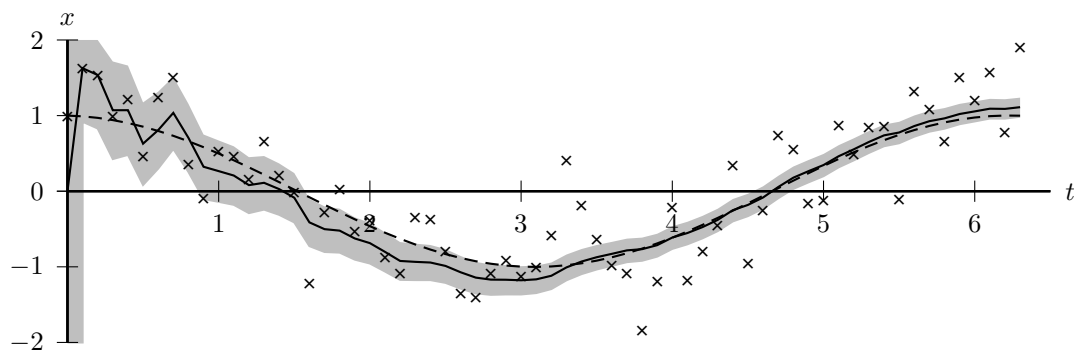


Introduction to Uncertainty Quantification

T. J. Sullivan
Free University of Berlin and Zuse Institute Berlin
Takustrasse 7, D-14195 Berlin-Dahlem, Germany
sullivan@zib.de



WARNING

These notes contain advanced mathematics, which may lead to headaches, frustration, euphoria, and enlightenment. To deter students from burning or hitting these notes, they have been printed on paper laced with 1-diaزيدocarbamoyl-5-azidotetrazole. No responsibility is accepted for loss of sanity, innocence, or limbs.

This page intentionally left almost blank.

Preface

This book is designed as a broad introduction to the mathematics of Uncertainty Quantification (UQ) at the fourth year (senior) undergraduate or beginning postgraduate level. It is aimed primarily at readers from a mathematical or statistical (rather than, say, engineering) background. The main mathematical prerequisite is familiarity with the language of linear functional analysis and measure / probability theory, and some familiarity with basic optimization theory. Chapters 2–5 of the text provide a review of this material, generally without detailed proof.

The aim of this book has been to give a survey of the main objectives in the field of UQ and a few of the mathematical methods by which they can be achieved. However, this book is no exception to the old saying that books are never completed, only abandoned. There are many more UQ problems and solution methods in the world than those covered here. For any grievous omissions, I ask for your indulgence, and would be happy to receive suggestions for improvements. With the exception of the preliminary material on measure theory and functional analysis, this book should serve as a basis for a course comprising 30–45 hours' worth of lectures, depending upon the instructor's choices in terms of selection of topics and depth of treatment.

The examples and exercises in this book aim to be simple but informative about individual components of UQ studies: practical applications almost always require some ad hoc combination of multiple techniques (e.g., Gaussian process regression plus quadrature plus reduced-order modelling). Such compound examples have been omitted in the interests of keeping the presentation of the mathematical ideas clean, and in order to focus on examples and exercises that will be more useful to instructors and students.

Each chapter concludes with a bibliography, the aim of which is threefold: to give sources for results discussed but not proved in the text; to give some historical overview and context; and, most importantly, to give students a jumping-off point for further reading and research. This has led to a large bibliography, but hopefully a more useful text for budding researchers.

I would like to thank Achi Dosanjh at Springer for her stewardship of this project, and the anonymous reviewers for their thoughtful comments, which prompted many improvements to the manuscript.

From initial conception to nearly finished product, this book has benefitted from interactions with many people: they have given support and encouragement, offered stimulating perspectives on the text and the field of UQ, and pointed out the inevitable typographical mistakes. In particular, I would like to thank Paul Constantine, Zach Dean, Charlie Elliott, Zydrunas Gimbutas, Calvin Khor, Ilja Klebanov, Han Cheng Lie, Milena Kremakova, David McCormick, Damon McDougall, Mike McKerns, Akil Narayan, Michael Ortiz, Houman Owjadi, Adwaye Rambojun, Asbjørn Nilsen Riseth, Clint Scovel, Colin Sparrow, Andrew Stuart, Florian Theil, Joy Tolia, Florian Wechsung, Thomas Whitaker, and Aimée Williams.

Finally, since the students on the 2013–14 iteration of the University of Warwick mathematics module MA4K0 Introduction to Uncertainty Quantification were curious and brave enough to be the initial 'guinea pigs' for this material, they deserve a special note of thanks.

The University of Warwick
Coventry, U.K.
July 2015

T.J.S.

Contents

Preface	i
Contents	i
1 Introduction	1
1.1 What is Uncertainty Quantification?	1
1.2 Mathematical Prerequisites	5
1.3 Outline of the Book	5
1.4 The Road Not Taken	6
2 Measure and Probability Theory	7
2.1 Measure and Probability Spaces	7
2.2 Random Variables and Stochastic Processes	11
2.3 Lebesgue Integration	12
2.4 Decomposition and Total Variation of Signed Measures	14
2.5 The Radon–Nikodým Theorem and Densities	15
2.6 Product Measures and Independence	16
2.7 Gaussian Measures	18
2.8 Interpretations of Probability	22
2.9 Bibliography	24
2.10 Exercises	24
3 Banach and Hilbert Spaces	27
3.1 Basic Definitions and Properties	27
3.2 Banach and Hilbert Spaces	30
3.3 Dual Spaces and Adjoints	33
3.4 Orthogonality and Direct Sums	34
3.5 Tensor Products	38
3.6 Bibliography	40
3.7 Exercises	40
4 Optimization Theory	42
4.1 Optimization Problems and Terminology	42
4.2 Unconstrained Global Optimization	43
4.3 Constrained Optimization	46
4.4 Convex Optimization	48
4.5 Linear Programming	51
4.6 Least Squares	52
4.7 Bibliography	55
4.8 Exercises	56

