m. From the previous section we have

$$\begin{aligned} \bar{N}_0 &= N_{in}e^{-2mR}, \\ E[\breve{f}_\phi(0,0)] &\approx \tilde{f}_\phi(0,0). \end{aligned}$$

Approximating the integral in (12.22) gives

$$\mathrm{Var}(\tilde{f}_\phi(0,0)) \approx \frac{\pi^2 d}{MN_{in}e^{-2mR}} \int_{-\infty}^{\infty} \phi^2(t)dt.$$

Using the Parseval formula and the assumption that $\hat{\phi}(\xi) \approx \chi_{[0,\Omega]}(|\xi|)|\xi|$ we obtain

$$\mathrm{Var}(\tilde{f}_\phi(0,0)) \approx \frac{\pi d\Omega^3}{6MN_{in}e^{-2mR}}.$$

The resolution $\delta \propto \Omega^{-1}$ hence the signal-to-noise ratio is

$$SNR \propto \sqrt{\delta^2 MN_{in}e^{-mR}}me^{-\frac{1}{2}mR}.$$

Let D denote the dosage of radiation absorbed by the center pixel in units of rads/cm^3. The photon density passing through the center pixel is proportional to $MN_{in}e^{-\mu_0 R}$. Assuming that the pixel size is proportional to the resolution, the number photons absorbed is proportional to $\delta MN_{in}e^{-\mu_0 R}$, If the thickness of the slice is also proportional to the resolution then

$$D \propto \frac{\delta MN_{in}e^{-mR}}{\delta^3}.$$

Using this in the formula for the signal-to-noise ratio leads to

$$\mathrm{SNR} \propto \sqrt{\delta^4 D}me^{-\frac{1}{2}mR}. \tag{12.23}$$

This shows that the signal-to-noise ratio has a very harsh dependence on the thickness and density of the object, i.e. $me^{-\frac{1}{2}mR}$ decays very rapidly as mR increases. It also demonstrates the "fourth power law" relating resolution to dosage: in order to increase the resolution by a factor of 2, i.e. $\delta \to \frac{1}{2}\delta$, keeping the signal-to-noise ratio constant, we need to increase the dosage by a factor of 16!

What is the importance of the signal-to-noise ratio? In a real physical measurement such as that performed in X-ray CT, the measured quantity assumes a definite range of values. In medical applications the absorption coefficient, quoted in Hounsfield units, takes

values between -1000 (air) and 1000 (bone), see table 2.1. The structures of interest are usually soft tissues and these occupy a tiny part of this range, about -50 to 60, or %5. The signal-to-noise ratio in the measurements determines the *numerical resolution* or accuracy in the reconstructed absorption coefficient. In imaging this is called *contrast*. Noise in the measurements interferes with the discrimination of low contrast objects, that is, contiguous objects with very similar absorption coefficients. In clinical applications the image is usually viewed on a monitor and the range of grays or colors available in the display is mapped to a certain part of the dynamic range of the reconstructed absorption coefficient. If, for example the features of interest lie between 0 and 100 Hounsfield units then everything below 0 is mapped to white and everything above 100 to black. If the absorption coefficient is reconstructed with an accuracy of 1/2% then a difference of 10 Hounsfield unit is meaningful and, by scaling the range of displayed values, should be discernible in the output. If on the other hand, the measured values are only accurate to %2 or 40 Hounsfield, then the scaled image will have a mottled appearance and contain little useful information.

The *accuracy* of the reconstruction should not be confused with the *spatial resolution*. In prosaic terms the accuracy is the number of significant digits in the values of the reconstructed absorption coefficient. The reconstructed values approximate spatial averages of the actual absorption coefficient over a pixel, or if the slice thickness is included, voxel of a certain size. The spatial resolution is a function of the dimensions of a voxel. This in turn, is largely determined by the beam width, sample spacing in the affine parameter and FWHM of the reconstruction algorithm. Increasing the resolution is essentially the same thing as decreasing the parameter δ in (12.23). If the dosage is fixed then this leads to a decrease in the SNR and a consequent decrease in in the contrast available in the reconstructed image. Joseph and Stockham give an interesting discussion of the relationship of contrast and resolution in CT images, see [38].

Remark 12.3.1. Our discussion of SNR is adapted from [4].

Appendix A

Background material

In applied subjects mathematics needs to be appreciated in three rather distinct ways: 1. In the abstract context of perfect and complete knowledge generally employed in mathematics itself, 2. In a less abstract context of fully specified, but incompletely known functions, this is the world of mathematical approximation. 3. In a realistic context of partially known functions and noisy, approximate data, this is closer to the real world of measurements. With these different perspectives in mind, we introduce some of the mathematical concepts underlying image reconstruction and signal processing. The bulk of this material is usually presented in undergraduate courses in linear algebra, analysis and functional analysis. Instead of a giving the usual development, which emphasizes mathematical rigor and proof techniques, we present this material from an engineering perspective. Many of the results are proved in exercises and examples are given to illustrate general phenomena. This material is intended to fill in background material and recast familiar material in a more applied framework; it should be referred to as needed.

A.1 Numbers

We begin by discussing numbers, beginning with the abstract concept of numbers and their arithmetic properties. Representations of number are then considered, leading to a comparison, between abstract numbers and the way numbers are actually used in computation.

A.1.1 Integers

Mathematicians think of numbers as a set which has two operations, addition and multiplication, which satisfy certain properties. The mathematical discussion of this subject always begins with the integers. We denote the set of integers by $\mathbb{Z}$ and the set of positive integers (the whole or natural numbers) by $\mathbb{N}$. There are two operations defined on the integers addition, $+$ and multiplication, $\times$. Associated to each of these operations is a special number: For addition that number is 0 it is *defined* by the property

$$n + 0 = 0 + n = n \text{ for every integer } n.$$

For multiplication that number is 1 and it is *defined* by the property

$$n \times 1 = 1 \times n = n \text{ for every integer } n.$$

The important axiomatic properties of addition and multiplication are

Commutative law:

$$n + m = m + n, \quad n \times m = m \times n, \text{ for every } m, n \in \mathbb{Z},$$

Associative law:

$$(n + m) + p = n + (m + p), \quad (n \times m) \times p = n \times (m \times p), \text{ for every } m, n, p \in \mathbb{Z},$$

Distributive law:

$$(m + n) \times p = m \times p + n \times p, \text{ for every } m, n, p \in \mathbb{Z}.$$

These rules are familiar from grade school and we use them all the time when we do computations by hand.

In mathematics numbers are treated in an axiomatic way. Neither a *representation* of numbers nor an *algorithm* to perform addition and multiplication has yet to be considered. We normally use the decimal representation, when working with numbers "by hand." To define a representation of numbers we first require some special symbols; for the decimal representation we use the symbols $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$ which represent the numbers zero through nine. We also introduce an additional symbol $-$ to indicate that a number is smaller than zero. The decimal representation of an integer is a string of numbers

$$a_m a_{m-1} \dots a_1 a_0 \text{ where } 0 \leq a_j \leq 9, \text{ for } j = 0, \dots, m.$$

What does this string of numbers mean? By definition

$$a_m a_{m-1} \dots a_1 a_0 = \sum_{j=0}^{m} a_j 10^j.$$

What appears on the right hand side of this formula is a mathematical number, what appears on the left is its decimal or base 10 representation. A negative number is represented by prepending the minus sign $-a_m \dots a_0$. For each positive integer $k > 1$ there is an analogous representation for integers called the base-k or k-ary expansion.

The usual algorithms for adding and multiplying numbers revolve around the decimal representation. To do addition we need to *know* how to do the sums $a + b$ for $0 \leq a, b \leq 9$, then we use "carrying" to add larger numbers. To do multiplication we need to *know* how to do the products $a \times b$ for $0 \leq a, b \leq 9$. The algorithms for these operations require an addition and multiplication table.

The normal human mind has no difficulty remembering these base 10 addition and multiplication tables. Especially in the early days, this was a large burden to place on a machine. It was found to be much easier to build a machine that uses a base 2 or binary representation to store and manipulate numbers. In a binary representation an integer is represented as a string of zeros and ones. By definition

$$b_m b_{m-1} \dots b_1 b_0 = \sum_{j=0}^{m} b_j 2^j \text{ where } b_j \in \{0, 1\} \text{ for } j = 0, \dots, m.$$

The analogous algorithms for adding and multiplying in base 2 only require a knowledge of $a + b, a \times b$ for $0 \leq a, b \leq 1$. That is a lot less to remember. On the other hand you need to do a lot more carrying, to add or multiply numbers of a given size.

Even in this very simply example we see that there is a trade off in efficiency of computation between the amount of memory utilized and the number of steps needed to do a certain computation. There is a second reason why binary representations are preferred for machine computation. For a machine to evaluate a binary digit, it only needs to distinguish between two possible states. This is easy to do, even with inexpensive hardware. To evaluate a decimal digit, a machine would need to distinguish between ten different possible states. This would require a much more expensive machine. Finally there is the issue of tradition. It might be cheaper and more efficient to use base 3 for machine computation, but the mere fact that so many base 2 machines already exist make it highly unlikely that we will soon have to start to learn to do arithmetic in base 3.

Because we have a conceptual basis for numbers, there is no limit to size of the numbers we can work with. Could a given number N be the largest number we can "handle?" It would be hard to see why, because if we could handle N then we could certainly $N+1$. In fact this is essentially the mathematical proof that there is no largest integer. The same cannot be said of a normally programmed computer, it has numbers of maximum size with which it can work.

Exercise A.1.1. Write algorithms to do addition and multiplication using the decimal representation of numbers.

Exercise A.1.2. Adding the symbols A, B, C, D, E to represent the decimal numbers $10, 11, 12, 13, 14, 15$ leads to the hexadecimal of base-16 representation of numbers. Work out the relationship between the binary and hexadecimal representations. Write out the addition and multiplication tables in hexadecimal.

A.1.2 Rational numbers

The addition operation also has an inverse operation which we call subtraction: given a number n there is a number $-n$ which has the property $n + (-n) = 0$. We are so used to this that it is difficult to see this as a "property," but note that, if we are only permitted to use integers then the multiplication operation does not have an inverse. This can be thought of in terms of solving equations: any equation of the form

$$x + m = n$$

where $m, n \in \mathbb{Z}$ has an integer solution $x = n - m$. On the other hand, for many choices of $m, n \in \mathbb{Z}$ the equation

$$n \times x = m \tag{A.1}$$

does not have an integer solution.

Again we learned in grade school how to handle this problem: we introduce fractions and then (A.1) has the solution

$$x = \frac{m}{n}.$$

This is just a *symbolic* formula and its meaning is a good deal more subtle than $x = n - m$. First of all if $n = 0$ then it means nothing. If $n \neq 0$ and p is another non-zero integer then the solution of the equation

$$p \times n \times x = p \times m \tag{A.2}$$

is the same as the solution to (A.1). This means that the *number* represented by the symbol $\frac{p\times m}{p\times n}$ is the same as the number represented by the symbol $\frac{m}{n}$. We now introduce rational numbers, $\mathbb{Q}$ as the set of symbols

$$\{\frac{m}{n} \ : \ m,n \in \mathbb{Z}\},$$

with the understanding that

(1). The denominator $n > 0$ and

(2). As numbers

$$\frac{m}{n} = \frac{p}{q}$$

if

$$m \times q = p \times n. \tag{A.3}$$

We have defined the set of rational numbers and now have to define the operations of addition and multiplication on them. Thus far, all we know is how to add and multiply integers. Our definitions for addition and multiplication of rational numbers have to be given in terms of these operations. Multiplication is relatively easy:

$$\frac{m}{n} \times \frac{p}{q} \stackrel{d}{=} \frac{m \times p}{n \times q}.$$

To define addition we use the familiar concept of a "common denominator" and set

$$\frac{m}{n} + \frac{p}{q} = \frac{m \times q + n \times p}{n \times q}. \tag{A.4}$$

The formula only involves operations that we have already defined, though it is not immediately obvious that this is actually an operation on *numbers* and not merely an operation on *symbols*. Now equation (A.1) can be solved for any $m, n \in \mathbb{Z}$ as long as $n \neq 0$. In fact we get a little more for our effort, the equations

$$p \times x = q$$

can be solved for any rational numbers q and $p \neq 0$.

There are two different ways to represent rational numbers: (1) as fractions or (2) as k-ary expansions analogous to those used for integers. Decimal representations of the form

$$a_m \ldots a_0 a_{-1} \ldots a_{-n} \stackrel{d}{=} \sum_{j=-n}^{m} a_j 10^{-j}, \text{ where } 0 \leq a_j \leq 9$$

represent rational numbers. It is easy to see that only fractions of the form

$$\frac{n}{10^k} \text{ for } n,k \in \mathbb{N},$$

have such a finite decimal representation. For some purposes the representation as fractions is more useful, it is certainly more efficient. For example using a fraction we have an exact representation of the number 1/3, using long division we find that

$$\frac{1}{3} = \sum_{j=1}^{\infty} \frac{3}{10^j}.$$

In other words, to *exactly* represent 1/3 as a decimal requires infinitely many decimal places. Thus far we have not even defined infinite sums but from the engineering point of view it is clear what this means.

Because the representation as fractions is not unique and because of the need to find common denominators for addition, fractions are not well adapted to machine computation. In a computer rational numbers are represented as strings of zeros and ones. Such a string of zeros and ones is called a *binary string*. Depending upon the application different numbers can be assigned to a given binary string. The simplest way to assign a number to a binary string with $2N+2$ entries or *bits* is to set

$$a_{N+1}a_N \ldots a_{-N} \stackrel{d}{=} (-1)^{a_{N+1}} 2^N \sum_{j=-N}^{N} a_j 2^j.$$

With this choice, the spacing between consecutive numbers is 1 and the maximum and minimum numbers which can be represented are $\pm(2^{2N+1}-1)$. This allows the representation of large numbers, but sacrifices accuracy. If we knew in advance that all our numbers would lie between -1 and +1 then we could use the same $2N+2$ bits to get more accurate representations for a smaller range of numbers by instead assigning the number

$$(-1)^{a_{N+1}} \frac{1}{2^N} \sum_{j=-N}^{N} a_j 2^j$$

to this binary string. Here the minimum spacing between numbers is 2^{-2N}.

Floating point numbers represents a compromise between these two extremes. The string of binary digits is divided into two parts, an exponent and a fractional part. Writing the string as $b_s e_s e_0 \ldots e_m f_1 \ldots f_n$, with $m+n=2N$, the corresponding number is

$$(-1)^{b_s} 2^{\left[(-1)^{e_s} \sum_{j=0}^{m} e_j 2^j\right]} \sum_{k=1}^{n} f_k 2^{-k}.$$

Using a floating point representation we can represent a much larger range of numbers. If, for example we let $m=n=N$ then with $2N+2$ bits we can represent numbers between $\pm 2^{2^N}$. The accuracy of the representation is *proportional* to the size of the number. For numbers between 2^{k-1} and 2^k the minimum spacing is 2^{k-N}. In applications this is a reasonable choice to make. Suppose a number x is the result of a measurement and its value is determined within Δx. The number Δx is called the *absolute error*, usually it is not a very interesting number. More useful is the ratio $\frac{\Delta x}{x}$ which is called the *relative error*. In a floating point representation the relative accuracy of the representation is

constant throughout the range of representable numbers. On the other hand it places subtle constraints on the kinds of computations that can accurately be done. For examples, subtracting numbers of vastly different sizes does not usually give a meaningful result.

Since we only have finitely many digits, computations done in a computer are essentially never exact. It is therefore very important to use algorithms that are not sensitive to repeatedly making small errors of approximation. In image reconstruction this is an important issue as the number of computations used to reconstruct a single image is usually in the millions. For a thorough discussion of treatment of numbers in machine computation see [78].

Exercise A.1.3. Show that the condition in (A.3) is the correct condition to capture the elementary concept that two fractions represent the same number.

Exercise A.1.4. Show that formula (A.4) defines an operation on rational numbers. That is if $\frac{m}{n} = \frac{m'}{n'}$ and $\frac{p}{q} = \frac{p'}{q'}$ then

$$\frac{m \times q + n \times p}{n \times q} = \frac{m' \times q' + n' \times p'}{n' \times q'}$$

as rational numbers.

Exercise A.1.5. Find the exact binary representation of 1/3.

Exercise A.1.6. What would it mean to represent a number in base 1? What numbers can be represented this way. Find as many problems with base 1 as you can. (Thanks to Dr. Fred Villars for suggesting this question)

Exercise A.1.7. Describe binary algorithms for addition, subtraction, multiplication and division.

Exercise A.1.8. ******* Exercises on floating point numbers.*******

A.1.3 Real numbers

In practice we can never use anything beyond rational numbers; indeed for machine computation we have at most a finite collection of numbers at our disposal. One could take the attitude that there is no point in considering numbers beyond rational numbers. Some people do, but it vastly limits the mathematical tools at our disposal. From a mathematical perspective, the rational numbers are inadequate. For example there is no rational number solving the equation

$$x^2 = 2.$$

In other words there are "holes" in the rational numbers. Calculus relies on the concept of a *continuum,* so it is necessary to fill these holes. It is well beyond the scope of this book to give an axiomatic development for the real numbers. Instead we assume that the real numbers exist and describe the essential difference between the real numbers and the rational numbers: the real numbers are *complete.* To define this concept this we need to define the limit of a sequence of numbers. Recall the absolute value function

$$|x| = \begin{cases} x & \text{for } x \geq 0, \\ -x & \text{for } x < 0. \end{cases}$$

The distance between two numbers x and y is defined to be

$$d(x,y) \stackrel{d}{=} |x-y|.$$

It is easy to see that this has the basic property of a distance, the *triangle inequality*

$$d(x,y) \leq d(x,z) + d(z,y). \tag{A.5}$$

This relation is called the triangle inequality by analogy with the familiar fact from Euclidean geometry: the shortest route between two points is the line segment between them, visiting a third point only makes the trip longer.

Sequences

A sequence of real numbers is an infinite, ordered list of numbers. Frequently the terms of a sequence are labeled or *indexed* by the positive integers $x_1, x_2, x_3, \ldots$ The notation $< x_n >$ refers to a sequence indexed by n. A sequence is *bounded* if there is a number M so that

$$|x_n| \leq M$$

for all choices of the index n. It is *monotone increasing* if $x_n \leq x_{n+1}$ for all n. The definition of limits and the completeness axiom for the real numbers are

Limits:

> If $< x_n >$ is a sequence of real numbers then we say that $< x_n >$ converges to x if the distances, $d(x_n, x)$ can be made arbitrarily small by taking the index sufficiently large. More technically, **given** a positive number $\epsilon > 0$ we can **find** an integer N so that
>
> $$d(x_n, x) < \epsilon \text{ provided that } n > N.$$
>
> In this case we say the "limit of the sequence $< x_n >$ is x" and write
>
> $$\lim_{n\to\infty} x_n = x.$$

Completeness Axiom:

> If $< x_n >$ is a monotone increasing, bounded sequence of real numbers then $< x_n >$ converges to limit, that is there exists an $x \in \mathbb{R}$ such that $\lim_{n\to\infty} x_n = x$.

From the completeness axiom it is easy to show that bounded, monotone decreasing sequences also converge. The completeness axiom is what distinguishes the real numbers from the rational numbers. It is, for example, not difficult to construct a bounded, monotone sequence of rational numbers $< x_n >$ which get closer and closer to $\sqrt{2}$, see exercise A.1.9 That is $d(x_n, \sqrt{2})$ can be made as small as one likes by taking n sufficiently large. But the $\sqrt{2}$ is not a rational number, showing that $< x_n >$ cannot converge to a rational number. The rational numbers are not complete!

Using the completeness axiom it is not difficult to show that every real number has a decimal expansion. That is, given a positive real number x we can find a (possibly infinite) sequence $< a_m, a_{m-1}, \cdots >$ of numbers such that $0 \leq a_j \leq 9$ and

$$x = \lim_{N\to\infty} \left[\sum_{j=-N}^{m} a_j 10^j \right].$$

In this context the index for the sequence $< a_j >$ is decreasing and tends to $-\infty$. If x has only finitely many non-zero terms in its decimal expansion then, by convention we set all the remaining digits to zero. To study such infinite decimal expansions it is useful to have a formula for the sum of a geometric series.

Proposition A.1.1. *If $r \in \mathbb{R}$ and $N \in \mathbb{N}$ then*

$$\sum_{j=0}^{N} r^j = \frac{r^{N+1} - 1}{r - 1}. \tag{A.6}$$

If $|r| < 1$ then the limit of this sum exists as $N \to \infty$, it is given by

$$\sum_{j=0}^{\infty} r^j = \frac{1}{1 - r}. \tag{A.7}$$

Because the digits in the decimal expansion are restricted to lie between zero and nine we can estimate the error in replacing x by a finite part of its decimal expansion

$$0 \leq x - \sum_{j=-N}^{m} a_j 10^j \leq \sum_{j=N+1}^{\infty} \frac{9}{10^j} = \frac{1}{10^N},$$

which agrees with our intuitive understanding of decimal representations. It tells us that real numbers can be approximated, with arbitrary accuracy by rational numbers. The addition and multiplication operations can therefore be extended by continuity to all real numbers: suppose that $< x_n >$ and $< y_n >$ are sequences of rational numbers converging to real numbers x and y then

$$\lim_{n\to\infty} (x_n + y_n) \stackrel{d}{=} x + y \text{ and } \lim_{n\to\infty} x_n \times y_n \stackrel{d}{=} x \times y.$$

Arguing in a similar way we can show that any positive number x has a binary representation, this is a (possibly infinite) binary sequence $< b_n, b_{n-1}, \cdots >$ such that

$$x = \lim_{N\to\infty} \left[\sum_{j=-N}^{n} b_j 2^j \right].$$

Note that if by a finite part of the binary expansion gives an estimate for x which satisfies

$$0 \leq x - \sum_{j=-N}^{n} b_j 2^j \leq \frac{1}{2^N}.$$

This introduction to real numbers suffices for our applications, a very good and complete introduction to this subject can be found in [11].

Remark A.1.1. Notational remark As is serves no further pedagogical purpose to use $\times$ to indicate multiplication of numbers we henceforth follow the standard notation of indicating multiplication of numbers by juxtaposition: If a, b are numbers then ab is the product of a and b.

Exercise A.1.9. Define the sequence by letting $x_0 = 2$ and

$$x_j = \frac{1}{2}(x_j + x_j^{-1}) \text{ for } j > 0.$$

Show that $< x_n >$ is a bounded, monotone decreasing sequence of rational numbers and explain why its limit must be $\sqrt{2}$. [**Extra credit:**] Show that there is a constant C such that

$$|x_j - \sqrt{2}| < C2^{-2^j}.$$

This shows that $< x_n >$ converges very quickly to $\sqrt{2}$.

Exercise A.1.10. ***** More exercises on sequences and convergence of sequences.*****

A.1.4 Cauchy sequences

In the previous section we discussed the properties of convergent sequences of numbers. Suppose that $< x_n >$ is a sequence of numbers, how do we decide if it has a limit or not? The definition of completeness only considers bounded monotone sequences; many convergent sequence are not monotone. In light of this it would be useful to have a more flexible criterion for a sequence to have a limit. If $< x_n >$ converges to x^* then, as n gets large, x_n gets closer to x^*. As an inevitable consequence of this, the distances between the terms of the sequence, $\{|x_n - x_m|\}$ must become small as *both* m and n get large. In order to converge, the terms of the sequence must cluster closer and closer to *each other* as the index gets large. A sequence with this latter property is called a Cauchy sequence.

Definition A.1.1. A sequence of real numbers $< x_n >$ is called a *Cauchy sequence* if given $\epsilon > 0$ there is an N so that

$$|x_n - x_m| < \epsilon \text{ whenever } m \text{ and } n > N. \tag{A.8}$$

This is called the **Cauchy criterion**.

The fundamental importance of this concept is contained in the following theorem.

Theorem A.1.1. *A sequence of real numbers converges if and only if it is a Cauchy sequence.*

Proof. To prove that a Cauchy sequence converges would take us too far afield. That a convergent sequence is Cauchy, is an elementary consequence of the triangle inequality. Suppose that $< x_n >$ is sequence which converges to x^* and let $\epsilon > 0$ be fixed. Because the sequence converges there exists an N so that

$$|x_n - x^*| < \frac{\epsilon}{2} \text{ if } n > N.$$

If both n and m are larger than N then we use the triangle inequality to obtain the desired result

$$|x_n - x_m| \leq |x_n - x^*| + |x^* - x_m| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

□

If one imagines "observing" a sequence of numbers, then it seems unlikely the one could directly observe its limit, if it exists. On the other hand, the clustering described in the Cauchy criterion is something which is readily observed.

Example A.1.1. Let $x_n = n^{-1}$, if $n < m$ then

$$|x_n - x_m| \leq \frac{1}{n}.$$

This shows that x_n is a Cauchy sequence.

Example A.1.2. Suppose that $< x_n >$ is a sequence and it is known that for any $\epsilon > 0$ there is an N so that $|x_n - x_{n+1}| < \epsilon$ if $n > N$. This does not imply that the sequence converges. For the sequence defined by

$$x_n = \sum_{j=1}^{n} \frac{1}{j},$$

the differences $x_{n+1} - x_n = (n+1)^{-1}$ go to zero as n tends to infinity. However $< x_n >$ is unbounded as n tends to infinity. This shows that it is not enough for the *successive* terms of a sequence to be close together. The Cauchy criterion requires that the differences $|x_n - x_m|$ be small for *all* sufficiently large values of m and n.

Exercise A.1.11. Suppose that $< x_n >$ is a sequence of real numbers such that

$$\lim_{N\to\infty} \sum_{j=1}^{N} |x_j - x_{j+1}| < \infty.$$

Show that $\lim_{n\to\infty} x_j$ exists.

A.2 Vector spaces

We now discuss the linear structure of Euclidean space, linear transformations and different ways to measure distances and angles. In the previous section we saw that numbers can be added and multiplied. This defines a *linear structure* on the set of real numbers which allows us to single out a special collection of functions.

Definition A.2.1. A function $f : \mathbb{R} \to \mathbb{R}$ is *linear* if it satisfies the following two conditions: For all pairs of real numbers x, y

$$\begin{aligned} f(x+y) &= f(x) + f(y), \\ f(xy) &= xf(y). \end{aligned} \tag{A.9}$$

From the definition it is clear that a linear function is determined by its value for any non-zero x. For suppose that we know $f(x)$ for any $x \neq 0$, any other number y can be written $y = x(yx^{-1})$, so (A.9) implies that

$$f(y) = (yx^{-1})f(x).$$

A linear function is therefore of the form $f(x) = ax$ for some $a \in \mathbb{R}$. This is very familiar but it is worth thinking over carefully as it encapsulates why calculus is such a powerful tool.

Recall the definition of the derivative, a function $f(x)$ has derivative $f'(x)$ at x provided that

$$f(x+h) - f(x) = f'(x)h + e(h) \tag{A.10}$$

where $e(h)$ goes to zero faster that $|h|$ as $h \to 0$. This formula tells us that replacing $f(x+h) - f(x)$ by the *linear function* $f'(x)h$ leads to an error of size smaller than $|h|$. If $f'(x) \neq 0$ then (A.10) gives a complete qualitative picture of f for arguments near to x, which is increasingly accurate as $h \to 0$. Of course if $f'(x) = 0$ then (A.10) says little beyond that f is not well approximated by a linear function near to x.

Geometrically $\mathbb{R}$ is usually represented by a straight line, the numbers are coordinates for this line. One can specify coordinates on a plane by choosing two, intersecting straight lines and coordinates in space are determined by choosing three lines which intersect in a point. Of course one can continue in this way. We denote the set of ordered pairs of real numbers by

$$\mathbb{R}^2 = \{(x, y) \mid x, y \in \mathbb{R}\}$$

and the set of ordered triples by

$$\mathbb{R}^3 = \{(x, y, z) \mid x, y, z \in \mathbb{R}\}.$$

These are known as the Euclidean 2-space and 3-space respectively. From a mathematical perspective there is no reason to stop at 3, for each $n \in \mathbb{N}$ we let $\mathbb{R}^n$ denote the set of ordered n-tuples $(x_1, x_2, \ldots, x_n)$ of real numbers. This is called the Euclidean n-space or just n-space for short. From a physical perspective we can think of n-space as giving (local) coordinates for a system with n-degrees of freedom. The physical space we occupy is 3-space, if we include time then this gives us 4-space. If we are studying the weather then we would want to know the temperature, humidity and barometric pressure at each point in space-time, so this requires 7 parameters (x, y, z, t, T, H, P). The more complicated the physical model the more dimensions one requires to describe it.

A.2.1 Euclidean n-space

All the Euclidean n-spaces have the structure of *linear or vector spaces*. This means that we know how to add two n-tuples of real numbers

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n)$$

and multiply an n-tuple of real numbers by a real number

$$a \cdot (x_1, \ldots, x_n) = (ax_1, \ldots ax_n).$$

These two operations are compatible in that

$$\begin{aligned} a \cdot (x_1, \ldots, x_n) + a \cdot (y_1, \ldots, y_n) &= a \cdot (x_1 + y_1, \ldots, x_n + y_n) \\ &= (a(x_1 + y_1), \ldots, a(x_n + y_n)). \end{aligned}$$

An ordered n-tuple of numbers is called an *n-vector* or *vector*. The first operation is called vector addition (or just addition) and the second operation is called *scalar* multiplication. For most values of n there is no way to define a compatible notion of "vector multiplication,"

however there are some special cases where this can be done (if n=2(complex numbers), n=3(cross product), n=4(quaternions), n=8(Cayley numbers)). It is often convenient to use a single letter to denote an n-tuple of numbers. In this book bold-face, Roman letters are used to denote vectors, that is

$$\mathbf{x} = (x_1, \ldots, x_n).$$

For the moment we also use $a \cdot \mathbf{x}$ to denote scalar multiplication. The compatibility of vector addition and scalar multiplication is then written as

$$a \cdot (\mathbf{x} + \mathbf{y}) = a \cdot \mathbf{x} + a \cdot \mathbf{y}.$$

There is a special vector all of whose entries are zero denoted by $\mathbf{0} = (0, \ldots, 0)$, it satisfies

$$\mathbf{x} + \mathbf{0} = \mathbf{x} = \mathbf{0} + \mathbf{x}$$

for any vector $\mathbf{x}$. It is also useful to single out a collection of n *coordinate vectors*. Let $\mathbf{e}_j \in \mathbb{R}^n$ denote the vector with all entries zero but for the j^{th}-entry which equals one. For example if $n = 3$ then the coordinate vectors are

$$\mathbf{e}_1 = (1, 0, 0), \quad \mathbf{e}_2 = (0, 1, 0), \quad \mathbf{e}_3 = (0, 0, 1).$$

These are called coordinate vectors because we can express any vector as a sum of these vectors, if $\mathbf{x} \in \mathbb{R}^n$ then

$$\mathbf{x} = x_1 \cdot \mathbf{e}_1 + \cdots + x_n \cdot \mathbf{e}_n = \sum_{j=1}^{n} x_j \cdot \mathbf{e}_j. \tag{A.11}$$

The n-tuple of numbers $(x_1, \ldots, x_n)$ are then the *coordinates* for the vector $\mathbf{x}$. The set of vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is also called the *standard basis* for $\mathbb{R}^n$.

As before, the linear structure singles out a special collection of functions $f : \mathbb{R}^n \to \mathbb{R}$,

Definition A.2.2. A function $f : \mathbb{R}^n \to \mathbb{R}$ is linear if it satisfies the following conditions: for any pair of vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $a \in \mathbb{R}$

$$\begin{aligned} f(\mathbf{x} + \mathbf{y}) &= f(\mathbf{x}) + f(\mathbf{y}), \\ f(a \cdot \mathbf{x}) &= a f(\mathbf{x}). \end{aligned} \tag{A.12}$$

In light of (A.11) it is clear that a linear function on $\mathbb{R}^n$ is completely determined by the n values $\{f(\mathbf{e}_1), \ldots, f(\mathbf{e}_n)\}$. For an arbitrary $\mathbf{x} \in \mathbb{R}^n$ (A.11) and (A.12) imply

$$f(\mathbf{x}) = \sum_{j=1}^{n} x_j f(\mathbf{e}_j).$$

On the other hand it is easy to see that given n-numbers $\{a_1, \ldots, a_n\}$ we can *define* a linear function on $\mathbb{R}^n$ by setting

$$f(\mathbf{x}) = \sum_{j=1}^{n} a_j x_j.$$

As in the one-dimensional case we therefore have an explicit knowledge of the collection of linear functions.

What measurements are required to determine a linear function? While it suffices, it is not actually necessary to measure $\{f(\mathbf{e}_1), \ldots, f(\mathbf{e}_n)\}$. To describes what is needed, requires a definition.

Definition A.2.3. If $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a collection of n vectors in $\mathbb{R}^n$ with the property that every vector $\mathbf{x}$ can be represented as

$$\mathbf{x} = \sum_{j=1}^{n} a_j \cdot \mathbf{v}_j, \tag{A.13}$$

for a collection of scalars $\{a_1, \ldots, a_n\}$ then we say that these vectors are a *basis* for $\mathbb{R}^n$. The coefficients are called the coordinates of $\mathbf{x}$ with respect to this basis.

Note that the standard bases defined above satisfy (A.13).

Example A.2.1. The standard basis for $\mathbb{R}^2$ is $\mathbf{e}_1 = (1, 0)$, $\mathbf{e}_2 = (0, 1)$. The vectors $\mathbf{v}_1 = (1, 1)$, $\mathbf{v}_2 = (0, 1)$ also define a basis for $\mathbb{R}^2$. To see this we observe that

$$\mathbf{e}_1 = \mathbf{v}_1 - \mathbf{v}_2 \text{ and } \mathbf{e}_2 = \mathbf{v}_2,$$

therefore if $\mathbf{x} = x_1 \cdot \mathbf{e}_1 + x_2 \cdot \mathbf{e}_2$ then

$$\mathbf{x} = x_1 \cdot (\mathbf{v}_1 - \mathbf{v}_2) + x_2 \cdot \mathbf{v}_2 = x_1 \cdot \mathbf{v}_1 + (x_2 - x_1) \cdot \mathbf{v}_2.$$

Proposition A.2.1. *A collection of n vectors in $\mathbb{R}^n$ $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ defines a basis if and only if the only n-tuple for which*

$$\sum_{j=1}^{n} a_j \cdot \mathbf{v}_j = 0$$

is the zero vector. This implies that the scalars appearing in (A.13) *are uniquely determined by* $\mathbf{x}$.

From the Proposition it is clear that the values,

$$\{f(\mathbf{v}_1), \ldots, f(\mathbf{v}_n)\},$$

for any basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, suffice to determine a linear function, f. On the other hand, given numbers $\{a_1, \ldots, a_n\}$ we can define a linear function f by setting

$$f(\mathbf{v}_j) = a_j \text{ for } 1 \leq j \leq n \tag{A.14}$$

and extending *by linearity*. This means that if

$$\mathbf{x} = \sum_{j=1} b_j \cdot \mathbf{v}_j$$

then

$$f(\mathbf{x}) = \sum_{j=1}^{n} b_j a_j. \tag{A.15}$$

From the standpoint of measurement, how are vectors in $\mathbb{R}^n$ distinguished from one another? Linear functions provide *an* answer to this question. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a basis and for each $1 \leq j \leq n$ we define the linear function f_j by the conditions

$$f_j(\mathbf{v}_j) = 1, \quad f_j(\mathbf{v}_i) = 1 \text{ for } i \neq j.$$

Suppose that we can build a machine whose output is $f_j(\mathbf{x})$. Two vectors $\mathbf{x}$ and $\mathbf{y}$ are equal if and only if $f_j(\mathbf{x}) = f_j(\mathbf{y})$ for $1 \leq j \leq n$. Linear functions are very useful in higher dimensions and play the same role in multi-variate calculus as they play in the single variable case.

Exercise A.2.1. Prove Proposition A.2.1.

Exercise A.2.2. Show that the function defined in (A.14) and (A.15) is well defined and linear.

Exercise A.2.3. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a non-zero linear function. Show that there is a basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ for $\mathbb{R}^n$ such that

$$f(\mathbf{v}_1) = 1 \text{ and } f(\mathbf{v}_j) = 0 \text{ for } 2 \leq j \leq n.$$

A.2.2 General vector spaces

As is often the case in mathematics it is useful to introduce an abstract concept which encompasses many special cases.

Definition A.2.4. Let V be a set, it is a *real vector space* if it has two operations:

Addition:
: Addition is a map from $V \times V \to V$. If $(\mathbf{v}_1, \mathbf{v}_2)$ is an element of $V \times V$ then we denote this by $(\mathbf{v}_1, \mathbf{v}_2) \mapsto \mathbf{v}_1 + \mathbf{v}_2$.

Scalar multiplication:
: Scalar multiplication is a map from $\mathbb{R} \times V \to V$. If $a \in \mathbb{R}$ and $\mathbf{v} \in V$ then we denote this by $(a, \mathbf{v}) \mapsto a \cdot \mathbf{v}$.

The operations have the following properties:

Commutative law:
: $\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{v}_2 + \mathbf{v}_1$,

Associative law:
: $(\mathbf{v}_1 + \mathbf{v}_2) + \mathbf{v}_3 = \mathbf{v}_1 + (\mathbf{v}_2 + \mathbf{v}_3)$,

Distributive law:
: $a \cdot (\mathbf{v}_1 + \mathbf{v}_2) = a \cdot \mathbf{v}_1 + a \cdot \mathbf{v}_2$.

Finally there is a special element $\mathbf{0} \in V$ such that

$$\mathbf{v} + \mathbf{0} = \mathbf{v} = \mathbf{0} + \mathbf{v} \text{ and } \mathbf{0} = 0 \cdot \mathbf{v},$$

this vector is called the *zero vector*.

Example A.2.2. For each $n \in \mathbb{N}$ the space $\mathbb{R}^n$ with the addition and scalar multiplication defined above is a vector space.

Example A.2.3. The set real valued functions defined on $\mathbb{R}$ is a vector space. We define addition by the rule $(f+g)(x) = f(x)+g(x)$, scalar multiplication is defined by $(a \cdot f)(x) = af(x)$. We denote the space of functions on $\mathbb{R}$ with these operations by $\mathcal{F}$.