5	Measures of Information and Uncertainty	57
5.1	The Existence of Uncertainty	57
5.2	Interval Estimates	58
5.3	Variance, Information and Entropy	59
5.4	Information Gain, Distances, and Divergences	61
5.5	Bibliography	66
5.6	Exercises	66
6	Bayesian Inverse Problems	69
6.1	Inverse Problems and Regularization	69
6.2	Bayesian Inversion in Banach Spaces	74
6.3	Well-Posedness and Approximation	77
6.4	Accessing the Bayesian Posterior Measure	80
6.5	Frequentist Consistency of Bayesian Methods	81
6.6	Bibliography	83
6.7	Exercises	84
7	Filtering and Data Assimilation	86
7.1	State Estimation in Discrete Time	87
7.2	Linear Kálmán Filter	89
7.3	Extended Kálmán Filter	95
7.4	Ensemble Kálmán Filter	96
7.5	Bibliography	97
7.6	Exercises	98
8	Orthogonal Polynomials and Applications	100
8.1	Basic Definitions and Properties	100
8.2	Recurrence Relations	105
8.3	Differential Equations	108
8.4	Roots of Orthogonal Polynomials	109
8.5	Polynomial Interpolation	111
8.6	Polynomial Approximation	114
8.7	Multivariate Orthogonal Polynomials	116
8.8	Bibliography	119
8.9	Exercises	119
8.10	Tables of Classical Orthogonal Polynomials	122
9	Numerical Integration	126
9.1	Univariate Quadrature	127
9.2	Gaussian Quadrature	129
9.3	Clenshaw–Curtis / Fejér Quadrature	132
9.4	Multivariate Quadrature	133
9.5	Monte Carlo Methods	136
9.6	Pseudo-Random Methods	142
9.7	Bibliography	146
9.8	Exercises	148
10	Sensitivity Analysis and Model Reduction	150
10.1	Model Reduction for Linear Models	150
10.2	Derivatives	153
10.3	McDiarmid Diameters	157
10.4	ANOVA/HDMR Decompositions	159
10.5	Active Subspaces	161
10.6	Bibliography	164

10.7 Exercises	166
11 Spectral Expansions	169
11.1 Karhunen–Loève Expansions	169
11.2 Wiener–Hermite Polynomial Chaos	177
11.3 Generalized Polynomial Chaos Expansions	180
11.4 Wavelet Expansions	185
11.5 Bibliography	188
11.6 Exercises	188
12 Stochastic Galerkin Methods	190
12.1 Weak Formulation of Nonlinearities	191
12.2 Random Ordinary Differential Equations	194
12.3 Lax–Milgram Theory and Random PDEs	197
12.4 Bibliography	206
12.5 Exercises	207
13 Non-Intrusive Methods	210
13.1 Non-Intrusive Spectral Methods	210
13.2 Stochastic Collocation	214
13.3 Gaussian Process Regression	219
13.4 Bibliography	222
13.5 Exercises	222
14 Distributional Uncertainty	224
14.1 Maximum Entropy Distributions	224
14.2 Hierarchical Methods	227
14.3 Distributional Robustness	227
14.4 Functional and Distributional Robustness	236
14.5 Bibliography	239
14.6 Exercises	240
Bibliography	242
Index	258

Chapter 1

Introduction

We must think differently about our ideas — and how we test them. We must become more comfortable with probability and uncertainty. We must think more carefully about the assumptions and beliefs that we bring to a problem.

*The Signal and the Noise: The Art of Science
and Prediction*
NATE SILVER

1.1 What is Uncertainty Quantification?

This book is an introduction to the mathematics of Uncertainty Quantification (UQ), but what is UQ? It is, roughly put, the coming together of probability theory and statistical practice with ‘the real world’. These two anecdotes illustrate something of what is meant by this statement:

- Until the early-to-mid 1990s, risk modelling for catastrophe insurance and re-insurance (i.e. insurance for property owners against risks arising from earthquakes, hurricanes, terrorism, etc., and then insurance for the providers of such insurance) was done on a purely statistical basis. Since that time, catastrophe modellers have tried to incorporate models for the underlying physics or human behaviour, hoping to gain a more accurate predictive understanding of risks by blending the statistics and the physics, e.g. by focussing on what is both statistically *and* physically reasonable. This approach also allows risk modellers to study interesting hypothetical scenarios in a meaningful way, e.g. using a physics-based model of water drainage to assess potential damage from rainfall 10% in excess of the historical maximum.
- Over roughly the same period of time, deterministic engineering models of complex physical processes began to incorporate some element of uncertainty to account for lack of knowledge about important physical parameters, random variability in operating circumstances, or outright ignorance about what the form of a ‘correct’ model would be. Again the aim is to provide more accurate predictions about systems’ behaviour.

Thus, a ‘typical’ UQ problem involves one or more mathematical models for a process of interest, subject to some uncertainty about the correct form of, or parameter values for, those models. Often, though not always, these uncertainties are treated probabilistically.

Perhaps as a result of its history, there are many perspectives on what UQ is, including at the extremes assertions like “UQ is just a buzzword for statistics” or “UQ is just error analysis”. These points of view are somewhat extremist, but they do contain a kernel of

truth: very often, the probabilistic theory underlying UQ methods is actually quite simple, but is obscured by the details of the application. However, the complications that practical applications present are also part of the essence of UQ: it is all very well giving an accurate prediction for some insurance risk in terms of an elementary mathematical object such as an expected value, but how will you actually go about evaluating that expected value when it is an integral over a million-dimensional parameter space? Thus, it is important to appreciate both the underlying mathematics and the practicalities of implementation, and the presentation here leans towards the former while keeping the latter in mind.

Typical UQ problems of interest include certification, prediction, model and software verification and validation, parameter estimation, data assimilation, and inverse problems. At its very broadest,

“UQ studies all sources of error and uncertainty, including the following: systematic and stochastic measurement error; ignorance; limitations of theoretical models; limitations of numerical representations of those models; limitations of the accuracy and reliability of computations, approximations, and algorithms; and human error. A more precise definition is UQ is the end-to-end study of the reliability of scientific inferences.” (U.S. Department of Energy, 2009, p. 135)

It is especially important to appreciate the “end-to-end” nature of UQ studies: one is interested in *relationships between pieces of information*, not the ‘truth’ of those pieces of information / assumptions, bearing in mind that they are only approximations of reality. There is always going to be a risk of ‘Garbage In, Garbage Out’. UQ cannot tell you that your model is ‘right’ or ‘true’, but only that, *if* you accept the validity of the model (to some quantified degree), *then* you must logically accept the validity of certain conclusions (to some quantified degree). In the author’s view, this is the proper interpretation of philosophically sound but somewhat unhelpful assertions like “Verification and validation of numerical models of natural systems is impossible” and “The primary value of models is heuristic” (Oreskes et al., 1994). UQ can, however, tell you that two or more of your modelling assumptions are mutually contradictory, and hence that your model is wrong, and a complete UQ analysis will include a meta-analysis examining the sensitivity of the original analysis to perturbations of the governing assumptions.

A prototypical, if rather over-used, example for UQ is an elliptic PDE with uncertainty coefficients:

Example 1.1. Consider the following elliptic boundary value problem on a connected Lipschitz domain $\mathcal{X} \subseteq \mathbb{R}^n$ (typically $n = 2$ or 3):

$$\begin{aligned} -\nabla \cdot (\kappa \nabla u) &= f && \text{in } \mathcal{X}, \\ u &= b && \text{on } \partial\mathcal{X}. \end{aligned} \tag{1.1}$$

Problem (1.1) is a simple but not overly naïve model for the pressure field u of some fluid occupying a domain \mathcal{X} . The domain \mathcal{X} consists of a material, and the tensor field $\kappa: \mathcal{X} \rightarrow \mathbb{R}^{n \times n}$ describes the permeability of this material to the fluid. There is a source term $f: \mathcal{X} \rightarrow \mathbb{R}$, and the boundary condition specifies the values $b: \partial\mathcal{X} \rightarrow \mathbb{R}$ that the pressure takes on the boundary of \mathcal{X} . This model is of interest in the earth sciences because Darcy’s law asserts that the velocity field v of the fluid flow in this medium is related to the gradient of the pressure field by

$$v = \kappa \nabla u.$$

If the fluid contains some kind of contaminant, then it may be important to understand where fluid following the velocity field v will end up, and when.

In a course on PDE theory, you will learn that, for each given positive-definite and essentially bounded permeability field κ , problem (1.1) has a unique weak solution u in the Sobolev space $H_0^1(\mathcal{X})$ for each forcing term f in the dual Sobolev space $H^{-1}(\mathcal{X})$. This is known as the *forward problem*. One objective of this book is to tell you that this is far

from the end of the story! As far as practical applications go, existence and uniqueness of solutions to the forward problem is only the beginning. For one thing, this PDE model is only an approximation of reality. Secondly, even if the PDE were a perfectly accurate model, the ‘true’ κ , f and b are not known precisely, so our knowledge about $u = u(\kappa, f, b)$ is also uncertain in some way. If κ , f and b are treated as random variables, then u is also a random variable, and one is naturally interested in properties of that random variable such as mean, variance, deviation probabilities etc. This is known as the *forward propagation of uncertainty*, and to perform it we must build some theory for probability on function spaces.

Another issue is that often we want to solve an *inverse problem*: perhaps we know something about f , b and u and want to infer κ via the relationship (1.1). For example, we may observe the pressure $u(x_i)$ at finitely many points $x_i \in \mathcal{X}$; This problem is hugely underdetermined, and hence ill-posed; ill-posedness is characteristic of many inverse problems, and is only worsened by the fact that the observations may be corrupted by observational noise. Even a prototypical inverse problem such as this one is of enormous practical interest: it is by solving such inverse problems that oil companies attempt to infer the location of oil deposits in order to make a profit, and seismologists the structure of the planet in order to make earthquake predictions. Both of these problems, the forward and inverse propagation of uncertainty, fall under the very general remit of UQ. Furthermore, in practice, the domain \mathcal{X} and the fields f , b , κ and u are all discretized and solved for numerically (i.e. approximately and finite-dimensionally), so it is of interest to understand the impact of these discretization errors.

Epistemic and Aleatoric Uncertainty. It is common to divide uncertainty into two types, *aleatoric* and *epistemic* uncertainty. Aleatoric uncertainty — from the Latin *alea*, meaning a die — refers to uncertainty about an inherently variable phenomenon. Epistemic uncertainty — from the Greek *ἐπιστήμη*, meaning knowledge — refers to uncertainty arising from lack of knowledge. If one has at hand a model for some system of interest, then epistemic uncertainty is often further subdivided into *model form* uncertainty, in which one has significant doubts that the model is even ‘structurally correct’, and *parametric* uncertainty, in which one believes that the form of the model reflects reality well, but one is uncertain about the correct values to use for particular parameters in the model.