Example A.2.4. If f_1 and f_2 are linear functions on $\mathbb{R}^n$ then define $f_1 + f_2$ as above:

$$(f_1 + f_2)(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathbb{R}^n$$

and $(a \cdot f)(\mathbf{x}) = af(\mathbf{x})$. A sum of linear functions is a linear function, as is a scalar multiple. Thus the set of linear functions on $\mathbb{R}^n$ is also a vector space. This vector space is called the *dual* vector space, it is denoted by $(\mathbb{R}^n)'$.

Example A.2.5. For each $n \in \mathbb{N} \cup \{0\}$ let $\mathcal{P}_n$ denote the set of polynomial functions on $\mathbb{R}$ of degree at most n. Since the sum of two polynomials of degree at most n is again a polynomial of degree at most n, as is a scalar multiple, it follows that $\mathcal{P}_n$ is a vector space.

Many natural mathematical objects have a vector space structure. Often a vector space is subset of a larger vector space.

Definition A.2.5. Let V be a vector space, a subset $U \subset V$ is a *subspace* if whenever $\mathbf{u}_1, \mathbf{u}_2 \in U$ then $\mathbf{u}_1 + \mathbf{u}_2 \in U$ and for every $a \in \mathbb{R}$, $a \cdot \mathbf{u}_1 \in U$ as well. Briefly, a subset U is a subspace if it is a vector space with the addition and scalar multiplication it inherits from V.

Example A.2.6. The subset of $\mathbb{R}^2$ consisting of the vectors $\{(x, 0) \mid x \in \mathbb{R}\}$ is a subspace.

Example A.2.7. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a linear function, the set of vectors $\{\mathbf{v} \in \mathbb{R}^n \mid f(\mathbf{v}) = 0\}$ is a subspace. This subspace is called the null-space of the linear function f.

Example A.2.8. The set of vectors $\{\mathbf{v} \in \mathbb{R}^n \mid f(\mathbf{v}) = 1\}$ in **not** a subspace. If $g : (x, y) \to \mathbb{R}$ is defined by $g(x, y) = x^2 - y$ then the set of vectors $\{(x, y) \in \mathbb{R}^2 \mid g(x, y) = 0\}$ is **not** a subspace.

Example A.2.9. The set of polynomials of degree at most 2 is a subspace of the set of polynomials of degree at most 3.

Definition A.2.6. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ be a collection of vectors in a vector space V. A vector of the form

$$\mathbf{v} = a_1 \cdot \mathbf{v}_1 + \cdots + a_m \cdot \mathbf{v}_m$$

is called a *linear combination* of the vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$. The *linear span* of these vectors is the set of all linear combinations

$$\text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_m) \stackrel{d}{=} \{a_1 \cdot \mathbf{v}_1 + \cdots + a_m \cdot \mathbf{v}_m \mid a_1, \ldots a_m \in \mathbb{R}\}.$$

Example A.2.10. The linear span of a collection of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\} \subset V$ is a subspace of V.

A basic feature of a vector space is its dimension. This is a precise mathematical formulation of the number of degrees of freedom. The vector space $\mathbb{R}^n$ has dimension n. The general concept of a *basis* is needed to define the dimension.

Definition A.2.7. Let V be a vector space, a set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset V$ is said to be *linearly independent* if

$$\sum_{j=1}^{n} a_j \cdot \mathbf{v}_j = 0$$

implies that $a_j = 0$ for $j = 1, \ldots, n$. This is another way of saying that it is not possible to write one of these vectors as a linear combination of the others. A finite set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset V$ is a *basis* for V if

(1). The vectors are linearly independent,

(2). Every vector in V is a linear combination of these vectors, that is

$$\operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_n) = V.$$

The definition of a basis given earlier for the vector spaces $\mathbb{R}^n$ is a special case of this definition. If a vector space V has a basis then every basis for V has the same number of elements. This fact allows makes it possible to define the dimension of a vector space.

Definition A.2.8. If a vector space V has a basis consisting of n vectors then the *dimension* of V is n. We write

$$\dim V = n.$$

If $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis for V then for every vector $\mathbf{v} \in V$ there is a unique point $(x_1, \ldots, x_n) \in \mathbb{R}^n$ such that

$$\mathbf{v} = x_1 \cdot \mathbf{v}_1 + \cdots + x_n \cdot \mathbf{v}_n. \tag{A.16}$$

A vector space V of dimension n has exactly the same number of degrees of freedom as $\mathbb{R}^n$. In fact, by choosing a basis we define an *isomorphism* between V and $\mathbb{R}^n$. This is because if $\mathbf{v} \leftrightarrow (x_1, \ldots, x_2)$ and $\mathbf{v}' \leftrightarrow (y_1, \ldots, y_n)$ in (A.16) then

$$\mathbf{v} + \mathbf{v}' = (x_1 + y_1) \cdot \mathbf{v}_1 + \cdots + (x_n + y_n) \cdot \mathbf{v}_n$$

and for $a \in \mathbb{R}$

$$a \cdot \mathbf{v} = (ax_1) \cdot \mathbf{v}_1 + \cdots + (ax_n) \cdot \mathbf{v}_n.$$

From this point of view, all vector spaces of dimension n are "the same." The abstract concept is still useful. Vector spaces often do not come with a *natural* choice of basis. Indeed the possibility of changing the basis, that is changing the identification of V with $\mathbb{R}^n$ is a very powerful tool. In applications one tries to choose a basis that is well adapted to the problem at hand. It is important to note that many properties of vector spaces are independent of the choice of basis.

Example A.2.11. The vector space $\mathcal{F}$ of all functions on $\mathbb{R}$ does not have a basis, that is we cannot find a finite collection of functions such that any function is a linear combination of these functions. The vector space $\mathcal{F}$ is infinite dimensional. The study of infinite dimensional vector spaces is called functional analysis, we return to this subject in section A.3.

Example A.2.12. For each n the set $\{1, x, \ldots, x^n\}$ is a basis for the $\mathcal{P}_n$. Thus the $\dim \mathcal{P}_n = n + 1$.

Exercise A.2.4. Show that $\mathcal{F}$, defined in example A.2.3 is a vector space.

Exercise A.2.5. Show that the set of polynomials $\{x^j(1-x)^{n-j} \mid 0 \leq j \leq n\}$ is a basis for $\mathcal{P}_n$.

Exercise A.2.6. Show that if a vector space V has a basis then any basis for V has the same number of vectors.

Exercise A.2.7. Let V be a vector space with $\dim V = n$ and let V' denote the set of linear functions on V. Show that V' is also a vector space with $\dim V' = n$.

Exercise A.2.8. Exercises on bases, dimensions, and vector spaces in general.

A.2.3 Linear Transformations and matrices

The fact that both $\mathbb{R}^n$ and $\mathbb{R}^m$ have linear structures allows us to single out a special class of maps between these spaces.

Definition A.2.9. A map $F : \mathbb{R}^n \to \mathbb{R}^m$ is called a *linear transformation* if for all pairs $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$ and $a \in \mathbb{R}$ we have

$$\begin{aligned} F(\mathbf{x}+\mathbf{y}) &= F(\mathbf{x}) + F(\mathbf{y}), \\ F(a \cdot \mathbf{x}) &= a \cdot F(\mathbf{x}). \end{aligned} \tag{A.17}$$

Comparing the definitions we see that a linear function is just the $m = 1$ case of a linear transformation. For each $n \in \mathbb{N}$ there is a special linear transformation of $\mathbb{R}^n$ to itself, called the identity map. It is defined by $\mathbf{x} \mapsto \mathbf{x}$ and denoted by Id_n .

If $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis for $\mathbb{R}^n$ then a linear transformation is determined by the values $\{F(\mathbf{v}_1), \ldots, F(\mathbf{v}_n)\}$. If $\mathbf{x} = a_j \cdot \mathbf{v}_1 + \cdots + a_n \cdot \mathbf{v}_n$ then (A.17) implies that

$$F(\mathbf{x}) = \sum_{j=1}^{n} a_j \cdot F(\mathbf{v}_j).$$

In this section, linear transformations are denoted by bold, upper case, Roman letters, e.g. $\mathbf{A}, \mathbf{B}$. The action of a linear

Connected to a linear transformation $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ are two natural subspaces.

Definition A.2.10. The set of vectors $\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = 0\}$ is called the kernel or null space of the linear transformation $\mathbf{A}$; we denote this subspace by $\ker \mathbf{A}$.

Definition A.2.11. The set of vectors $\{\mathbf{A}\mathbf{x} \in \mathbb{R}^m \mid \mathbf{x} \in \mathbb{R}^n\}$ is called the **image** of the linear transformation $\mathbf{A}$; we denote this $\mathrm{Im}\, \mathbf{A}$.

The kernel and image of a linear transformation are basic examples of subspaces of a vector space which are defined *without reference* to a basis. There is, in general no natural choice of a basis for either subspace.

As above let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a basis for $\mathbb{R}^n$, if we also choose a basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ for $\mathbb{R}^m$ then there is a collection of mn-numbers $\{a_{ij}\}$ so that for each j

$$\mathbf{A}(\mathbf{v}_j) = \sum_{i=1}^{m} a_{ij}\mathbf{u}_i.$$

Such a collection of numbers, labeled with two indices is called a *matrix*. Once bases for the domain and range of $\mathbf{A}$ are fixed, the matrix determines and is determined by the linear transformation. If $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^n$ then one usually selects a single basis $\{\mathbf{v}_j\}$ and uses it to represent vectors in both the domain and range of $\mathbf{A}$. Often times it is implicitly understood that the bases are the standard bases.

Example A.2.13. If $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis for $\mathbb{R}^n$ then $\mathrm{Id}_n(\mathbf{v}_j) = \mathbf{v}_j$. The matrix for Id_n, with respect to any basis is denoted by

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{fi } i \neq j. \end{cases}$$

Once a pair of bases is fixed then one can identify the set of linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$ with the collection of $m \times n$-arrays (read m by n) of numbers. If we think of (a_{ij}) as a rectangular array of numbers then the first index, i labels the rows and the second index, j labels the columns.

$$\begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix} \tag{A.18}$$

A vector in $\mathbb{R}^n$ can be thought of as either a row vector, that is an $1 \times n$-matrix or a column vector, that is an $n \times 1$-matrix. An $m \times n$-matrix has n columns consisting of $m \times 1$ vectors,

$$\mathbf{a} = (\mathbf{a}_1 \ldots \mathbf{a}_n)$$

or m rows consisting of $1 \times n$ vectors

$$\mathbf{a} = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_m \end{pmatrix}.$$

Precisely how one wishes to think about a matrix depends on the situation at hand.

We can define a notion of multiplication between column vectors and matrices.

Definition A.2.12. Let $\mathbf{a}$ be an $m \times n$ matrix with entries $a_{ij}, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n$ and $\mathbf{x}$ be an n-vector with entries $x_j, 1 \leq j \leq n$ then we define the product $\mathbf{a} \cdot \mathbf{x}$ to be the m-vector $\mathbf{y}$ with entries

$$y_i = \sum_{j=1}^{n} a_{ij} x_j, \quad i = 1, \ldots, m.$$

Concisely this is written $\mathbf{y} = \mathbf{a} \cdot \mathbf{x}$. In this section we use lower case, bold Roman letters to denote matrices, e.g. $\mathbf{a}, \mathbf{b}$.

Proposition A.2.2. *Let* $\mathbf{a}$ *be an* $m \times n$*-matrix,* $\mathbf{x}_1, \mathbf{x}_2 n$*-vectors and* $a \in \mathbb{R}$ *then*

$$\mathbf{a} \cdot (\mathbf{x}_1 + \mathbf{x}_2) = \mathbf{a} \cdot \mathbf{x}_1 + \mathbf{a} \cdot \mathbf{x}_2 \text{ and } \mathbf{a} \cdot (a \cdot \mathbf{x}) = a \cdot (\mathbf{a} \cdot \mathbf{x}).$$

These conditions show that the map $\mathbf{x} \mapsto \mathbf{a} \cdot \mathbf{x}$ is a linear transformation of $\mathbb{R}^n$ to $\mathbb{R}^m$.

We can also define multiplication between matrices with compatible dimensions. Let $\mathbf{a}$ be an $m \times n$ matrix and $\mathbf{b}$ be an $l \times m$ matrix. If $\mathbf{x}$ is an n-vector then $\mathbf{a} \cdot \mathbf{x}$ is an m-vector so $\mathbf{b} \cdot (\mathbf{a} \cdot \mathbf{x})$ is defined. As the composition is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^l$, this defines $\mathbf{c} = \mathbf{b} \cdot \mathbf{a}$ as an $l \times n$-matrix. If (a_{ij}) are the entries of $\mathbf{a}$ and (b_{pq}), the entries of $\mathbf{b}$ then the entries of the product $\mathbf{c}$ are given by

$$c_{pj} = \sum_{i=1}^{m} b_{pi} a_{ij}.$$

This shows that we can multiply an $m \times n$ matrix by an $l \times m$ matrix and the result is an $l \times n$ matrix. If $\mathbf{a}$ and $\mathbf{b}$ are both $n \times n$-matrices then both products $\mathbf{a} \cdot \mathbf{b}, \mathbf{b} \cdot \mathbf{a}$ are defined. In general they are **not** equal. One says that matrix multiplication is *non-commutative.* The product of an $m \times n$-matrix and an n-vector is the special case of multiplying an $n \times 1$ matrix by a $m \times n$-matrix, as expected the result if an $m \times 1$-matrix or an m-column vector.

Suppose that $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ are both bases for $\mathbb{R}^n$. The definition of a basis implies that there are $n \times n$-matrices $\mathbf{a} = (a_{ij})$ and $\mathbf{b} = (b_{ij})$ so that

$$\mathbf{v}_i = \sum_{j=1}^{n} a_{ji} \cdot \mathbf{u}_j \text{ and } \mathbf{u}_i = \sum_{j=1}^{n} b_{ji} \cdot \mathbf{v}_j.$$

These are called *change of basis* matrices. If $\mathbf{x} \in \mathbb{R}^n$ then there are vectors $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$ so that

$$\mathbf{x} = \sum_{j=1}^{n} a_j \cdot \mathbf{v}_j \text{ and also } \mathbf{x} = \sum_{j=1}^{n} b_j \cdot \mathbf{u}_j.$$

Substituting our expression for the $\{\mathbf{v}_j\}$ in terms of the $\{\mathbf{u}_j\}$ gives

$$\begin{aligned} \mathbf{x} &= \sum_{j=1}^{n} a_j \cdot \left[\sum_{k=1}^{n} a_{kj} \cdot \mathbf{u}_k \right] \\ &\sum_{k=1}^{n} \left[\sum_{j=1}^{n} a_{kj} a_j \right] \cdot \mathbf{u}_k. \end{aligned} \tag{A.19}$$

Comparing (A.19) with our earlier formula we see that

$$b_k = \sum_{j=1}^{n} a_{kj} a_j \text{ for } k = 1, \ldots, n.$$

This explains $\mathbf{a}$ is called the change of basis matrix.

Suppose that $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation and we select bases $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ for $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. Let (a_{ij}) denote the matrix of this linear transformation with respect to this choice of bases. How does the matrix change if these bases are replaced by a different pair of bases? We consider what it means for "(a_{ij}) to be the matrix representing $\mathbf{A}$ with respect to the bases $\{\mathbf{v}_j\}$ and $\{\mathbf{u}_i\}$" by putting into words

the computations performed above: Suppose that $\mathbf{x}$ is a vector in $\mathbb{R}^n$ with coordinates $(x_1, \ldots, x_n)$ with respect to the basis $\{\mathbf{v}_j\}$, then the coordinates of $\mathbf{y} = \mathbf{A}\mathbf{x}$ with respect to $\{\mathbf{u}_i\}$ are

$$y_i = \sum_{j=1}^{n} a_{ij} x_j, \quad i = 1, \ldots, m.$$

The fact to keep in mind is that we are dealing with different representations of fixed (abstract) vectors $\mathbf{x}$ and $\mathbf{A}\mathbf{x}$.

Suppose that $\{\mathbf{v}'_j\}$ and $\{\mathbf{u}'_i\}$ are new bases for $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively and let (b_{lj}) and (c_{ki}) be change of basis matrices, that is

$$\mathbf{v}'_j = \sum_{l=1}^{n} b_{lj} \cdot \mathbf{v}_l \text{ and } \mathbf{u}_i = \sum_{k=1}^{m} c_{ki} \cdot \mathbf{u}'_k.$$

Let a'_{ij} be the matrix of $\mathbf{A}$ with respect to $\{\mathbf{v}'_j\}$ and $\{\mathbf{u}'_i\}$. If $(x'_1, \ldots, x'_n)$ are the coordinates of $\mathbf{x}$ with respect to $\{\mathbf{v}'_j\}$ and $(y'_1, \ldots, y'_m)$ the coordinates of $\mathbf{A}\mathbf{x}$ with respect to $\{\mathbf{u}'_i\}$ then

$$y'_i = \sum_{j=1}^{n} a'_{ij} x'_j.$$

Formula (A.19) tells us that

$$x_j = \sum_{l=1}^{n} b_{jl} x'_l$$

and therefore

$$\begin{aligned} y_i &= \sum_{j=1}^{n} a_{ij} \left[\sum_{l=1}^{n} b_{jl} x'_l \right] \\ &= \sum_{l=1}^{n} \left[\sum_{j=1}^{n} a_{ij} b_{jl} \right] x'_l \end{aligned} \tag{A.20}$$

gives the expression for $\mathbf{A}\mathbf{x}$ with respect to the $\{\mathbf{u}_i\}$. To complete our computation we only need to re-express $\mathbf{A}\mathbf{x}$ with respect to the basis $\{\mathbf{u}'_i\}$. To that end we apply (A.19) one more time to obtain that

$$y'_i = \sum_{k=1}^{m} c_{ik} y_k.$$

Putting this into (A.20) and reordering the sums we obtain that

$$y'_i = \sum_{j=1}^{n} \left[\sum_{k=1}^{m} \sum_{l=1}^{n} c_{ik} a_{kl} b_{lj} \right] x'_j.$$

This shows that

$$a'_{ij} = \sum_{k=1}^{m} \sum_{l=1}^{n} c_{ik} a_{kl} b_{lj}.$$

Using $\mathbf{a}, \mathbf{a}', \mathbf{b}, \mathbf{c}$ to denote the matrices defined above and $\mathbf{x}, \mathbf{x}', \mathbf{y}, \mathbf{y}'$, the column vectors of coordinates, we can rewrite these expressions more concisely as

$$\begin{aligned} \mathbf{x} = \mathbf{b} \cdot \mathbf{x}', \quad \mathbf{y}' = \mathbf{c} \cdot \mathbf{y}, \\ \mathbf{a}' = \mathbf{c} \cdot \mathbf{a} \cdot \mathbf{b}. \end{aligned} \tag{A.21}$$

The reader should be aware that this formula differs slightly from that usually given in textbooks, this is because $\mathbf{b}$ changes from $\mathbf{x}'$ to $\mathbf{x}$ whereas $\mathbf{c}$ changes from $\mathbf{y}$ to $\mathbf{y}'$. transformation $\mathbf{A}$ on a vector $\mathbf{x}$ is often denoted by $\mathbf{Ax}$.

Exercise A.2.9. Show that that if $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{B} : \mathbb{R}^m \to \mathbb{R}^l$ are linear transformations then the composition $\mathbf{B} \circ \mathbf{A}(\mathbf{x}) \stackrel{d}{=} \mathbf{B}(\mathbf{A}(\mathbf{x}))$ is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^l$.

Exercise A.2.10. Let $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation, show that $\ker \mathbf{A}$ is a subspace of $\mathbb{R}^n$.

Exercise A.2.11. Let $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation, show that $\operatorname{Im} \mathbf{A}$ is a subspace of $\mathbb{R}^m$.

Exercise A.2.12. Suppose that we use a basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ for the domain and $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ for the range, what is the matrix for Id_n?

Exercise A.2.13. Prove Proposition A.2.2.

Exercise A.2.14. If

$$\mathbf{a} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

then show that $\mathbf{a} \cdot \mathbf{b} \neq \mathbf{b} \cdot \mathbf{a}$.

Exercise A.2.15. Show that if $\mathbf{a}$ is the matrix of a linear transformation $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{b}$ is the matrix of a linear transformation $\mathbf{B} : \mathbb{R}^m \to \mathbb{R}^l$ then $\mathbf{b} \cdot \mathbf{a}$ is the matrix of their composition $\mathbf{B} \circ \mathbf{A} : \mathbb{R}^n \to \mathbb{R}^l$.

Exercise A.2.16. Show that $\partial_x : \mathcal{P}_n \to \mathcal{P}_n$ is a linear transformation. It is defined without reference to a basis. Find the basis for ∂_x in terms of the basis $\{1, x, \ldots, x^n\}$. Find bases for $\ker \partial_x$ and $\operatorname{Im} \partial_x$.

Exercise A.2.17. Show that the space of linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$ is a vector space with addition defined by

$$(\mathbf{A} + \mathbf{B})\mathbf{x} \stackrel{d}{=} \mathbf{Ax} + \mathbf{Bx} \text{ for all } \mathbf{x} \in \mathbb{R}^n$$

and scalar multiplication defined by

$$(a \cdot \mathbf{A})(\mathbf{x}) \stackrel{d}{=} a \cdot (\mathbf{Ax}).$$

Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ be bases for $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. For $1 \leq i \leq m$ and $1 \leq j \leq n$ define the linear transformations $\mathbf{l}_{ij}$ by letting

$$\mathbf{l}_{ij}(\mathbf{v}_j) = \mathbf{u}_i \text{ and } l_{ij}(\mathbf{v}_k) = 0 \text{ if } k \neq j.$$

Show that the $\{\mathbf{l}_{ij} \mid 1 \leq i \leq m, \quad 1 \leq j \leq n\}$ are a basis for this vector space. This shows that the space of linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$ is isomorphic to $\mathbb{R}^{mn}$.

A.2.4 Norms and Metrics

In the previous section we concentrated on algebraic properties of vector spaces. In applications of linear algebra to physical problems it is also important to have a way to measure distances. In part, this is because measurements and computations are inaccurate and one needs a way to quantify the errors. Measurement of distance in a vector space usually begins with a notion of *length*. The distance between two vectors $\mathbf{x}$ and $\mathbf{y}$ is then defined as the length of $\mathbf{x} - \mathbf{y}$, again taking advantage of the underlying linear structure.

There are many reasonable ways to measure length in a vector space. Let us consider the case of $\mathbb{R}^n$. The usual way to measure length is to define

$$\text{the length of } (x_1, \ldots, x_n) = \sqrt{\sum_{j=1}^{n} x_j^2}.$$

Here are two other reasonable ways to define the length of a vector

$$\begin{aligned} \|(x_1, \ldots, x_n)\|_1 &= \sum_{j=1}^{n} |x_j|, \\ \|(x_1, \ldots, x_n)\|_\infty &= \max\{|x_1|, \ldots, |x_n|\}. \end{aligned} \tag{A.22}$$

What makes a notion of length reasonable? There are three basic properties that a reasonable notion of length should have. Let l denote a real valued function defined on a vector space V; it defines a reasonable notion of length if it satisfies the following conditions:

Non-degeneracy:
: For every $\mathbf{v} \in V$, $l(\mathbf{v}) \geq 0$ and $l(\mathbf{v}) = 0$ if and only if $\mathbf{v} = \mathbf{0}$. In other words, every vector has non-negative length and only the zero vector has length zero.

Homogeneity:
: If $a \in \mathbb{R}$ and $\mathbf{v} \in V$ then $l(a \cdot \mathbf{v}) = |a| l(\mathbf{v})$. If we scale a vector, its length gets multiplied by the scaling factor.

Triangle inequality:
: If $\mathbf{v}, \mathbf{v}' \in V$ then
$$l(\mathbf{v} + \mathbf{v}') \leq l(\mathbf{v}) + l(\mathbf{v}'). \tag{A.23}$$

Definition A.2.13. A function $l : V \to \mathbb{R}$ which satisfies these three conditions is called a *norm.* A vector space V with a *choice* of norm is called a normed vector space.

The functions defined in (A.22) satisfy these conditions and therefore define norms.

Example A.2.14. If p is a real number with $1 \leq p$ then the function

$$\|(x_1, \ldots, x_n)\|_p = \left[\sum_{j=1}^{n} |x_j|^p \right]^{\frac{1}{p}} \tag{A.24}$$

defines a norm on $\mathbb{R}^n$. The case $p = \infty$ is given above, it is called the *sup-norm.* The standard Euclidean norm is usually denote by $\|\mathbf{x}\|_2$.

Using a norm $\|\cdot\|$ we can define a notion of distance between two vectors by setting

$$d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|.$$

For any choice of norm this function has the following properties

Non-degeneracy:
$d(\mathbf{x},\mathbf{y}) \geq 0$ with equality if and only if $\mathbf{x}=\mathbf{y}$.

Symmetry:
$d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$.

Triangle inequality:
For any 3 points $\mathbf{x},\mathbf{y},\mathbf{z}$ we have that

$$d(\mathbf{x},\mathbf{z}) \leq d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z}).$$

Any function $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with these properties is called a *metric*. There are metrics on $\mathbb{R}^n$ which are not define by norms.

A metric gives a way to measure distances and therefore a way to define the convergence of sequences.

Definition A.2.14. Suppose that $d(\cdot,\cdot)$ is a metric defined by a norm and that $< \mathbf{x}_j > \subset \mathbb{R}^n$ is a sequence of vectors. The sequence converges to $\mathbf{x}$ in the d-sense if

$$\lim_{j\to\infty} d(\mathbf{x}_j,\mathbf{x}) = 0.$$

Proposition A.2.3. *Suppose that $\|\cdot\|$ and $\|\cdot\|'$ are two norms on $\mathbb{R}^n$ with associated metrics d and d' then there is a positive constant C so that*

$$C^{-1}\|\mathbf{x}\|' \leq \|\mathbf{x}\| \leq C\|\mathbf{x}\|' \text{ for all } \mathbf{x} \in \mathbb{R}^n.$$

The proof is left as an exercise. Given a notion of distance it is also possible to define Cauchy sequences.

Definition A.2.15. Suppose that d is a metric on $\mathbb{R}^n$ and $< \mathbf{x}_n >$ is a sequence. It is a *Cauchy sequence* with respect to d if, for any $\epsilon > 0$, there exists an N so that

$$d(\mathbf{x}_n,\mathbf{x}_m) < \epsilon \text{ provided that } m \text{ and } n > N.$$

The importance of this concept is contained in the following theorem.

Theorem A.2.1. *Let d be a metric on $\mathbb{R}^n$. A sequence $< \mathbf{x}_n >$ converges in the d-sense if and only if it is a Cauchy sequence.*

The choice of which norm to use in a practical problem is often dictated by physical considerations. For example if we have a system whose state is described by a point in $\mathbb{R}^n$ and we allow the same uncertainty in each of our measurements then it would be reasonable to use the sup-norm ,i.e. $\|\cdot\|_\infty$. If on the other hand we can only tolerate a certain fixed

aggregate error, but it is not important how this error is distributed among the various measurements, then it would be reasonable to use $\|\cdot\|_1$ to define the norm. If the errors are expected to follow a Gaussian distribution then one would usually use the Euclidean norm.

There are also computational considerations which can dictate the choice of a norm. If $\mathbf{a}$ is an $m \times n$ matrix with $m > n$ then the system of linear equations $\mathbf{ax} = \mathbf{y}$ is over-determined. For most choices of $\mathbf{y}$ it has no solution. A way to handle such equations is to look for a vector such that the "size" of the error $\mathbf{ax} - \mathbf{y}$ is minimized. To do this, one needs to choose a norm on $\mathbb{R}^m$ to measure the size of the error. It turns out that among all possible choices the Euclidean norm leads to the simplest minimization problems. The vector $\bar{\mathbf{x}}$ such that $\|\mathbf{a}\bar{\mathbf{x}} - \mathbf{y}\|_2$ is minimal is called the *least squares solution.*

In exercise A.2.17 it is shown that the space of linear transformations from$\mathbb{R}^n$ to $\mathbb{R}^m$ is a vector space. When discussing numerical methods for solving linear equations it is very useful to have a way to measure the size of linear transformation which is connected to its geometric properties as a map. We can use norms on the domain and range to define a notion of size for a linear transformation $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$. Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ and $\|\cdot\|'$ be a norm on $\mathbb{R}^m$. The *operator norm* of $\mathbf{A}$ is defined by setting

$$|\!|\!|\mathbf{A}|\!|\!| = \max_{\mathbf{x}\in\mathbb{R}^n\setminus\{0\}} \frac{\|\mathbf{Ax}\|'}{\|\mathbf{x}\|}. \tag{A.25}$$

This norm gives a measure of how much $\mathbf{A}$ changes the lengths of vectors. For all $\mathbf{x} \in \mathbb{R}^n$ we have the estimate

$$\|\mathbf{Ax}\|' \leq |\!|\!|\mathbf{A}|\!|\!|\|\mathbf{x}\|. \tag{A.26}$$

The estimate (A.26) implies that a linear transformation for $\mathbb{R}^n$ to $\mathbb{R}^m$ is always continuous. This is because $\mathbf{Ax}_1 - \mathbf{Ax}_2 = \mathbf{A}(\mathbf{x}_1 - \mathbf{x}_2)$. Thus we see that

$$\|\mathbf{Ax}_1 - \mathbf{Ax}_2\|' = \|\mathbf{A}(\mathbf{x}_1 - \mathbf{x}_2)\| \leq |\!|\!|\mathbf{A}|\!|\!|\|\mathbf{x}_1 - \mathbf{x}_2\| \tag{A.27}$$

There are other ways to define norms on linear transformations. If we fix bases in the domain and range then norms defined on Euclidean spaces, $\mathbb{R}^{mn}$ can be used to define norms on the set of linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$. If (a_{ij}) is the matrix of a linear transformation $\mathbf{A}$ then we can, for example define

$$|\!|\!|\mathbf{A}|\!|\!|_p = \left[\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^p\right]^{\frac{1}{p}}.$$

These norms are not as closely connected to the geometric properties of the map. If $p \neq 2$ then it is **not** generally true that $\|\mathbf{Ax}\|_p \leq |\!|\!|\mathbf{A}|\!|\!|_p\|\mathbf{x}\|_p$. Result like that in exercise A.2.24 are also generally false for these sorts of norms. In physical applications a linear transformation or matrix often models a measurement process: if $\mathbf{x}$ describes the state of a system then $\mathbf{Ax}$ is the result of performing measurements on the system. The appropriate notion of size for $\mathbf{A}$ may then determined by the sorts of errors which might arise in the model.

Exercise A.2.18. Suppose that $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$ and that $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$. If

$$d(\mathbf{x}, \mathbf{z}) = d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}).$$

then show that the three points lie along a line in the indicated order. Is this true if we use $\|\cdot\|_p$ with $p \neq 2$ to define the metric?

Exercise A.2.19. Prove Proposition A.2.3. Hint: use the fact that $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$.

Exercise A.2.20. Let $\|\cdot\|$ and $\|\cdot\|'$ be two norms on $\mathbb{R}^n$ and d, d' the corresponding metrics. Show that a sequence $< \mathbf{x}_j >$ converges to $\mathbf{x}$ in the d-sense if and only if it converges in the d'-sense. This shows that the notion of limits on Euclidean spaces is independent of the choice of norm. Hint: Use Proposition A.2.3.

Exercise A.2.21. Let $\|\cdot\|$ and $\|\cdot\|'$ be two norms on $\mathbb{R}^n$ and d, d' the corresponding metrics. Show that a sequence $< \mathbf{x}_j >$ converges to $\mathbf{x}$ in the d-sense if and only if it converges in the d'-sense. This shows that the notion of limits on Euclidean spaces is independent of the choice of norm. Hint: Use Proposition A.2.3.

Exercise A.2.22. Suppose that w_1, w_2 are positive numbers show that

$$l_w((x_1, x_2)) = \sqrt{w_1 x_1^2 + w_2 x_2^2}$$

defines a norm on $\mathbb{R}^2$. What physical considerations might lead to using a norm like l_w instead of the standard Euclidean norm?

Exercise A.2.23. Use estimate (A.27) to show that if a sequence $< \mathbf{x}_n >$ converges to $\mathbf{x}$ in $\mathbb{R}^n$ then $< \mathbf{A}\mathbf{x}_n >$ also converges to $\mathbf{A}\mathbf{x}$ in $\mathbb{R}^m$. This is just the statement that $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ is continuous.

Exercise A.2.24. Let $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{B} : \mathbb{R}^m \to \mathbb{R}^l$. Choose norms $\|\cdot\|, \|\cdot\|'$ and $\|\cdot\|''$ for $\mathbb{R}^n, \mathbb{R}^m$ and $\mathbb{R}^l$ respectively and let $|\!|\!| \cdot |\!|\!|_{n\to m}$, $|\!|\!| \cdot |\!|\!|_{m\to l}$ and $|\!|\!| \cdot |\!|\!|_{n\to l}$ denote the operator norms they define. Show that

$$|\!|\!|\mathbf{B} \circ \mathbf{A}|\!|\!|_{n\to l} \leq |\!|\!|\mathbf{A}|\!|\!|_{n\to m} |\!|\!|\mathbf{B}|\!|\!|_{m\to l}.$$

Exercise A.2.25. Let $A : \mathbb{R}^n \to \mathbb{R}^n$ have matrix (a_{ij}) with respect to the standard basis. Show that

$$\|\mathbf{A}\mathbf{x}\|_2 \leq |\!|\!|\mathbf{a}|\!|\!|_2 \|\mathbf{x}\|_2.$$

Remark A.2.1 *(***Important notational remark)**. From this point on, we no longer use $\cdot$ to denote the operations of scalar multiplication or multiplication of a vector by a matrix. That is for $\mathbf{x} \in \mathbb{R}^n$, $a \in \mathbb{R}$ the notation $a\mathbf{x}$ indicates scalar multiplication and for $\mathbf{a}$ and $\mathbf{b}$ matrices, $\mathbf{ab}$ is matrix product of $\mathbf{a}$ and $\mathbf{b}$.

A.2.5 Inner product structure

The notions of distance considered in the previous section do not allow for the measurement of angles between vectors. Recall the formula for the dot product in $\mathbb{R}^2$

$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2 = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos\theta,$$

where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$. We can generalize the notion of the dot-product to n-dimensions by setting

$$\mathbf{x} \cdot \mathbf{y} = \sum_{j=1}^{n} x_j y_j.$$

This is sometimes denoted by $\langle \mathbf{x}, \mathbf{y} \rangle$, it is also called an *inner product*.

Proposition A.2.4. *If $\mathbf{x}, \mathbf{y}, \mathbf{z}$ are vectors in $\mathbb{R}^n$ and $a \in \mathbb{R}$ then*

$$\begin{aligned} \langle \mathbf{x}, \mathbf{y} \rangle &= \langle \mathbf{y}, \mathbf{x} \rangle, \\ \langle (\mathbf{x} + \mathbf{y}), \mathbf{z} \rangle &= \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle \ \textit{and} \\ \langle a\mathbf{x}, \mathbf{y} \rangle &= a \langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, a\mathbf{y} \rangle. \end{aligned} \tag{A.28}$$

The inner product is connected with the Euclidean norm by the relation

$$\langle \mathbf{x}, \mathbf{x} \rangle = \|\mathbf{x}\|_2^2.$$

Most of the special properties of the Euclidean norm stem from this fact. There is a very important estimate which also connects these two objects called the *Cauchy-Schwarz inequality*

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2. \tag{A.29}$$

It is proved in exercise A.2.31. It implies that

$$-1 \leq \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} \leq 1. \tag{A.30}$$

In light of (A.30) we can **define** the angle θ between two non-zero vectors in $\mathbb{R}^n$ by the formula

$$\cos\theta = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}.$$

An important special case is an angle of 90° which is the case if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. The vectors $\mathbf{x}$ and $\mathbf{y}$ are said to be *orthogonal*.

Suppose that $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis for $\mathbb{R}^n$. In order to make practical use of this basis it is necessary to be able to determine the coordinates of a vector with respect to it. Suppose that $\mathbf{x}$ is a vector, we would like to find scalars $\{a_j\}$ so that

$$\mathbf{x} = \sum_{j=1}^{n} a_j \mathbf{v}_j.$$

Expressing the basis vectors and $\mathbf{x}$ in terms of the standard basis, $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{v}_j = (v_{1j}, \ldots, v_{nj})$ this can be re-expressed as a system of linear equations,

$$\sum_{j=1}^{n} v_{ij} a_j = x_i \text{ for } i = 1, \ldots, n.$$

In general this system can be quite difficult to solve, however there is a special case when it is very easy to write down a formula for the solution.

Suppose that the basis vectors are of Euclidean length one and pairwise orthogonal, that is

$$\|\mathbf{v}_j\|_2 = 1 \text{ for } j = 1, \ldots, n \text{ and } \langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0 \text{ if } i \neq j.$$

Such a basis is called an *orthonormal basis.* The standard basis is an orthonormal basis. In this case

$$a_j = \langle \mathbf{x}, \mathbf{v}_j \rangle;$$

that is, the coordinates of $\mathbf{x}$ with respect to $\{\mathbf{v}_j\}$ can be computing by simply evaluating these inner products, hence

$$\mathbf{x} = \sum_{j=1}^{n} \langle \mathbf{x}, \mathbf{v}_j \rangle \mathbf{v}_j. \tag{A.31}$$

A consequence of (A.31) is that, in any orthonormal basis $\{\mathbf{v}_j\}$, the Pythagorean theorem holds

$$\|\mathbf{x}\|_2^2 = \sum_{j=1}^{n} |\langle \mathbf{x}, \mathbf{v}_j \rangle|^2 \tag{A.32}$$

An immediate consequence of (A.32) is that the individual coordinates of a vector, with respect to an orthonormal basis are bounded by the Euclidean length of the vector.

Orthonormal bases are often preferred in applications because they display stability properties not shared by arbitrary bases.

Example A.2.15. If $\epsilon \neq 0$ then the vectors $\mathbf{v} = (1, 0)$ and $\mathbf{u}_\epsilon = (1, \epsilon)$ are a basis for $\mathbb{R}^2$. If ϵ is small then the angle between these vectors is very close to zero. The representation of $(0, 1)$ with respect to $\{\mathbf{v}, \mathbf{u}_\epsilon\}$ is

$$(0, 1) = \frac{1}{\epsilon}\mathbf{u}_\epsilon - \frac{1}{\epsilon}\mathbf{v}.$$

The coefficients blow up as ϵ goes to zero. For non-orthonormal bases, it can be difficult to estimate the sizes of the coefficients in terms of the length of the vector.

The Gram-Schmidt method

The problem then arises of how to construct orthonormal bases; this problem is solved using the *Gram-Schmidt* method. Beginning with an arbitrary basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$, the Gram-Schmidt method produces an orthonormal basis. It has the special property that for each $1 \leq j \leq n$ the linear span of $\{\mathbf{u}_1, \ldots, \mathbf{u}_j\}$ is the same as the linear span of $\{\mathbf{v}_1, \ldots, \mathbf{v}_j\}$. This method is important for both theoretical and practical applications.

We describe the Gram-Schmidt method as an algorithm:

Step 1 Replace $\mathbf{u}_1$ with the vector

$$\mathbf{v}_1 = \frac{\mathbf{u}_1}{\|\mathbf{u}_1\|_2}.$$

The span of $\mathbf{v}_1$ clearly agrees with that of $\mathbf{u}_1$.

Step 2 For a $1 \leq j < n$ suppose that we have found orthonormal vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_j\}$ such that the linear span of $\{\mathbf{v}_1, \ldots, \mathbf{v}_j\}$ is the same as that of $\{\mathbf{u}_1, \ldots, \mathbf{u}_j\}$. Set

$$\mathbf{v}'_{j+1} = \mathbf{u}_{j+1} + \sum_{k=1}^{j} \alpha_k \mathbf{v}_k$$

where $\alpha_k = -\langle \mathbf{u}_{j+1}, \mathbf{v}_k \rangle$. An calculation shows that

$$\langle \mathbf{v}'_{j+1}, \mathbf{v}_k \rangle = 0 \text{ for } 1 \leq k \leq j.$$

Step 3 Since $\{\mathbf{u}_i\}$ is a basis and $\{\mathbf{v}_1, \ldots, \mathbf{v}_j\}$ are in the linear span of $\{\mathbf{u}_1, \ldots, \mathbf{u}_j\}$ it follows that $\mathbf{v}'_{j+1} \neq 0$, thus we can set

$$\mathbf{v}_{j+1} = \frac{\mathbf{v}'_{j+1}}{\|\mathbf{v}'_{j+1}\|_2}.$$

Step 4 If $j = n$ we are done otherwise return to Step 2.

This algorithms shows that there are many orthonormal bases.

Linear functions and the inner product

The inner product also gives a way to represent linear functions. A vector $\mathbf{y} \in \mathbb{R}^n$ defines a linear function $l_{\mathbf{y}}$ by the rule

$$l_{\mathbf{y}}(\mathbf{x}) = \langle \mathbf{x}, \mathbf{y} \rangle. \tag{A.33}$$

Proposition A.2.5. *The linear function $l_{\mathbf{y}}$ is zero if and only if $\mathbf{y} = \mathbf{0}$. moreover*

$$l_{\mathbf{y}_1 + \mathbf{y}_2} = l_{\mathbf{y}_1} + l_{\mathbf{y}_2}$$

and if $a \in \mathbb{R}$ then $l_{a\mathbf{y}_1} = a l_{\mathbf{y}_1}$.

The proposition shows that the map $\mathbf{y} \mapsto l_{\mathbf{y}}$ defines an isomorphism between $\mathbb{R}^n$ and $(\mathbb{R}^n)'$. The map is clearly linear; because $l_{\mathbf{y}} = 0$ if and only if $\mathbf{y} = 0$ it follows (from (A.36) below) that the image of the map is all of $(\mathbb{R}^n)'$. In other words, every linear function on $\mathbb{R}^n$ has a representation as $l_{\mathbf{y}}$ for a unique $\mathbf{y} \in \mathbb{R}^n$.

Let $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation, if $\mathbf{y} \in \mathbb{R}^m$ then $\mathbf{x} \mapsto \langle \mathbf{A}\mathbf{x}, \mathbf{y} \rangle$ is a linear function on $\mathbb{R}^n$. This means that there is a vector $\mathbf{z} \in \mathbb{R}^n$ such that

$$\langle \mathbf{A}\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle, \text{ for all } \mathbf{x} \in \mathbb{R}^n.$$

We denote this vector by $\mathbf{A}^t\mathbf{y}$. It is not difficult to show that the map $\mathbf{y} \mapsto \mathbf{A}^t\mathbf{y}$ is a linear transformation from $\mathbb{R}^m$ to $\mathbb{R}^n$.

Proposition A.2.6. *If* $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ *has matrix* (a_{ij}) *with respect to the standard bases then* $\mathbf{A}^t$ *has matrix* (a_{ji}) *with respect to the standard bases,*

$$(\mathbf{A}^t\mathbf{y})_i = \sum_{j=1}^{m} a_{ji}y_j.$$

The linear transformation from $\mathbf{A}^t : \mathbb{R}^m \to \mathbb{R}^n$ is called the *transpose* (or *adjoint*) of $\mathbf{A}$. Note that while the matrices representing $\mathbf{A}$ and its transpose are simply related, the transpose is defined without reference to a particular basis, for by definition

$$\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle_m = \langle \mathbf{x}, \mathbf{A}^t\mathbf{y}\rangle_n$$

for all $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. In order to avoid confusion, we have used $\langle \cdot, \cdot \rangle_n$ (resp. $\langle \cdot, \cdot \rangle_m$) to denote the inner product on $\mathbb{R}^n$ (resp. $\mathbb{R}^m$).

We close this section by placing these considerations in a slightly more abstract framework.

Definition A.2.16. Let V be a vector space, a function $b : V \times V \to \mathbb{R}$ which satisfies the conditions $b(\mathbf{v}, \mathbf{v}) \geq 0$ with $b(\mathbf{v}, \mathbf{v}) = 0$ if and only if $\mathbf{v} = 0$ and for all $\mathbf{v}, \mathbf{w}, \mathbf{z} \in V$ and $a \in \mathbb{R}$

$$\begin{aligned} b(\mathbf{v}, \mathbf{w}) &= b(\mathbf{w}, \mathbf{v}), \\ b(\mathbf{v} + \mathbf{w}, \mathbf{z}) &= b(\mathbf{v}, \mathbf{z}) + b(\mathbf{w}, \mathbf{z}) \text{ and} \\ b(a\mathbf{v}, \mathbf{w}) &= ab(\mathbf{v}, \mathbf{w}) = b(\mathbf{v}, a\mathbf{w}) \end{aligned} \tag{A.34}$$

defines an **inner product** on V. A function with the properties in (A.34) is called a **bilinear function**.

Example A.2.16. Let $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^n$ be a linear transformation with $\ker \mathbf{A} = \{\mathbf{0}\}$ then

$$\langle \mathbf{x}, \mathbf{y}\rangle_{\mathcal{A}} = \langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{y}\rangle$$

defines an inner product on $\mathbb{R}^n$.

Example A.2.17. Let $\mathcal{P}_n$ be the real valued polynomials of degree at most n then

$$b_n(p, q) = \int_{-1}^{1} p(x)q(x)dx$$

defines an inner product on $\mathcal{P}_n$.

Exercise A.2.26. Let b be an inner product on a vector space V. Using the same idea as used above to prove the Cauchy-Schwarz inequality, show that

$$|b(\mathbf{v}_1, \mathbf{v}_2)| \leq \sqrt{b(\mathbf{v}_1, \mathbf{v}_1)b(\mathbf{v}_2, \mathbf{v}_2)}.$$

Exercise A.2.27. Show that the Gram-Schmidt procedure can be applied to an arbitrary vector space with an inner product.

Exercise A.2.28. Apply the Gram-Schmidt process to the basis $\{1, x, x^2\}$ with the inner product given in example A.2.17 to find an orthonormal basis for $\mathcal{P}_2$.

Exercise A.2.29. Prove Proposition A.2.4.

Exercise A.2.30. If $\mathbf{a}$ is an $m \times n$-matrix and $\mathbf{x} \in \mathbb{R}^n$ then we can use the inner product to express the matrix product $\mathbf{ax}$. Show that if we write $\mathbf{a}$ in terms of its rows

$$\mathbf{a} = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_m \end{pmatrix}$$

then

$$\mathbf{ax} = \begin{pmatrix} \langle \mathbf{a}_1, \mathbf{x} \rangle \\ \vdots \\ \langle \mathbf{a}_m, \mathbf{x} \rangle \end{pmatrix}. \tag{A.35}$$

Exercise A.2.31. Calculus can be used to proved (A.29). Let $\mathbf{x}$ and $\mathbf{y}$ be vectors in $\mathbb{R}^n$ and define the function

$$f(t) = \langle \mathbf{x} + t\mathbf{y}, \mathbf{x} + t\mathbf{y} \rangle = \|\mathbf{x} + t\mathbf{y}\|_2^2.$$

This function satisfies $f(t) \geq 0$ for all $t \in \mathbb{R}$. Use calculus to locate the value of t where f assumes it minimum. By evaluating f at its minimum and using the fact that $f(t) \geq 0$ show that (A.29) holds.

Exercise A.2.32. Prove formula (A.31).

Exercise A.2.33. Prove Proposition A.2.5.

Exercise A.2.34. Prove Proposition A.2.6.

Exercise A.2.35. Suppose that $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{B} : \mathbb{R}^m \to \mathbb{R}^l$ show that

$$(\mathbf{B} \circ \mathbf{A})^t = \mathbf{A}^t \circ \mathbf{B}^t.$$

Express this relation in terms of the matrices for these transformations with respect to the standard bases.

Exercise A.2.36. Show that $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}}$ is an inner product. Why do we need to assume that $\ker \mathbf{A} = \{0\}$?

Exercise A.2.37. Prove that b_n defined in example A.2.17 is an inner product.

A.2.6 Linear transformations and linear equations

Linear transformations give a geometric way to think about linear equations. Suppose that $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation. The kernel of $\mathbf{A}$ is nothing more that the set of solutions to the equation

$$\mathbf{Ax} = \mathbf{0}.$$

This is sometimes called the *homogeneous equation*. The system of equations

$$\mathbf{Ax} = \mathbf{y}$$

has a solution if and only if $\mathbf{y}$ belongs to the image of $\mathbf{A}$. Theorem 1.3.2 relates the dimensions of the kernel and image of a linear transformation $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$; they satisfy the relation

$$\dim \ker \mathbf{A} + \dim \operatorname{Im} \mathbf{A} = n. \tag{A.36}$$

If $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^n$ and $\dim \ker \mathbf{A} = 0$, then formula (A.36) implies that $\dim \operatorname{Im} \mathbf{A} = n$ and therefore for every $\mathbf{y} \in \mathbb{R}^n$ there is a *unique* $\mathbf{x} \in \mathbb{R}^n$ such that

$$\mathbf{Ax} = \mathbf{y}.$$

A linear transformation with this property is called **invertible,** we let $\mathbf{A}^{-1}$ denote the *inverse* of $\mathbf{A}$. It is also a linear transformation. A linear transformation and its inverse satisfy the relations

$$\mathbf{A}^{-1}\mathbf{A} = \mathrm{Id}_n = \mathbf{A}\mathbf{A}^{-1}.$$

If (a_{ij}) is the matrix of $\mathbf{A}$ with respect to a basis and (b_{ij}) is the matrix for $\mathbf{A}^{-1}$ then these relations imply that

$$\sum_{j=1}^{n} b_{ij}a_{jk} = \delta_{ik} = \sum_{j=1}^{n} a_{ij}b_{jk}. \tag{A.37}$$

From a purely mathematical standpoint the problem of solving the linear equation $\mathbf{Ax} = \mathbf{y}$ is simply a matter of computing $\mathbf{A}^{-1}\mathbf{y}$. Cramer's rule gives an explicit formula for $\mathbf{A}^{-1}$, though it is very unusual to solve linear equations this way. The direct computation of $\mathbf{A}^{-1}$ is enormously expensive, computationally and also unstable. Less direct, computationally more stable and efficient methods are usually employed.

Definition A.2.17. An $n \times n$ matrix (a_{ij}) is called *upper triangular* if

$$a_{ij} = 0 \text{ if } j < i.$$

A system of equations is upper triangular if its matrix of coefficients is upper triangular.

Upper triangular systems are very easy to solve. Suppose that (a_{ij}) is an upper triangular matrix with all of its diagonal entries $\{a_{ii}\}$ non-zero. The system of equations $\mathbf{ax} = \mathbf{y}$ becomes

$$\sum_{j=i}^{n} a_{ij}x_j = y_i \text{ for } i = 1, \ldots, n.$$

It is easily solved using the *back substitution* algorithm:

Step 1 Let $x_n = a_{nn}^{-1} y_n$.

Step 2 For a $1 < j < n$ assume we know $(x_{j+1}, \ldots, x_n)$ and let

$$x_{j+1} = \frac{y_{j+1} - \sum_{k=j+1}^{n} a_{(j+1)k} x_k}{a_{(j+1)(j+1)}}.$$

Step 3 If $j = n$ we are done otherwise return to step 2.

Another important class of matrices has orthonormal rows (and columns).

Definition A.2.18. A matrix $\mathbf{a} = (a_{ij})$ is **orthogonal** if

$$\sum_{j=1}^{n} a_{ij} a_{kj} = \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{if } i \neq k. \end{cases}$$

In terms of matrix multiplication this condition is expressed by

$$\mathbf{a}\mathbf{a}^t = \mathrm{Id}_n = \mathbf{a}^t\mathbf{a}.$$

Hence a matrix is orthogonal if $\mathbf{a}^t$ is the inverse of $\mathbf{a}$.

Let $\mathbf{a}$ be an orthogonal matrix and let $\{\mathbf{a}_j \mid i = 1, \ldots n\}$ denote its columns thought of $n \times 1$-vectors. The solution to the equation $\mathbf{a}\mathbf{x} = \mathbf{y}$ is given by

$$x_j = \langle \mathbf{a}_j, \mathbf{y} \rangle \text{ for } j = 1, \ldots, n.$$

We have found two classes of linear equations which are computationally simple to solve. Using the Gram-Schmidt algorithm one can prove the following statement.

Theorem A.2.2. *Suppose that* $\mathbf{a}$ *is an invertible* $n \times n$ *matrix, then there exists an upper triangular matrix* $\mathbf{r}$ *and an orthogonal matrix* $\mathbf{q}$ *such that*

$$\mathbf{a} = \mathbf{q}\mathbf{r}.$$

Once a matrix is expressed in this form, the system of equations $\mathbf{a}\mathbf{x} = \mathbf{y}$ is easily solved in two steps: multiplying by $\mathbf{q}^t$ gives the upper triangular system $\mathbf{r}\mathbf{x} = \mathbf{q}^t\mathbf{y}$ which is then solved by back substitution. There is an enormous literature devoted to practical implementations of this and similar results. A good starting point for further study is [78].

Exercise A.2.38. Show that if a_{ij} is an upper triangular matrix with $a_{ii} = 0$ for some i then there is a non-zero vector $(x_1, \ldots, x_n)$ such that

$$\sum_{j=i}^{n} a_{ij} x_j = 0.$$

In other words the homogeneous equation has a non-trivial solution.

Exercise A.2.39. Let $\mathbf{a}$ be an invertible upper triangular matrix, show that $\mathbf{a}^{-1}$ is also upper triangular.

Exercise A.2.40. Show that if $\mathbf{a}$ and $\mathbf{b}$ are upper triangular matrices then so it $\mathbf{a}\mathbf{b}$.

Exercise A.2.41. Prove Theorem A.2.2.

A.2.7 Linear algebra with uncertainties

In applications of linear algebra a vector $\mathbf{x}$ often represents the (unknown) state of a system, a matrix $\mathbf{a}$ models a measurement process and

$$\mathbf{y} = \mathbf{a}\mathbf{x}$$

are the (known) results of the measurements. The simple minded problem is then to solve this system of linear equations. In reality things are more involved. The model for the measurements is only an approximation and therefore it is perhaps more reasonable to think of the measurement matrix as $\mathbf{a} + \delta\mathbf{a}$. Here $\delta\mathbf{a}$ represents an aggregation of errors in the model. The measurements are themselves subject to error and therefore should also be considered to have the form $\mathbf{y} + \delta\mathbf{y}$. A more realistic problem is therefore to solve the system of equations

$$(\mathbf{a} + \delta\mathbf{a})\mathbf{x} = \mathbf{y} + \delta\mathbf{y}. \tag{A.38}$$

But what does this mean?

We consider only the simplest case where $\mathbf{a}$ is an $n \times n$, invertible matrix. Let $\|\cdot\|$ denote a norm on $\mathbb{R}^n$ and $|\!|\!|\cdot|\!|\!|$ the operator norm defined as in (A.25). Suppose that we can bound the uncertainty in both the model and the measurements in the sense that we have constants $\epsilon > 0$ and $\eta > 0$ such that

$$\|\delta\mathbf{y}\| < \epsilon \text{ and } |\!|\!|\delta\mathbf{a}|\!|\!| < \eta.$$

In the absence of more detailed information about the systematic errors, "the solution" to (A.38) should be defined as the set of vectors

$$\{\mathbf{x} : |\ (\mathbf{a} + \delta\mathbf{a})\mathbf{x} = \mathbf{y} + \delta\mathbf{y} \text{ for some choice of } \delta\mathbf{a}, \delta\mathbf{y} \text{ with } \|\delta\mathbf{y}\| < \epsilon, |\!|\!|\delta\mathbf{a}|\!|\!| < \eta\}.$$

This is a little cumbersome. In practice one finds a vector which satisfies

$$\mathbf{a}\mathbf{x} = \mathbf{y}$$

and a bound for the error one makes in asserting that the actual state of the system is $\mathbf{x}$.

To proceed with this analysis we assume that all the possible model matrices, $\mathbf{a} + \delta\mathbf{a}$ are invertible. If $|\!|\!|\delta\mathbf{a}|\!|\!|$ is sufficiently small then this condition is satisfied. As $\mathbf{a}$ is invertible the number

$$\mu = \min_{\mathbf{x}\neq 0} \frac{\|\mathbf{a}\mathbf{x}\|}{\|\mathbf{x}\|}$$

is a positive. If $|\!|\!|\delta\mathbf{a}|\!|\!| < \mu$ then $\mathbf{a} + \delta\mathbf{a}$ is also invertible. If it were not then there would be a vector $\mathbf{v} \neq 0$ such that

$$(\mathbf{a} + \delta\mathbf{a})\mathbf{v} = 0.$$

Because $|\!|\!|\cdot|\!|\!|$ is an operator norm we can use (A.26) and the triangle inequality to see that

$$\begin{aligned} 0 = \|(\mathbf{a} + \delta\mathbf{a})\mathbf{v}\| &\geq \|\mathbf{a}\mathbf{v}\| - \|\delta\mathbf{a}\mathbf{v}\| \\ &\geq \mu\|\mathbf{v}\| - |\!|\!|\delta\mathbf{a}|\!|\!|\|\mathbf{v}\| \\ &\geq (\mu - |\!|\!|\delta\mathbf{a}|\!|\!|)\|\mathbf{v}\|. \end{aligned} \tag{A.39}$$

Because $(\mu - \|\delta\mathbf{a}\|)$ and $\mathbf{v}$ were assumed to be positive, the first and last lines are in contradiction. This shows that if $\mathbf{a}$ is an invertible matrix then so is $\mathbf{a}+\delta\mathbf{a}$, for *sufficiently small* perturbations $\delta\mathbf{a}$. Note that the definition of small depends on $\mathbf{a}$. In the remainder of this discussion we assume that η, the bound on the uncertainty in the model is smaller than μ.

An estimate on the error in $\mathbf{x}$ is found in two steps. First, fix the model and consider only errors in measurement. Suppose that $\mathbf{a}\mathbf{x} = \mathbf{y}$ and $\mathbf{a}(\mathbf{x}+\delta\mathbf{x}) = \mathbf{y}+\delta\mathbf{y}$. Taking the difference of these two equations gives

$$\mathbf{a}\delta\mathbf{x} = \delta\mathbf{y}$$

and therefore $\delta\mathbf{x} = \mathbf{a}^{-1}\delta\mathbf{y}$. Using (A.26) again we see that

$$\|\delta\mathbf{x}\| \leq \|\mathbf{a}^{-1}\|\|\delta\mathbf{y}\|.$$

This is a bound on the absolute error; it is more meaningful to bound the relative error $\|\delta\mathbf{x}\|/\|\mathbf{x}\|$. To that end observe that

$$\|\mathbf{y}\| \leq \|\mathbf{a}\|\|\mathbf{x}\|$$

and therefore

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|\mathbf{a}\|\|\mathbf{a}^{-1}\|\frac{\|\delta\mathbf{y}\|}{\|\mathbf{y}\|}. \tag{A.40}$$

This is a very useful estimate: it estimates the relative uncertainty in the state in terms of the relative uncertainty in the measurements. The coefficient

$$c_{\mathbf{a}} = \|\mathbf{a}\|\|\mathbf{a}^{-1}\| \tag{A.41}$$

is called the *condition number* of the matrix $\mathbf{a}$. It is very useful measure of the stability of a model of this type.

To complete our analysis we need to incorporate errors in the model. Suppose that $\mathbf{x}+\delta\mathbf{x}$ solves

$$(\mathbf{a}+\delta\mathbf{a})(\mathbf{x}+\delta\mathbf{x}) = \mathbf{y}+\delta\mathbf{y}.$$

Subtracting this from $\mathbf{a}\mathbf{x} = \mathbf{y}$ gives

$$(\mathbf{a}+\delta\mathbf{a})\delta\mathbf{x} = \delta\mathbf{y} - \delta\mathbf{a}\mathbf{x}.$$

Proceeding as before we see that

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|(\mathbf{a}+\delta\mathbf{a})^{-1}\|\|\mathbf{a}\|\frac{\|\delta\mathbf{y}\|}{\|\mathbf{y}\|} + \|(\mathbf{a}+\delta\mathbf{a})^{-1}\|\|\delta\mathbf{a}\|. \tag{A.42}$$

If $\delta\mathbf{a}$ is very small (relative to μ) then

$$(\mathbf{a}+\delta\mathbf{a})^{-1} = \mathbf{a}^{-1} - \mathbf{a}^{-1}\delta\mathbf{a}\mathbf{a}^{-1} + O([\delta\mathbf{a}]^2).$$

The triangle inequality implies that

$$\|(\mathbf{a}+\delta\mathbf{a})^{-1}\| \lessapprox \|\mathbf{a}^{-1}\| + \|\mathbf{a}^{-1}\delta\mathbf{a}\mathbf{a}^{-1}\|.$$

Ignoring quadratic error terms this gives the estimate

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq c_{\mathbf{a}} \left[\frac{\|\delta\mathbf{y}\|}{\|\mathbf{y}\|} + \frac{|||\delta\mathbf{a}|||}{|||\mathbf{a}|||} \right]. \tag{A.43}$$

Once again, it is the condition number of $\mathbf{a}$ which relates the relative error in the predicted state to the relative errors in the model and measurements.

This analysis considers a very special case, but it indicates how gross features of the model constrain the accuracy of its predictions. We have discussed neither the effects of using a particular algorithm to solve the system of equations or round-off error, a similar analysis applies to study these problems. A very good reference for this material is [78].

Exercise A.2.42. Show that the condition number is given by the following ratio

$$c_{\mathbf{a}} = \frac{\max_{\mathbf{x}\neq 0} \frac{\|\mathbf{ax}\|}{\|\mathbf{x}\|}}{\min_{\mathbf{x}\neq 0} \frac{\|\mathbf{ax}\|}{\|\mathbf{x}\|}}. \tag{A.44}$$

This shows that the condition number of any matrix is at least 1.

A.2.8 The least squares method

As described above, one often considers over-determined systems of linear equations. These are of the form

$$\mathbf{ax} = \mathbf{y}$$

where $\mathbf{a}$ is an $m \times n$-matrix with $m > n$. If $\mathbf{x}$ describes the state of a physical system and $\mathbf{a}$ models a measurement process it is reasonable to assume that the columns of this matrix are linearly independent. If this is false then our model has redundant state variables in the sense that there are distinct states which our measurement process cannot distinguish. More concretely there are vectors $\mathbf{x}_1 \neq \mathbf{x}_2$ such that

$$\mathbf{ax}_1 = \mathbf{ax}_2.$$

In an realistic situation the measurements $\mathbf{y}$ contain errors and this means that one does not expect to find an exact solution to this system of equations. A very common method for associating a particular state to a given set of real measurements is to look for the vector $\bar{\mathbf{x}}$ such that

$$\|\mathbf{a}\bar{\mathbf{x}} - \mathbf{y}\|_2 = \min\{\|\mathbf{ax} - \mathbf{y}\|_2 \mid \mathbf{x} \in \mathbb{R}^n\}. \tag{A.45}$$

Under the assumption that the columns of $\mathbf{a}$ are linearly independent, it is not difficult to show that the vector $\bar{\mathbf{x}}$ is uniquely determined by (A.45). This vector is called the least squares solution.

Suppose that $\bar{\mathbf{x}}$ is the least squares and $\mathbf{v}$ is any other vector in $\mathbb{R}^n$. The function

$$f(t) = \|\mathbf{a}(\bar{\mathbf{x}} + t\mathbf{v}) - \mathbf{y}\|_2^2 = < (\mathbf{a}(\bar{\mathbf{x}} + t\mathbf{v}) - \mathbf{y}), \mathbf{a}(\bar{\mathbf{x}} + t\mathbf{v}) - \mathbf{y}) >$$

assumes it minimum value at $t = 0$; therefore $f'(0) = 0$. Using the bilinearity and symmetry of the inner product we see that

$$f(t) = \|\mathbf{a}\bar{\mathbf{x}} - \mathbf{y}\|_2^2 + 2t\langle \mathbf{av}, (\mathbf{ax} - \mathbf{y})\rangle + t^2\langle \mathbf{av}, \mathbf{av}\rangle.$$

Computing $f'(0)$ we find

$$0 = f'(0) = 2\langle \mathbf{av}, (\mathbf{a}\bar{\mathbf{x}} - \mathbf{y})\rangle = 2\langle \mathbf{v}, (\mathbf{a}^t\mathbf{a}\bar{\mathbf{x}} - \mathbf{a}^t\mathbf{y})\rangle.$$

Since this must hold for all vectors $\mathbf{v} \in \mathbb{R}^n$ it follows that

$$\mathbf{a}^t\mathbf{ax} = \mathbf{a}^t\mathbf{y}. \tag{A.46}$$

This is called the system of *normal equations*. The matrix $\mathbf{a}^t\mathbf{a}$ is $n \times n$ and under our assumptions it is invertible. If it were not invertible then there would be a vector $\mathbf{x}_0 \neq 0$ so that $\mathbf{a}^t\mathbf{ax}_0 = \mathbf{0}$. This would imply that

$$0 = \langle \mathbf{a}^t\mathbf{ax}_0, \mathbf{x}_0\rangle = \langle \mathbf{ax}_0, \mathbf{ax}_0\rangle.$$

Hence $\mathbf{ax}_0 = \mathbf{0}$, but this means that the columns of $\mathbf{a}$ are not linearly independent. In terms of measurements, the state $\mathbf{x}_0 \neq \mathbf{0}$ cannot be distinguished from $\mathbf{0}$. This shows that the normal equations have a unique solution for any set of measurements $\mathbf{y}$. The matrix $\mathbf{a}^t\mathbf{a}$ is a special type of matrix, called a *positive definite, symmetric* matrix. This means that

$$\langle \mathbf{a}^t\mathbf{ax}, \mathbf{x}\rangle > 0 \text{ if } \mathbf{x} \neq \mathbf{0} \text{ and } (\mathbf{a}^t\mathbf{a})^t = \mathbf{a}^t\mathbf{a}.$$

There are many special algorithms for solving a system of equations with a positive definite coefficient matrix, for example steepest descent or the conjugate gradient method, see [43] or [19]. These considerations explain, in part why the Euclidean norm is usually chosen to measure the error in an over determined linear system.

A.2.9 Complex numbers and the Euclidean plane

Thus far we have only considered real numbers and vector spaces over the real numbers. While the real numbers are complete in the sense that there are no holes, they are not complete from an algebraic standpoint. There is no real number which solves the algebraic equation

$$x^2 = -1.$$

To remedy this we simply introduce a new symbol i which is *defined* by the condition that

$$i^2 = -1.$$

It is called the *imaginary unit*.

Definition A.2.19. The **complex numbers** are defined to be the collection of symbols

$$\{x + iy \mid x, y \in \mathbb{R}\}$$

with the addition operation defined by

$$(x_1 + iy_1) + (x_2 + iy_2) \stackrel{d}{=} (x_1 + x_2) + i(y_1 + y_2)$$

and multiplication defined by

$$(x_1 + iy_1)(x_2 + iy_2) \stackrel{d}{=} (x_1x_2 - y_1y_2) + i(x_1y_2 + x_2y_1).$$

The set of complex numbers is denoted by $\mathbb{C}$.

Note that addition and multiplication are defined in terms of the addition and multiplication operations on real numbers. The complex number $0 = 0 + i0$ satisfies $0 + (x + iy) = x + iy$ and the complex number $1 = 1 + i0$ satisfies $(1 + i0)(x + iy) = x + iy$. We often use the letters z or w to denote complex numbers. The sum is denoted by $z + w$ and the product by zw.

Proposition A.2.7. *The addition and multiplication defined for complex numbers satisfy the commutative, associative and distributive laws. That is, if z_1, z_2, z_3 are three complex numbers then*

$$\begin{aligned} z_1 + z_2 = z_2 + z_1 \quad & (z_1 + z_2) + z_3 = z_1 + (z_2 + z_3), \\ z_1 z_2 = z_2 z_1 \quad & (z_1 z_2) z_3 = z_1 (z_2 z_3), \\ & z_1(z_2 + z_3) = z_1 z_2 + z_1 z_3. \end{aligned} \tag{A.47}$$

Definition A.2.20. If $z = x + iy$ then the *real part* of z is x and the *imaginary part* of z is y. These functions are written symbolically as

$$\operatorname{Re} z = x, \quad \operatorname{Im} z = y.$$

The set underlying the complex numbers is $\mathbb{R}^2$,

$$x + iy \leftrightarrow (x, y);$$

addition of complex numbers is the same as vector addition. Often the set of complex numbers is called the *complex plane.* The Euclidean norm is used to define the *absolute value* of a complex number

$$|x + iy| = \sqrt{x^2 + y^2}.$$

There is a very useful operation defined on complex numbers called *complex conjugation.* If $z = x + iy$ then its complex conjugate is $\bar{z} = x - iy$.

Exercise A.2.44 shows that the multiplication defined above has all of the expected properties of a product. The real numbers sit inside the complex numbers as the set $\{z \in \mathbb{C} \mid \operatorname{Im} z = 0\}$. It is easy to see that the two definitions of addition and multiplication agree on this subset. Complex conjugation is simply reflection across the real axis. The problem we were trying to solve by introducing i was that of solving polynomial equations. In this regard the complex numbers are a complete success.

Theorem A.2.3 (The Fundamental Theorem of Algebra). *If $p(z)$ is a polynomial of degree n with complex coefficients, i.e.*

$$p(z) = z^n + a_{n-1} z^{n-1} + \ldots a_1 z + a_0, \text{ for } a_j \in \mathbb{C}, j = 0, \ldots, n-1$$

then there are n complex numbers $\{z_1, \ldots, z_n\}$ such that

$$p(z_j) = 0 \text{ for } j = 1, \ldots, n.$$

In other words there is no possibility of further extending the concept of number by solving polynomial equations. The main structural difference between the real and complex numbers is that the real numbers have a natural order relation and the complex numbers

do not. If x and y are real numbers then we say that $x < y$ if $y - x > 0$. This relation has many familiar properties: if $x < y$ and s is another real number then $x + s < y + s$; if $s > 0$ then $xs < ys$ as well. In other words, the order relation is compatible with the arithmetic operations. It is not difficult to show that no such compatible order relation exists on the complex numbers.

It is useful to understand the multiplication of complex numbers geometrically. For this purpose we represent points in the complex plane using polar coordinates. The radial coordinate is simply $r = |z|$. The ratio $\omega = z|z|^{-1}$ is a number of unit length and therefore has a representation as $\omega = \cos\theta + i\sin\theta$ so that

$$z = r(\cos\theta + i\sin\theta) \tag{A.48}$$

The angle θ is called the *argument* of z, which is denoted by $\arg(z)$. It is only determined up to multiples of 2π. If z and w are two complex numbers then they can be expressed in polar form as

$$z = r(\cos\theta + i\sin\theta), \quad w = \rho(\cos\phi + i\sin\phi).$$

Computing their product we find that

$$\begin{aligned} zw =& r\rho([\cos\theta\cos\phi - \sin\theta\sin\phi] + i(\cos\theta\sin\phi + \sin\theta\cos\phi]) \\ =& r\rho(\cos(\theta+\phi) + i\sin(\theta+\phi)). \end{aligned} \tag{A.49}$$

In the second line we used the sum formulæ for the sine and cosine. This shows us that complex multiplication of w by z can be understood geometrically as scaling the length of w by $|z|$ and rotating it in the plane through an angle $\arg(z)$.

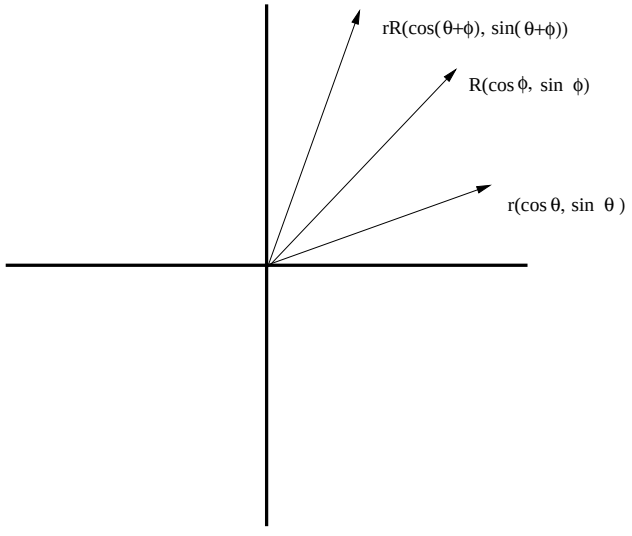


Figure A.1: Multiplication of complex numbers

Using the notion of distance on $\mathbb{C}$ defined above we can define the concept of convergence for sequences of complex numbers.

Definition A.2.21. Let $< z_n >$ be a sequence of complex numbers. The sequence converges to z^* if

$$\lim_{n\to\infty} |z_n - z^*| = 0.$$

In this case case z^* is called the limit of $< z_n >$ and we write

$$\lim_{n\to\infty} z_n = z^*.$$

Exercise A.2.43. Prove Proposition A.2.7.

Exercise A.2.44. Let z and w be complex numbers show that

$$\begin{gathered}\overline{(z+w)} = \bar{z} + \bar{w}, \quad \overline{zw} = \bar{z}\bar{w}, \\ z\bar{z} = |z|^2.\end{gathered} \tag{A.50}$$

Using the second condition show that if $z \neq 0$ then the complex number defined by

$$z^{-1} = \frac{\bar{z}}{|z|^2}$$

satisfies $zz^{-1} = 1 = z^{-1}z$.

Exercise A.2.45. Let $< z_n >$ be a sequence of complex numbers. Show that z_n converges to z^* if and only if $\operatorname{Re} z_n$ converges to $\operatorname{Re} z^*$ and $\operatorname{Im} z_n$ converges to $\operatorname{Im} z^*$.

A.2.10 Complex vector spaces

The collection of ordered n-tuples of complex numbers is denoted by $\mathbb{C}^n$. It is a vector space with

$$(z_1, \ldots, z_n) + (w_1, \ldots, w_n) = (z_1 + w_1, \ldots, z_n + w_n).$$

Since the entries of the vectors are now complex numbers we can define scalar multiplication for $w \in \mathbb{C}$ by

$$w \cdot (z_1, \ldots, z_n) = (wz_1, \ldots, wz_n).$$

For this reason $\mathbb{C}^n$ is called a *complex vector space*. The Euclidean norm on $\mathbb{C}^n$ is defined by

$$\|(z_1, \ldots, z_n)\|_2 = \sqrt{\sum_{j=1}^{n} |z_j|^2}.$$

An inner product is defined by setting

$$\langle (z_1, \ldots, z_n), (w_1, \ldots, w_n) \rangle = \sum_{j=1}^{n} z_j \bar{w}_j.$$

We let $\mathbf{w}$ and $\mathbf{z}$ denote vectors in $\mathbb{C}^n$. One easily sees that $\|\mathbf{z}\|_2^2 = \langle \mathbf{z}, \mathbf{z} \rangle$. It is also not difficult to prove the Cauchy-Schwarz inequality

$$|\langle \mathbf{z}, \mathbf{w} \rangle| \leq \|\mathbf{z}\|_2 \|\mathbf{w}\|_2,$$

keeping in mind that $|\cdot|$ is the absolute value of a complex number.

Proposition A.2.8. *If $w \in \mathbb{C}$ and $\mathbf{z} \in \mathbb{C}^n$ then $\|w\mathbf{z}\| = |w|\|\mathbf{z}\|$. Moreover, if $\mathbf{z}, \mathbf{w} \in \mathbb{C}^n$ then*

$$\langle \mathbf{z}, \mathbf{w} \rangle = \overline{\langle \mathbf{w}, \mathbf{z} \rangle} \tag{A.51}$$

and if $w \in \mathbb{C}$ then

$$\langle w\mathbf{z}, \mathbf{w} \rangle = w\langle \mathbf{z}, \mathbf{w} \rangle \text{ and } \langle \mathbf{z}, w\mathbf{w} \rangle = \bar{w}\langle \mathbf{z}, \mathbf{w} \rangle. \tag{A.52}$$

The Euclidean inner product on $\mathbb{C}^n$ is not symmetric, but rather *hermitian symmetric*.

The theory of complex vector spaces is very similar to that of real vector spaces. The concepts of linear functions, transformations, bases, matrices and matrix multiplication all carry over without change. One simply allows the scalars to be complex numbers. For example we consider the linear functions from $\mathbb{C}^n$ to $\mathbb{C}$. As before every such function has a unique representation as $l_{\mathbf{w}}(\mathbf{z}) = \langle \mathbf{z}, \mathbf{w} \rangle$. As is evident from (A.51) and (A.52) some small differences arise when dealing with inner products on complex vector spaces.