To a certain extent, the distinction between epistemic and aleatoric uncertainty is an imprecise one, and repeats the old debate between frequentist and subjectivist (e.g. Bayesian) statisticians. Someone who was simultaneously a devout Newtonian physicist and a devout Bayesian might argue that the results of dice rolls are not aleatoric uncertainties — one simply doesn’t have complete enough information about the initial conditions of die, the material and geometry of the die, any gusts of wind that might affect the flight of the die, and so forth. On the other hand, it is usually clear that some forms of uncertainty are epistemic rather than aleatoric: for example, when physicists say that they have yet to come up with a Theory of Everything, they are expressing a lack of knowledge about the laws of physics in our universe, and the correct mathematical description of those laws. In any case, regardless of one’s favoured interpretation of probability, the language of probability theory is a powerful tool in describing uncertainty.

Some Typical UQ Objectives. Many common UQ objectives can be illustrated in the context of a system, F , that maps inputs X in some space \mathcal{X} to outputs $Y = F(X)$ in some space \mathcal{Y} . Some common UQ objectives include:

- The *forward propagation* or *push-forward problem*. Suppose that the uncertainty about the inputs of F can be summarized in a probability distribution μ on \mathcal{X} . Given this, determine the induced probability distribution $F_*\mu$ on the output space \mathcal{Y} , as defined by

$$(F_*\mu)(E) := \mathbb{P}_\mu(\{x \in \mathcal{X} \mid F(x) \in E\}) \equiv \mathbb{P}_\mu[F(X) \in E].$$

This task is typically complicated by μ being a complicated distribution, or F being

non-linear. Because $(F_*\mu)$ is a very high-dimensional object, it is often more practical to identify some specific outcomes of interest and settle for a solution of the following problem:

- The *reliability* or *certification problem*. Suppose that some set $\mathcal{Y}_{\text{fail}} \subseteq \mathcal{Y}$ is identified as a ‘failure set’, i.e. the outcome $F(X) \in \mathcal{Y}_{\text{fail}}$ is undesirable in some way. Given appropriate information about the inputs X and forward process F , determine the failure probability,

$$\mathbb{P}_\mu[F(X) \in \mathcal{Y}_{\text{fail}}].$$

Furthermore, in the case of a failure, how large will the deviation from acceptable performance be, and what are the consequences?

- The *prediction problem*. Dually to the reliability problem, given a maximum acceptable probability of error $\varepsilon > 0$, find a set $\mathcal{Y}_\varepsilon \subseteq \mathcal{Y}$ such that

$$\mathbb{P}_\mu[F(X) \in \mathcal{Y}_\varepsilon] \geq 1 - \varepsilon.$$

i.e. the prediction $F(X) \in \mathcal{Y}_\varepsilon$ is wrong with probability at most ε .

- An *inverse problem*, such as *state estimation* (often for a quantity that is changing in time) or *parameter identification* (usually for a quantity that is not changing, or is non-physical model parameter). Given some observations of the output, Y , which may be corrupted or unreliable in some way, attempt to determine the corresponding inputs X such that $F(X) = Y$. In what sense are some estimates for X more or less reliable than others?
- The *model reduction* or *model calibration problem*. Construct another function F_h (perhaps a numerical model with certain numerical parameters to be *calibrated*, or one involving far fewer input or output variables) such that $F_h \approx F$ in an appropriate sense. Quantifying the accuracy of the approximation may itself be a certification or prediction problem.

Sometimes a UQ problem consists of several of these problems coupled together: for example, one might have to solve an inverse problem to produce or improve some model parameters, and then use those parameters to propagate some other uncertainties forwards, and hence produce a prediction that can be used for decision support in some certification problem.

Typical issues to be confronted in addressing these problems include the high dimension of the parameter spaces associated with practical problems; the approximation of integrals (expected values) by numerical quadrature; the cost of evaluating functions that often correspond to expensive computer simulations or physical experiments; and non-negligible epistemic uncertainty about the correct form of vital ingredients in the analysis, such as the functions and probability measures in key integrals.


The aim of this book is to provide an introduction to the fundamental mathematical ideas underlying the basic approaches to these types of problems. Practical UQ applications almost always require some ad hoc combination of multiple techniques, adapted and specialized to suit the circumstances, but the emphasis here is on basic ideas, with simple illustrative examples. The hope is that interested students or practitioners will be able to generalize from the topics covered here to their particular problems of interest, with the help of additional resources cited in the bibliographic discussions at the end of each chapter. So, for example, while Chapter 12 discusses intrusive (Galerkin) methods for UQ with an implicit assumption that the basis is a polynomial chaos basis, one should be able to adapt these ideas to non-polynomial bases.



A Word of Warning. UQ is not a mature field like linear algebra or single-variable complex analysis, with stately textbooks containing well-polished presentations of classical theorems bearing august names like Cauchy, Gauss and Hamilton. Both because of its youth as a field and its very close engagement with applications, UQ is much more about problems,

methods, and ‘good enough for the job’. There are some very elegant approaches *within* UQ, but as yet no single, general, over-arching theory *of* UQ.

1.2 Mathematical Prerequisites

Like any course or text, this book has some prerequisites. The perspective on UQ that runs through this book is strongly (but not exclusively) grounded in probability theory and Hilbert spaces, so the main prerequisite is familiarity with the language of linear functional analysis and measure / probability theory. As a crude diagnostic test, read the following sentence: 

Given any σ -finite measure space $(\mathcal{X}, \mathcal{F}, \mu)$, the set of all \mathcal{F} -measurable functions $f: \mathcal{X} \rightarrow \mathbb{C}$ for which $\int_{\mathcal{X}} |f|^2 d\mu$ is finite, modulo equality μ -almost everywhere, is a Hilbert space with respect to the inner product $\langle f, g \rangle := \int_{\mathcal{X}} \bar{f}g d\mu$.

None of the symbols, concepts or terms used or implicit in that sentence should give prospective students or readers any serious problems. Chapters 2 and 3 give a recap, without proof, of the necessary concepts and results, and most of the material therein should be familiar territory. In addition, Chapters 4 and 5 provide additional mathematical background on optimization and information theory respectively. It is assumed that readers have greater prior familiarity with the material in Chapters 2 and 3 than the material in Chapters 4 and 5; this is reflected in the way that results are presented mostly without proof in Chapters 2 and 3, but with proof in Chapters 4 and 5.

If, in addition, students or readers have some familiarity with topics such as numerical analysis, ordinary and partial differential equations, and stochastic analysis, then certain techniques, examples and remarks will make more sense. None of these are essential prerequisites, but, some ability and willingness to implement UQ methods — even in simple settings — in e.g. C/C++, Mathematica, Matlab, or Python is highly desirable. (Some of the concepts covered in the book will be given example numerical implementations in Python.) Although the aim of this book is to give an overview of the mathematical elements of UQ, this is a topic best learned in the doing, not through pure theory. However, in the interests of accessibility and pedagogy, none of the examples or exercises in this book will involve serious programming legerdemain.

1.3 Outline of the Book

The first part of this book lays out basic and general mathematical tools for the later discussion of UQ. Chapter 2 covers measure and probability theory, which are essential tools given the probabilistic description of many UQ problems. Chapter 3 covers some elements of linear functional analysis on Banach and Hilbert spaces, and constructions such as tensor products, all of which are natural spaces for the representation of random quantities and fields. Many UQ problems involve a notion of ‘best fit’, and so Chapter 4 provides a brief introduction to optimization theory in general, with particular attention to linear programming and least squares. Finally, although much of the UQ theory in this book is probabilistic, and is furthermore an L^2 theory, Chapter 5 covers more general notions of information and uncertainty.

The second part of the book is concerned with mathematical tools that are much closer to the practice of UQ. We begin in Chapter 6 with a mathematical treatment of inverse problems, and specifically their Bayesian interpretation; we take advantage of the tools developed in Chapters 2 and 3 to discuss Bayesian inverse problems on function spaces, which are especially important in PDE applications. In Chapter 7, this leads to a specific class of inverse problems, filtering and data assimilation problems, in which data and unknowns are decomposed in a sequential manner. Chapter 8 introduces orthogonal polynomial theory, a classical area of mathematics that has a double application in UQ: orthogonal

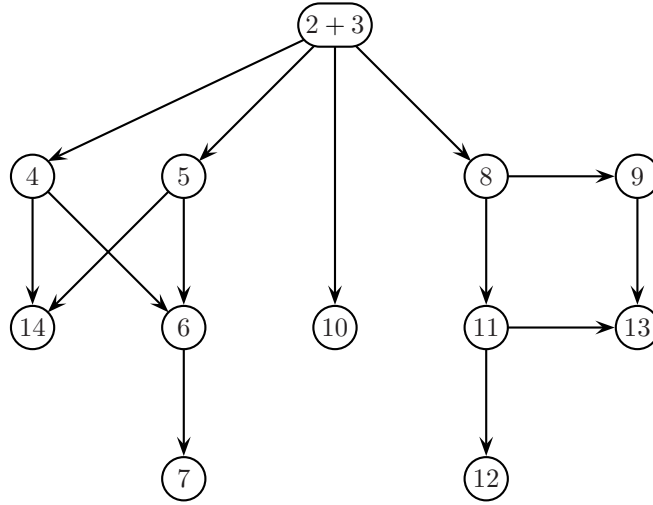


Figure 1.1: Outline of the book (Leitfaden). An arrow from m to n indicates that Chapter n substantially depends upon material in Chapter m .

polynomials are useful basis functions for the representation of random processes, and form the basis of powerful numerical integration (quadrature) algorithms. Chapter 9 discusses these quadrature methods in more detail, along with other methods such as Monte Carlo. Chapter 10 covers one aspect of forward uncertainty propagation, namely sensitivity analysis and model reduction, i.e. finding out which input parameters are influential in determining the values of some output process. Chapter 11 introduces spectral decompositions of random variables and other random quantities, including but not limited to polynomial chaos methods. Chapter 12 covers the intrusive (or Galerkin) approach to the determination of coefficients in spectral expansions; Chapter 13 covers the alternative non-intrusive (sample-based) paradigm. Finally, Chapter 14 discusses approaches to probability-based UQ that apply when even the probability distributions of interest are uncertain in some way.

1.4 The Road Not Taken

There are many topics relevant to UQ that are either not covered or discussed only briefly here, including: detailed treatment of data assimilation beyond the confines of the Kálmán filter and its variations; accuracy, stability and computational cost of numerical methods; details of numerical implementation of optimization methods; stochastic homogenization and other multiscale methods; optimal control and robust optimization; machine learning; issues related to ‘big data’; and the visualization of uncertainty.

Chapter 2

Measure and Probability Theory

To be conscious that you are ignorant is a great step to knowledge.

Sybil
BENJAMIN DISRAELI

Probability theory, grounded in Kolmogorov's axioms and the general foundations of measure theory, is an essential tool in the quantitative mathematical treatment of uncertainty. Of course, probability is not the only framework for the discussion of uncertainty: there is also the paradigm of interval analysis, and intermediate paradigms such as Dempster–Shafer theory, as discussed in Section 2.8 and Chapter 5.

This chapter serves as a review, without detailed proof, of concepts from measure and probability theory that will be used in the rest of the text. Like Chapter 3, this chapter is intended as a review of material that should be understood as a prerequisite before proceeding; to an extent, Chapters 2 and 3 are interdependent and so can (and should) be read in parallel with one another.