Example A.2.18. The space $\mathbb{C}^n$ can be regarded as either a real on complex vector space. If it is thought of as a complex vector space then the vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ are a basis. Thinking of it as a real vector space means that we only allow scalar multiplication by real numbers. In this case n-vectors do not suffice to define a basis. Instead one could use

$$\{\mathbf{e}_1, \ldots, \mathbf{e}_n, i\mathbf{e}_1, \ldots, \mathbf{e}_n\}.$$

Example A.2.19. The set of continuous, complex valued functions defined on $[0, 1]$ is a complex vector space.

Exercise A.2.46. Prove Proposition A.2.8.

Exercise A.2.47. Prove that every linear function on $\mathbb{C}^n$ has a representation as $l_{\mathbf{w}}$ for a unique $\mathbf{w} \in \mathbb{C}^n$. Explain how to find $\mathbf{w}$.

A.3 Functions, theory and practice

The idea of a function is familiar from calculus. A real valued function on $\mathbb{R}$ is a *rule* for assigning to each $x \in \mathbb{R}$ a unique value $y \in \mathbb{R}$. Usually we write something like $y = f(x)$. In this context what is meant by a "rule?" The simplest functions are described by explicit formulæ involving a variable and arithmetic operations. For example

$$\begin{aligned} f_1(x) &= 1, \\ f_2(x) &= 2 + x^3, \\ f_3(x) &= \frac{7 + 3x + 6x^3 + 17x^9}{3 + 4x^2 + 5x^4}. \end{aligned} \tag{A.53}$$

The functions one gets this way are called *rational functions*, these are functions that can be expressed as ratios of polynomials. These functions have the following considerable virtue: if we "know" what x is and we can "do" arithmetic then we can actually compute (in finite time) the value of $f(x)$ for any given rational function f. No *infinite processes* are required to evaluate a rational function. This is a concept we will consider in some detail so we give it a name.

Definition A.3.1. A function f is a *real computable function* if its value can be determined for any $x \in \mathbb{R}$ by doing a finite number of *feasible* operations. We refer to these functions as **computable functions**, for more on this concept see [85].

What are the "feasible operations?" Provisionally the feasible operations are those which only require the ability to do arithmetic and to determine if a number is non-negative. These are the operations which can be done *approximately* by a computer. One can give an analogous definition for computable functions defined on $\mathbb{R}^n$ or on subsets of $\mathbb{R}^n$, $n \geq 1$.

Rational functions are evidently computable functions, but there are other types of computable functions. If $[a, b] \subset \mathbb{R}$ is an *interval*, that is

$$[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$$

then we define the characteristic function of an interval by the rule

$$\chi_{[a,b]}(x) = \begin{cases} 1 \text{ if } & x \in [a, b], \\ 0 \text{ if } & x \notin [a, b]. \end{cases}$$

Again if we know the exact value of x then to compute $\chi_{[a,b]}(x)$ we only need to perform feasible operations to check if $0 \leq x - a$ and $0 \leq b - x$.

Proposition A.3.1. *Suppose that $f(x)$ and $g(x)$ are two computable functions then $f + g, fg, f - g, f/g$ and $f \circ g$ are also computable functions.*

The set of computable functions is, in essence the set of functions that are actually available for computational purposes. They are the functional analogue of floating point numbers. However it is very easy to define functions, quite explicitly which do not fall into this class. The function $f(x) = x^3$ is a computable function and it is one-to-one and onto. That is $f(x_1) = f(x_2)$ implies that $x_1 = x_2$. For every $y \in \mathbb{R}$ there is an x (evidently unique) so that $f(x) = y$. This means there is a function $g(y)$ which inverts $f(x)$, that is $g(f(x)) = x$. Of course this is just the cube root function. Much less evident is how to compute $g(y)$ for an arbitrary value of y.

A function can also be defined implicitly via a functional relation. For example we think of y as a function of x defined by the computable relation

$$x^2 + y^2 = 1.$$

Evaluating y as a function of x entails solving this equation, formally we can write

$$y_{\pm}(x) = \pm\sqrt{1 - x^2}.$$

The relation actually defines two functions, which is not a serious difficulty, however to compute either $y_+(x)$ or $y_-(x)$ requires the ability to calculate a square root. In this case there is trick which effectively avoids the computation of the square root. If

$$x(t) = \frac{1 - t^2}{1 + t^2} \text{ and } y(t) = \frac{2t}{1 + t^2}$$

then

$$x(t)^2 + y(t)^2 = 1.$$

Both of the functions $x(t)$ and $y(t)$ are computable and so we see that, at the expense of expressing both x and y in terms of an auxiliary variable, t we are able to solve $x^2+y^2=1$. For only slightly more complicated equations in two variables it is known that no such trick exists. So the problem of solving non-linear equations in one or several variables leads to non- computable functions.

Probably the most important examples of non- computable functions are the solutions of linear, ordinary differential equations. For example the $\sin x$, $\cos x$, $\exp x$ all arise in this context as well as the Bessel functions, Legendre functions, etc. Such functions are called transcendental functions. For many purposes these functions are regarded as completely innocuous. They are however not computable, except for very special values of x. The reason that these functions are not greeted with horror is that they are all well approximated by computable functions in a precise sense: For each of these functions there are computable approximations and estimates for the differences between the actual functions and their approximations. As machine computation is always approximate, it is not necessary (or even possible) to evaluate functions exactly. It is only necessary to be able to evaluate functions to within a *specified* error.

Exercise A.3.1. Prove Proposition A.3.1.

Exercise A.3.2. Give a definition for computable functions of several variables. Show that linear functions are always computable functions. Show moreover that the solution $\mathbf{x}$ of a system of linear equations

$$\mathbf{ax}=\mathbf{y}$$

is a computable function of $\mathbf{y}$.

A.3.1 Infinite series

Most of the functions that one encounters can be represented as infinite sums. Elementary arithmetic defines any finite sum of numbers but a further definition is needed to define an infinite sum. This is clear because not every infinite sum makes sense, for example what value should assigned to the sums,

$$\sum_{j=1}^{\infty}(-1)^j \text{ or } \sum_{j=1}^{\infty}\frac{1}{j} \text{ or } \sum_{j=1}^{\infty}\frac{(-1)^j}{j}?$$

Some of these make sense and others do not.

Definition A.3.2. Let $< a_j >$ be a sequence of complex numbers, the **partial sums** for the **series**

$$\sum_{j=1}^{\infty}a_j \tag{A.54}$$

is the sequence of complex numbers defined by

$$s_n \overset{d}{=} \sum_{j=1}^{n}a_j.$$

If the sequence $< s_n >$ has a limit then the sum in (A.54) **converges,** otherwise the sum diverges. If the sum converges then *by definition*

$$\sum_{j=1}^{\infty} a_j \stackrel{d}{=} \lim_{n\to\infty} s_n.$$

In the first example above the partial sums are given by $s_n = (-1)^n$, so this series diverges; in the second example one can show that $s_n > \log n$ and therefore this series also diverges. In the last case, the terms are decreasing, alternating in sign and converge to zero, so the alternating series test (B.4.7) applies to show that the sum converges. A sum can diverge in two different ways: the partial sums can tend to $\pm\infty$ or simply fail to approach a finite limit. In the former case we sometimes write $\sum_{j=1}^{\infty} a_j = \infty$.

Note that if the order of terms in the third sum is changed then the value of the sum can also be changed. Perhaps the most dramatic way to see this is to first add up the positive terms and then the negative terms. But observe that

$$\sum_{j=1}^{\infty} \frac{(-1)^j}{j} \neq \sum_{j=1}^{\infty} \frac{1}{2j} - \sum_{j=1}^{\infty} \frac{1}{2j-1}.$$

The two sums appearing on the right hand side are infinite and there is no way to define $\infty - \infty$. Hence an infinite series can converge for two different reasons:

(1). If the terms of the sequence $\{a_j\}$ go to zero fast enough so that the series

$$\sum_{j=1}^{\infty} |a_j| \tag{A.55}$$

converges then we say that the series *converges absolutely.* In this case the value of the sum is independent of the order in which the terms are added.

(2). If the sum in (A.55) diverges, but

$$\sum_{j=1}^{\infty} a_j$$

converges then we say that the series *converges conditionally.* In this case the value of the sum depends very much on the order in which the terms are added. If a series converges conditionally but not absolutely then both

$$\sum_{j=1}^{\infty} \max\{0, a_j\} = \infty \text{ and } \sum_{j=1}^{\infty} \min\{0, a_j\} = -\infty.$$

The sum is therefore converging because of subtle cancelations between the positive and negative terms.

This distinction is important to understand because conditionally convergent series arise frequently in imaging applications. As might be expected, conditionally convergent series require more care to approximate than absolutely convergent sums.

There are two very useful tests for convergence that can often be applied in practice. These tests are called the integral test and the alternating series test. The first applies to show that certain series are absolutely convergent while the second can be used to study special, conditionally convergent series.

Proposition A.3.2 (The integral test). *Suppose that $f(x)$ is a function defined on $[1, \infty)$ which is non-negative and monotone decreasing, that is $0 \leq f(x) \leq f(y)$ if $x > y$. If $a_n = f(n)$ then the series*

$$\sum_{n=1}^{\infty} a_n$$

converges if and only if

$$\lim_{R \to \infty} \int_1^R f(x)dx < \infty.$$

If the series converges then

$$\left| \sum_{n=1}^{\infty} a_n - \sum_{n=1}^{N} a_n \right| \leq \int_N^{\infty} f(x)dx. \tag{A.56}$$

The test which can be applied to series with non-positive terms is

Proposition A.3.3 (Alternating series test). *Let $< a_n >$ be a sequence that satisfies the three conditions*

$$\begin{gathered} \lim_{n \to \infty} a_n = 0, \\ a_n a_{n+1} < 0 \text{ and } |a_{n+1}| \leq |a_n| \text{ for all } n. \end{gathered} \tag{A.57}$$

Then the series

$$\sum_{n=1}^{\infty} a_n$$

converges and

$$\left| \sum_{n=1}^{\infty} a_n - \sum_{n=1}^{N} a_n \right| \leq |a_{N+1}|. \tag{A.58}$$

These tests not only give criteria for certain infinite series to converge, but also give estimates for the errors made by replacing the infinite sum by s_N for any value of N.

Exercise A.3.3. Suppose that the series

$$\sum_{j=1}^{\infty} a_j$$

converges. Show that the $\lim_{j \to \infty} a_j = 0$. Note the converse statement is false: the $\sum j^{-1} = \infty$ even though $\lim_{j \to \infty} j^{-1} = 0$.

Exercise A.3.4. Show that $a_j = j^{-p}$ converges if $p > 1$ and diverges if $p \geq 1$.

Exercise A.3.5. Prove the alternating series test.

A.3.2 Partial summation

One of the most useful tools for working with integrals is the integration by parts formula. If f and g are differentiable functions on an interval $[a, b]$ then

$$\int_a^b f'(x)g(x)dx = f(b)g(b) - f(a)g(a) - \int_a^b f(x)g'(x)dx.$$

There is an analogue of this formula which is very important in the study of non-absolutely convergent series. It is called the *summation by parts formula*.

Proposition A.3.4 (Summation by Parts Formula). *Let $< a_n >$ and $< b_n >$ be sequences of numbers. For each n let*

$$B_n = \sum_{k=1}^{n} b_k$$

then

$$\sum_{n=1}^{N} a_n b_n = a_N B_N - \sum_{n=1}^{N-1} (a_{n+1} - a_n) B_n. \tag{A.59}$$

Using this formula it is often possible to replace a conditionally convergent sum by an absolutely convergent sum.

Example A.3.1. Let $\alpha = e^{2\pi i x}$ where $x \notin \mathbb{Z}$, so that $\alpha \neq 1$. For any such α, the series

$$\sum_{n=1}^{\infty} \frac{\alpha^n}{\sqrt{n}}$$

converges. To prove this observe that

$$B_n = \sum_{k=1}^{n} \alpha^k = \frac{\alpha^{n+1} - 1}{\alpha - 1}$$

is a uniformly bounded sequence and

$$\frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n+1}} \leq \frac{1}{n^{\frac{3}{2}}}.$$

The summation by parts formula gives

$$\sum_{n=1}^{N} \frac{\alpha^n}{\sqrt{n}} = \left(\frac{1}{\sqrt{N}} - \frac{1}{\sqrt{N+1}} \right) B_N - \sum_{n=1}^{N-1} \left(\frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n+1}} \right) B_n.$$

The boundary term on the right goes to zero as $N \to \infty$ and the sum is absolutely convergent. This shows how the summation by parts formula can be used to convert a conditionally convergent sum into an absolutely convergent sum.

A.3.3 Power series

A special sub-class of infinite series are called *power series*, it is the infinite series generalization of a polynomial and involves the powers of a variable.

Definition A.3.3. Let $< a_j >$ be a sequence of complex numbers. The *power series* with these coefficients is the infinite series

$$\sum_{j=0}^{\infty} a_j z^j, \tag{A.60}$$

z is a complex number.

As it stands a power series is a formal expression. The theory of convergence of power series is relatively simple. Roughly speaking, a power series converges for a complex argument z provided that $\lim_{j\to\infty} |a_j z^j| = 0$. The exact result is given in the following theorem.

Theorem A.3.1. *Suppose that* $r \geq 0$ *and*

$$\lim_{j\to\infty} |a_j| r^j = 0 \tag{A.61}$$

then the power series (A.60) *converges absolutely for all complex numbers* z *with* $|z| < r$.

The supremum of the numbers which satisfy (A.61) is called the *radius of convergence* of the power series; we denote it by r_{conv}. For values of z with $|z| < r_{\text{conv}}$ the power series converges absolutely, if $|z| = r_{\text{conv}}$ then the question of convergence or divergence of the series is again quite subtle.

Example A.3.2. If $a_j = j^{-1}$ then $r_{\text{conv}} = 1$. For $|z| < 1$ the series

$$\sum_{j=1}^{\infty} \frac{z^j}{j}$$

converges absolutely. If $z = 1$ then the series diverges, while if $z = -1$ the series converges.

Example A.3.3. Suppose that $a_j = j^j$, then for any number $r > 0$ we have that

$$a_j r^j = (jr)^j.$$

If $jr > 2$ then $a_j r^j > 2^j$ and this shows that the radius of convergence of the power series with these coefficients is 0. In general, if the coefficients grow too quickly, then the series does not converge for any non-zero value of z. While such series do not, strictly speaking, define functions they often appear in applications as *asymptotic expansions* for functions.

In the set

$$D_{\text{conv}} = \{z \,|\, |z| < r_{\text{conv}}\}$$

the series (A.60) defines a function of z with many of the properties of polynomials. Let

$$f(z) = \sum_{j=0}^{\infty} a_j z^j, \tag{A.62}$$

and suppose that the radius of convergence is $r_{\text{conv}} > 0$. Formally differentiating gives a new power series,

$$f_1(z) = \sum_{j=1}^{\infty} j a_j z^{j-1}.$$

It is not difficult to show that the radius of convergence of this series is also r_{conv} and in fact $f'(z) = f_1(z)$, see [1]. This can of course be repeated over and over. These observations are summarized in the following theorem.

Theorem A.3.2. *Suppose that the radius of convergence of the power series* (A.62) *is* $r_{\text{conv}} > 0$; *the function,* $f(z)$ *it defines in* D_{conv} *is infinitely differentiable. For each* $k \geq 0$

$$f^{[k]}(z) = \sum_{j=k}^{\infty} a_j j(j-1)\dots(j-k+1) z^{j-k}$$

also has radius of convergence r_{conv}. *Note in particular that*

$$f^{[k]}(0) = k! a_k.$$

Example A.3.4. The functions $\sin(z)$, $\cos(z)$, $\exp(z)$ are defined as the solutions of differential equations. The sine and cosine satisfy

$$f'' + f = 0$$

while the exponential solves

$$f' - f = 0.$$

Assuming that these functions have power series expansions, we find by substituting into the differential equations that

$$\begin{aligned} \sin(x) &= -\sum_{j=0}^{\infty} \frac{(-x)^{2j+1}}{(2j+1)!}, \\ \cos(x) &= \sum_{j=0}^{\infty} \frac{(-x)^{2j}}{(2j)!}, \\ \exp(x) &= \sum_{j=0}^{\infty} \frac{x^j}{j!}. \end{aligned} \tag{A.63}$$

Here we have use the facts that $\sin(0) = 0$ and $\cos(0) = 1$. From these formulæ it is not difficult to see that the radii of convergence of these series are infinite and that each of these functions satisfies the appropriate differential equation. Recall that a power series is defined for complex numbers, if we substitute $z = ix$ into the series for exp we learn that

$$\exp(ix) = \cos(x) + i\sin(x). \tag{A.64}$$

This is a very useful fact both theoretically and for computation; it is called *Euler's formula*.

The exponential function is given as an infinite series with positive coefficients and therefore $x > 0$ implies that $\exp(x) > 0$. Since $\exp(x)\exp(-x) = 1$ this holds for any real number. Combining this observation with the fact that $\partial_x \exp(x) = \exp(x)$ shows that the exponential is strictly monotone increasing on the real line. Thus exp has an inverse function $l(y)$, defined for positive real numbers y, which satisfies

$$\exp(l(y)) = y \text{ and } l(\exp(x)) = 0.$$

Note that $l(1) = 0$. This function is called the logarithm (or natural logarithm). Following standard practice we use the notation $\log(y)$ for $l(y)$. As the derivative of exp is non-vanishing its inverse is also differentiable. Using the chain rule we obtain that

$$\log'(y) = \frac{1}{y}. \tag{A.65}$$

The differential equation and $\log(1) = 0$ imply that $\log(y)$ can be expressed as an integral:

$$\log(y) = \int_1^y \frac{ds}{s}. \tag{A.66}$$

Because the log is not defined at 0 this function does not have a convergent power series expansion about $x = 0$.

Except for the log, the elementary transcendental functions are given by power series which converge in the whole complex plane. While a function defined as an infinite sum is not in general a computable function, a power series is computable to any *given precision*. Suppose that $f(z)$ is a power series, (A.62) with a positive radius of convergence r_{conv}. If we specify an $\epsilon > 0$ and an argument z with $|z| < r_{\text{conv}}$ then there is a N such that

$$|f(z) - \sum_{j=0}^{N} a_j z^j| < \epsilon.$$

If we have a formula for the coefficients $\{a_j\}$, as is usually the case then one can compute N as a function of ϵ and $|z|/r_{\text{conv}}$.

Example A.3.5. The series defining the sine and cosine are alternating series, this means that

$$\left|\sin(x) + \sum_{j=0}^{N} \frac{(-x)^{2j+1}}{(2j+1)!}\right| \leq \frac{|x|^{2N+3}}{(2N+3)!} \text{ and } \left|\cos(x) - \sum_{j=0}^{N} \frac{(-x)^{2j}}{(2j)!}\right| \leq \frac{|x|^{2N+2}}{(2N+2)!}, \tag{A.67}$$

see (A.58). Because $\sin(x) = (-1)^k \sin(x + k\pi)$, it suffices to consider values of x between $-\pi/2$ and $\pi/2$. For such x, to compute $\sin(x)$ with a error less than $\epsilon > 0$ requires an N which satisfies

$$\frac{\pi^{2N+3}}{2^{2N+3}(2N+3)!} < \epsilon.$$

Using this formula we obtain the following table.

This gives an effective algorithm to compute the sine (or cosine) to a given accuracy. In actual applications, computing the partial sums of the power series is not used because much faster algorithms can be obtained by using the multiple angle formulæ.

N	Maximum error
4	10^{-4}
6	10^{-8}
8	10^{-12}
10	10^{-16}

Table A.1: Errors approximating $\sin(x)$ by partial sums of its Taylor series.

Exercise A.3.6. Using Euler's formula deduce that

$$\cos(x) = \frac{e^{ix} + e^{-ix}}{2}, \quad \sin(x) = \frac{e^{ix} - e^{-ix}}{2i}. \tag{A.68}$$

Exercise A.3.7. Using the uniqueness theorem for ordinary differential equations prove that for real numbers x and y

$$\exp(x+y) = \exp(x)\exp(y), \quad \exp(-x) = [\exp(x)]^{-1}. \tag{A.69}$$

The function $g(x) = \exp(ix)$ satisfies the ODE $g' - ig = 0$ and therefore the argument shows that these relations hold with x and y replaced by ix and xpiy. Deduce the multiple angle formulæ

$$\cos(x+y) = \cos(x)\cos(y) - \sin(x)\sin(y), \quad \sin(x+y) = \sin(x)\cos(y) + \sin(y)\cos(x). \tag{A.70}$$

Exercise A.3.8. Euler's formula shows that a complex number has a *polar representation* in the form $z = re^{i\theta}$ where r and θ are real numbers, compare with (A.48). If $w = \rho e^{i\phi}$ show that

$$zw = r\rho e^{i(\theta+\phi)}. \tag{A.71}$$

Exercise A.3.9. Using the integral formula, (A.66) prove that

$$\log(xy) = \log(x) + \log(y). \tag{A.72}$$

Exercise A.3.10. Show that the log has a convergent power series expansion about $y = 1$ given by

$$\log(1+t) = -\sum_{j=1}^{\infty} \frac{(-t)^j}{j}.$$

For what values of t does this series converge?

A.3.4 Binomial formula

The elementary binomial formula gives the expansion for $(x+y)^n$ where n is a positive integer,

$$(x+y)^n = \sum_{j=0}^{n} \binom{n}{j} x^j y^{n-j}. \tag{A.73}$$

The coefficients are the *binomial* coefficients given by

$$\binom{n}{j} = \frac{n!}{j!(n-1)!}.$$

One of the earliest uses of power series was the generalization of this formula to arbitrary values of n. If n is not a positive integer the result is an infinite series. For a real number α we have the formula

$$(x+y)^\alpha = y^\alpha \left[1 + \alpha\left(\frac{x}{y}\right) + \frac{\alpha(\alpha-1)}{2!}\left(\frac{x}{y}\right)^2 + \cdots + \frac{\alpha(\alpha-1)\ldots(\alpha-k+1)}{k!}\left(\frac{x}{y}\right)^k + \ldots\right]. \tag{A.74}$$

The infinite sum converges so long as $|x/y| < 1$. This formula can be used to compute approximations to the roots of numbers $x^{\frac{1}{n}}$. Choose y to be the smallest number of the form $k^n, k \in \mathbb{N}$ which is larger than x. The general formula then gives

$$x^{\frac{1}{n}} = (k^n + x - k^n)^{\frac{1}{n}} = k\left[1 - \frac{1}{n}\left(1 - \frac{x}{k^n}\right) + \frac{n-1}{2n^2}\left(1 - \frac{x}{k^n}\right)^2 - \ldots\right].$$

Again, in principle we have an usable algorithm for computing the roots of positive numbers to any desired accuracy. In practice there are more efficient algorithms than those arising from the power series representation.

By directly multiplying the power series for the exponential function one can show that (A.69) holds for any pair of complex numbers. This gives a way to compute roots of complex numbers. Let $z = re^{i\theta}$ then (A.69) and (A.71) imply that

$$\zeta_n = r^{\frac{1}{n}} e^{i\frac{\theta}{n}}$$

is a n^{th}-root of z. Using (A.64) we can rewrite this as

$$\zeta_n = r^{\frac{1}{n}}(\cos(\frac{\theta}{n}) + i\sin(\frac{\theta}{n})).$$

This reduces the problem of approximating roots of complex numbers to problems we have already solved. Note that if $r = \exp(x)$ for x a real number then

$$r^{\frac{1}{n}} = \exp\left(\frac{x}{n}\right),$$

gives another way to approximate roots of real numbers. It reduces the problem of approximating roots to that of approximating the log-function.

If x is a large real number then $\exp(-x)$ is a very small, positive number, it is given as an infinite sum by

$$\exp(-x) = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} - \ldots,$$

whereas

$$\exp(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots.$$

Note that the numbers which appear in these two sums are identical, only the signs are different. The first sum is a very small positive number and the second a very large positive number. This means that there is a lot of rather subtle cancelation occurring in the first sum. Using floating point arithmetic it is very difficult to compute such a sum accurately. A much more accurate computation of $\exp(-x)$ is obtained by first computing an approximation $y \simeq \exp(x)$ and then setting $\exp(-x) \simeq y^{-1}$. We can compute the relative error, first suppose

$$y = e^x + \epsilon$$

then a calculation shows that

$$\frac{1}{y} - e^{-x} = \frac{\epsilon e^{-x}}{y}.$$

This shows that the *relative* error we make in setting e^{-x} equal to y^{-1} is

$$\frac{|y^{-1} - e^{-x}|}{e^{-x}} = \frac{|\epsilon|}{y} \simeq |\epsilon| e^{-x} = \frac{|y - e^x|}{e^x}.$$

Thus if we compute e^x with a given relative error then the relative error in using y^{-1} for e^{-x} is the same.

A.3.5 The Gamma function

Perhaps the most important "higher transcendental function" is the *Gamma function*. For complex numbers z with $\operatorname{Re} z > 0$ it is defined by the formula

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt. \tag{A.75}$$

From the formula it is clear that $\Gamma(z)$ is an analytic function in the right half plane. The Γ-function is characterized by the functional equation it satisfies.

Proposition A.3.5. *For any z with $\operatorname{Re} z > 0$ the Gamma function satisfies the relation:*

$$\Gamma(z+1) = z\Gamma(z). \tag{A.76}$$

The proof is a simple integration by parts. Using the functional equation, the Γ-function can be extended to be a meromorphic function on the whole complex plane with poles at the non-positive integers. For z with $-n < \operatorname{Re} z$, $\Gamma(z)$ is defined by

$$\Gamma(z) = \frac{\Gamma(z+n)}{z(z+1)\cdots(z+n-1)}.$$

In many applications it is important to understand the behavior of $\Gamma(x)$ as x tends to infinity. Stirling's formula states that

$$\Gamma(x) = \sqrt{2\pi} x^{x-\frac{1}{2}} e^{-x} (1 + O(\frac{1}{x})). \tag{A.77}$$

We give an outline of the derivation of Stirling's formula. It is a special case of *Laplace's method* for obtaining asymptotics for functions of the form

$$f(x) = \int e^{x\phi(s)}\psi(s)ds.$$

The idea is very simple: only the global maxima of the exponent $\phi(s)$ contribute, asymptotically to $f(x)$ as x tends to infinity. A fuller account of this method can be found in [58]. We begin by setting $t = s(x-1)$ in (A.75) to obtain,

$$\Gamma(x) = (x-1)^x \int_0^\infty e^{(x-1)(\log s - s)}ds.$$

The function in the exponent $\log s - s$ has a unique maximum at $s = 1$ where it assumes the value -1. This implies that for any small $\delta > 0$ we have the asymptotic formula

$$\int_0^\infty e^{(x-1)(\log s - s)}ds = \int_{1-\delta}^{1+\delta} e^{(x-1)(\log s - s)}ds + O(e^{-(x-1)(1+\frac{\delta^2}{2})}). \tag{A.78}$$

The second derivative of $\log s - s$ at $s = 1$ is -1 which means that the function

$$v = \begin{cases} \sqrt{(2(u - \log(1+u))} \text{ for } u > 0, \\ - \sqrt{(2(u - \log(1+u))} \text{ for } u < 0 \end{cases}$$

is smooth and invertible in an open interval around $u = 0$. Using v as the variable of integration, the integral becomes

$$\int_{1-\delta}^{1+\delta} e^{(x-1)(\log s - s)} = e^{-(x-1)} \int_{-\delta'}^{\delta''} e^{-(x-1)v^2}h'(v)dv.$$

Here δ' and δ'' are positive numbers.

As $h'(v)$ is a smooth function and $h'(0) = 1$ it is not difficult to prove, that as x tends to infinity

$$\int_{-\delta'}^{\delta''} e^{-(x-1)v^2}h'(v)dv = \frac{1}{\sqrt{x-1}} \int_{-\infty}^{\infty} e^{-\tau^2}d\tau(1 + O(\frac{1}{x})). \tag{A.79}$$

Collecting the pieces gives

$$\begin{aligned} \Gamma(x) &= x^x \left[1 - \frac{1}{x}\right]^x \frac{\sqrt{2\pi}e^{1-x}}{\sqrt{x-1}}(1 + O(\frac{1}{x})) \\ &= \sqrt{2\pi}x^{(x-\frac{1}{2})}e^{-x}(1 + O(\frac{1}{x})). \end{aligned} \tag{A.80}$$

As a special case, Stirling's formula gives an asymptotic evaluation of $n!$ as n tends to infinity

$$n! = \Gamma(n+1) \approx \sqrt{2\pi n}\left[\frac{n}{e}\right]^n.$$

Exercise A.3.11. Prove the functional equation (A.76). Deduce that

$$\Gamma(z+n) = z(z+1)\cdots(z+n-1)\Gamma(z).$$

Exercise A.3.12. Show that for a positive integer n, $\Gamma(n+1) = n!$

Exercise A.3.13. For m a non-positive integer compute the limit

$$\lim_{z\to m}(z-m)\Gamma(z).$$

Exercise A.3.14. Prove formula A.78.

Exercise A.3.15. Prove that v is a smooth, invertible function of u for u in an interval about 0 and that if $u = h(v)$ then $h'(0) = 1$.

Exercise A.3.16. Fill in the details of the last step in the derivation of Stirling's formula.

Exercise A.3.17. Prove that if x and y are positive real numbers then

$$\int_0^1 t^{x-1}(1-t)^{y-1}dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}. \tag{A.81}$$

A.3.6 Bessel functions

For a complex number ν, a *Bessel function* of *order* ν is any solution of the ordinary differential equation

$$\frac{d^2f}{dz^2} + \frac{1}{z}\frac{df}{dz} + \left[1 - \frac{\nu^2}{z^2}\right] f = 0. \tag{A.82}$$

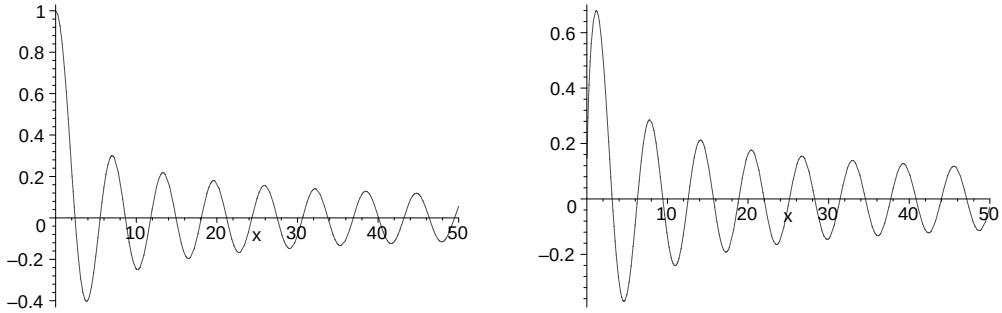
The *J-Bessel* functions, of integral and half-integral orders are important in Fourier analysis. If $\nu \in \mathbb{C} \setminus \{-2, -3, \dots\}$ then $J_\nu(z)$ is defined by the power series

$$J_\nu(z) = \left[\frac{z}{2}\right]^\nu \sum_{k=0}^{\infty}(-1)^k \frac{z^{2k}}{2^{2k}k!\Gamma(\nu+k+1)}. \tag{A.83}$$

The infinite sum converges in the whole complex plane. If ν is not an integer then $J_\nu(z)$ is defined for z with $|\arg(z)| < \pi$ be setting

$$x^\nu = e^{\nu \log x} \text{ for } x \in (0, \infty),$$

here $\log x$ is taken to be real for positive real values of x. Graphs of J_0 and $J_{\frac{1}{2}}$ are shown in figure A.2.



(a) J_0. (b) $J_{\frac{1}{2}}$.

Figure A.2: Some J-Bessel functions.

The J-Bessel functions have a variety of integral representations. Their connection with Fourier analysis is a consequence of the formula,

$$J_\nu(z) = \frac{\left[\frac{z}{2}\right]^\nu}{2\Gamma\left(\nu+\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)} \int_0^{2\pi} e^{iz\cos\theta}\sin^{2\nu}\theta d\theta, \tag{A.84}$$

valid if 2ν is a non-negative integer. The J-Bessel function is also the one dimensional, Fourier transform of a function, if $\operatorname{Re}\nu > -\frac{1}{2}$ then

$$J_\nu(z) = \frac{\left[\frac{z}{2}\right]^\nu}{\Gamma\left(\nu+\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)} \int_{-1}^{1} (1-x^2)^{\nu-\frac{1}{2}} e^{izx} dx. \tag{A.85}$$

The J-Bessel functions have very simple asymptotic behavior as $|z|$ tends to infinity,

$$J_\nu(z) = \sqrt{\frac{2}{\pi z}}\left[\cos(z-\frac{\pi\nu}{2}-\frac{\pi}{4}) - (\nu^2-\frac{1}{4})\sin(z-\frac{\pi\nu}{2}-\frac{\pi}{4}) + O(\frac{1}{z^2})\right] \text{ if } |\arg z| < \pi. \tag{A.86}$$

This formula is proved using (A.84) and the method of "stationary phase." Indeed the Bessel functions have complete asymptotic expansions. Several additional facts are outlined in the exercises. A thorough treatment of Bessel functions is given in [81]. Many useful relations and definite integrals involving Bessel functions can be found in [20].

Exercise A.3.18. Show that the function defined in (A.83) satisfies (A.82).

Exercise A.3.19. Show that the function defined in (A.84) satisfies (A.82).

Exercise A.3.20. Show how to deduce (A.85) from (A.84).

Exercise A.3.21. Derive the power series expansion, (A.83). from (A.85). Hint: Write the exponential as a power series and use (A.81).

Exercise A.3.22. Bessel functions with half integral order can be expressed in terms of trigonometric functions. Show that

(1). $J_{\frac{1}{2}}(z) = \sqrt{\frac{2}{\pi z}} \sin z,$

(2). $J_{\frac{3}{2}}(z) = \sqrt{\frac{2}{\pi z}} \left[\frac{\sin z}{z} - \cos z \right].$

Exercise A.3.23. Show that

$$\frac{d}{dz} J_0(z) = -J_1(z).$$

Exercise A.3.24. Use the Parseval formula to compute

$$\int_{-\infty}^{\infty} |J_\nu(x)|^2 |x|^{-2\nu} dx,$$

for ν a non-negative integer or half integer.

Exercise A.3.25. By considering the differential equation (A.82), for large z explain the asymptotic expansion of the Bessel function. In particular, why is the rate of decay of J_ν independent of ν?

A.4 Spaces of functions

In mathematics functions are usually grouped together into vector spaces, according to their smoothness properties or rates of decay. Norms or metrics are defined on these vector spaces which incorporate these properties.

A.4.1 Examples of function spaces

A basic example is the space $\mathcal{C}^0([0,1])$; it is the set of continuous functions defined on the interval $[0,1]$. A function $f(x)$ belongs to $\mathcal{C}^0([0,1])$ if, for every $x \in [0,1]$ $\lim_{y \to x} f(y) = f(x)$, at the endpoints we need to use one-sided limits

$$\lim_{y \to 0^+} f(y) = 0, \quad \lim_{y \to 1^-} f(y) = f(1).$$

A scalar multiple of a continuous function is continuous, as is a sum of two continuous functions. Thus the set $\mathcal{C}^0([0,1])$ is a vector space. Define a norm on this vector space by setting

$$\|f\|_{\mathcal{C}^0} = \max_{x \in [0,1]} |f(x)|.$$

Notice that the expression on the right is defined for any bounded function defined on $[0,1]$.

A sequence of functions $< f_n >$ converges in this norm to a function f provided

$$\lim_{n\to\infty} \|f_n - f\|_{\mathcal{C}^0} = 0.$$

It is a non-trivial result in analysis that if $< f_n >$ converges to f, in this sense, then f is also a continuous function. This norm is sometimes called the *uniform norm* and convergence in this norm is called *uniform convergence*. The vector space $\mathcal{C}^0([0,1])$ is *complete* with respect to this norm.

For each $k \in \mathbb{N}$ we let $\mathcal{C}^k([0,1])$ denote the space of functions defined on $[0,1]$ with k continuous derivatives. We define a norm on this space by setting

$$\|f\|_{\mathcal{C}^k} = \sum_{j=0}^{k} \|f^{[j]}\|_{\mathcal{C}^0}. \tag{A.87}$$

This norm defines a notion of convergence for k-times differentiable functions. As before, if a sequence $< f_n >$ converges to f in this sense then f is also a function with k continuous derivatives. The basic result in analysis used to study these function spaces is:

Theorem A.4.1. *Let $< f_n >$ be a sequence of k-times differentiable functions defined on $[a,b]$. If $< f_n >$ converges* uniformly *to f and the sequences of derivatives $< f_n^{[j]} >$, $j = 1, \ldots, k$ converge* uniformly *to functions $g_1, \ldots, g_k$ then f is k-times, continuously differentiable and*

$$f^{[j]} = g_j \text{ for } j = 1, \ldots, k.$$

A proof of this theorem is given in [67].

Let $\mathcal{C}^\infty([0,1])$ denote the vector space of functions, defined on $[0,1]$ with infinitely many continuous derivatives. The expression on the right hand side of (A.87) makes no sense if $k = \infty$. In fact there is no way to define a norm on the vector space $\mathcal{C}^\infty([0,1])$. We can however define a metric on $\mathcal{C}^\infty([0,1])$ by setting

$$d(f,g) = \sum_{j=0}^{\infty} 2^{-j} \frac{\|f-g\|_{\mathcal{C}^j}}{1 + \|f-g\|_{\mathcal{C}^j}}.$$

A sequence of functions $< f_n > \subset \mathcal{C}^\infty([0,1])$ converges to f if

$$\lim_{n\to\infty} d(f_n, f) = 0.$$

Analogous spaces are defined with $[0,1]$ replaced with other sets, for example $\mathcal{C}^0(\mathbb{R}^n)$ or $\mathcal{C}^\infty(\mathbb{S}^1 \times \mathbb{R})$, etc.

The foregoing examples are defined by considering the smoothness properties of functions. Other types of spaces are defined by considering rates of decay at infinity and blow-up at finite points. For example the space $L^2([0,1])$ consists of functions for which

$$\|f\|_2 = \sqrt{\int_0^1 |f(x)|^2 dx} < \infty.$$

Such a function is said to be *square integrable*. It is not necessary for f to be continuous in order for it to belong to $L^2([0,1])$, only that this integral makes sense. The function $|x-\frac{1}{2}|^{-\frac{1}{4}}$ is not even bounded, but it belongs to $L^2([0,1])$ because

$$\int \frac{1}{\sqrt{|x-\frac{1}{2}|}} dx < \infty.$$

Example A.4.1. Let $0 \leq a < b < 1$ then the functions $\chi_{[a,b]}(x)$ belong to $L^2([0,1])$. Note the following

$$\chi_{[a,b]}(x) - \chi_{(a,b)}(x) = \begin{cases} 1 & \text{if } x = a \text{ or } b, \\ 0 & \text{otherwise.} \end{cases}$$

The L^2-norm cannot detect the difference between these two functions

$$\|\chi_{[a,b]} - \chi_{(a,b)}\|_2 = 0.$$

This is a general feature of norms defined by integrals: they do not distinguish functions which differ on very small sets. The technical term for these very small sets is *sets of measure zero*. This property does not create significant difficulties, but is important to keep in mind. Indeed, this is also a feature of physical measurements and explains, in part the relevance of integral norms in practical applications.

Once again if $< f_n > \subset L^2([0,1])$ is a sequence of functions with converge to a function f in the sense that

$$\lim_{n\to\infty} \|f_n - f\|_2 = 0$$

then f also belongs to $L^2([0,1])$. This type of convergence is often called *convergence in the mean*; we use the special notation

$$\underset{n\to\infty}{LIM} f_n = f.$$

The behavior of L^2-convergent sequences is quite different from that of $\mathcal{C}^0$-convergent sequences. Examples best illustrate this point.

Example A.4.2. Let $f_n(x) = x^n$, if $x \in [0,1)$ then

$$\lim_{n\to\infty} f_n(x) = 0$$

whereas $\lim_{n\to\infty} f_n(1) = 1$. For each x, $f_n(x)$ converges, but the limit function

$$f(x) = \begin{cases} 0 & \text{if } 0 \leq x < 1, \\ 1 & \text{if } x = 1 \end{cases}$$

is **not** continuous. This means that $\|f_n - f\|_{\mathcal{C}^0}$ cannot go to zero as $n \to \infty$. On the other hand

$$\int_0^1 |x^n|^2 dx = \frac{1}{2n+1},$$

and therefore $\lim_{n\to\infty} \|f_n - 0\|_2 = 0$. So the sequence $< f_n >$ does converge in the L^2-norm to the function which is identically zero. Note that the pointwise limit, $f(x)$ cannot be distinguished from the zero function by the L^2-norm. Note also the related fact: the L^2-convergence of a sequence $< f_n >$ to a function f does **not** require that $\lim_{n\to\infty} f_n(x) = f(x)$ for all x.

Example A.4.3. Define a sequence of functions

$$f_n(x) = \begin{cases} 0 & \text{for } x \in [0, \frac{n-1}{2n}], \\ nx - \frac{n-1}{2} & \text{for } x \in (\frac{n-1}{2n}, \frac{n+1}{2n}), \\ 1 & \text{for } x \in [\frac{n+1}{2n}, 1]. \end{cases}$$

Each of these functions is continuous and it is not difficult to show that

$$f(x) = \lim_{n\to\infty} f_n(x) = \begin{cases} 0 & \text{for } x \in [0, \frac{1}{2}), \\ \frac{1}{2} & \text{for } x \in (\frac{1}{2}, 1]. \end{cases}$$

Once again the limit function is not continuous, and it is easy to see that

$$\|f_n - f\|_{C^0} = \frac{1}{2}$$

for every $n \in \mathbb{N}$. On the other hand it is also not hard to show that

$$\lim_{n\to\infty} \int_0^1 |f_n(x) - f(x)|^2 dx = 0.$$

Spaces of functions are generally infinite dimensional. Introducing a basis for a *finite dimensional*, real vector space establishes an isomorphism between that vector space and $\mathbb{R}^n$, for some n. In exercise A.2.3 it is shown that the notion of convergence for a sequence in $\mathbb{R}^n$ is independent of the choice of norm. This is *not* true for infinite dimensional vector spaces. There are many non-isomorphic vector spaces and different norms lead to different convergent sequences. By analogy to the norms, $\|\cdot\|_p$ defined on $\mathbb{R}^n$ the L^p-norms are defined for functions defined on $[0, 1]$ by setting

$$\|f\|_{L^p} = \left[\int_0^1 |f(x)|^p dx \right]^{\frac{1}{p}}. \tag{A.88}$$

If $1 \leq p$ then this defines a norm; the restriction on p is needed to establish the triangle inequality,

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p. \tag{A.89}$$

We can also let $p = \infty$ by defining

$$\|f\|_\infty = \max\{|f(x)| \ : \ x \in [0, 1]\}.$$

Define the vector space $L^p([0, 1])$ to be those locally integrable functions f such that $\|f\|_p < \infty$. The various L^p-spaces are related by a fundamental inequality.

Theorem A.4.2 (Hölder's inequality). *Let $1 \leq p \leq \infty$ and define q to be*

$$q = \begin{cases} \frac{p}{p-1} & \text{if } p \neq 1, \infty, \\ 1 & \text{if } p = \infty, \\ \infty & \text{if } p = 1. \end{cases} \tag{A.90}$$

If $f \in L^p([0,1])$ and $g \in L^q([0,1])$ then

$$\int_0^1 |f(x)g(x)|dx \leq \|f\|_{L^p}\|g\|_{L^q}. \tag{A.91}$$

In particular the product fg belongs to $L^1([0,1])$.

The analogous result holds with $[0,1]$ replaced by $\mathbb{R}$.

Exercise A.4.2 shows that a sequence of functions $< f_n >$ which belongs to $L^2([0,1])$ also belongs to $L^1([0,1])$. The next example shows that a bounded sequence in $L^2([0,1])$ need not have a limit in L^2-norm even though it does have a limit in the L^1-norm.

Example A.4.4. Define a sequence of functions

$$f_n(x) = \begin{cases} n & \text{if } x \in [0, \frac{1}{n^2}], \\ 0 & \text{if } n \in (\frac{1}{n^2}, 1]. \end{cases}$$

Note that if $x \neq 0$ then $\lim_{n\to\infty} f_n(x) = 0$, on the other hand, for all $n \in \mathbb{N}$ we have that

$$\|f_n\|_2 = 1.$$

This shows that this sequence is bounded in $L^2([0,1])$ but does not converge to anything. Note that $\lim_{n\to\infty} \|f_n\|_1 = 0$ and therefore $< f_n >$ does converge to zero in the L^1-norm.

Exercise A.4.1. In example A.4.2 n find the maximum value of $f(x) - f_n(x)$, for each n and show that this does not go to zero as $n \to \infty$.

Exercise A.4.2. Use Hölder's inequality to show that if $f \in L^p([0,1])$ and $1 \leq p' < p$ then $f \in L^p([0,1])$ as well. Hint: Take $g = 1$.

Exercise A.4.3. Show that the function

$$f_\alpha(x) = \frac{1}{x^\alpha}$$

belongs to $L^p([0,1])$ if $\alpha < p^{-1}$ and does **not** belong to $L^p([0,1])$ if $\alpha \geq p^{-1}$.

A.4.2 Completeness

In the finite dimensional case we introduced the concept of a Cauchy sequence as a way of describing which sequences *should* converge. The real power of this idea only becomes apparent in the infinite dimensional context.

Definition A.4.1. Let $(V, \|\cdot\|)$ be a normed vector space. A sequence $< v_n > \subset V$ is a *Cauchy sequence* if, for any $\epsilon > 0$, there exists an N so that

$$\|v_n - v_m\| < \epsilon \text{ provided that } m \text{ and } n > N.$$

Reasoning by analogy, the Cauchy sequences are the ones which "should converge." However, because there are many different norms which can be used on an infinite dimensional space, this is quite a subtle question.

Example A.4.5. Let V be the continuous functions on $[0,1]$ and use for a norm

$$\|f\| = \int_0^1 |f(x)|dx.$$

Define a sequence $< f_n > \subset V$ by setting

$$f_n(x) = \begin{cases} 0 & \text{for } 0 \le x \le \frac{1}{2} - \frac{1}{n}, \\ n(x - \frac{1}{2}) & \text{for } \frac{1}{2} - \frac{1}{n} \le x \le \frac{1}{2}, \\ 1 & \text{for } \frac{1}{2} \le x \le 1. \end{cases}$$

The distances between the terms of the sequence satisfy the estimates

$$\|f_n - f_m\| \le \frac{1}{2}(\frac{1}{n} + \frac{1}{m}).$$

This implies that $< f_n >$ is a Cauchy sequence. Pointwise $< f_n >$ converges to

$$f = \begin{cases} 0 & \text{for } 0 \le x < \frac{1}{2}, \\ 1 & \text{for } \frac{1}{2} \le x \le 1. \end{cases}$$

Indeed it is not difficult to show that

$$\lim_{n\to\infty} \|f_n - f\| = 0.$$

The only difficulty is that f is not a continuous function. This is an example of a Cauchy sequence which does not converge.

This sort of example leads to the following definition.

Definition A.4.2. A normed vector space $(V, \|\cdot\|)$ is said to be *complete* if every Cauchy sequence $< v_n > \subset V$ converges to a limit *in* V.

Note that completeness is a property of a *normed vector* space. It makes no sense to say "the set of continuous functions on $[0,1]$ is complete." Rather one must say that "the set of continuous functions on $[0,1]$, with the sup-norm is complete." Completeness is a very important property for a normed linear space and most of the spaces we consider have this property.

Theorem A.4.3. *For $1 \le p < \infty$ the normed linear spaces $L^p([0,1])$ (or $L^p(\mathbb{R}^n)$) are complete. For any non-negative integer k, the normed linear spaces $\mathcal{C}^k([0,1])$ (or $\mathcal{C}^k(\mathbb{R}^n)$) are complete.*

Exercise A.4.4. In example A.4.5 show that $< f_n >$ is not a Cauchy sequence in the usual norm on $\mathcal{C}^0([0,1])$.

A.4.3 Linear functionals

For finite dimensional vector spaces the concept of a linear function is given by the purely algebraic conditions (A.12). For infinite dimensional vector more care is required because linear functions may not be continuous.

Example A.4.6. Let V be the set of once differentiable functions on $[0,1]$. Instead of using the the usual $\mathcal{C}^1$-norm we use the $\mathcal{C}^0$-norm. With this choice of norm a sequence of functions $< f_n > \subset V$ converges to $f \in V$ if

$$\lim_{n\to\infty} \|f_n - f\|_{\mathcal{C}^0} = 0.$$

Suppose $< f_n >$ is a sequence which converges to 0 in this sense and that $l : V \to \mathbb{R}$ is a linear function. If l is continuous then

$$\lim_{n\to\infty} l(f_n) = 0.$$

Define a function on V by setting

$$l(f) = f'(\frac{1}{2}).$$

The usual rules of differentiation show that this is a linear function. It is however, not continuous. Define a sequence of functions in V by letting

$$f_n(x) = \begin{cases} 0 & \text{if } x \notin (\frac{n-1}{2n}, \frac{n+3}{2n}), \\ \frac{1}{\sqrt{n}}(1 - [n(x - \frac{1}{2n})]^2)^2 & \text{if } x \in (\frac{n-1}{2n}, \frac{n+3}{2n}). \end{cases}$$

It is not difficult to show that $f_n \in V$ for each n and that

$$f_n(x) \leq \frac{1}{\sqrt{n}} \text{ for } x \in [0,1].$$

This shows that $< f_n >$ converges to $f(x) \equiv 0$ in the sense defined above. However a calculation gives that

$$l(f_n) = f_n'(\frac{1}{2}) = -\frac{3}{2}\sqrt{n}.$$

In other words $\lim_{n\to\infty} l(f_n) = -\infty$, even though $< f_n >$ converges to zero. If we use the $\mathcal{C}^1$-norm instead, then l is indeed a continuous linear function. The reason this does not contradict the example above is that the sequence $< f_n >$ does **not** converge to zero in the $\mathcal{C}^1$-norm.

In light of this example, it is clear that additional care is needed in the study of linear functions on infinite dimensional vector spaces.

Definition A.4.3. Let V be a vector space with norm $\|\cdot\|$. A linear function $l : V \to \mathbb{R}$ is called a *linear functional* if it is continuous with respect to the norm. That is, if $< f_n >, f \subset V$ and

$$\lim_{n\to\infty} \|f - f_n\| = 0$$

then

$$\lim_{n\to\infty} l(f_n) = l(f).$$

We denote the set of linear functionals by V', as before it is a vector space called the *dual* vector space. It has a naturally defined norm given by

$$\|l\|' = \sup_{V \ni f \neq 0} \frac{|l(f)|}{\|f\|}. \tag{A.92}$$

For the normed vector spaces of greatest interest one can give a complete description of the dual vector space. Let $1 \leq p \leq \infty$ and let q be defined by (A.90). Choose a function $g \in L^q([0,1])$, then Hölder's inequality implies that for every $f \in L^p([0,1])$ the function fg is integrable and

$$\left| \int_0^1 f(x)g(x)dx \right| \leq \|f\|_{L^p} \|g\|_{L^q}.$$

The real valued function

$$l_g(f) = \int_0^1 f(x)g(x)dx \tag{A.93}$$

is therefore well defined for all $f \in L^p([0,1])$. The elementary properties of the integral imply that it is linear and Hölder's inequality implies that it is continuous. Suppose that $< f_n > \subset L^p([0,1])$ which converges in the L^p-sense to f. We see that

$$|l_g(f_n) - l(f)| = |l_g(f_n - f)| \leq \|f_n - f\|_{L^p} \|g\|_{L^q}.$$

This shows that $\lim_{n\to\infty} l_g(f_n) = l_g(f)$.

In fact all linear functionals on these normed vector spaces are of this form.

Theorem A.4.4 (Riesz Representation Theorem 1). *If $1 \leq p < \infty$ and q is given by* (A.90) *then $L^q([0,1])$ is the dual space to $L^p([0,1])$. That is every continuous linear function on $L^p([0,1])$ is given by l_g for some $g \in L^q([0,1])$.*

Remark A.4.1. Note that the case $p = \infty$ is not included in the above theorem. The space $L^\infty([0,1])$ turns out to be considerably more complicated as a normed vector space than $L^p([0,1])$ for $1 \leq p < \infty$. As the details of this space are not needed in the sequel we do not pursue the matter further.

Starting with n-tuples of numbers and the $\|\cdot\|_p$-norm defined in (A.24) leads to another collection of infinite dimensional analogues.

Definition A.4.4. For $1 \leq p \leq \infty$ let l^p denote the collection of sequences $< a_j >$ such that

$$\| < a_j > \|_p = \left[\sum_{j=1}^{\infty} |a_j|^p \right]^{\frac{1}{p}} < \infty. \tag{A.94}$$

These are *complete* normed vector spaces.

Example A.4.7. The space l^1 consists of sequences which define absolutely convergent sums, that is $< a_j > \in l^1$ if and only if

$$\sum_{j=1}^{\infty} |a_j| < \infty.$$

If $p < p'$ then it is clear that $l^p \subset l^{p'}$. There is also a version of the Hölder inequality. Let $1 \leq p \leq \infty$ and q by given by (A.90), for $< a_j > \in l^p$ and $< b_j > \in l^q$ the sequence $< a_j b_j > \in l^1$ and

$$\sum_{j=1}^{\infty} |a_j b_j| \leq \| < a_j > \|_p \| < b_j > \|_q. \tag{A.95}$$

This inequality shows that if $\mathbf{b} =< b_j > \in l^q$ then we can define a bounded linear functional on l^p by setting

$$l_{\mathbf{b}}(\mathbf{a}) = \sum_{j=1}^{\infty} a_j b_j.$$

This again gives all bounded functionals provided p is finite.

Theorem A.4.5 (Riesz Representation Theorem 2). *If $1 \leq p < \infty$ and q is given by* (A.90) *then l^q is the dual space to l^p. That is every continuous linear function on l^p is given by $l_{\mathbf{b}}$ for some $\mathbf{b} \in l^q$.*

Exercise A.4.5. Prove that $l^p \subset l^{p'}$.

A.4.4 Measurement, linear functionals and weak convergence

Suppose that the state of a system is described by a function $f \in L^p([0,1])$. In this case the measurements that one can make are often modeled as the evaluation of linear functionals. That is we have a collection of functions $\{g_1, \ldots, g_k\} \subset L^q([0,1])$ and our measurements are given by

$$m_j(f) = \int_0^1 f(x) g_j(x) dx, \quad j = 1, \ldots, k.$$

From the point of view of measurement this suggests a different, perhaps more reasonable, notion of convergence. In so far as these measurements are concerned, a sequence of states $< f_n >$ would appear to converge to a state f if

$$\lim_{n \to \infty} \int_0^1 f_n(x) g_j(x) dx = \int_0^1 f(x) g_j(x) dx, \text{ for } j = 1, \ldots k.$$

Since we are only considering finitely many measurements on an infinite dimensional state space this is clearly a much weaker condition than the condition that $< f_n >$ converge to f in the L^p-sense.

Of course if $< f_n >$ converges to f in the L^p-sense then, for any $g \in L^q([0,1])$ $\lim_{n\to\infty} l_g(f_n) = l_g(f)$. However the L^p-convergence is not required for these conditions to hold. It is a very important observation that the condition

$$\lim_{n\to\infty} \int_0^1 f_n(x)g(x)dx = \int_0^1 f(x)g(x)dx \text{ for } \textbf{every} \text{ function } g \in L^q([0,1])$$

is a much weaker condition than L^p-convergence.

Definition A.4.5. Suppose that $(V, \|\cdot\|)$ is a normed vector space and $< \mathbf{v}_n >$ is a sequence of vectors in V. If there exists a vector $\mathbf{v} \in V$ such that for **every** continuous linear function l we have that

$$\lim_{n\to\infty} l(\mathbf{v}_n) = l(\mathbf{v})$$

then we say that $\mathbf{v}_n$ *converges weakly* to $\mathbf{v}$. This is sometimes denoted by

$$\mathbf{v}_n \rightharpoonup \mathbf{v}.$$

From the point of view of measurement, weak convergence is often the appropriate notion. Unfortunately it *cannot* be defined by a norm and a sequence does not exert very much control over the properties of its weak limit. For example it is not in general true that

$$\lim_{n\to\infty} \|\mathbf{v}_n\| = \|\mathbf{v}\|$$

for a weakly convergent sequence. This is replaced by the statement

$$\text{If } \mathbf{v}_n \rightharpoonup \mathbf{v} \text{ then } \limsup_{n\to\infty} \|\mathbf{v}_n\| \geq \|\mathbf{v}\|. \tag{A.96}$$

Example A.4.8. The sequence of functions $< f_n >$ defined in example A.4.4 is a sequence with

$$\|f_n\|_{L^2} = 1$$

for all n. On the other hand if $x \in (0,1]$ then

$$\lim_{n\to\infty} f_n(x) = 0.$$

These two facts allow the application of standard results from measure theory to conclude that

$$\lim_{n\to\infty} \int_0^1 f_n(x)g(x)dx = 0,$$

for every function $g \in L^2([0,1])$. In other words the sequence $< f_n >$ converges *weakly* to zero even though it does not converge to anything in the L^2-sense.

Example A.4.9. Let $< \mathbf{a}_n > \subset l^2$ be the sequence defined by

$$\mathbf{a}_n(j) = \begin{cases} 1 & \text{if } j = n, \\ 0 & \text{if } j \neq n. \end{cases}$$

Since $\mathbf{a}_n(j) = 0$ if $j < n$ it is clear that if $< \mathbf{a}_n >$ were to converge to $\mathbf{a}$, in the l^2-sense then $\mathbf{a} = \mathbf{0}$. On the other hand $\|\mathbf{a}_n\|_{l^2} = 1$ for all n and this shows that $\mathbf{a}_n$ cannot converge in the l^2-sense. On the other hand if $\mathbf{b} \in l^2$ then

$$\langle \mathbf{a}_n, \mathbf{b} \rangle_{l^2} = \mathbf{b}(n).$$

Because $\|\mathbf{b}\|_{l^2} < \infty$ it is clear that

$$\lim_{n \to \infty} \mathbf{b}(n) = 0$$

and therefore $\mathbf{a}_n$ converges weakly to $\mathbf{0}$.

Exercise A.4.6. Suppose that $< f_n > \subset L^2([0,1])$ and $< f_n >$ has a weak limit, show that it is unique.

A.4.5 The L^2-case

Of particular note is the case $p = 2$, for in this case (and only this case) $p = q$. That is $L^2([0,1])$ is its own dual vector space. This distinction is already familiar from the finite dimensional case. The space $L^2([0,1])$ has an inner product which defines its metric. It is given by

$$\langle f, g \rangle_{L^2} = \int_0^1 f(x) g(x) dx.$$

Hölder's inequality in this case is just the infinite dimensional analogue of the Cauchy-Schwarz inequality,

$$|\langle f, g \rangle_{L^2}| \leq \|f\|_{L^2} \|g\|_{L^2}. \tag{A.97}$$

An inner product can be defined on every finite dimensional vector space. This is **false** in infinite dimensions. Among the L^p-spaces, L^2 is the only space which has an inner product defining its norm

As before (A.97) leads to a definition of the angle, θ between two vectors, f, g by *defining*

$$\cos\theta = \frac{\langle f, g \rangle_{L^2}}{\|f\|_{L^2} \|g\|_{L^2}}.$$

We say that two vectors f, g are *orthogonal* if $\langle f, g \rangle_{L^2} = 0$. In this special case we can extend the concept of an orthonormal basis.

Definition A.4.6. A set of vectors $\{e_j\} \subset L^2([0,1])$ is **orthonormal** if

$$\langle e_i, e_j \rangle_{L^2} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{A.98}$$

The set $\{e_j\}$ is an *orthonormal basis* if for every vector $f \in L^2([0,1])$ there is a sequence of numbers $< a_j >$ so that

$$\lim_{N\to\infty} \|f - \sum_{j=1}^{N} a_j e_j\|_{L^2} = 0. \tag{A.99}$$

In this case we write

$$f = \sum_{j=1}^{\infty} a_j e_j. \tag{A.100}$$

It follows from (A.98) and (A.99) that

$$\|f\|_{L^2}^2 = \sum_{j=1}^{\infty} |a_j|^2. \tag{A.101}$$

This shows that $L^2([0,1])$ is a reasonable, infinite dimensional analogue of Euclidean space with its Euclidean norm.

Example A.4.10. To prove that a set of functions defines an orthonormal basis for $L^2([0,1])$ is a highly non-trivial matter. The functions

$$\{1\} \cup \{\sqrt{2}\cos(n\pi x) \mid n = 1, \dots\}$$

define an orthonormal basis for $L^2([0,1])$ as do

$$\{\sqrt{2}\sin(n\pi x) \mid n = 1, 2, \dots\}$$

and

$$\{\exp(2\pi i n x) \mid n \in \mathbb{Z}\}.$$

These facts are the foundation of Fourier analysis.

If $\{e_j\}$ is an orthonormal basis for $L^2([0,1])$ then the coefficients $< a_j >$, appearing in (A.100), are computed by taking the inner products of both sides of this equation with the vectors $\{e_j\}$. Applying (A.98) we deduce that

$$a_j = \langle f, e_j \rangle_{L^2}, \tag{A.102}$$

just as in the finite dimensional case. By introducing an orthonormal basis we can replace a function by an infinite sequence of numbers. This introduces problems in computation that are quite similar to those introduced by infinite decimal expansions. We need to find a way to approximate functions by finite sequences of numbers. A obvious choice is to say that

$$f \simeq \sum_{j=1}^{N} \langle f, e_j \rangle_{L^2} e_j.$$

What is the error made replacing f by a partial sum of it expansion in terms of this basis? Because we are working with functions in L^2 the most reasonable way to measure the error is in terms of the L^2-norm. Using (A.98) and (A.101) we see that

$$\|f - \sum_{j=1}^{N} \langle f, e_j \rangle_{L^2} e_j\|_{L^2}^2 = \sum_{j=N+1}^{\infty} |\langle f, e_j \rangle_{L^2}|^2. \tag{A.103}$$

Since the sum in (A.101) is finite, this sum can be made as small as one likes by choosing N sufficiently large. How large N must be clearly depends on f, in our study of Fourier series we examine this question carefully.

A.4.6 Generalized functions on $\mathbb{R}$

Within mathematics and also in its applications the fact that many functions are not differentiable can be a serious difficulty. Within the context of *linear analysis, generalized functions* or *distributions* provides a very comprehensive solution to this problem. Though it is more common in the mathematics literature, we avoid the term "distribution," because there are so many other things in imaging that go by this name. In this section we outline the theory of generalized functions and give many examples. The reader wishing to attain a degree of comfort with these ideas is strongly urged to do the exercises at the end of the section.

Let $\mathcal{C}_c^\infty(\mathbb{R})$ denote infinitely differentiable functions defined on $\mathbb{R}$ which vanish outside of bounded sets. These are sometimes called *test functions*.

Definition A.4.7. A generalized function on $\mathbb{R}$ is a linear function, l defined on the set of test functions such that there is a constant C and an integer k so that, for every $f \in \mathcal{C}_c^\infty(\mathbb{R})$ we have the estimate

$$|l(f)| \leq C \sup_{x\in\mathbb{R}} \left[(1+|x|)^k \sum_{j=0}^{k} |\partial_x^j f(x)| \right] \tag{A.104}$$

These are linear functions on $\mathcal{C}_c^\infty(\mathbb{R})$ which are, in a certain sense continuous. The constants C and k in (A.104) depend on l but do not depend on f. The expression on the right hand side defines a norm on $\mathcal{C}_c^\infty(\mathbb{R})$, for convenience we let

$$\|f\|_k = \sup_{x\in\mathbb{R}} \left[(1+|x|)^k \sum_{j=0}^{k} |\partial_x^j f(x)| \right].$$

If $f \in \mathcal{C}_c^\infty(\mathbb{R})$ then it easy to show that $\|f\|_k$ is finite for every $k \in \mathbb{N} \cup \{0\}$.

A few examples of generalized function should help clarify the definition.

Example A.4.11. The most famous generalized function of all is the Dirac δ-function. If is defined by

$$\delta(f) = f(0).$$

It is immediate from the definition that $f \mapsto \delta(f)$ is linear and

$$|\delta(f)| \leq \|f\|_0,$$

so the δ-function *is* a generalized function. For $j \in \mathbb{N}$ define

$$\delta^{(j)}(f) = \partial_x^j f(0).$$

Since differentiation is linear, these also define linear functions on $\mathcal{C}_c^\infty(\mathbb{R})$ which satisfy the estimates

$$|\delta^{(j)}(f)| \leq \|f\|_j.$$

Hence these are also generalized functions.

Example A.4.12. Let $\varphi(x)$ be a function which is integrable on any finite interval and such that

$$C_\varphi = \int_{-\infty}^{\infty} |\varphi(x)|(1+|x|)^{-k} \leq \infty$$

for some non-negative integer k. Any such function defines a generalized function

$$l_\varphi(f) = \int_{-\infty}^{\infty} f(x)\varphi(x)dx.$$

Because f has bounded support the integral converges absolutely. The linearity of the integral implies that $f \mapsto l_\varphi(f)$ is linear. To prove the estimate we observe that

$$|f(x)| \leq \frac{\|f\|_k}{(1+|x|)^k}$$

and therefore

$$|l_\varphi(f)| \leq \int_{-\infty}^{\infty} \|f\|_k \frac{|\varphi(x)|}{(1+|x|)^k} dx = C_\varphi \|f\|_k.$$

Thus l_φ is also a generalized function. This shows that every function in $\mathcal{C}_c^\infty(\mathbb{R})$ defines a generalized function, so that, in a reasonable sense, a generalized function is a generalization of a function!

Example A.4.13. Recall that the Cauchy principal value integral is defined, when the limit exists, by

$$\text{P. V.} \int_{-\infty}^{\infty} f(x)dx = \lim_{\epsilon \downarrow 0} \left[\int_{-\infty}^{-\epsilon} f(x)dx + \int_{\epsilon}^{\infty} f(x)dx \right],$$

A generalized function is defined by

$$l_{1/x}(f) = \text{P. V.} \int_{-\infty}^{\infty} \frac{f(x)dx}{x}.$$

Because $1/x$ is not integrable in any neighborhood of 0 the ordinary integral $f(x)/x$ is not defined. The principal value is well defined for any test function and defines a generalized function. To prove this observe that, for any $\epsilon > 0$,

$$\int_{-1}^{-\epsilon} \frac{f(x)dx}{x} + \int_{\epsilon}^{1} \frac{f(x)dx}{x} = \int_{-1}^{-\epsilon} \frac{(f(x)-f(0))dx}{x} + \int_{\epsilon}^{1} \frac{(f(x)-f(0))dx}{x}.$$

This is because $1/x$ is an odd function and the region of integration is symmetric about 0. The ratio $(f(x)-f(0))/x$ is a smooth bounded function in a neighborhood of 0 and

therefore the limit exists as $\epsilon \to 0$. This shows that

$$l_{1/x}(f) = \int_{-1}^{1} \frac{(f(x) - f(0))dx}{x} + \int_{|x|\geq 1} \frac{f(x)dx}{x}.$$

It is left as an exercise to show that

$$|l_{1/x}(f)| \leq C\|f\|_1. \tag{A.105}$$

As noted in example A.4.12 the map $f \mapsto l_f$ identifies every smooth function with a unique generalized function. If the world were very simple then *every* generalized function would be of this form for some locally integrable function. But this is not true! It is not hard to show that the δ-function is not of this form: Suppose that $\delta = l_\varphi$ for some locally integrable function φ. One can show that $\varphi(x)$ must vanish for all $x \neq 0$, this is because $\delta(f)$ only depends on $f(0)$. But an integrable function supported at one point has integral 0 so

$$l_\varphi(f) = 0$$

for all $f \in \mathcal{C}^\infty_{\mathrm{c}}(\mathbb{R})$.

Recall that our goal is to extend the notion of differentiability. The clue to how this should be done is given by the integration by parts formula. Let f and g be test functions, then

$$\int_{-\infty}^{\infty} \partial_x f(x)g(x)dx = -\int_{-\infty}^{\infty} f(x)\partial_x g dx \tag{A.106}$$

Thinking of f as a generalized function, this formula can be rewritten as

$$l_{\partial_x f}(g) = l_f(-\partial_x g). \tag{A.107}$$

The right hand side of (A.106) *defines* a generalized function which we identify as the derivative of the l_f,

$$[\partial_x l_f]\,(g) \stackrel{d}{=} -l_f(\partial_x g). \tag{A.108}$$

This equation is really just notation, the following proposition shows that the underlying idea can be used to define the derivative of any generalized function.

Proposition A.4.1. *If l is a generalized function defined on $\mathbb{R}$ then*

$$l'(f) = l(\partial_x f)$$

is also a generalized function.

Proof. Because ∂_x maps $\mathcal{C}^\infty_{\mathrm{c}}(\mathbb{R})$ to itself the linear function, l' is well defined we only need to prove that it satisfies an estimate of the form (A.104). As l is a generalized function there is a C and k so that

$$|l(f)| \leq C\|f\|_k.$$

From the definition of l' it is clear that

$$|l'(f)| \leq C\|\partial_x f\|_k.$$

The proof is completed by showing that

$$\|\partial_x f\|_k \leq \|f\|_{k+1}.$$

This is left as an exercise. □

With this proposition we can now define the derivative of a generalized function. It is very important to keep in mind that the derivative of a generalized function is another generalized function! To distinguish this concept of derivative from the classical one, the derivative of a generalized function is called a *weak derivative*.

Definition A.4.8. Let l be a generalized function. The *weak derivative* of l is the generalized function $l^{[1]}$ defined by

$$l^{[1]}(f) \stackrel{d}{=} -l(\partial_x f). \tag{A.109}$$

If $l = l_f$ for a smooth function f, then $l^{[1]} = l_{\partial_x f}$, so this definition *extends* the usual definition of derivative. Because every generalized function is differentiable and its weak derivative is another generalized function, it follows that every generalized function is twice differentiable. Indeed arguing recursively it follows that every generalized function is *infinitely* differentiable. Let $\{l^{[j]}\}$ denote the successive weak derivatives of l. It is left as an exercise for the reader to prove the general formula

$$l^{[j]}(f) \stackrel{d}{=} (-1)^j l(\partial_x^j f). \tag{A.110}$$

Example A.4.14. The weak derivative of the δ-function is just $-\delta^{(1)}$ as already defined in example A.4.11. The definition states that

$$\delta^{[1]}(f) = -\delta(\partial_x f) = -\partial_x f(0).$$

It is clear that the weak derivative of $\delta^{[1]}$ is $-\delta^{(2)}$ and so on.

Example A.4.15. Let $\varphi(x) = \chi_{[0,\infty)}(x)$. Since φ is bounded and piecewise continuous is clearly defines a generalized function. This function also has a classical derivative away from $x = 0$, but is not even continuous at 0. Nonetheless, it has a weak derivative as a generalized function. To find it we apply the definition:

$$\begin{aligned} l^{[1]}_{\chi_{[0,\infty)}}(f) &= -l_{\chi_{[0,\infty)}}(\partial_x f) \\ &= -\int_0^\infty \partial_x f(x)dx = f(0). \end{aligned} \tag{A.111}$$

This shows that $l^{[1]}_{\chi_{[0,\infty)}} = \delta$. This is an example of an ordinary function, whose weak derivative, as a generalized function, is not represented by an ordinary function. However, we have accomplished exactly what we set out to do, because now the function $\chi_{[0,\infty)}$ has a derivative.

Example A.4.16. If $f(x)$ is a smooth function, with bounded support then the previous example generalizes to shows that

$$l^{[1]}_{f\chi_{[0,\infty)}} = f(0)\delta + l_{\partial_x f}. \tag{A.112}$$

The set of generalized functions is a vector space. If l and k are generalized functions then so is the sum

$$(l+k)(f) \stackrel{d}{=} l(f) + k(f)$$

as well as scalar multiples

$$(al)(f) \stackrel{d}{=} a(l(f)) \text{ for } a \in \mathbb{R}.$$

Differentiation is a linear operation with respect to this vector space structure, i.e.

$$(l+k)^{[1]} = l^{[1]} + k^{[1]} \text{ and } (al)^{[1]} = al^{[1]}.$$

The notion of weak convergence is perfectly adapted to generalized functions.

Definition A.4.9. A sequence $\{l_n\}$ of generalized functions converges weakly to a generalized function l if, for every test function f,

$$\lim_{n\to\infty} l_n(f) = l(f).$$

Weak derivatives of generalized functions behave very nicely under weak limits.

Proposition A.4.2. *If $< l_n >$ is a sequence of generalized functions which converge weakly to a generalized function l then for every $j \in \mathbb{N}$ the sequence of generalized functions $< l_n^{[j]} >$ converges weakly to $l^{[j]}$.*

Generalized functions seem to have many nice properties and they provide a systematic way to define derivatives of all functions, though the derivatives are, in general **not** functions. Multiplication is the one basic operation that cannot be done with generalized functions. Indeed, it is a theorem in mathematics that there is no way to define a product on generalized functions so that $l_f \cdot l_g = l_{fg}$. However if f is a test function and l is a generalized function then the product $f \cdot l$ is defined, it is

$$(f \cdot l)(g) \stackrel{d}{=} l(fg).$$

This product satisfies the usual Leibniz formula

$$(f \cdot l)^{[1]} = f \cdot l^{[1]} + \partial_x f \cdot l. \tag{A.113}$$

This is generalized slightly in exercise A.4.16.

We close this brief introduction to the idea of a generalized function with a proposition that gives a fairly concrete picture of the "general" generalized function in term of easier to imagine examples.

Proposition A.4.3. *If l is a generalized function then there is a sequence of test functions $< f_n >$ such that l is the weak limit of the sequence of generalized functions $< l_{f_n} >$.*

In other words, any generalized function is the weak limit of generalized functions defined by integration.

Example A.4.17. Let $\varphi(x)$ be a smooth non-negative function with support in $(-1,1)$ normalized so that

$$\int_{-\infty}^{\infty} \varphi(x)dx = 1.$$

For each $n \in \mathbb{N}$ define $\varphi_n(x) = n\varphi(nx)$, then δ is the limit of l_{φ_n}.

The generalized functions considered in this section are usually called *tempered distributions* in the mathematics literature. This is because they have "tempered growth" at infinity. A more systematic development and proofs of the results in this section can be found in [5]. The theory of generalized functions extends essentially verbatim to $\mathbb{R}^n$. A very complete treatment of this subject including its higher dimensional generalizations is given in [28].

Exercise A.4.7. Show that if $\varphi = e^{|x|}$ then l_φ is **not** a generalized function.

Exercise A.4.8. Prove (A.105).

Exercise A.4.9. Suppose that $f \in \mathcal{C}_c^\infty(\mathbb{R})$ show that

$$\|\partial_x f\|_k \leq \|f\|_{k+1}.$$

Exercise A.4.10. Prove (A.110).

Exercise A.4.11. Compute the derivative of $l_{1/x}$.

Exercise A.4.12. Let $\varphi(x) = (1-|x|)\chi_{[-1,1]}(x)$ and, for $n \in \mathbb{N}$ set

$$\varphi_n(x) = n\varphi(nx).$$

Prove that l_{φ_n} converges to δ. Show by direct computation that $l^{[1]}_{\varphi_n}$ converges to $\delta^{[1]}$.

Exercise A.4.13. Prove (A.112).

Exercise A.4.14. Prove Proposition A.4.2.

Exercise A.4.15. Prove (A.113).

Exercise A.4.16. Let l be a generalized function and $f\mathcal{C}^\infty(\mathbb{R})$ a function with *tempered growth*. This means that there is a $k \in \mathbb{N}$ and constants $\{C_j\}$ so that

$$|\partial_x^j f(x)| \leq C_j(1+|x|)^k.$$

Show that $(f \cdot l)(g) \stackrel{d}{=} l(fg)$ defines a generalized function.

Exercise A.4.17. Show that any polynomial is a function of tempered growth. Show that a smooth periodic function is a function of tempered growth.

A.4.7 Generalized functions on $\mathbb{R}^n$.