2.1 Measure and Probability Spaces

The basic objects of measure and probability theory are sample spaces, which are abstract sets; we distinguish certain subsets of these sample spaces as being ‘measurable’, and assign to each of them a numerical notion of ‘size’. In probability theory, this size will always be a real number between 0 and 1, but more general values are possible, and indeed useful.

Definition 2.1. A *measurable space* is a pair $(\mathcal{X}, \mathcal{F})$, where

- (a) \mathcal{X} is a set, called the *sample space*; and
- (b) \mathcal{F} is a σ -algebra on \mathcal{X} , i.e. a collection of subsets of \mathcal{X} containing \emptyset and closed under countable applications of the operations of union, intersection, and complementation relative to \mathcal{X} ; elements of \mathcal{F} are called *measurable sets* or *events*.

Example 2.2. (a) On any set \mathcal{X} , there is a *trivial σ -algebra* in which the only measurable sets are the empty set \emptyset and the whole space \mathcal{X} .

- (b) On any set \mathcal{X} , there is also the *power set σ -algebra* in which every subset of \mathcal{X} is measurable. It is a fact of life that this σ -algebra contains too many measurable sets to be useful for most applications in analysis and probability.

- (c) When \mathcal{X} is a topological — or, better yet, metric or normed — space, it is common to take \mathcal{F} to be the *Borel σ -algebra* $\mathcal{B}(\mathcal{X})$, the smallest σ -algebra on \mathcal{X} so that every open set (and hence also every closed set) is measurable.
-

- Definition 2.3.** (a) A *signed measure* (or *charge*) on a measurable space $(\mathcal{X}, \mathcal{F})$ is a function $\mu: \mathcal{F} \rightarrow \mathbb{R} \cup \{\pm\infty\}$ that takes at most one of the two infinite values, has $\mu(\emptyset) = 0$, and, whenever $E_1, E_2, \dots \in \mathcal{F}$ are pairwise disjoint with union $E \in \mathcal{F}$, then $\mu(E) = \sum_{n \in \mathbb{N}} \mu(E_n)$. In the case that $\mu(E)$ is finite, we required that the series $\sum_{n \in \mathbb{N}} \mu(E_n)$ converges absolutely to $\mu(E)$.
- (b) A *measure* is a signed measure that does not take negative values.
- (c) A *probability measure* is a measure such that $\mu(\mathcal{X}) = 1$.

The triple $(\mathcal{X}, \mathcal{F}, \mu)$ is called a *signed measure space*, *measure space*, or *probability space* as appropriate. The sets of all signed measures, measures, and probability measures on $(\mathcal{X}, \mathcal{F})$ are denoted $\mathcal{M}_\pm(\mathcal{X}, \mathcal{F})$, $\mathcal{M}_+(\mathcal{X}, \mathcal{F})$, and $\mathcal{M}_1(\mathcal{X}, \mathcal{F})$ respectively.

- Example 2.4.** (a) The *trivial measure* can be defined on any set \mathcal{X} and σ -algebra: $\tau(E) := 0$ for every $E \in \mathcal{F}$.
- (b) The *unit Dirac measure* at $a \in \mathcal{X}$ can also be defined on any set \mathcal{X} and σ -algebra:

$$\delta_a(E) := \begin{cases} 1, & \text{if } a \in E, E \in \mathcal{F}, \\ 0, & \text{if } a \notin E, E \in \mathcal{F}. \end{cases}$$

- (c) Similarly, we can define *counting measure*:

$$\kappa(E) := \begin{cases} n, & \text{if } E \in \mathcal{F} \text{ is a finite set with exactly } n \text{ elements,} \\ +\infty, & \text{if } E \in \mathcal{F} \text{ is an infinite set.} \end{cases}$$

- (d) *Lebesgue measure* on \mathbb{R}^n is the unique measure on \mathbb{R}^n (equipped with its Borel σ -algebra $\mathcal{B}(\mathbb{R}^n)$, generated by the Euclidean open balls) that assigns to every rectangle its n -dimensional volume in the ordinary sense. To be more precise, Lebesgue measure is actually defined on the completion $\mathcal{B}_0(\mathbb{R}^n)$ of $\mathcal{B}(\mathbb{R}^n)$, which is a larger σ -algebra than $\mathcal{B}(\mathbb{R}^n)$. The rigorous construction of Lebesgue measure is a non-trivial undertaking.
- (e) Signed measures / charges arise naturally in the modelling of distributions with positive and negative values, e.g. $\mu(E)$ = the net electrical charge within some measurable region $E \subseteq \mathbb{R}^3$. They also arise naturally as differences of non-negative measures: see Theorem 2.24 later on.

Remark 2.5. Probability theorists usually denote the sample space of a probability space by Ω ; PDE theorists often use the same letter to denote a domain in \mathbb{R}^n on which a partial differential equation is to be solved. In UQ, where the worlds of probability and PDE theory often collide, the possibility of confusion is clear. Therefore, this book will tend to use Θ for a probability space and \mathcal{X} for a more general measurable space, which may happen to be the spatial domain for some PDE.

Definition 2.6. Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space.

- (a) If $N \subseteq \mathcal{X}$ is a subset of a measurable set $E \in \mathcal{F}$ such that $\mu(E) = 0$, then N is called a *μ -null set*.
- (b) If the set of $x \in \mathcal{X}$ for which some property $P(x)$ does not hold is μ -null, then P is said to hold *μ -almost everywhere* (or, when μ is a probability measure, *μ -almost surely*).
- (c) If every μ -null set is in fact an \mathcal{F} -measurable set, then the measure space $(\mathcal{X}, \mathcal{F}, \mu)$ is said to be *complete*.