The theory of generalized function extends essentially verbatim to functions of several variables. We give a very brief sketch. For each non-negative integer k define a semi-norm on $CIc(\mathbb{R}^n)$ by setting

$$\|f\|_k = \sup_{x\in\mathbb{R}^n}\left[(|1+\|x\||)^k \sum_{\boldsymbol{\alpha}\ |\boldsymbol{\alpha}|\le k} |\partial_x^{\boldsymbol{\alpha}} f(x)|\right].$$

Here $\boldsymbol{\alpha}$ is an n-multi-index, i.e. an n-tuple of non-negative integers, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)$ with

$$\partial_x^{\boldsymbol{\alpha}} \stackrel{d}{=} \partial_{x_1}^{\alpha_1} \cdots \partial_{x_n}^{\alpha_n} \text{ and } |\boldsymbol{\alpha}| \stackrel{d}{=} \alpha_1 + \cdots + \alpha_n.$$

Definition A.4.10. A linear function $l : \mathcal{C}_c^\infty(\mathbb{R}^n) \to \mathbb{R}$ is a generalized function if there exists a $k \in \mathbb{N} \cup \{0\}$ and a constant C so that

$$|l(f)| \le C\|f\|_k.$$

As before the set of generalized functions is a vector space. A sequence of generalized functions, $< l_n >$ converges weakly to a generalized function l provided that

$$\lim_{n\to\infty} l_n(f) = l(f) \text{ for every } f \in \mathcal{C}_c^\infty(\mathbb{R}^n).$$

Example A.4.18. The Dirac δ-function is defined in n-dimensions by

$$\delta(f) = f(0).$$

It satisfies the estimate $|\delta(f)| \le \|f\|_0$ and is therefore a generalized function.

Example A.4.19. If φ is a locally integrable function of tempered growth, i.e. there is a $k \ge 0$ and a constant so that

$$|\varphi(x)| \le C(1+\|x\|)^k$$

then

$$l_\varphi(f) = \int_{\mathbb{R}^n} \varphi(x) f(x) dx$$

satisfies

$$|l_\varphi(f)| \le C'\|f\|_{n+1+k}. \tag{A.114}$$

This shows that l_φ is a generalized function.

If $\boldsymbol{\alpha}$ is an n-multi-index and $f \in \mathcal{C}_c^\infty(\mathbb{R}^n)$ then $\partial_x^{\boldsymbol{\alpha}} f$ is also in $\mathcal{C}_c^\infty(\mathbb{R}^n)$ and satisfies the estimates

$$\|\partial_x^{\boldsymbol{\alpha}} f\|_k \le \|f\|_{k+|\boldsymbol{\alpha}|}.$$

As before this allows us to extend the notion of partial derivatives to generalized functions.

Definition A.4.11. If l is a generalized function then, for $1 \le j \le n$ the weak j^{th}-partial derivative of l is the generalized function defined by

$$[\partial_{x_j} l](f) = (-1) l(\partial_{x_j} f). \tag{A.115}$$

Since $\partial_{x_j} l$ is a generalized function as well, it also partial derivatives. To make a long story short, for an arbitrary multi-index $\boldsymbol{\alpha}$ the weak $\boldsymbol{\alpha}^{\text{th}}$-partial derivative of the generalized function l is defined by

$$[\partial_x^{\boldsymbol{\alpha}} l](f) = (-1)^{|\boldsymbol{\alpha}|} l(\partial_x^{\boldsymbol{\alpha}} f). \tag{A.116}$$

If $f, g \in \mathcal{C}_{\text{c}}^{\infty}(\mathbb{R}^n)$ then the n-dimensional integration by parts formula states that

$$\int\limits_{\mathbb{R}^n} [\partial_{x_j} f(x)] g(x) dx = -\int\limits_{\mathbb{R}^n} [\partial_{x_j} g(x)] f(x) dx. \tag{A.117}$$

Applying this formula recursively gives the integration by parts for higher order derivatives

$$\int\limits_{\mathbb{R}^n} [\partial_x^{\boldsymbol{\alpha}} f(x)] g(x) dx = (-1)^{|\boldsymbol{\alpha}|} \int\limits_{\mathbb{R}^n} [\partial_x^{\boldsymbol{\alpha}} g(x)] f(x) dx. \tag{A.118}$$

It therefore follows that if $f \in \mathcal{C}_{\text{c}}^{\infty}(\mathbb{R}^n)$ then the definition of the weak partial derivatives of l_f is consistent with the classical definition of the derivatives of f in that

$$\partial_x^{\boldsymbol{\alpha}} l_f = l_{\partial_x^{\boldsymbol{\alpha}} f} \text{ for all } \boldsymbol{\alpha}. \tag{A.119}$$

Finally we remark that every generalized function on $\mathbb{R}^n$ is a weak limit of "nice" generalized functions.

Proposition A.4.4. *If l is a generalized function on $\mathbb{R}^n$ then there is a sequence of functions $< f_n > \subset \mathcal{C}_c^{\infty}(\mathbb{R}^n)$ so that*

$$l(g) = \lim_{n \to \infty} l_{f_n}(g) \text{ for all } g \in \mathcal{C}_c^{\infty}(\mathbb{R}^n).$$

Exercise A.4.18. Prove that if $f \in \mathcal{C}_{\text{c}}^{\infty}(\mathbb{R}^n)$ then and $j \leq k$ then

$$|\partial_x^{\boldsymbol{\alpha}} f(x)| \leq \frac{\|f\|_k}{(1 + \|x\|)^k}$$

provided $|\boldsymbol{\alpha}| \leq k$.

Exercise A.4.19. Prove (A.114).

Exercise A.4.20. Let $f \in \mathcal{C}_{\text{c}}^{\infty}(\mathbb{R}^2)$ show that

$$l(f) = \int\limits_{-\infty}^{\infty} f(x, 0) dx$$

defines a generalized function.

Exercise A.4.21. Prove that (A.115) defines a generalized function.

Exercise A.4.22. Show that the right hand side of (A.116) defines a generalized function.

Exercise A.4.23. By writing the integrals over $\mathbb{R}^n$ as iterated 1-dimensional integrals, prove (A.117). Deduce (A.118).

Exercise A.4.24. Prove (A.119).

Exercise A.4.25. Let $\varphi(x,y) = \chi_{[0,\infty)}(x) \cdot \chi_{[0,\infty)}(y)$. Show that

$$\partial_x \partial_y l_\varphi = \delta.$$

Exercise A.4.26. Let $\varphi(x,y) = \frac{1}{4}\chi_{[-1,1]}(x)\chi_{[-1,1]}(y)$ and $\varphi_n(x,y) = n^2\varphi(nx,ny)$. Prove that

$$\lim_{n\to\infty} l_{\varphi_n}(f) = \delta(f) \text{ for every } f \in \mathcal{C}_c^\infty(\mathbb{R}^2).$$

A.5 Bounded linear operators

The state of a physical system is often described by a function or collection of functions. Measurements which can be performed are then modeled by operations performed on the state. The simplest such operations are linear operations.

Example A.5.1. Suppose that the system is a heated plate (of infinite extent, for simplicity) and that we would like to determine the temperature distribution. While it is tempting to say that we should "just measure the temperature" at each point, this is not something which can in practice be done. Real measurement processes always involve some sort of average. Let $T(x,y)$ be the temperature at point (x,y). A reasonable mathematical model for what we can measure is an average of the temperatures over some fixed region. Averages are described mathematically as weighted integrals. Let $w(x,y)$ be the weight function, assume that it is non-negative and is normalized to satisfy

$$\int_{\mathbb{R}^2} w(x,y)dxdy = 1.$$

The weight function provides a mathematical model for the measuring instrument.

For example a uniform, circularly symmetric weight is defined by

$$w(x,y) = \begin{cases} \frac{1}{\pi r^2} & \text{if } x^2+y^2 \le r^2, \\ 0 & \text{if } x^2+y^2 > r^2. \end{cases}$$

The output of the thermometer is then a uniformly weighted average of the temperatures over a disk of radius r. What is actually measured is therefore

$$\begin{aligned} M(T)(x,y) &= \int_{(s-x)^2+(t-y)^2\le r^2} \frac{T(s,t)}{\pi r^2} dsdt \\ &= \int_{\mathbb{R}^2} T(s,t)w(x-s,y-t)dsdt. \end{aligned} \tag{A.120}$$

The measurement is linearly related to the state of the system. That is if T_1 and T_2 are two states then $M(T_1+T_2) = M(T_1)+M(T_2)$ and if $a \in \mathbb{R}$ then $M(aT) = aM(T)$.

From the example we see that if the state of the system is described by a function then an idealized, model measurement is described by a function as well. The measurement should be thought of as a *function* of the state. The measurement process should therefore be thought of as a map. To determine the state of the system from measurements we need to invert this map. From our experience with finite dimensional problems we know that the easiest case to study is that of a linear mapping. In the finite dimensional case we gave a complete theory for solvability of linear equations. This theory is entirely *algebraic* which means that it does not require a way to measure distances.

In infinite dimensions it remains true that linear maps are the simplest to analyze. The most important difference between the finite and infinite dimensional cases is that there is **no** effective way to study the behavior of linear maps without **first** introducing norms. This is quite similar to what we have already seen for linear functions.

Definition A.5.1. Let $(X, \|\cdot\|)$ and $(Y, \|\cdot\|')$ be normed linear spaces. A linear map $A : X \to Y$ is **bounded** if there is a constant M such that, for all $\mathbf{x} \in X$ we have

$$\|A\mathbf{x}\|' \leq M\|\mathbf{x}\|.$$

Such maps are often called **bounded linear operators**.

As in the finite dimensional case this estimate implies that the map is continuous, for

$$\|A\mathbf{x}_1 - A\mathbf{x}_2\|' = \|A(\mathbf{x}_1 - \mathbf{x}_2)\|' \leq M\|\mathbf{x}_1 - \mathbf{x}_2\|.$$

If $(X, \|\cdot\|)$ and $(Y, \|\cdot\|')$ are *complete* normed linear spaces then continuity is equivalent to boundedness, see [16] or [66]. In the finite dimensional case a linear map is invertible if it is onto and its null-space consists of the zero vector. This of course remains true in the infinite dimensional case as well. But in this case, we also need to ask if the inverse is continuous. For the case of complete normed linear spaces this question has a very satisfactory answer.

Theorem A.5.1 (Open Mapping Theorem). *Suppose that $(X, \|\cdot\|)$ and $(Y, \|\cdot\|')$ are complete normed linear spaces and $A : X \to Y$ is a continuous linear map. Then A has a continuous inverse if and only if it is both one-to-one and onto.*

Suppose that $A : X \to Y$ is an invertible linear map between complete normed linear spaces and let B denote its inverse. Because B is continuous we see that there is a constant M so that

$$\|By\| \leq M\|y\|'.$$

For $y = Ax$ this can be rewritten

$$\|Ax\|' \geq \frac{1}{M}\|x\|. \tag{A.121}$$

In the finite dimensional case (A.121) is a consequence of the assumption that the null-space of A equals $\{0\}$. In infinite dimensions it gives a necessary condition for a linear map to be invertible.

Corollary A.5.1. *In order for a linear map $A : X \to Y$ to be invertible it is necessary that A satisfy* (A.121) *for some $M > 0$.*

We now consider examples which illustrate some differences between the finite and infinite dimensional cases.

Example A.5.2. Let $X = l^2$ and $Y = l^2$, we define a linear map by setting

$$A < a_j >=< \frac{a_j}{j} > .$$

Because $j \geq 1$ we see that

$$\|A\mathbf{a}\|_2^2 = \sum_{j=1}^{\infty} \frac{|a_j|^2}{j^2} \leq \sum_{j=1}^{\infty} |a_j|^2.$$

In other words

$$\|A\mathbf{a}\|_2 \leq \|\mathbf{a}\|_2,$$

so $A : l^2 \to l^2$ is a bounded linear map. It is also evident that the only vector $\mathbf{a}$ for which $A\mathbf{a} = 0$ is the zero vector.

Example A.5.3. Let $X = l^2$ and define a linear map $B < a_j >=< ja_j >$. A moments thought shows that this map does not take values in l^2; in order for $B < a_j > \in l^2$ it is necessary that

$$\sum_{j=1}^{\infty} j^2 |a_j|^2 < \infty.$$

For any $< a_j > \in l^2$ the sequence $< ja_j >$ is well defined however, it is not generally a sequence in l^2. Note that $B \circ A\mathbf{a} = \mathbf{a}$ and so B is *formally* equal to A^{-1}. On the other hand B is not a bounded operator, since $B\mathbf{a} \notin l^2$ for many vectors $\mathbf{a} \in l^2$. This shows that, even though $A : l^2 \to l^2$ is continuous and its null space equals $\mathbf{0}$, it is *not* invertible, in the sense that its inverse is not a bounded operator. This is in marked contrast to the finite dimensional case.

The operator A turns out to be a reasonable model for a measurement process. The fact that its inverse is not bounded is typical of the difficulties which arise in trying to determine the exact state of a system from realistic measurements. Suppose that the actual state of our system is given by the vector $\mathbf{a}$, so the exact measurements would be $\mathbf{b} = A\mathbf{a}$. Let $\delta\mathbf{b}$ be the uncertainty in our measurements. In many situations, the uncertainty can be made small in the sense that

$$\|\delta\mathbf{b}\|_2^2 = \sum_{j=1}^{\infty} |\delta b_j|^2 << 1.$$

However this is not adequate because we actually need to have

$$\sum_{j=1} j^2 |\delta b_j|^2 << 1$$

in order to be able to reliably determine the state of our system from the measurements. As this is not usually possible in practice it is necessary to use a regularized approximate inverse.

Example A.5.4. The operator A in example A.5.2 is not invertible because it does not satisfy the estimate (A.121) for any M. For each N define the operator

$$B_N < a_j >= \begin{cases} ja_j & \text{for } j \leq N, \\ 0 & \text{for } j > N. \end{cases}$$

This is an approximate left inverse for A in the sense that

$$B_N \circ A < a_j >= \begin{cases} a_j & \text{for } j \leq N, \\ 0 & \text{for } j > N. \end{cases}$$

Thus we have the identity

$$\|\mathbf{a} - B_N \circ A\mathbf{a}\|_2^2 = \sum_{j=N+1}^{\infty} |a_j|^2.$$

In applications, the coefficients $(a_{N+1}, a_{N+2}, \dots)$ would represent the "high frequency" information in $< a_j >$ which is attenuated by the measurement process and corrupted by noise. As such it cannot be measured reliably. For an appropriate choice of N, B_N provides an approximate inverse to our measurement process A which captures all the reliable data that is present in the measurements and discards the parts of the measurement that are not usable. The choice of N depends on the resolution and accuracy of the measuring device.

Example A.5.5. We consider another bounded operator defined on l^p for any $p \geq 1$. It is defined by setting

$$S : (a_1, a_2, a_3, \dots) \mapsto (0, a_1, a_2, a_3, \dots).$$

This is called a shift operator. For any $p \geq 1$ we see that

$$\|S\mathbf{a}\|_p = \|\mathbf{a}\|_p.$$

Clearly $S\mathbf{a} = 0$ if an only if $\mathbf{a} = \mathbf{0}$. However we see that the image of S is not all of l^p. A vector $\mathbf{b} = (b_1, b_2, b_3, \dots)$ is in the image of S if and only if $b_1 = 0$. This is different way that a linear transformation of infinite dimensional space having only the zero vector in its null space can fail to be invertible.

Linear transformations of function spaces are often written as integrals. For example if $k(x, y)$ is a function defined on $[0, 1] \times [0, 1]$ then

$$Kf(x) = \int_0^1 k(x, y) f(y) dy$$

defines a linear transformation. It can be subtle to decide whether or not this is a bounded operator. Here is a simple criterion which implies that $K : L^2([0, 1]) \to L^2([0, 1])$ is bounded.

Proposition A.5.1. *Suppose that $k(x,y)$ satisfies*

$$\iint\limits_{[0,1]\times[0,1]} |k(x,y)|^2 dxdy < \infty,$$

then the linear operator $K : L^2([0,1]) \to L^2([0,1])$ is bounded.

Proof. The proof is an application of the Hölder inequality. We need to show that there is a constant M so that

$$\int\limits_0^1 |Kf(x)|^2 dx \leq M \int\limits_0^1 |f(x)|^2 dx$$

for every $f \in L^2([0,1])$. If we write out the left hand side and use the Hölder inequality we see that

$$\begin{aligned}
\int\limits_0^1 |Kf(x)|^2 dx &= \int\limits_0^1 | \int\limits_0^1 k(x,y) f(y) dy|^2 dx \\
&\leq \int\limits_0^1 \left[\int\limits_0^1 |k(x,y)|^2 dy \int\limits_0^1 |f(y)|^2 dy \right] dx \qquad \text{(A.122)} \\
&= \left[\int\limits_0^1 |f(y)|^2 dy \right] \int\limits_0^1 \int\limits_0^1 |k(x,y)|^2 dy dx.
\end{aligned}$$

From (A.122) we see that

$$\int\limits_0^1 |Kf(x)|^2 dx \leq \left[\int\limits_0^1 \int\limits_0^1 |k(x,y)|^2 dy dx \right] \int\limits_0^1 |f(x)|^2 dx,$$

which establishes the needed estimate with

$$M = \sqrt{\int\limits_0^1 \int\limits_0^1 |k(x,y)|^2 dy dx}.$$

□

Exercise A.5.1. Define an operator

$$Kf(x) = \int\limits_0^x (x-y) f(y) dy.$$

Show that K is a bounded operator on $L^2([0,1])$ and that

$$\partial_x^2 Kf = f \text{ and } Kf(0) = \partial_x Kf(0) = 0.$$

Exercise A.5.2. Let $k(x,y)$ be a function defined on $[0,1]\times[0,1]$ for which there is a constant M such that

$$\max_{x\in[0,1]}\int_0^1 |k(x,y)|dy \leq M \text{ and } \max_{y\in[0,1]}\int_0^1 |k(x,y)|dx \leq M.$$

Show that the operator $f\mapsto Kf$ defined by $k(x,y)$ is a bounded operator from $L^2([0,1])\to L^2([0,1])$.

Exercise A.5.3. Define an operator $A: L^1(\mathbb{R})\to \mathcal{C}^0(\mathbb{R})$ by letting

$$Af(x) = \int_x^{x+1} f(s)ds.$$

Prove that A is bounded and that $Af=0$ implies that $f=0$. Is A invertible? Why or why not?

Exercise A.5.4. In example A.5.2 show directly that there is no constant $M>0$ so that A satisfies the estimate (A.121) for all $\mathbf{a}\in L^2$.

A.6 Functions in the real world

In section A.4 we considered functions from the point of view of a mathematician. In this approach, functions are described by abstract *properties* such as differentiability or integrability. Using these properties functions are grouped together into normed vector spaces. The principal reason for doing this is to study the mapping properties of linear transformations. It is a very abstract situation because we do not, even in principle, have a way to compute most of the functions under consideration: they are described by their properties and not defined by rules or formulæ. This level of abstraction even leads to difficulties in the mathematical development of the subject. In practice we can only *approximately* measure a function, in most circumstances, at a finite collection of values. What mediates between these two, very different views of functions? This is a question with many different answers, but the basic ideas involve the concepts of *approximation*, *sampling* and *interpolation*.

A.6.1 Approximation

The basic problem of approximation theory is to begin with a function from an abstract class, for example continuous functions and approximate it, in an appropriate sense, by functions from a more concrete class, for example polynomials. We begin with the basic theorem in this subject.

Theorem A.6.1 (The Weierstrass Approximation Theorem). *Given a function $f\in\mathcal{C}^0([0,1])$ and an $\epsilon>0$ there is a polynomial p such that*

$$\|f-p\|_{\mathcal{C}^0} < \epsilon. \tag{A.123}$$

The set of polynomial functions has an analogous relationship to the set of continuous functions as the set of finite decimal expansions has to the set of real numbers. For the purposes of approximate computations (even with a specified error) it suffices to work with polynomials. This theorem is the prototype for many other such results.

A very useful result for L^p-spaces, uses approximation by *step functions*. Recall that if E is an subset of $\mathbb{R}$ then its *characteristic function* is defined by

$$\chi_E(x) = \begin{cases} 1 & \text{if } x \in E, \\ 0 & \text{if } x \notin E. \end{cases}$$

Definition A.6.1. A function f is called a **step function** if there is a finite collection of intervals $\{[a_i, b_i) \,:\, i = 1, \ldots, N\}$ and constants $\{c_i\}$ so that

$$f(x) = \sum_{i=1}^{N} c_i \chi_{[a_i, b_i)}(x).$$

Step functions are computable functions.

Theorem A.6.2 (L^p-approximation Theorem). *Suppose that $1 \leq p < \infty$, $f \in L^p(\mathbb{R})$ and $\epsilon > 0$ is given. There exists a step function F such that*

$$\|f - F\|_{L^p} < \epsilon. \tag{A.124}$$

Note that $p = \infty$ is excluded, the theorem is false in this case. The proof of this theorem uses the definition of the Lebesgue integral and the structure of Lebesgue measurable sets. It is beyond the scope of this text but can be found in [16]. It has a very useful corollary.

Corollary A.6.1. *Suppose that $1 \leq p < \infty$, $f \in L^p(\mathbb{R})$ and $\epsilon > 0$ is given. There exists a continuous function G such that*

$$\|f - G\|_{L^p} < \epsilon. \tag{A.125}$$

Proof. Theorem A.6.2 gives the existence of a step function F so that $\|f - F\|_{L^p} < \epsilon/2$. This means that it suffices to find a continuous function G so that

$$\|F - G\|_{L^p} < \frac{\epsilon}{2}.$$

In light of exercise A.6.1 there is a sequence $a = a_0 < a_1 < \cdots < a_m = b$ and constants $\{c_j\}$ so that

$$F(x) = \sum_{j=1}^{m} c_j \chi_{[a_{j-1}, a_j)}(x).$$

Such a function is easily approximated, in the L^p-norm by continuous, piecewise linear functions. Fix an $\eta > 0$, so that

$$2\eta < \min\{a_i - a_{i-1} \,:\, i = 1, \ldots, m\}.$$

For each $1 \geq j < m$ define the piecewise linear function

$$l_j(x) = \begin{cases} 0 & \text{if } |x - a_j| > \eta, \\ c_j \frac{x - (a_j - \eta)}{2\eta} + c_{j-1} \frac{(a_j + \eta - x)}{2\eta} & \text{if } |x - a_j| \leq \eta. \end{cases}$$

For $j = 0$ or m we let

$$l_0(x) = \begin{cases} 0 & \text{if } |x - a_0| > \eta, \\ c_j \frac{x-(a_0-\eta)}{2\eta} & \text{if } |x - a_0| \le \eta, \end{cases} \quad l_m(x) = \begin{cases} 0 & \text{if } |x - a_m| > \eta, \\ c_m \frac{(a_m+\eta-x)}{2\eta} & \text{if } |x - a_m| \le \eta. \end{cases}$$

A continuous, piecewise linear function is defined by

$$G = \sum_{j=0}^{m} l_j(x) + \sum_{j=1}^{m} \chi_{[a_{j-1}+\eta, a_j-\eta)}(x).$$

The L^p-norm of the difference $F - G$ is estimated by

$$\|F - G\|_{L^p}^p \le 2\eta \sum_{j=0}^{m+1} |c_j - c_{j-1}|^p.$$

Here $c_{-1} = c_{m+1} = 0$. As η can be made arbitrarily small, this proves the corollary. □

While the precise statements of these three results are quite different, their structures are identical: In each case we have a normed vector space $(V, \|\cdot\|)$ and a subspace $S \subset V$ consisting of computable functions. The theorems assert that, if the norm $\|\cdot\|$ is used to measure the error then the set S is dense in V. Neither theorem addresses the problem of finding the approximating function, though the usual proofs of these theorems provide algorithms, at least in principle. Analogous statement hold with $\mathbb{R}$ replaced by $\mathbb{R}^n$ or finite intervals.

Let us return to the problem of polynomial approximation for continuous functions. We can ask a more precise question: how well can a given continuous function f be approximated by a polynomial of degree n. Let $\mathcal{P}_n$ denote the polynomials of degree at most n and define

$$E_n(f) \stackrel{d}{=} \min\{\|f - p\|_{\mathcal{C}^0} \ : \ p \in \mathcal{P}_n\}. \tag{A.126}$$

Weierstrass' theorem implies that for any $f \in \mathcal{C}^0([0,1])$

$$\lim_{n\to\infty} E_n(f) = 0.$$

This suggests two questions:

(1). Is there an element $p_n \in \mathcal{P}_n$ for which

$$\|f - p_n\|_{\mathcal{C}^0} = E_n(f)?$$

(2). Is there an estimate for the rate at which $E_n(f)$ goes to zero?

The answer to the first question yes: If f is a continuous function then there is a unique polynomial $p_n \in \mathcal{P}_n$ such that

$$\|f - p_n\|_{\mathcal{C}^0} = E_n(f).$$

The answer to the second question turns out to depend on the smoothness of f.

Theorem A.6.3 (Jackson's Theorem). *If $f \in \mathcal{C}^k([0,1])$ for a $k \in \mathbb{N}$ then*

$$E_n(f) \leq \frac{C}{n^k}.$$

The smoother the function, the faster the sequence of "best" approximating polynomials converges to it. These facts suggest another question: Can the "best" approximating polynomial be found? The answer, at present is no. One might then ask if another approximation $q_n \in \mathcal{P}_n$ can be found such that $\|f - q_n\|_{\mathcal{C}^0}$ goes to zero at about same rate as $E_n(f)$? The answer to this question is yes but it is generally quite complicated to do this in practice. The interested reader is referred to [63]. Below we give a an effective method for finding a sequence $\{q_n\}$ such that $\|f - q_n\|_{\mathcal{C}^0}$ goes to zero at nearly the optimal rate.

We close our discussion of approximation in the $\mathcal{C}^0$-norm with a formula for an approximating sequence of polynomials that works for any continuous function.

Theorem A.6.4 (Bernstein's formula). *Let $f \in \mathcal{C}^0([0,1])$ and define the polynomial of degree n by*

$$B_n(f;x) = \sum_{j=0}^{n} f\left(\frac{j}{n}\right)\binom{j}{n} x^j (1-x)^{n-j}.$$

This sequence converges to f in the $\mathcal{C}^0$-norm, that is

$$\lim_{n\to\infty} \|B_n(f) - f\|_{\mathcal{C}^0} = 0.$$

If f is once differentiable then there is a constant M so that

$$|B_n(f;x) - f(x)| \leq \frac{M}{\sqrt{n}}.$$

Bernstein's formula gives a sequence of polynomials that always works and gives somewhat better results if the function is smoother. However, even for very smooth functions, the Bernstein polynomials do not behave like the best approximants.

The difficulties encountered with finding the best polynomial approximations are mostly a result of using the $\mathcal{C}^0$-norm to measure the error. A much easier approximation problem results from measuring the error in the L^2-norm. Indeed, allowing a slightly more general norm does not introduce any additional difficulties. Let $w(x)$ be a non-negative, integrable function defined on $[0,1]$ and define the L^2-norm *with weight* w to be

$$\|f\|_{2,w} = \int_0^1 |f(x)|^2 w(x) dx.$$

An infinite dimensional generalization of the Gram-Schmidt method leads to a very simple solution to the following problem

$$\text{Find } p_n \in \mathcal{P}_n \text{ such that } \|f - p_n\|_{2,w} = \min\{\|f - q\|_{2,w} \ : \ q \in \mathcal{P}_n\}.$$

Let $\langle \cdot, \cdot \rangle_w$ denote the inner product associated with this weighted L^2-norm

$$\langle f, g \rangle_w = \int_0^1 f(x)g(x)w(x)dx.$$

We use the linearly independent functions $\{1, x, x^2, \dots\}$ to define a sequence of polynomials $\{P_n\}$ by the properties

(1). $\deg P_n = n$,

(2). $\langle P_n, P_m \rangle_w = \delta_{mn}$,

The algorithm is exactly the same as the finite dimensional case:

Step 1 Let $P_0 = [\langle 1, 1 \rangle_w]^{-\frac{1}{2}}$.

Step 2 Suppose that we have found $\{P_0, \dots, P_j\}$. Let

$$\widetilde{P}_{j+1} = x^{j+1} + \sum_{i=0}^{j} \alpha_i P_i$$

where

$$\alpha_i = -\langle x^{j+1}, P_i \rangle_w.$$

This function is orthogonal to $\{P_0, \dots, P_j\}$.

Step 3 Set

$$P_{j+1} = \frac{\widetilde{P}_{j+1}}{\sqrt{\|\widetilde{P}_{j+1}\|_{2,w}}}.$$

Because the set $\{x^j \ : \ j \geq 0\}$ is infinite this procedure does not terminate. Observe that any polynomial p of degree n has a unique expression in the form

$$p = \sum_{j=0}^{n} a_j P_j$$

and that

$$\|p\|_{2,w}^2 = \sum_{j=0}^{n} |a_j|^2.$$

Theorem A.6.5. *If f is function with $\|f\|_{2,w} < \infty$ then the polynomial*

$$p_n = \sum_{j=0}^{n} \langle f, P_j \rangle_w P_j$$

satisfies

$$\|f - p_n\|_{2,w} = \min\{\|f - q\|_{2,w} \ : \ q \in \mathcal{P}_n\}.$$

This function is called the best, weighted, least squares approximation to f of degree n. Thus the best polynomial approximations with respect to these weighted L^2-norms are easy to find.

If f is continuous how does $\|f - p_n\|_{\mathcal{C}^0}$ behave? Do these give better approximations if the function f is smoother? The answers to these questions depend, in part on the weight function. For the case

$$w(x) = [x(1-x)]^{-\frac{1}{2}}$$

the answer happens to be very simple.

Theorem A.6.6. *Suppose that $f \in \mathcal{C}^0([0,1])$ and p_n is the best, weighted, least squares approximation of degree n then*

$$\|f - p_n\|_{\mathcal{C}^0} \leq \left[4 + \frac{4}{\pi^2}\log n\right] E_n(f).$$

The proof of this theorem can be found in [63]. Combining this result with Jackson's theorem we see that if $f \in \mathcal{C}^k([0,1])$ then there is a constant C so that

$$\|f - p_n\|_{\mathcal{C}^0} \leq \frac{C \log n}{n^k}.$$

In other words, the easily found least squares approximants give almost the optimal rate of decrease for the error *measured in the $\mathcal{C}^0$-norm*. This is quite a remarkable and useful fact.

Exercise A.6.1. Suppose that F is a step function then there exists a finite increasing sequence $a_0 < a_1 < \cdots < a_m$ and constants $\{c_1, \ldots, c_m\}$ so that

$$F(x) = \sum_{j=1}^{m} c_j \chi_{[a_{j-1}, a_j)}(x).$$

Exercise A.6.2. Because the computations are easier we work on the interval $[-1,1]$. For each $k \in \mathbb{N}$ show that there is a polynomial

$$T_k(x) = \sum_{j=0}^{k} a_{kj} x^j$$

so that

$$\cos(k\theta) = T_k(\cos(\theta)).$$

Show that the polynomials $\{T_k\}$ satisfy the relations

$$\int_{-1}^{1} \frac{T_k(x) T_l(x) dx}{\sqrt{1-x^2}} = 0 \text{ if } j \neq k.$$

Hint: Use the change of variables $x = \cos(\theta)$. These polynomials are called the Chebyshev polynomials. They have many remarkable properties and are often used in approximation theory, see [62]. Note that setting $y = \frac{1}{2}(1+x)$ maps $[-1,1]$ into $[0,1]$ and

$$y(1-y) = \frac{1-x^2}{4}.$$

A.6.2 Sampling and Interpolation

Suppose that we have a system whose state is described by a function f of a variable t, $f(t)$. The simplest way to model a measurement is as *evaluation* of this function. That is we have a sequence of "times" $< t_j >$ and the measurement consists in evaluating $f(t_j)$. The sequence of numbers $< f(t_j) >$ are called the *samples* of f at the times $< t_j >$. The sample times are usually labeled in a monotone fashion, that is

$$t_j < t_{j+1}.$$

The differences $\Delta t_j = t_j - t_{j-1}$ are called the *sample spacings*. If they are all equal to a single value Δt then we say f is *uniformly sampled with sample spacing* Δt. In a real application we can measure at most finitely many samples and of course we can only measure them with finite precision. In analyzing measurement processes it is often useful to assume that we can evaluate f along an infinite sequence and that the measurements are exact.

A question of primary interest is to decide what the samples $< f(t_j) >$ tell us about the value of $f(t)$ for times t not in our sample set. The answer of course depends on how close together the sample times are **and** *a priori* knowledge of the smoothness of f. Such information is usually incorporated implicitly into a model. Suppose that we sample a *differentiable* function $f(t)$ at the points $\{t_j\}$. Let t lie between t_j and t_{j+1} then the mean value theorem implies that there is a point $\tau \in (t_j, t_{j+1})$ such that

$$f(t) = f(t_j) + f'(\tau)(t - t_j).$$

If the points are close together and the derivative is continuous then

$$f'(\tau) \simeq \frac{f(t_j) - f(t_{j+1})}{t_j - t_{j+1}}.$$

Thus we define

$$F(t) = f(t_j) + \left[\frac{f(t_j) - f(t_{j+1})}{t_j - t_{j+1}}\right](t - t_j), \text{ for } t \in [t_j, t_{j+1}]. \tag{A.127}$$

This is a continuous, piecewise linear function with $F(t_j) = f(t_j), F(t_{j+1}) = f(t_{j+1})$; in general $F(t)$ is not differentiable. We say that F is a piecewise linear function, interpolating f at the points $\{t_j\}$. For a smooth function the error $\|f - F\|_{\mathcal{C}^0}$ goes to zero as the sample spacing goes to zero. However the approximating function is not differentiable. This means that we cannot use F effectively to compute approximate values of $f'(t)$ and the graph of F has "corners" even though the graph of f is smooth.

Example A.6.1. Suppose that $f(t)$ is the price of a stock as a function of time. The value of f varies in a discontinuous and random fashion. Suppose that we sample the price of the stock each day at closing time. Let $< t_j >$ denote that sequence of times. For a time t near to a closing time t_j with $t < t_j$ we can be reasonably confident that $f(t)$ is close to $f(t_j)$. This is because we know something about how prices on the stock market are determined. It is also the case that for times t much earlier in the day $f(t)$ might be quite different from $f(t_j)$. Two schemes for "predicting" the value of $f(t)$ would be: 1. Setting $f(t) = f(t_j)$ for all times t on the same day. 2. Joining the points on the graph $\{(t_j, f(t_j))\}$ by straight

lines. Because of the very random nature of this type of measurement, neither scheme gives very good results. Indeed, if one could find a reliable way to predict the price of a stock, even ten minutes hence, he or she could easily become a very rich person!

If we know that $f(t)$ is a polynomial function of t then a finite number of samples determines f completely. If $f(t)$ is a polynomial of degree 0, in other words a constant, then a single sample determines f. If the degree is 1 then 2 samples are required and if the degree is n then $n+1$-samples suffice. Indeed there are simple explicit formulæ to reconstruct a polynomial from such data. For example, if the sample points are $\{t_1, t_2\}$ then a linear polynomial is reconstructed as follows

$$f(t) = f(t_1)\frac{t-t_2}{t_1-t_2} + f(t_2)\frac{t-t_1}{t_2-t_1}.$$

More generally if f is of degree n and the sample points are $\{t_1, \ldots, t_{n+1}\}$ then

$$f(t) = \sum_{j=1}^{n+1} f(t_j)\frac{\prod_{k\neq j}(t-t_k)}{\prod_{k\neq j}(t_j-t_k)}. \tag{A.128}$$

The expression on the right hand side of (A.128) is called the Lagrange interpolation formula.

If $f(t)$ is a continuous function which we sample at the $n+1$-points $\{t_j\}$ then we can define an n^{th}-degree polynomial using (A.128)

$$F(t) = \sum_{j=1}^{n+1} f(t_j)\frac{\prod_{k\neq j}(t-t_k)}{\prod_{k\neq j}(t_j-t_k)}. \tag{A.129}$$

This polynomial has the property that $F(t_j) = f(t_j)$ for $j = 1, \ldots, n+1$. We say that F is the n^{th}-degree interpolant for f at the points $\{t_1, \ldots, t_{n+1}\}$. The question of principal interest is how well $F(t)$ approximates $f(t)$ for $t \neq t_j$. Perhaps somewhat surprisingly, the answer to this question is that, in general F does a very poor job. Figure A.3 shows graphs of the function $f(t) = |t - \frac{1}{2}|$ along with degree 2, 6 and 12 polynomial interpolants found using equally spaced samples. Note that, as the degree increases, the polynomial provides a worse and worse approximation to f, away from the sample points. For this reason it is unusual to use a high degree polynomial to interpolate the values of a function. This does not contradict the results of the previous subsection on the existence of accurate, high degree polynomial approximations to continuous functions. It only demonstrates that such approximations cannot, in general be found by simply interpolating.

How then can good approximations to sampled functions be found? One answer lies in using functions which are piecewise polynomials of low degree. We consider only the simplest case. Suppose that $f(t)$ is a differentiable function on $[0,1]$ and that we sample it at the points $T_n = \{0 = t_0, \ldots, t_n = 1\}$. Using a piecewise cubic polynomial we can find a function $F(t)$ which interpolates f at the sample points and is itself twice differentiable differentiable.

Definition A.6.2. For T_n a set of points as above define $S(T_n)$ to be the subset of $\mathcal{C}^2([0,1])$ with the property that for each $i \in \{0, n-1\}$ the restrictions of $f \in S(T_n)$ to the intervals $[t_i, t_{i+1}]$ are given by cubic polynomials. Such a function is called a *cubic spline* with *nodes* $\{t_0 < \cdots < t_n\}$.

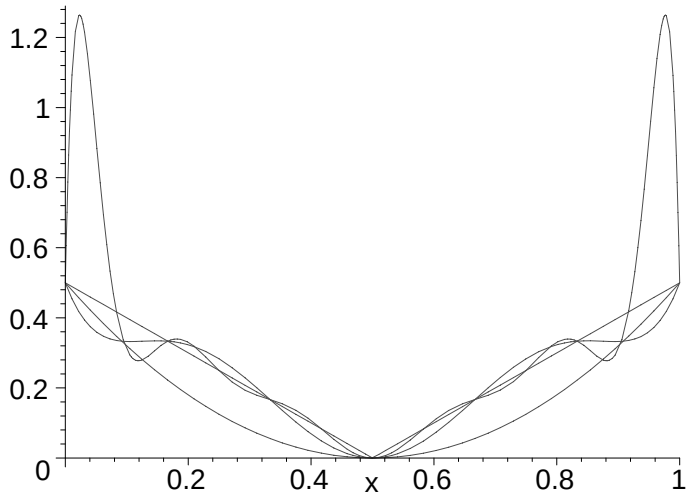


Figure A.3: Polynomial interpolants for $|x - \frac{1}{2}|$

The basic approximation result is the following.

Theorem A.6.7. *Given numbers* $\{f_0, \ldots, f_n\}$ *and* a_0, a_1 *there is a unique cubic spline* $F \in S(T_n)$ *such that*

$$\begin{aligned} F(t_i) &= f_i \text{ for } i = 0, \ldots, n, \\ F'(0) &= a_0 \text{ and } F'(1) = a_1. \end{aligned} \tag{A.130}$$

The theorem tells us that once we fix values for the derivatives at the points 0 and 1 there is a unique cubic spline which interpolates f at the given sample points. The values $\{f(t_j\}$ do not determine the cubic spline interpolant for f, the numbers a_0, a_1 also need to be specified. If we know or can reasonably approximate $f'(0)$ and $f'(1)$ then these give reasonable choices for a_0 and a_1. If this data is not known then another common way to pick a cubic spline F to interpolate f is to require that $F''(0) = F''(1) = 0$. This is sometimes called the *natural cubic spline* interpolating $\{f(t_i)\}$.

The problem of finding cubic splines is easily reduced to a system of linear equations. We give this reduction for the case considered in the theorem with the additional assumption that $t_i - t_{i-1} = h$ for all i. Let $f_i = f(t_i)$, to define the basic building blocks set

$$\begin{aligned} c_i(t) = &\left[\frac{(t-t_{i+1})^2}{h^2} + \frac{2(t-t_i)(t-t_{i+1})^2}{h^3}\right] f_i + \\ &\left[\frac{(t-t_i)^2}{h^2} - \frac{2(t-t_{i+1})(t-t_i)^2}{h^3}\right] f_{i+1} + \\ &\frac{(t-t_i)(t-t_{i+1})^2}{h^2} a_i + \frac{(t-t_{i+1})(t-t_i)^2}{h^2} a_{i+1}. \end{aligned} \tag{A.131}$$

Evaluating this function gives

$$\begin{aligned} c_i(t_{i+1}) &= c_{i+1}(t_{i+1}) = f_{i+1}, \\ c_i'(t_{i+1}) &= c_{i+1}'(t_{i+1}) = a_{i+1}. \end{aligned} \tag{A.132}$$

In other words, for any choice of the values of $\{a_1, \ldots, a_{n-1}\}$ these functions piece together to define a continuously differentiable function, interpolating the values of f. To find the spline with these properties we need to select these coefficients so that the resultant function also has a continuous second derivative. Evaluating the second derivatives and comparing at the adjacent endpoints we derive the relations

$$a_i + 4a_{i+1} + a_{i+2} = \frac{3}{h}(f_{i+2} - f_i), \text{ for } i = 0, \ldots, n-2.$$

These, in turn lead to an invertible system of linear equations for $\{a_1, \ldots, a_{n-1}\}$ which are "tridiagonal." After solving for these coefficients, set

$$F(t) = c_i(t) \text{ for } t \in [t_i, t_{i+1}], \quad i = 0, n-1.$$

Splines have many desirable properties leading to interpolants which have, in a certain sense, the minimal oscillation among all twice differentiable functions which interpolate the given values. If f is twice differentiable then the first derivative of the spline derived above is also a good approximation to $f'(t)$. This discussion is adapted from that given in [63] and [45].

A.7 Numerical techniques for differentiation and integration

In calculus we learn a variety of rules for computing derivatives and integrals of functions given by formulæ. In applications we need to have ways to approximate these operations for measured data. These are discussed in two stages: first we consider the problems of approximating integration and differentiation for *a priori* known functions and then the same questions for noisy sampled data.

If $f(t)$ is a function defined on $[0, 1]$ then its integral can be defined as the following limit

$$\int_0^1 f(t)dt = \lim_{N\to\infty} \sum_{j=0}^{N-1} \frac{1}{N} f\left(\frac{j}{N}\right).$$

Using this sum for a fixed value of N gives an way to approximate an integral, called a *Riemann sum* approximation. For functions with more smoothness, there are better approximations. If f is differentiable then its derivative is also defined as a limit

$$f'(t) = \lim_{\Delta t\to 0} \frac{f(t+\Delta t) - f(t)}{\Delta t}.$$

Using this formula with positive values of Δt leads to approximations for the derivative called *finite differences*.

If $f(t)$ is a continuous function and we set

$$m_N = \max_{0\leq j\leq N-1} \max_{t\in[\frac{j}{N},\frac{j+1}{N}]} |f(t) - f(j/N)|$$

then

$$\left| \int_0^1 f(t)dt = \sum_{j=0}^{N-1} \frac{1}{N} f\left(\frac{j}{N}\right) \right| \leq m_N.$$

To find the analogous estimate for the approximate derivative we let

$$M_N = \max_{|s-t|\leq \frac{1}{N}} |f'(t) - f'(s)|$$

then applying the mean value theorem we see that

$$\left| \frac{f(t+\frac{1}{N}) - f(t)}{N^{-1}} - f'(t) \right| \leq M_N.$$

Comparing these formulæ we see that the accuracy of the approximate integral is controlled by the size of m_N, while that of the approximate derivative is controlled by M_N. If f is differentiable then $m_N \propto N^{-1}$, whereas this implies no estimate for M_N. In order to know that $M_N \propto N^{-1}$ we would need to know that f is *twice* differentiable. This indicates why, in principle it is harder to approximate derivatives than integrals.

For real data, approximate integration is in general much simpler and more accurate than approximate differentiation. The reason for this lies in the nature of noise. Suppose that $f(t)$ represents the "actual" state of our system. One often aggregates various (possibly unknown) sources of error and uncertainty into a single function $n(t)$ which we call *noise*. What is measured are then are actually samples of $f(t) + n(t)$.

Intuitively noise is a random process so it goes up and down unpredictably, such a function is typically not differentiable. This means that $|n(t + \Delta t) - n(t)|$ may well be large compared to Δt. On the other hand if we have a good enough model, then the noise term should be equally likely to be positive or negative, even on small time scales. What is meant by this is that averages of $n(t)$ over small intervals should be, on average small. Symbolically

$$\begin{aligned} \left| \frac{n(t+\Delta t) - n(t)}{\Delta t} \right| &>> 1, \\ \left| \frac{1}{\Delta} \int_t^{t+\Delta t} n(s)ds \right| &<< 1. \end{aligned} \tag{A.133}$$

Exercise A.7.1. Explain why formula (A.128) gives the correct answer if $f(t)$ is known to be a polynomial of degree n.

A.7.1 Numerical integration

In addition to the (right) Riemann sum formula,

$$R_N(f) = \frac{1}{N}\sum_{j=1}^{N} f(\frac{j}{N})$$

there are two other commonly used formulæ for numerical integration, the trapezoidal rule and Simpson's rule. Suppose that $f(t)$ is a continuous function on $[0,1]$, the trapezoidal approximation to the integral of f with $N+1$-points is given by

$$T_N(f) = \frac{1}{2N}(f(0)+f(1)) + \frac{1}{N}\sum_{j=1}^{N-1} f\left(\frac{j}{N}\right). \tag{A.134}$$

If f is twice differentiable then we have the estimate for the error

$$\left| T_N(f) - \int_0^1 f(t)dt \right| \leq \frac{\max_{t\in[0,1]} |f''(t)|}{12N^2}. \tag{A.135}$$

Simpson's rule also uses the midpoints, it is given by

$$S_N(f) = \frac{1}{6N}\left[f(0)+f(1)+2\sum_{j=1}^{N-1} f\left(\frac{j}{N}\right) + 6\sum_{j=0}^{N-1} f\left(\frac{2j+1}{2N}\right)\right]. \tag{A.136}$$

Though Simpson's rule only requires about twice as much computation it gives a much smaller error if f is *four* times differentiable. The error estimate is

$$\left| S_N(f) - \int_0^1 f(t)dt \right| \leq \frac{\max_{t\in[0,1]} |f^{[iv]}(t)|}{2880N^4}. \tag{A.137}$$

This explains why Simpson's rule is used so often in applications. It is very important to note that the error estimate for the trapezoidal rule assumes that f is twice differentiable and for Simpson's rule that f is four times differentiable. If this is not true then these "higher order" integration schemes do not produce such precise results. For real data, higher order schemes are often not used because the noise present in the data has the effect of making the integrand non-differentiable. This means that the higher rates of convergence are not realized. These approximate integration techniques are examples of relatively elementary methods. Each method for approximating functions leads, via integration, to a method for approximating integrals. These go under the general rubric of *quadrature methods*. What distinguishes these methods, at least theoretically is the set of functions for which the approximate formula gives the correct answer. The Riemann sum is correct for constant functions, the trapezoidal rule gives an exact result for linear functions and Simpson's rule gives the exact answer for cubic polynomials. A more complete discussion of this rich and important subject can be found in [26].

N	Simpson	Riemann
16	.0025504	.0064717
64	.0003176	.0008113
128	.0001122	.0002870
144	.0000940	.0002405

Table A.2: Comparison of errors in different approximations to the integral of $\sqrt{x(1-x)}$.

N	Simpson	Riemann
16	167	.104
48	2596	.06
80	9304	.05
112	21571	.04
144	40426	.034

Table A.3: Comparison of rates of convergence in different approximations to the integral of $\sqrt{x(1-x)}$.

Example A.7.1. We consider the results of using Riemann sums and Simpson's rule to approximately compute

$$\int_0^1 \sqrt{x(1-x)}dx = \frac{\pi}{8}.$$

This function is not differentiable at the endpoints of the interval. The results are summarized in the table Since this function vanishes at the endpoints, the trapezoidal rule and the Riemann sum give the same result. Note that while Simpson's rule give a better result for a given number of samples it improves much more slowly than expected. Indeed the ratio of the error and the expected rates of decay (i.e. N^{-4} for Simpson's rule and N^{-1} for Riemann sums) are summarized in the next table.

As this table shows, the order of convergence for Simpson's rule is much slower than it would be if f had the requisite four derivatives.

Example A.7.2. Now we consider these integration schemes applied to a "random" piecewise linear function taking values between -1 and $+1$. The graph of such a function is shown in the figure. In the table we give the errors made using Riemann sums and the trapezoidal rule for various values of N, and these errors rescaled by the expected rate of decrease for sufficiently differentiable data. Here $I = -.06972...$ is the actual value of the integral. Note that in absolute terms, the result from the trapezoidal rule does not improve as we increase N. Relative to the expected error things are consequently degenerating rapidly.

Exercise A.7.2. The log-function is defined by an integral

$$\log(x) = \int_1^x \frac{ds}{s}.$$

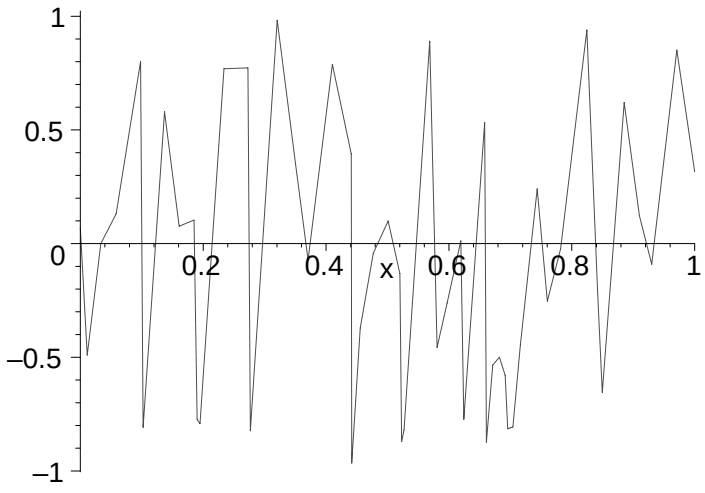


Figure A.4: A random piecewise linear function.

N	$\lvert R_N - I\rvert$	$\lvert R_N - I\rvert N$	$\lvert T_N - I\rvert$	$\lvert T_N - I\rvert N^2$
16	.254	.41	.063	16
40	.017	.69	.068	109
56	.021	1.15	.069	216
80	.007	.64	.069	441

Table A.4: Rates of convergence and errors for approximations to the integral of a random function.

By using numerical integration techniques, approximate values for $\log(x)$ can be obtained. For the Riemann sum, trapezoidal rule and Simpson's rule how large a value of N is needed to compute $\log(2)$ with 10 digits of accuracy?

Exercise A.7.3. Use the functional equation $\log(xy) = \log(x)+\log(y)$ to devise an efficient method for approximately computing the logarithms of numbers between .01 and 100.

A.7.2 Numerical differentiation

If $f(x)$ is a differentiable function, defined on $[0,1]$ sampled at the points $\{\frac{j}{N} \,|\, j = 0, \ldots, N\}$ then we can use the finite difference formula to approximate the derivatives of f at the points in the sample set. We could use the left difference, right difference or centered difference

to obtain approximations to $f'(\frac{j}{N})$ (So long as $j \neq 0, N$):

$$\begin{aligned} D_l f(\frac{j}{N}) &= \frac{f(\frac{j}{N}) - f(\frac{j-1}{N})}{N^{-1}} \qquad \textbf{(left)} \ , \\ D_r f(\frac{j}{N}) &= \frac{f(\frac{j+1}{N}) - f(\frac{j}{N})}{N^{-1}} \qquad \textbf{(right)} \ , \\ D_c f(\frac{j}{N}) &\simeq \frac{f(\frac{j+1}{N}) - f(\frac{j-1}{N})}{2N^{-1}} \qquad \textbf{(centered).} \end{aligned} \tag{A.138}$$

For a twice differentiable function, Taylor's formula gives the error estimates

$$|D_r f(\frac{j}{N}) - f'(\frac{j}{N})| \leq \frac{M_2}{N}, \quad |D_l f(\frac{j}{N}) - f'(\frac{j}{N})| \leq \frac{M_2}{N}, \tag{A.139}$$

where M_2 is proportional to the maximum of $|f^{[2]}(x)|$. If f is three time differentiable then

$$|D_c f(\frac{j}{N}) - f'(\frac{j}{N})| \leq \frac{M_3}{N^2}, \tag{A.140}$$

where M_3 is proportional to the maximum of $|f^{[3]}(x)|$.

There are other ways to assign approximate values to $f'(\frac{j}{N})$ given sampled data $\{f(\frac{j}{N})\}$. For example if $F_N(x)$ is a cubic spline interpolating these values then $F_N(x)$ is a twice differentiable function. A reasonable way to approximate the derivatives of f is to use $\{F'_N(\frac{j}{N})\}$. Since F_N is defined for all $x \in [0,1]$ one can even use $F'_N(x)$ as an approximate value for $f'(x)$ for any $x \in [0,1]$. If f is twice differentiable and $F_N(x)$ is the cubic spline defined using the endpoint data,

$$a_0 = D_r f(0), \quad a_N = D_l f(1)$$

then $< F_N >$ converges to f in $\mathcal{C}^1([0,1])$. That is

$$\lim_{N\to\infty} [\|f - F_N\|_{\mathcal{C}^0} + \|f' - F'_N\|_{\mathcal{C}^0}] = 0.$$

Therefore, for sufficiently large N, $F'_N(x)$ is a good approximation to $f'(x)$ for any $x \in [0,1]$. The rate of convergence depends on the size of the second derivatives of f. Fourier series and integrals provide other ways to approximate the derivatives of a function.

We close this section by considering approximation of derivatives in a more realistic situation. Suppose that we are trying to sample a function $f(t)$. Ordinarily one models the actual data as samples of $f(t) + \epsilon n(t)$ where $n(t)$ is a "random" noise function. Here we scale things so the $|n(t)| \leq 1$ for all t and ϵ is then the amplitude of the noise. The noise is a random function in two different senses. In the first place we cannot say with certainty what the function is, so one usually thinks of this function as being *randomly selected* from some family. It is also random in the sense that functions in these families do not vary smoothly. Thus for n a fixed member of the family, the value of $n(t)$ at a given time t is itself *random*. For instance one could use the family of piecewise constant functions, or the family of all piecewise linear functions as models for the noise. The graphs of such a function is shown in figure (A.5).

N	.25	.5	.75
10	-7.41	16.36	3.35
100	-68.8	82.7	-8.3
1000	296.2	143.4	-280.5

Table A.5: Finite differences for a random piecewise linear function.

If the sampled data is of the form $\{f(\frac{j}{N}) + \epsilon n(\frac{j}{N})\}$ then the right difference approximation to the first derivative is

$$f'(\frac{f}{N}) \simeq D_r f(\frac{j}{N}) + \epsilon N[n(\frac{j+1}{N}) - n(\frac{j}{N})].$$

Due to the random nature of the noise (in the second sense) there is no reason why the difference $n(\frac{j+1}{N}) - n(\frac{j}{N})$ should be small. The contribution of noise to the error, $|f'(\frac{j}{N}) - D_r f(\frac{j}{N})|$ can only be bounded by $2\epsilon N$. In order to get a good approximate value for f', using a finite difference, it is necessary to choose N, so that $N\epsilon$ remains small.

Example A.7.3. The table shows the finite differences evaluated at the points $x \in \{.25, .5, .75\}$ with the indicated values of N for the random, piecewise linear function shown in figure A.5.

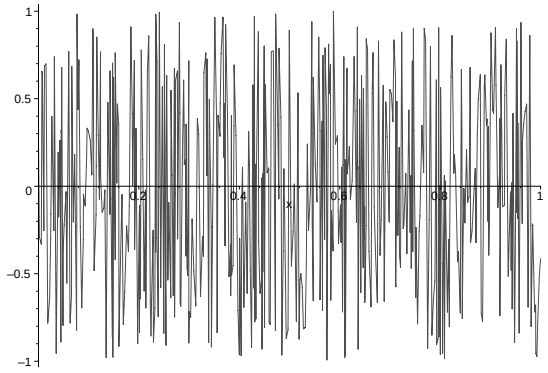


Figure A.5: A fairly random function

In the forgoing example the measurement process is modeled as functional evaluation. Actual measurements are always some sort of an average, to measure the value of a function at a single moment of time would require infinite energy. Owing to the fact that random functions often have small averages, this fact actually works in our favor when trying to approximate derivatives. The simplest model for an average is a uniform average, instead

of evaluating a function at the arguments $\{\frac{j}{N}\}$ we actually measure its average over an interval $[\frac{j}{N}, \frac{j+\delta}{N}]$. Our samples are therefore

$$f_j = \delta^{-1} \int_{\frac{j}{N}}^{\frac{j+\delta}{N}} [f(s) + n(s)]ds.$$

The finite difference then gives

$$\frac{f_{j+1} - f_j}{N^{-1}} = \delta^{-1} \int_{\frac{j}{N}}^{\frac{j+\delta}{N}} \frac{(f(s + N^{-1}) - f(s))ds}{N^{-1}} + \frac{\epsilon N}{\delta} \int_{\frac{j}{N}}^{\frac{j+\delta}{N}} [n(s + N^{-1}) - n(s)]ds. \quad \text{(A.141)}$$

Using the mean value theorem it follows that for each $s \in [\frac{j}{N}, \frac{j+\delta}{N}]$ there is an ξ_s in this interval so that

$$\frac{f(s + N^{-1}) - f(s)}{N^{-1}} = f'(\xi_s).$$

Thus the first term in (A.141) is an average of values of $f'(t)$ over $[\frac{j}{N}, \frac{j+\delta}{N}]$.

Because of the randomness of n there is no reason for the differences $[n(s+N^{-1}) - n(s)]$ to be small, however, for a large enough δ, the individual averages

$$\frac{1}{\delta} \int_{\frac{j}{N}}^{\frac{j+\delta}{N}} n(s)ds, \quad \frac{1}{\delta} \int_{\frac{j+1}{N}}^{\frac{j+1+\delta}{N}} n(s)ds,$$

should themselves be small. This would in turn make the second term in (A.141) small. This illustrates a familiar dichotomy between noise reduction and resolution: by increasing δ we can diminish the effect of the noise, both in the measured values of f and in the finite difference approximations to f'. On the other hand increasing δ also smears out the values of f'. The price for reducing for the noise component of a measurement is decreasing its resolution.

Example A.7.4. Using the same function considered in example A.7.3 we compute the finite differences for averaged data. In order to be able to make comparisons we fix $N^{-1} = .1$ and consider the results obtained with $\delta \in \{.1, .02, .01, .005\}$. The table bears out the prediction that averaging the data diminishes the effect of noise on the computation of finite differences, with longer averaging intervals generally producing a larger effect. However, there is also some failure of this to occur. This is because the experiment is performed on a "random" piecewise linear function, which is in some sense, not especially random.

Exercise A.7.4. Show how to use Taylor's formula to derive (A.139) and (A.140) these error estimates and give formulæ for M_2 and M_3.

Exercise A.7.5. Show that for the function $f(t) = |t|$ and all t the centered differences converge as $N \to \infty$. Is this limit always equal to $f'(t)$?

δ	.25	.5	.75
0	-7.41	16.36	3.35
.1	.269	.593	2.57
.02	5.97	1.64	.281
.01	3.99	.174	5.2
.005	6.74	-1.129	5.50

Table A.6: Finite differences for an averaged, random piecewise linear function.

Appendix B

Basic analysis

This appendix contains some of the facts from analysis that are used in this book. Many good treatments of this material are available, for example [67] or [73].

Definition B.0.1. A subset S of the real numbers is bounded from below if there is some number m so that

$$m \leq x, \forall x \in S,$$

and bounded from below if there is a number M such that

$$M \geq x, \forall x \in S.$$

If a set is bounded from above and below, then we say it is bounded. If a set is bounded from below then we define $\inf S$ as the largest number m such that $m \leq x \quad \forall x \in S$ and if S is bounded from above we define $\sup S$ to be the smallest number such that $x \leq M \quad \forall x \in S$.

B.1 Sequences

The most important idea in introductory analysis is the concept of a *sequence*. A sequence is a function from the positive integers (natural numbers = $\mathbb{N}$.) to some set. The simplest examples are sequences of real numbers. Using standard functional notation we could denote such a function as

$$x : \mathbb{N} \longrightarrow \mathbb{R},$$

then the n^{th} term of the sequence would be denoted $x(n)$. It is customary not to use functional notation but rather to use subscripts, so that the n^{th} term is denoted by x_n. We also consider sequences of functions, for example we could consider a sequence of functions defined on $[0, 1]$, $f_n(x)$ denotes the value of the n^{th} term in the sequence at the point $x \in [0, 1]$. Almost as important as the concept of a sequence is the concept of a *subsequence*. Given a sequence $\{x_n\}$, a subsequence is defined by selecting a subset of $\{x_n\}$ and keeping them in the same order as they appear in $\{x_n\}$. In practice this amounts to defining a function from $\mathbb{N}$ to itself. We denote this function by n_j. It must have the following properties $n_j < n_{j+1}$. The j^{th} term of the subsequence is given by x_{n_j}. As an example, consider the sequence $x_n = (-1)^n n$; the mapping $n_j = 2j$ defines the subsequence $x_{n_j} = (-1)^{2j} 2j$.

Definition B.1.1. A sequence of real numbers, $\{x_n\}$ has a *limit* if there is a number L such that given any $\epsilon > 0$ there exists a $N > 0$ such that

$$|x_n - L| < \epsilon \text{ whenever } n > N.$$

A sequence with a limit is called a *convergent sequence*, we then write

$$\lim_{n\to\infty} x_n = L.$$

The limit, when it exists is unique. A sequence may fail to have limit but it may have a subsequence which does. In this case the sequence is said to have a *convergent subsequence*. For example $x_n = (-1)^n \frac{1}{n}$ is not convergent but the subsequence defined by $n_j = 2j$ is.

B.2 Rules for Limits

There are rules for computing limits of algebraic combinations of convergent sequences.

Theorem B.2.1 (Algebraic Rules for Limits). *Suppose that $\{x_n\}, \{y_n\}$ are convergent sequences of real numbers then*

$$\begin{aligned}
\lim_{n\to\infty} ax_n &\text{ exists and equals } a \lim_{n\to\infty} x_n, \text{ for all } a \in \mathbb{R}, \\
\lim_{n\to\infty} (x_n + y_n) &\text{ exists and equals } \lim_{n\to\infty} x_n + \lim_{n\to\infty} y_n, \\
\lim_{n\to\infty} (x_n y_n) &\text{ exists and equals } (\lim_{n\to\infty} x_n)(\lim_{n\to\infty} y_n), \\
\lim_{n\to\infty} \frac{x_n}{y_n} &\text{ exists, provided } \lim_{n\to\infty} y_n \neq 0, \text{ and equals } \frac{\lim_{n\to\infty} x_n}{\lim_{n\to\infty} y_n}.
\end{aligned} \tag{B.1}$$

In this theorem the non–trivial claim is that the limits exist, once this is clear it is easy to show what they must be.

B.3 Existence of limits

A problem of fundamental importance is to decide whether or not a sequence has a limit. A sequence $\{x_n\}$ is bounded if there is a number M such that

$$|x_n| < M \text{ for all } n.$$

A sequence is non–increasing if

$$x_n \geq x_{n+1} \text{ for all } n,$$

and non–decreasing if

$$x_n \leq x_{n+1} \text{ for all } n.$$

The *completeness axiom* of the real numbers states *A bounded non–increasing or non–decreasing sequence of real numbers has a limit.* If a bounded sequence is neither non–decreasing, nor non–increasing then the only general theorem about convergence is

Theorem B.3.1 (Bolzano–Weierstrass Theorem). *A bounded sequence of real numbers has a convergent subsequence.*

Note that this does not assert that any bounded sequence converges but only that any bounded sequence has a convergent subsequence.

Definition B.3.1. In general, if $S \subset \mathbb{R}$ then the set of points which can be obtained as limits of sequences $\{x_n\} \subset S$ is called the set of accumulation points of S.

A subset S is *dense* in an interval I, if I is a subset of the set of accumulation points of S. For example the rational numbers $\mathbb{Q}$ are dense in every interval. The following two lemmas are very useful

Lemma B.3.1. *If x_n, y_n, z_n are sequences of real numbers such that*

$$x_n \leq y_n \leq z_n$$

and x_n and z_n are convergent with

$$L = \lim_{n\to\infty} x_n = \lim_{n\to\infty} z_n$$

then y_n converges with

$$\lim_{n\to\infty} y_n = L.$$

Lemma B.3.2. *If $x_n \geq 0$ is convergent then*

$$\lim_{n\to\infty} x_n \geq 0.$$

In the above discussion of limits we always assumed that the limit is known in advance. There is a criterion, due to Cauchy which implies that a given sequence has a limit but makes no reference to the limit itself.

Theorem B.3.2 (Cauchy Criterion for Sequences). *If $\{x_n\}$ is a sequence of real numbers such that given $\epsilon > 0$ there exists an N for which*

$$|x_n - x_m| < \epsilon \text{ whenever both } n \text{ and } m \text{ are greater than } N,$$

then the sequence is convergent.

A sequence satisfying this condition is called a *Cauchy sequence.*

B.4 Series

A series is the sum of a sequence, it is denoted by

$$\sum_{n=1}^{\infty} x_n.$$

Definition B.4.1. A *series* converges if the *sequence* of partial sums

$$s_k = \sum_{n=1}^{k} x_n,$$

converges. In this case

$$\sum_{n=1}^{\infty} x_n \stackrel{d}{=} \lim_{k\to\infty} s_k.$$

If a series does not converge then it diverges.

Definition B.4.2. A series converges *absolutely* if the sum of the absolute values

$$\sum_{n=1}^{\infty} |x_n|$$

converges.

The following theorem describes the elementary properties of series.

Theorem B.4.1 (Theorem on Series). *Suppose that x_n, y_n are sequences. Suppose that $\sum_{n=1}^{\infty} x_n$, $\sum_{n=1}^{\infty} y_n$ converge then*

$$\begin{aligned}
&\sum_{n=1}^{\infty}(x_n + y_n) \text{ converges and } \sum_{n=1}^{\infty}(x_n + y_n) = \sum_{n=1}^{\infty} x_n + \sum_{n=1}^{\infty} y_n, \\
&\text{If } a \in \mathbb{R} \sum_{n=1}^{\infty} a x_n = a \sum_{n=1}^{\infty} x_n, \\
&\text{If } x_n \geq 0 \text{ for all } n, \text{ then } \sum_{n=1}^{\infty} x_n \geq 0.
\end{aligned} \tag{B.2}$$

There are many criterion that are used to determine if a given series converges. The most important is the comparison test

Theorem B.4.2 (Comparison Test). *Suppose that x_n, y_n are sequences such that $|x_n| \leq y_n$ if $\sum_{n=1}^{\infty} y_n$ converges then so does $\sum_{n=1}^{\infty} x_n$. If $0 \leq y_n \leq x_n$ and $\sum_{n=1}^{\infty} y_n$ diverges then so does $\sum_{n=1}^{\infty} x_n$.*

To apply this test we need to have some series which we know converge or diverge. The simplest case is a geometric series. This is because there is a formula for the partial sums:

$$\sum_{n=0}^{k} a^k = \frac{a^{k+1} - 1}{a - 1}.$$

From this formula we immediately conclude

Theorem B.4.3 (Convergence of Geometric Series). *A geometric converges if and only if* $|a| < 1$.

The root and ratio tests are really special cases of the comparison test where the series are comparable to geometric series.

Theorem B.4.4 (Ratio Test). *If* x_n *is a sequence with*

$$\limsup_{n\to\infty} \left| \frac{x_{n+1}}{x_n} \right| = \alpha$$

then the series

$$\sum_{n=1}^{\infty} x_n \begin{cases} \text{converges if} & \alpha < 1 \\ \text{diverges if} & \alpha > 1. \end{cases}$$

The test gives no information if $\alpha = 1$.

We also have

Theorem B.4.5 (Root Test). *If* x_n *is a sequence with*

$$\limsup_{n\to\infty} |x_n|^{\frac{1}{n}} = \alpha$$

then the series

$$\sum_{n=1}^{\infty} x_n \begin{cases} \text{converges if} & \alpha < 1 \\ \text{diverges if} & \alpha > 1. \end{cases}$$

The test gives no information if $\alpha = 1$.

If $\alpha < 1$ in the ratio or root tests then the series converge absolutely.

Another test is obtained by comparing a series to an integral

Theorem B.4.6 (Integral Test). *If* $f(x)$ *is an integrable function defined for* $x > 0$ *which satisfies*

$$f(x) \geq 0, \quad f(x) \leq f(y) \text{ if } y < x,$$

then

$$\sum_{n=1}^{\infty} f(n) \text{ converges if and only if } \lim_{n\to\infty} \int_1^n f(x)\,dx \text{ exists.}$$

Using this test we can easily show that the sum,

$$\sum_{n=1}^{\infty} \frac{1}{n^p}$$

converges if and only if $p > 1$. A final test is sometimes useful for showing that a series with terms that alternate in sign converges.

Theorem B.4.7 (Alternating Series Test). *Suppose that x_n is a sequence such that the sign alternates, the $\lim_{n\to\infty} x_n = 0$ and $|x_{n+1}| \leq |x_n|$ then*

$$\sum_{n=1}^{\infty} x_n$$

converges.

Note that this test requires that the signs alternate, the absolute value of the sequence is monotonely decreasing, and the sequence tends to zero. If any of these conditions are not met the series may fail to converge.

B.5 Limits of Functions and Continuity

The next thing to consider is the behavior of functions defined on intervals in $\mathbb{R}$. Suppose that $f(x)$ is defined for $x \in (a, c) \cup (c, b)$. This is called a punctured neighborhood of c.

Definition B.5.1. We say that $f(x)$ has a limit, L as x approaches c if given $\epsilon > 0$ there exists $\delta > 0$ such that

$$|f(x) - L| < \epsilon \text{ provided } 0 < |x - c| < \delta$$

and we write

$$\lim_{x\to c} f(x) = L.$$

Note that in this definition nothing is said about the value of $f(x)$ at $x = c$. This has no bearing at all on whether the limit exists.

Definition B.5.2. If $f(c)$ is defined and we have that

$$\lim_{x\to c} f(x) = f(c)$$

then we say that $f(x)$ is continuous at $x = c$. If $f(x)$ is continuous for all $x \in (a, b)$ then we say that $f(x)$ is continuous on (a, b).

In addition to the ordinary limit, we also define one sided limits. If $f(x)$ is defined in (a, b) and there exists an L such that given $\epsilon > 0$ there exists δ such that

$$|f(x) - L| < \epsilon \text{ provided } 0 < x - a < \delta \text{ then } \lim_{x\to a^+} f(x) = L.$$

If instead

$$|f(x) - L| < \epsilon \text{ provided } 0 < b - x < \delta \text{ then } \lim_{x\to b^-} f(x) = L.$$

The rules for dealing with limits of functions are very similar to the rules for handling limits of sequences

Theorem B.5.1 (Algebraic Rules for Limits of Functions). *Suppose that $f(x), g(x)$ are defined in a punctured neighborhood of c and that*

$$\lim_{x\to c} f(x) = L, \quad \lim_{x\to c} g(x) = M.$$

Then

$$\begin{aligned} \lim_{x\to c}(af(x)) &\text{ exists and equals } aL \text{ for all } a \in \mathbb{R}, \\ \lim_{x\to c}(f(x)+g(x)) &\text{ exists and equals } L+M, \\ \lim_{x\to c}(f(x)g(x)) &\text{ exists and equals } LM, \\ \lim_{x\to c}\frac{f(x)}{g(x)} &\text{ exists, provided } M\neq 0 \text{ and equals } \frac{L}{M}. \end{aligned} \tag{B.3}$$

From this we deduce the following results about continuous functions

Theorem B.5.2 (Algebraic Rules for Continuous Functions). *If $f(x), g(x)$ are continuous at $x = c$ then so are $af(x)$, $f(x) + g(x)$, $f(x)g(x)$. If $g(c) \neq 0$ then $f(x)/g(x)$ is also continuous at $x = c$.*

For functions there is one further operation which is very important, composition.

Theorem B.5.3 (Continuity of Compositions). *Suppose that $f(x), g(y)$ are two functions such that $f(x)$ is continuous at $x = c$ and $g(y)$ is continuous at $y = f(c)$ then the composite function, $g \circ f(x)$ is continuous at $x = c$.*

Definition B.5.3. A function defined on an interval $[a, b]$ is said to be uniformly continuous if given $\epsilon > 0$ there exists δ such that

$$|f(x) - f(y)| < \epsilon, \forall x, y \in [a, b] \text{ with } |x - y| < \delta.$$

The basic proposition is

Proposition B.5.1. *A continuous function on a closed, bounded interval is uniformly continuous.*

Using similar arguments we can also prove

Proposition B.5.2 (Max–Min theorem for Continuous Functions). *If $f(x)$ is continuous on a closed bounded interval, $[a, b]$ then there exists $x_1 \in [a, b]$ and $x_2 \in [a, b]$ which satisfy*

$$f(x_1) = \sup_{x\in[a,b]} f(x), \quad f(x_2) = \inf_{x\in[a,b]} f(x).$$

As a final result on continuous functions we have Intermediate Value Theorem

Theorem B.5.4 (Intermediate Value Theorem). *Suppose that $f(x)$ is continuous on $[a, b]$ and $f(a) < f(b)$ then given $y \in (f(a), f(b))$ there exists $c \in (a, b)$ such that $f(c) = y$.*

B.6 Differentiability

A function defined in a neighborhood of a point c is said to be *differentiable* at c if the function

$$g(x) = \frac{f(x) - f(c)}{x - c},$$

defined in a deleted neighborhood of c has a limit as $x \to c$. This limit is called the derivative of f at c; we denote it by $f'(c)$. A function which is differentiable at every point of an interval is said to be differentiable in the interval. If the derivative is itself continuous then the function is said to be *continuously differentiable*. As with continuous functions we have algebraic rules for differentiation.

Proposition B.6.1 (Rules for Differentiation). *Suppose that $f(x), g(x)$ are differentiable at $x = c$ then so are $af(x), (f(x) + g(x)), f(x)g(x)$. If $g(c) \neq 0$ then so is $f(x)/g(x)$. The derivatives are given by*

$$\begin{aligned} (af)'(c) &= a(f'(c)), \\ (f+g)'(c) &f'(c) + g'(c), \\ (fg)'(c) &= f'(c)g(c) + f(c)g'(c), \\ \left(\frac{f}{g}\right)'(c) &= \frac{f'(c)g(c) - f(c)g'(c)}{g(c)^2}. \end{aligned} \tag{B.4}$$

In addition we can also differentiate a composition

Proposition B.6.2 (The Chain Rule). *If $f(x)$ is differentiable at $x = c$ and $g(y)$ is differentiable at $y = f(c)$ then $g \circ f(x)$ is differentiable at $x = c$; the derivative is*

$$g \circ f'(c) = g'(f(c))f'(c).$$

It is often useful to be able to compare the size of two functions $f(x), g(x)$ near a point $x = c$ without being too specific.

Definition B.6.1. When we write

$$f(x) = o(g(x)) \text{ near to } x = c$$

it means that

$$\lim_{x \to c} \frac{|f(x)|}{g(x)} = 0.$$

If we write that

$$f(x) = O(g(x)) \text{ near to } x = c$$

this means that we can find an M and an $\epsilon > 0$ such that

$$|f(x)| < Mg(x) \text{ provided } |x - c| < \epsilon.$$

For example a function $f(x)$ is differentiable at $x = c$ if and only if there exists a number L for which

$$f(x) = f(c) + L(x - c) + o(|x - c|).$$

Of course $L = f'(c)$. From this observation it follows that a function which is differentiable at a point is also continuous at that point. The converse statement is false: a function may be continuous at a point without being differentiable, for example $f(x) = |x|$ is continuous at $x = 0$ but not differentiable.

B.7 Higher Order Derivatives and Taylor's Theorem

If the first derivative of function, $f'(x)$ happens itself to be differentiable then we say that $f(x)$ is twice differentiable. The second derivative is denoted by $f''(x)$. Inductively if the k^{th} derivative happens to be differentiable then we say that f is $(k+1)$–times differentiable. We denote the k^{th} derivative by $f^{[k]}(x)$. For a function which has n derivatives we can find a polynomial which agrees with $f(x)$ to order $n-1$ at a point.