Example 2.7. Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space, and let $f: \mathcal{X} \rightarrow \mathbb{R}$ be some function. If $f(x) \geq t$ for μ -almost every $x \in \mathcal{X}$, then t is an *essential lower bound* for f ; the greatest such t is called the *essential infimum* of f :

$$\text{ess inf } f := \sup \{t \in \mathbb{R} \mid f \geq t \text{ } \mu\text{-almost everywhere}\}.$$

Similarly, if $f(x) \leq t$ for μ -almost every $x \in \mathcal{X}$, then t is an *essential upper bound* for f ; the least such t is called the *essential supremum* of f :

$$\text{ess sup } f := \inf \{t \in \mathbb{R} \mid f \leq t \text{ } \mu\text{-almost everywhere}\}.$$

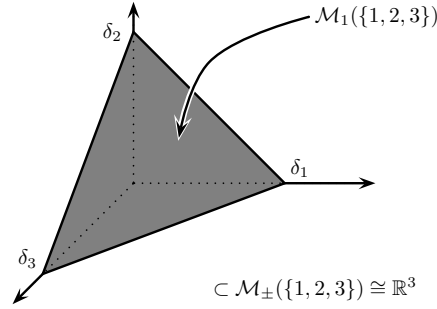


Figure 2.1: The probability simplex $\mathcal{M}_1(\{1, 2, 3\})$, drawn as the triangle spanned by the unit Dirac masses δ_i , $i \in \{1, 2, 3\}$, in the vector space of signed measures on $\{1, 2, 3\}$.

It is so common in measure and probability theory to need to refer to the set of all points $x \in \mathcal{X}$ such that some property $P(x)$ holds true that an abbreviated notation has been adopted: simply $[P]$. Thus, for example, if $f: \mathcal{X} \rightarrow \mathbb{R}$ is some function, then

$$[f \leq t] := \{x \in \mathcal{X} \mid f(x) \leq t\}.$$

As noted above, when the sample space is a topological space, it is usual to use the Borel σ -algebra (i.e. the smallest σ -algebra that contains all the open sets); measures on the Borel σ -algebra are called *Borel measures*. Unless noted otherwise, this is the convention followed here.

Definition 2.8. The *support* of a measure μ defined on a topological space \mathcal{X} is

$$\text{supp}(\mu) := \bigcap \{F \subseteq \mathcal{X} \mid F \text{ is closed and } \mu(\mathcal{X} \setminus F) = 0\}.$$

That is, $\text{supp}(\mu)$ is the smallest closed subset of \mathcal{X} that has full μ -measure. Equivalently, $\text{supp}(\mu)$ is the complement of the union of all open sets of μ -measure zero, or the set of all points $x \in \mathcal{X}$ for which every neighbourhood of x has strictly positive μ -measure.

Especially in Chapter 14, we shall need to consider the set of all probability measures defined on a measurable space. $\mathcal{M}_1(\mathcal{X})$ is often called the *probability simplex* on \mathcal{X} . The motivation for this terminology comes from the case in which $\mathcal{X} = \{1, \dots, n\}$ is a finite set equipped with the power set σ -algebra, which is the same as the Borel σ -algebra for the discrete topology on \mathcal{X} .^[2.1] In this case, functions $f: \mathcal{X} \rightarrow \mathbb{R}$ are in bijection with column vectors

$$\begin{bmatrix} f(1) \\ \vdots \\ f(n) \end{bmatrix}$$

and probability measures μ on the power set of \mathcal{X} are in bijection with the $(n-1)$ -dimensional set of row vectors

$$[\mu(\{1\}) \quad \dots \quad \mu(\{n\})]$$

such that $\mu(\{i\}) \geq 0$ for all $i \in \{1, \dots, n\}$ and $\sum_{i=1}^n \mu(\{i\}) = 1$. As illustrated in Figure 2.1, the set of such μ is the $(n-1)$ -dimensional simplex in \mathbb{R}^n that is the convex hull of the n points $\delta_1, \dots, \delta_n$,

$$\delta_i = [0 \quad \dots \quad 0 \quad 1 \quad 0 \quad \dots \quad 0],$$

^[2.1]It is an entertaining exercise to see what pathological properties can hold for a probability measures on a σ -algebra other than the power set of a finite set \mathcal{X} .

with 1 in the i^{th} column. Looking ahead, the expected value of f under μ (to be defined properly in Section 2.3) is exactly the matrix product:

$$\mathbb{E}_\mu[f] = \sum_{i=1}^n \mu(\{i\})f(i) = \langle \mu | f \rangle = [\mu(\{1\}) \quad \cdots \quad \mu(\{n\})] \begin{bmatrix} f(1) \\ \vdots \\ f(n) \end{bmatrix}.$$

It is useful to keep in mind this geometric picture of $\mathcal{M}_1(\mathcal{X})$ in addition to the algebraic and analytical properties of any given $\mu \in \mathcal{M}_1(\mathcal{X})$. As poetically highlighted by Sir Michael Atiyah (2004, Paper 160, p. 7):

“Algebra is the offer made by the devil to the mathematician. The devil says: ‘I will give you this powerful machine, it will answer any question you like. All you need to do is give me your soul: give up geometry and you will have this marvellous machine.’”

Or, as is traditionally but perhaps apocryphally said to have been inscribed over the entrance to Plato’s Academy:

ΑΓΕΩΜΕΤΡΗΤΟΣ ΜΗΔΕΙΣ ΕΙΣΙΤΩ

In a sense that will be made precise in Chapter 14, for any ‘nice’ space \mathcal{X} , $\mathcal{M}_1(\mathcal{X})$ is the simplex spanned by the collection of unit Dirac measures $\{\delta_x \mid x \in \mathcal{X}\}$. Given a bounded, measurable function $f: \mathcal{X} \rightarrow \mathbb{R}$ and $c \in \mathbb{R}$,

$$\{\mu \in \mathcal{M}(\mathcal{X}) \mid \mathbb{E}_\mu[f] \leq c\}$$

is a half-space of $\mathcal{M}(\mathcal{X})$, and so a set of the form

$$\{\mu \in \mathcal{M}_1(\mathcal{X}) \mid \mathbb{E}_\mu[f_1] \leq c_1, \dots, \mathbb{E}_\mu[f_m] \leq c_m\}$$

can be thought of as a polytope of probability measures.