Theorem B.7.1 (Taylor's Theorem). *Suppose that $f(x)$ has n derivatives at a point c then*

$$f(x) - \sum_{j=0}^{n-1} \frac{f^{[j]}(c)(x-c)^j}{j!} + R_n(x)$$

where

$$R_n(x) = O(|x-c|^n).$$

There are many different formulæ for the error term $R_n(x)$. One from which all the others can be derived is given by

$$R_n(x) = \int_c^x f^{[n]}(t)(x-t)^{n-1}dt.$$

An important special case of Taylor's Theorem is the mean value theorem

Theorem B.7.2 (Mean Value Theorem). *Suppose that $f(x)$ is continuous on $[a,b]$ and differentiable on (a,b) then there exists a $c \in (a,b)$ such that*

$$f(b) - f(a) = f'(c)(b-a).$$

B.8 Integration

The inverse operation to differentiation is integration. Suppose that $f(x)$ is a bounded, function defined on a finite interval $[a,b]$. An increasing sequence $P = \{a = x_0 < x_1 < \cdots < x_N = b\}$ defines a *partition* of the interval. The mesh size of the partition is defined to be

$$|P| = \max\{|x_i - x_{i-1}| \ : \ i = 1, \ldots, N\}.$$

To each partition we associate two approximations of the "area under the graph of f," by the rules

$$\begin{aligned} U(f,P) &= \sum_{j=1}^{N} \sup_{x \in [x_{j-1}, x_j]} f(x)(x_j - x_{j-1}), \\ L(f,P) &= \sum_{j=1}^{N} \inf_{x \in [x_{j-1}, x_j]} f(x)(x_j - x_{j-1}). \end{aligned} \tag{B.5}$$

These are called the upper and lower Riemann sums, observe that

$$U(f,P) \geq L(f,P). \tag{B.6}$$

If P and P' are partitions with the property that every point in P is also a point in P' then we say that P' is a *refinement* of P and write $P < P'$. If P_1 and P_2 are two partitions then, by using the union of the points in the two underlying sets, we can define a new partition P_3 with the property that

$$P_1 < P_3 \text{ and } P_2 < P_3.$$

A partition with this property is called a *common refinement* of P_1 and P_2. From the definitions it is clear that if $P < P'$ then

$$U(f,P) \geq U(f,P') \text{ and } L(f,P) \leq L(f,P'). \tag{B.7}$$

We define the upper Riemann integral of f to be

$$\overline{\int_a^b} f(x)dx = \inf_P U(f,P),$$

and the lower Riemann integral to be

$$\underline{\int_a^b} f(x)dx = \sup_P L(f,P).$$

In light of (B.6) it is not hard to show that

$$\overline{\int_a^b} f(x)dx \geq \underline{\int_a^b} f(x)dx.$$

Definition B.8.1. A bounded function f defined on an interval $[a,b]$ is *Riemann integrable* if

$$\overline{\int_a^b} f(x)dx \underline{\int_a^b} f(x)dx.$$

In this case we denote the common value by

$$\int_a^b f(x)dx.$$

Most "nice" functions are Riemann integrable. For example we have the following basic result

Theorem B.8.1. *Suppose that f is a piecewise continuous function defined on $[a,b]$ then f is Riemann integrable and*

$$\int_a^b f(x)dx = \lim_{N\to\infty} \sum_{j=1}^{N} f(a + \frac{j}{N}(b-a))\frac{b-a}{N}.$$

The proof of this theorem is not difficult, relying primarily on the uniform continuity of a continuous function on a closed, bounded interval and (B.7). The formula for the integral is true for any Riemann integrable function but is more difficult to prove in this generality. The formula for the integral as a limit of Riemann sums can also be greatly generalized, allowing any sequence of partitions $\{P_j\}$ for which $\lim_{j\to\infty} |P_j| = 0$.

The integral has several important properties

Theorem B.8.2. *Suppose that f and g are Riemann integrable functions then $f+g$ and fg are integrable as well. If $c \in \mathbb{R}$ then*

$$\int_a^b (f(x)+g(x))dx = \int_a^b f(x)dx + \int_a^b g(x)dx \text{ and } \int_a^b cf(x)dx = c\int_a^b f(x)dx.$$

In other words the integral is a linear map from integrable functions to the real numbers.

Theorem B.8.3. *Suppose that f is Riemann integrable on $[a,b]$ and that $c \in [a,b]$ then f is Riemann integrable on $[a,c]$ and $[c,b]$, moreover*

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx. \tag{B.8}$$

There is also a mean value theorem for the integral, similar to Theorem B.7.2.

Theorem B.8.4. *Suppose that f is a continuous function and w is a non-negative integrable function. There exists a point $c \in (a,b)$ so that*

$$\int_a^b f(x)w(x)dx = f(c)\int_a^b w(x)dx.$$

The basic method for calculating integrals comes from the fundamental theorem of calculus. To state this result we need to think of the integral in a different way. As described above the integral associates a number to a function defined on a fixed interval. Suppose instead that f is defined and Riemann integrable on $[a,b]$. Theorem B.8.3 states that for each $x \in [a,b]$ f is also Riemann integrable on $[a,x]$. The new idea to use the integral to define a new function on $[a,b]$:

$$F(x) = \int_a^x f(y)dy.$$

This function is called the *indefinite integral* or anti-derivative of f. In this context one often refers to $\int_a^b f(x)dx$ as the *definite integral* of f.

Theorem B.8.5 (The Fundamental Theorem of Calculus). *If f is a continuous function on $[a,b]$ then $F'(x) = f(x)$. If $f \in \mathcal{C}^1([a,b])$ then*

$$\int_a^b f'(x)dx = f(b) - f(a).$$

There are two further basics tools needed to compute and manipulate integrals. The first is called *integration by parts*, it is a consequence of the product rule for derivatives, see Proposition B.6.1.

Proposition B.8.1 (Integration by parts). *If $f, g \in \mathcal{C}^1([a,b])$ then*

$$\int_a^b f'(x)g(x)dx = f(b)g(b) - f(a)g(a) - \int_a^b f(x)g'(x)dx.$$

The other formula follows from the chain rule, Proposition B.6.2.

Proposition B.8.2 (Change of variable). *Let g be a monotone increasing, differentiable function defined $[a,b]$ with $g(a) = c$, $g(b) = d$ and let f be a Riemann integrable function on $[c,d]$. The following formula holds*

$$\int_c^d f(y)dy = \int_a^b f(g(x))g'(x)dx.$$

There is analogous treatment for the integration of functions of several variables. A result which is especially important in applications is Fubini's theorem. Only a very special case is usually required. The statement of the special case requires only the definition of the one-dimensional Riemann integral. Suppose that $f(x,y)$ is a continuous function on the rectangle $[a,b] \times [c,d]$ then for each fixed valued of y, $f(\cdot, y)$ is an integrable function on $[a,b]$ and similarly for each fixed x, $f(x,\cdot)$ is an integrable function on $[c,d]$. Performing these integrals leads to two new functions

$$g(y) = \int_a^b f(x,y)dx, \quad h(x) = \int_c^d f(x,y)dy$$

which are themselves integrable on the appropriate intervals.

Theorem B.8.6 (Fubini's theorem). *Let $f(x,y)$ be a continuous function on $[a,b]\times[c,d]$ then*

$$\int_c^d \left(\int_a^b f(x,y)dx \right) dy = \int_a^b \left(\int_c^d f(x,y)dy \right) dx.$$

More colloquially we say that we can *change the order of the integrations*.

B.9 Improper integrals

In the previous section we defined the Riemann integral for *bounded* functions on *bounded* intervals. In applications both of these restrictions need to be removed. This leads to various notions of *improper integrals*. The simplest situation is that of a function $f(x)$

defined on $[0,\infty)$ and integrable on $[0,R]$ for every $R > 0$. We say that the improper integral,

$$\int_0^\infty f(x)dx$$

exists if the limit,

$$\lim_{R\to\infty}\int_0^R f(x)dx \tag{B.9}$$

exists. In this case the improper integral is given by the limiting value. By analogy with the theory of infinite series there are two distinct situations in which the improper integral exists. If the improper integral of $|f|$ exists then we say that f is *absolutely integrable* on $[0,\infty)$.

Example B.9.1. The function $(1+x^2)^{-1}$ is absolutely integrable on $[0,\infty)$. Indeed we see that if $R < R'$ then

$$\begin{aligned} 0 \le \int_0^{R'} \frac{dx}{1+x^2} - \int_0^R \frac{dx}{1+x^2} &= \int_R^{R'} \frac{dx}{1+x^2} \\ &\le \int_R^{R'} \frac{dx}{x^2} \\ &\le \frac{1}{R}. \end{aligned} \tag{B.10}$$

This shows that

$$\lim_{R\to\infty}\int_0^R \frac{dx}{1+x^2}$$

exists.

Example B.9.2. The function $\sin(x)/x$ is integrable on $[0,\infty)$ but not absolutely integrable. The part of the integral from say 0 to 1 is no problem. Using integration by parts we find that

$$\int_1^R \frac{\sin(x)dx}{x} = \frac{\cos(x)}{x}\Big|_1^R - \int_1^R \frac{\cos(x)dx}{x^2}.$$

Using this formula and the previous example it is not difficult to show that

$$\lim_{R\to\infty}\int_0^R \frac{\sin(x)dx}{x}$$

exists. On the other hand because

$$\int_1^R \frac{dx}{x} = \log R$$

it is not difficult to show that

$$\int_0^R \frac{|\sin(x)|dx}{x}$$

grows like $\log R$ and therefore diverges as R tend to infinity.

There are similar definitions for the improper integrals

$$\int_{-\infty}^{0} f(x)dx \text{ and } \int_{-\infty}^{\infty} f(x)dx.$$

The only small subtlety is that we say that the improper integral exists in the second case only when both the improper integrals

$$\int_{-\infty}^{0} f(x)dx, \quad \int_{0}^{\infty} f(x)dx$$

exist separately. Similar definitions apply to functions defined on bounded intervals (a, b) which are integrable on any subinterval $[c, d]$. We say that the improper integral

$$\int_{a}^{b} f(x)dx$$

exists if the limits

$$\lim_{c\to a^+} \int_{c}^{e} f(x)dx \text{ and } \lim_{c\to b^-} \int_{e}^{c} f(x)dx$$

both exist. Here e is any point in (a, b); the existence or non-existence of these limits is clearly independent of which (fixed) point we use. Because the improper integrals are defined by limits of proper integrals they have the same linearity properties as integrals. For example:

Proposition B.9.1. *Suppose that f and g are improperly integrable on $[0, \infty)$ then $f + g$ is as well and*

$$\int_{0}^{\infty} (f(x) + g(x))dx = \int_{0}^{\infty} f(x)dx + \int_{0}^{\infty} g(x)dx,$$

for $a \in \mathbb{R}$, af is improperly integrable and

$$\int_{0}^{\infty} af(x)dx = a \int_{0}^{\infty} f(x)dx.$$

The final case that requires consideration is that of a function f defined on a deleted interval $[a, b) \cup (b, c]$ and integrable on subintervals of the form $[a, e]$ and $[f, c]$ where $a \leq e < b$ and $b < f \leq c$. If both limits

$$\lim_{e\to b^-} \int_{a}^{e} f(x)dx \text{ and } \lim_{f\to b^+} \int_{f}^{c} f(x)dx$$

exist then we say that f is improperly integrable on $[a, b]$. For example the function $f(x) = x^{-\frac{1}{3}}$ is improperly integrable on $[-1, 1]$. On the other hand the function $f(x) = x^{-1}$ is not improperly integrable because

$$\lim_{e\to 0^+} \int_{e}^{1} \frac{dx}{x} = \infty \text{ and } \lim_{f\to 0^-} \int_{-1}^{f} \frac{dx}{x} = -\infty.$$

There is however a further extension of the notion of integrability that allows us to assign a meaning to

$$\int_{-1}^{1} \frac{dx}{x}.$$

This is called a *principal value integral* The observation is that for any $\epsilon > 0$

$$\int_{-1}^{-\epsilon} \frac{dx}{x} + \int_{\epsilon}^{1} \frac{dx}{x} = 0,$$

so the limit of this sum exists as ϵ goes to zero.

Definition B.9.1. Suppose that f is defined on the deleted interval $[a, b) \cup (b, c]$ and is integrable on any subinterval $[a, e]$, $a \leq e < b$ or $[f, c]$, $b < f \leq c$. If the limit

$$\lim_{\epsilon \to 0} \int_{a}^{b-\epsilon} f(x)dx + \int_{b+\epsilon}^{c} f(x)dx$$

exists then we say that f has a *principal value integral* on $[a, c]$. We denote the limit by

$$\text{P. V.} \int_{a}^{c} f(x)dx.$$

For a function which is **not** (improperly) integrable, the principal value integral exists because of a subtle cancelation between the divergences of the two parts of the integral. It is crucial to observe that the approach to the singular point in symmetric. Both the existence of the limit and its value depend crucially on this fact.

Example B.9.3. We observed that the function x^{-1} has a principal value integral on $[-1, 1]$ and its value is zero. To see the importance of symmetry in the definition of the principal value integral observe that

$$\int_{-1}^{-\epsilon} \frac{dx}{x} + \int_{2\epsilon}^{1} \frac{dx}{x} = -\log 2$$

and

$$\int_{-1}^{-\epsilon} \frac{dx}{x} + \int_{\epsilon^2}^{1} \frac{dx}{x} = -\log \epsilon.$$

In the first case we get a different limit and in the second case the limit diverges.

The following proposition indicates why principal value integrals are important in applications.

Proposition B.9.2. *Let $f \in \mathcal{C}^1([-1, 1])$ then the function $f(x)/x$ has a principal value integral on $[-1, 1]$.*

Proof. The proof of this result is very instructive. Because x^{-1} is an odd function we have, for any positive ϵ the following identity

$$\int_{-1}^{-\epsilon} \frac{f(x)dx}{x} + \int_{\epsilon}^{1} \frac{f(x)dx}{x} = \int_{-1}^{-\epsilon} \frac{(f(x) - f(0))dx}{x} + \int_{\epsilon}^{1} \frac{(f(x) - f(0))dx}{x}.$$

The function f is continuously differentiable and therefore the function

$$g(x) = \begin{cases} \frac{f(x)-f(0)}{x} & \text{if } x \neq 0, \\ f'(0) & \text{if } x = 0, \end{cases}$$

is continuous on $[-1, 1]$. As a continuous function is integrable this completes the proof. □

Indeed from the proof we get the formula

$$\text{P. V.} \int_{-1}^{1} \frac{f(x)dx}{x} = \int_{-1}^{1} \frac{(f(x) - f(0))dx}{x}.$$

The material in this chapter is usually covered in an undergraduate course in mathematical analysis. The proofs of these results and additional material can be found in [67] and [73].

Bibliography

[1] Lars V. Ahlfors, *Complex analysis*, McGraw-Hill, New York, 1979.

[2] R.E. Alvarez and A. Macovski, *Energy selective reconstructions in x-ray computerized tomography*, Phys. Med. Biol. **21** (1976), 733–744.

[3] R.J. Barlow, *Statistics, a guide to the use of statistical methods in the physical sciences*, The Manchester Physics Series, John Wiley& Sons, 1989.

[4] Harrison H. Barrett and William Swindell, *Radiological imaging*, Academic Press, 1981.

[5] R. Beals, *Advanced mathematical analysis*, Graduate Texts in Mathematics, vol. 119, Springer Verlag, 1988.

[6] George B. Benedek and Felix M.H. Villars, *Physics with Illustrative Examples from Medicine and Biology, Electricity and Magnetism, 2nd ed.*, AIP Press and Springer Verlag, New York, 2000.

[7] William E. Boyce and Richard C. DiPrima, *Elementary differential equations, 6th ed.*, Wiley, New York, 1997.

[8] Robert Grover Brown, *Introduction to random signal analysis and Kalman filtering*, John Wiley & Sons, New York, 1983.

[9] Yair Censor and Gabor T. Herman, *On some optimization techniques in image reconstruction from projections*, Appl. Numer. Math. **3** (1987), 365–391.

[10] A.M.J. Cormack, *Representation of a function by its line integrals, with some radiological applications i., ii.*, J. Applied Physics **34,35** (1963,1964), 2722–2727, 195–207.

[11] John D'Angelo and Douglas West, *Mathematical thinking, problem-solving and proofs, 2nd ed.*, Prentice Hall, Upper Saddle River, NJ, 2000.

[12] J.L. Doob, *Stochastic processes*, John Wiley & Sons, Inc., New York, 1953.

[13] Edward R. Dougherty, *Random processes for image and signal processing*, SPIE/IEEE series on imaging science and engineering, IEEE press, Piscataway, NJ, 1999.

[14] Charles L. Epstein and Bruce Kleiner, *Spherical means in annular regions*, CPAM **44** (1993), 441–451.

[15] W. Feller, *Introduction to probability theory and its applications, I and II*, John Wiley & Sons, New York, 1968, 1971.

[16] G.B. Folland, *Real analysis, modern techniques and their applications*, John Wiley and sons, New York, NY, 1984.

[17] ———, *Introduction to partial differential equation, 2nd ed.*, Princeton Univ. Press, Princeton, NJ, 1995.

[18] Arnold Lent Gabor T. Herman and Stuart Rowland, *Art: Mathematics and applications*, J. Theo. Bio. **42** (1973), 1–32.

[19] Gene H. Golub and Charles F. Van Loan, *Matrix computations, 3rd ed.*, The Johns Hopkins University Press, Baltimore, 1996.

[20] I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series and Products*, Academic Press, New York, 1980.

[21] A.V. Lakshminarayanan G.T. Herman and A. Naparstek, *Reconstruction using divergent ray shadowgraphs*, In Ter-Pergossian [77], pp. 105–117.

[22] G.H. Hardy and J.E. Littlewood, *Some properties of fractional integrals*, Math. Zeit. **27** (1928), 565–606.

[23] S. Helgason, *The Radon transform, 2nd ed.*, Birkhäuser, Boston, 1999.

[24] Gabor T. Herman, *Image reconstruction from projections*, Academic Press, New York, 1980.

[25] Gabor T. Herman and Dewey Odhner, *Performance evaluation of an iterative image reconstruction algorithm for positron emmision tomography*, IEEE Trans. on Med. Im. **10** (1991), 336–346.

[26] F.B. Hildebrand, *Introduction to numerical analysis*, McGraw-Hill, New York, 1956.

[27] M. Holschneider, *Wavelets, an analysis tool*, Clarendon Press, Oxford, 1995.

[28] L. Hörmander, *The analysis of linear partial differential operators*, vol. 1, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1983.

[29] ———, *The analysis of linear partial differential operators*, vol. 3, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1985.

[30] G.N. Hounsfield, *Computerised transverse axial scanning (tomography i. description of system*, Br. J. Radiology **46** (1973), 1016–1022.

[31] International Commission on Radiation Units and Measurement, Bethesda, MA, *Tissue substitute in radiation dosimetry and measurement Report 44*, 1898, Available at http://physics.nist.gov/PhysRefData/XrayMassCoef/cover.html.

[32] Bernd Jähne, *Digital image processing, concepts, algorithms and scientific applications, third ed.*, Springer Verlag, Berlin, Heidelberg, 1995.

[33] Peter M. Joseph, *Image noise and smoothing in computed tomography (CT) scanners*, SPIE – Optical Instrumentation in Medicine VI **127** (1977), 43–49.

[34] ———, *The influence of gantry geometry on aliasing and other geometry dependent errors*, IEEE Transactions on Nuclear Science **NS-27** (1980), 1104–1111.

[35] ———, *Artifacts in computed tomography*, Technical Aspects of Computed Tomgraphy, vol. 5 (St. Louis) (M. D . Thomas H. Newton and M. D . D. Gordon Potts, eds.), The C.V. Mosby Company, 1981, pp. 3956–3992.

[36] Peter M. Joseph and Raymond A. Schulz, *View sampling requirements in fan beam computed tomography*, Med. Phys. **7** (1980), 692–702.

[37] Peter M. Joseph and Robin D. Spital, *A method for corrrecting bone induced artifacts in computer tomography scanners*, Journal of Computer Assisted Tomography **2** (1978), 100–108.

[38] Peter M. Joseph and Charles D. Stockham, *The influence of modulation transfer function shape on computer tomographic image quality*, Radiology **145** (1982), 179–185.

[39] Avinash Kak and Malcolm Slaney, *Principles of Computerized Tomographic Imaging*, IEEE press, 1988.

[40] Yitzhak Katznelson, *An introduction to harmonic analysis*, Dover, New York, NY, 1976.

[41] Joseph B. Keller, *Inverse problems*, American Math. Monthly **83** (1976), 107–118.

[42] Reinhard Klette and Piero Zamperoni, *Handbook of image processing operators*, John Wiley and Sons, Chichester, 1996.

[43] P.D. Lax, *Linear algebra*, Wiley, New York, 1997.

[44] Peter D. Lax and Ralph S. Phillips, *The Paley-Wiener theorem for the Radon transform*, CPAM **23** (1970), 409–424.

[45] Peter Linz, *Theoretical numerical analysis*, John Wiley, New York, NY, 1979.

[46] B.F. Logan, *The uncertainty principle in reconstructing functions from projections*, Duke Math. Journal **42** (1975), 661–706.

[47] B.F. Logan and L.A. Shepp, *Optimal reconstruction of a function from its projections*, Duke Math. Journal **42** (1975), 645–659.

[48] Donald Ludwig, *The Radon transform on Euclidean space*, CPAM **19** (1966), 49–81.

[49] Wilhelm Magnus and tr. from the German by John Wermer Fritz Oberhettinger, *Formulas and theorems for the special functions of mathematical physics*, Chelsea, New York, 1949.

[50] Frank Natterer, *Mathematics of medical imaging, 2nd ed.*, SIAM, Philadelphia, 2001.

[51] Frank Natterer and Frank Wübbelling, *Mathematical methods in image reconstruction*, SIAM, Philadelphia, 2001.

[52] Zeev Nehari, *Conformal mapping*, Dover, New York, 1952.

[53] Alan V. Oppenheim and Ronald W. Schafer, *Digital signal processing*, Prentice Hall, 1975.

[54] Sidney C. Port Paul G. Hoel and Charles J. Stone, *Introduction to Stochastic Processes*, Houghton Mifflin, Boston, Ma, 1972.

[55] Isaac Pesenson, *A sampling theorem on homogeneous manifolds*, Trans. of the Amer. Math. Soc. **352** (2000), 4257–4269.

[56] Robin D. Spital Peter M. Joseph and Charles D. Stockham, *The effects of sampling on CT-images*, Computerized Tomography **4** (1980), 189–206.

[57] Mark A. Pinsky and Michael E. Taylor, *Pointwise Fourier inversion: a wave equation approach*, The Journal of Fourier Analysis and Applications **3** (1997), 647–703.

[58] G. Pólya and G. Szegő, *Problems and theorems in analysis, I*, Springer-Verlag, New York, 1972.

[59] Johan Radon, *über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten*, Ber. Sachs.Akad. Wiss., Leipzig **69** (1917), 262–267.

[60] Matthew Sands Richard P. Feynman, Robert B. Leighton, *The Feynman Lectures on Physics, vol. 2*, Addison-Wesley, Reading, Mass., 1964.

[61] F. Riesz and B. Sz.-Nagy, *Functional analysis*, Fredrick Ungar, New York, 1955.

[62] Theodore J. Rivlin, *The Chebyshev polynomials*, New York, NY, Wiley, 1974.

[63] ———, *An introduction to the approximation theory of functions*, New York, NY, Dover, 1981.

[64] ed. Robert J. Marks II, *Advanced topics in Shannon sampling and interpolation theory*, Springer-Verlag, New York, Berlin, Heidelberg, 1993.

[65] H.L. Royden, *Real Analysis, second edition*, Macmillan, New York, NY, 1968.

[66] Walter Rudin, *Real and complex analysis, second edition*, McGraw-Hill, New York, 1974.

[67] ———, *Principles of mathematical analysis, 3rd ed.*, McGraw-Hill, New York, 1976.

[68] Hermann Schomberg and Jan Timmer, *The gridding method for image reconstruction by Fourier transformation*, IEEE Journal on Medical Imaging **14** (1995), no. 3, 596–607.

[69] L.A. Shepp and J.B. Kruskal, *Computerized tomography: The new medical X-ray technology*, Amer. Math. Monthly (1978), 421–439.

[70] L.A. Shepp and B.F. Logan, *The Fourier reconstruction of a head section*, IEEE Trans. Nuc. Sci. **NS-21** (1990), 21–43.

[71] L.A. Shepp and J.A. Stein, *Simulated reconstruction artifacts in computerized X-ray tomography*, In Ter-Pergossian [77], pp. 33–48.

[72] Elias M. Stein and Guido Weiss, *Introduction to Fourier analysis on Euclidean spaces*, Princeton Press, Princeton, NJ, 1971.

[73] Robert Strichartz, *The way of analysis*, Jones and Bartlett, Boston,MA, 1995.

[74] Kunio Tanabe, *Projection method for solving a singular system of linear equations and its applications*, Num. Math. **17** (1971), 203–214.

[75] M.E. Taylor, *Pseudodifferential Operators*, Princeton Mathematical Series, vol. 34, Princeton University Press, Princeton, NJ, 1981.

[76] ———, *Partial differential equations, vol. 2*, Applied Mathematical Sciences, vol. 116, Springer Verlag, New York, 1996.

[77] M.M. et al. Ter-Pergossian (ed.), *Reconstruction tomography in diagnostic radiology and nuclear medicine*, Baltimore, University Park Press, 1977.

[78] Lloyd N. Trefethen and III David Bau, *Numerical linear algebra*, Philadelphia,PA, SIAM, 1997.

[79] S.R.S. Varadhan, *Stochastic processes*, Courant Institute of Mathematical Sciences, New York, 1968.

[80] E. Mark Haacke, Robert W. Brown, Michael R. Thompson, Ramesh Venkatesan, *Magnetic Resonance Imaging*, Wiley-Liss, New York, 1999.

[81] G.N. Watson, *A treatise on the theory of Bessel functions, 2nd edition*, Cambridge University Press, Cambridge, 1948.

[82] E.T. Whittaker and G.N. Watson, *A Course of Modern Analysis, fourth ed.*, Cambridge University Press, London, 1935.

[83] Harold Widom, *Lectures on integral equations*, Van Nostrand-Reinhold Co., New York, 1969.

[84] Jr. Wilbur B. Davenport and William L. Root, *An introduction to the theory of random signals and noise*, Lincoln Laboratory Publications, McGraw Hill Co., New York, 1958.

[85] Yu. L. Ershov, S. S. Goncharov, A. Nerode, J. B. Remmel and V. W. Marek, *Handbook of recursive mathematics. vol. 2 recursive algebra, analysis and combinatorics*, Studies in Logic and the Foundations of Mathematics, vol. 138, North-Holland, Amsterdam, 1998.

[86] Joie P. Jones Zang-Hee Cho and Manbir Singh, *Foundations of medical imaging*, John Wiley & Sons, New York, 1993.

Index

L^2-derivative
    higher derivatives, 99
    higher dimensions, 127
    one dimension, 98
    periodic case, 193
    periodic case, higher dimensions, 216
$L^2([0,1)$, 580
$L^2([0,1])$, 570
$\mathcal{C}^0([0,1])$, 569
$\mathcal{C}^j(\mathbb{R})$, 74
$\mathcal{C}^k([0,1])$, 570
δ_{ij}, 532
σ-algebra, 425
l^p spaces, 576

Abel transform
    definition, 56
    inversion formula, 59
absolute convergence, 558
absolute value, 551
absolutely convergent series, 616
accumulation points, 615
addition
    vector, 526
aliasing, 229
alternating series test, 558, 618
amplitude, 249, 253
apodizing
    filter, 258
    function, 258, 350
approximation
    step function, 595
approximation problems, 192

back substitution, 545
back-projection formula, 56
bandwidth
    effective, 282
basis, 527, 530
Bayes' law, 438
Bayesian, 417
beam hardening, 45
beam profile, 367
Beer's law, 36
Bernoulli detector, 466
Bessel function, 567
    asymptotic expansion, 568
    integral formula, 568
    power series expansion, 567
Bessel's inequality, 192
bilinear function, 543
binary representation, 516
binary string, 519
binomial formula
    elementary, 564
    general, 564
Borel sets, 426
bounded linear operator, 590
Brownian motion, 482

capacitor, 267
carrier frequency, 234
Cauchy criterion, 523
Cauchy sequence, 523, 615
    normed vector space, 574
    on $\mathbb{R}^n$, 537
Cauchy-Schwarz inequality
    $L^2(\mathbb{R})$, 80
    $\mathbb{R}^n$, 540
centered moment, 443
Central limit theorem, 459
Central slice theorem, 135
    higher dimensions, 171
change of variable formula, 624
characteristic function, 445
characteristic polynomial, 78
Chebyshev inequality, 443
collimator, 332
common refinement, 622
comparison test, 616
completeness, 521, 574
    axiom, 615
complex conjugation, 551
complex exponential, 68
    higher dimensions, 114
complex numbers, 550

complex plane, 551
condition number, 548
conditional convergence, 558
convergence
  generalized functions, 585
  in the mean, 571
  uniform, 570
  with respect to a metric, 537
convergent sequence, 614
convergent subsequence, 614
convex region, 8
convolution
  and Fourier series, 198
  definition $\mathbb{R}^n$, 122
  definition in higher dimensional periodic case, 215
  definition in periodic case, 197
  derivatives and, 124
  Fourier transform of, 85
  of sequences, 198
  one dimension, 84
coordinate vectors, 526
correlation
  coefficient, 448
  matrix, 453
covariance, 448
  matrix, 453
cross-correlation function, 495
cumulative distribution, 441

decay
  rate of, 74
decimal representation, 516, 522
δ-function, 92
density
  probability, 442
derivative
  classical, 620
  generalized function, 584
differentiation
  rules of computation, 620
dimension, 530
Dirichlet kernel, 199
disk of radius r, 48
distance function, 537
distribution function, 442
dot product, 540
dual vector space, 529
dynamic range, 34

effectively bandlimited, 232, 282
empty set, 426
ensemble average, 434
equivalent width, 261
Euclidean n-space, 525
event, 424
  allowable, 424
expected value, 440
exponential polynomials, 191
extend by linearity, 528

fan beam scanner, 342
Fejer kernel, 206
Fejer means, 206
Fejer's theorem, 207
filter, 243
  bandpass, 257
  cascade, 263
  causal, 256
  comb, 279
  commuting, 248
  high pass, 257
  input, 243
  inverse, 281
  isotropic, 288
  linear, 245
  low pass, 232, 257
  multiplication, 248
  non-linear, 245
  output, 243
  passive linear, 267
  shift invariant, 248
filter mask, 320
filtered backprojection, 143, 145
finite difference, 603, 608
finite dimensional distributions, 476
finite Fourier transform, 293
Fourier coefficients, 178
Fourier series
  higher dimensional, 213
  inversion formula, 179, 197
  inversion formula in higher dimensions, 213
  localization principle, 211
  partial sums, 179
  partial sums in higher dimensions, 214
Fourier transform
  definition $\mathbb{R}^1$, 69
  definition $\mathbb{R}^n$-case, 113
  derivatives, 75
  differential equations, 78
  functional notation $\mathcal{F}$, 70
  generalized functions, 108
  inversion formula $\mathbb{R}^1$, 69

inversion formula, higher dimensions, 115
on L^2, 81
fractional derivative, 62
L^2, 100
classical, 100
fractional integral, 62
frequency space description, 252
Fubini's theorem, 624
full width half maximum, 94
higher dimensions, 289
function
L-periodic, 197
rect, 257
absolutely integrable, 69
bandlimited, 220
continuous, 618
differentiable, 620
Riemann integrable, 622
fundamental theorem of algebra, 551
Fundamental Theorem of Calculus, 624
FWHM, *see* full width half maximum

Gamma function, 565
Gaussian
focal spot, 368
Fourier transform, 71
Gaussian focal spot, 331
generalized function, 106, 581
n-dimensions, 587
generating function, 446
geometric distortion, 310
geometric series, 617
sum, 522
Gibbs number, 205
Gibbs phenomenon, 201
Gram-Schmidt, 546
infinite dimensions, 598

Hölder continuity, 62
Hölder space, 62
Hölder's inequality, 573
Hölder-$\frac{1}{2}$ function, 98
Hanning window, 258
Hausdorff-Young inequality, 84
Heaviside function, 255
Heisenberg uncertainty principle, 104
Hilbert transform
as principal value integral, 150
definition, 144
homogeneous equation, 545
Hounsfield units, 34

ideal circuit elements, 267
identity filter, 256
ill-conditioned, 27
ill-posed problem, 65
image, 531
imaginary part, 551
impedence, 267
improper integral, 625
impulse response, 286
inconsistent measurements, 399
independent events, 436
independent increments, 482
independent random variables, 448
inductor, 267
inner product, 540, 543
$\mathbb{R} \times S^1$, 141
integers, 516
integrable function, 431
integral
Lebesgue-Stieltjes, 433
simple function, 430
integral operator, 28
integral test, 558, 617
integration by parts, 74, 624
Intermediate Value Theorem, 619
interpolation
generalized Shannon-Whittaker, 223
spline, 602

Jackson's Theorem, 597
Johnson noise, 497
joint distribution, 447

Kaczmarz method, 412
kernel, 23, 531
kernel function, 63, 247
Kirchoff's laws, 269

Lagrange interpolation, 601
Laplace operator, 135, 148
fractional powers, 148
Laplace's method, 566
Law of large numbers, 460
least squares solution, 411
limit in the mean, 190
limits, 521
complex sequences, 553
for functions, 618
infinite sums, 557, 616
real sequences, 614
rules for computation, 614
linear
function, 524

function, $\mathbb{R}^n$, 526
linear combination, 529
linear equations
finite dimensions, 23
linear functional, 576
linear model, 24
linear operator, 28
linear span, 529
linear system
determined, 25
overdetermined, 25
underdetermined, 26
linear transformation, 531
linearly independent, 530
little 'o' and big 'O' notation, 620
logarithm
complex, 68

mathematical phantom, 374
matrix, 532
change of basis, 533
multiplication, 532
positive definite, 550
sparse, 410
mean
of a random variable, 440
Mean Value Theorem for integrals, 623
measurable set, 425
measure
Lebesgue-Stieltjes, 433
measure space, 424
measure zero, 50
mesh size, 621
method of projections, 412
metric, 537
modulation transfer function, 252
moment conditions, 157
moments
of a random variable, 443
of Radon transform, 158
MRI
sampling in, 227
multi-index notation, 117
multiplication
scalar, 526
mutually exclusive events, 427

Neumann series, 63
noise
quantization, 238
quantum, 501
non-measurable set, 427
norm, 536
normal equations, 411, 550
null space, 23, 531
Nyquist rate, 222
Nyquist width, 262
Nyquist's theorem
noise, 498
sampling, 220

operator norm, 538
oriented hyperplanes, 170
oriented lines
$\mathbb{R}^2$, 15
orthogonal complement, 188
orthogonal matrix, 546
orthogonal projection, 188
orthogonal vectors, 540
oversampling, 222
overshoot, 201

parallel beam scanner, 342
Parseval formula
Fourier series, 188, 197
Fourier series in higher dimensions, 216
Fourier transform, 80
partial sums, 556
partition, 621
passband, 257
passive circuit elements, 269
periodic convolution, 294
phantom, 374
phase, 249
phase shift, 253
picture element, 319
pixel, 319, 340
point source, 254
two dimensions, 38
point spread function, 254
Poisson summation formula, 226
n-dimensions, 241
dual, 228
polynomial approximation
Bernstein polynomials, 597
Weierstrass theorem, 595
power series, 560
principal value integral, 627
prior information, 417
probability, 427
probability measure, 427
probability space, 427
Pythagoras Theorem
infinite dimensional, 190

QR factorization, 546
quantization, 237

Radon inversion formula
  $\mathbb{R}^2$, 138
Radon transform, 47
  adjoint, 141
  and the Laplace operator, 172
  and the wave equation, 172
  convolution property, 132
  definition in higher dimensions, 170
  inverse for radial functions, 57
  inversion formula in higher dimensions, 171
  natural domain, 47, 135
  Parseval formula, 137
  radial functions, 48
random process, 474
  Bernoulli, 476
  continuous parameter, 475
  discrete parameter, 475
  independent, 476
  stationary, 478
  weak sense stationary, 478
random variable, 439
  Bernoulli, 455
  binomial, 455
  complex, 439
  Gaussian, 442, 457
  Poisson, 456
rank value filtering, 322
Ratio test, 617
ray, 343
real computable function, 555
real part, 551
reconstruction algorithm, 337
reconstruction grid, 340
rectifier, 245
regularized inverse, 154
relative error, 548
relaxation parameters, 419
resistor, 267
Riemann Lebesgue Lemma, 73
  Fourier series, 181
  Fourier series in higher dimension, 214
Riemann sum, 603
  upper and lower, 622
Riesz Representation Theorem, 576
Root test, 617

sample path, 475
sample points, 220
sample space, 424
sample spacing, 220, 600
sampling, 600
sampling rate, 220
Schwartz class, 106
semi-norm, 107
sequences
  real, 613
series, 616
shadow function, 16
side lobes, 94
signal-to-noise ratio, 456
simple function, 430
Simpson's rule, 605
sinc function, 95
sinogram, 346
SIRT, 420
smoothness principle, 394
Sobolev embedding theorem, 128
spectral density, 479
spectral function, 399
splines, 602
square integrable, 571
standard basis, 526
standard deviation, 444
state variables, 3
stationary increments, 482
  wide sense, 482
step function, 595
strip integral, 367
subsequence, 613
subspace, 529
summation by parts formula, 559
support line, 8
support of a function, 35

Taylor's theorem, 621
tempered distributions, 586
tempered growth, 586
test function, 97, 127
thermal noise, 497
time average, 434
tomography, 34
transfer function, 252, 286
transpose, 23, 543
trapezoidal rule, 605
triangle inequality, 521

undersampling, 222
uniform continuity, 619
uniform sampling, 600
unit impulse, 92, 254

upper triangular, 545

variance, 444
vector, 526
vector space
    complete normed, 574
    complex, 553
    dual, 576
    normed, 536
    real, 528
view
    fourth generation machine, 346
    parallel beam, 343
    third generation fan beam, 346
Volterra operator, 63
voxel, 340

wave equation, 172
weak convergence, 578
weak derivative, 60, 97, 584
    higher dimensions, 127
    partial, 588
white noise, 491
Wiener filter, 499
Wiener process, 482

zero padding, 298
    higher dimensional, 303

Version 1.0: October 18, 2001 ; Run: November 6, 2001