One operation on probability measures that must frequently be performed in UQ applications is conditioning, i.e. forming a new probability measure $\mu(\cdot|B)$ out of an old one μ by restricting attention to subsets of a measurable set B . Conditioning is the operation of supposing that B has happened, and examining the consequently updated probabilities for other measurable events.

Definition 2.9. If $(\Theta, \mathcal{F}, \mu)$ is a probability space and $B \in \mathcal{F}$ has $\mu(B) > 0$, then the *conditional probability measure* $\mu(\cdot|B)$ on (Θ, \mathcal{F}) is defined by

$$\mu(E|B) := \frac{\mu(E \cap B)}{\mu(B)} \quad \text{for } E \in \mathcal{F}.$$

The following theorem on conditional probabilities is fundamental to subjective (Bayesian) probability and statistics (q.v. Section 2.8):

Theorem 2.10 (Bayes’ rule). *If $(\Theta, \mathcal{F}, \mu)$ is a probability space and $A, B \in \mathcal{F}$ have $\mu(A), \mu(B) > 0$, then*

$$\mu(A|B) = \frac{\mu(B|A)\mu(A)}{\mu(B)}.$$

Both the definition of conditional probability and Bayes’ rule can be extended to much more general contexts (including cases in which $\mu(B) = 0$) using advanced tools such as regular conditional probabilities and the disintegration theorem. In Bayesian settings, $\mu(A)$ represents the ‘prior’ probability of some event A , and $\mu(A|B)$ its ‘posterior’ probability, having observed some additional data B .

2.2 Random Variables and Stochastic Processes

Definition 2.11. Let $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$ be measurable spaces. A function $f: \mathcal{X} \rightarrow \mathcal{Y}$ generates a σ -algebra on \mathcal{X} by

$$\sigma(f) := \sigma(\{[f \in E] \mid E \in \mathcal{G}\}),$$

and f is called a *measurable function* if $\sigma(f) \subseteq \mathcal{F}$. That is, f is measurable if the pre-image $f^{-1}(E)$ of every \mathcal{G} -measurable subset E of \mathcal{Y} is an \mathcal{F} -measurable subset of \mathcal{X} . A measurable function whose domain is a probability space is usually called a *random variable*.

Remark 2.12. Note that if \mathcal{F} is the power set of \mathcal{X} , or if \mathcal{G} is the trivial σ -algebra $\{\emptyset, \mathcal{Y}\}$, then every function $f: \mathcal{X} \rightarrow \mathcal{Y}$ is measurable. At the opposite extreme, if \mathcal{F} is the trivial σ -algebra $\{\emptyset, \mathcal{X}\}$, then the only measurable functions $f: \mathcal{X} \rightarrow \mathcal{Y}$ are the constant functions. Thus, in some sense, the sizes of the σ -algebras used to define measurability provide a notion of how well- or ill-behaved the measurable functions are.

Definition 2.13. A measurable function $f: \mathcal{X} \rightarrow \mathcal{Y}$ from a measure space $(\mathcal{X}, \mathcal{F}, \mu)$ to a measurable space $(\mathcal{Y}, \mathcal{G})$ defines a measure $f_*\mu$ on $(\mathcal{Y}, \mathcal{G})$, called the *push-forward* of μ by f , by

$$(f_*\mu)(E) := \mu([f \in E]), \quad \text{for } E \in \mathcal{G}.$$

When μ is a probability measure, $f_*\mu$ is called the *distribution* or *law* of the random variable f .

Definition 2.14. Let S be any set and let $(\Theta, \mathcal{F}, \mu)$ be a probability space. A function $U: S \times \Theta \rightarrow \mathcal{X}$ such that each $U(s, \cdot)$ is a random variable is called an \mathcal{X} -valued *stochastic process* on S .

Whereas measurability questions for a single random variable are discussed in terms of a single σ -algebra, measurability questions for stochastic processes are discussed in terms of families of σ -algebras; when the indexing set S is linearly ordered, e.g. by the natural numbers, or by a continuous parameter such as time, these families of σ -algebras are increasing in the following sense:

Definition 2.15. (a) A *filtration* of a σ -algebra \mathcal{F} is a family $\mathcal{F}_\bullet = \{\mathcal{F}_i \mid i \in I\}$ of sub- σ -algebras of \mathcal{F} , indexed by an ordered set I , such that

$$i \leq j \text{ in } I \implies \mathcal{F}_i \subseteq \mathcal{F}_j.$$

(b) The *natural filtration* associated with a stochastic process $U: I \times \Theta \rightarrow \mathcal{X}$ is the filtration \mathcal{F}_\bullet^U defined by

$$\mathcal{F}_i^U := \sigma(\{U(j, \cdot)^{-1}(E) \subseteq \Theta \mid E \subseteq \mathcal{X} \text{ is measurable and } j \leq i\}).$$

(c) A stochastic process U is *adapted* to a filtration \mathcal{F}_\bullet if $\mathcal{F}_i^U \subseteq \mathcal{F}_i$ for each $i \in I$.

Measurability and adaptedness are important properties of stochastic processes, and loosely correspond to certain questions being ‘answerable’ or ‘decidable’ with respect to the information contained in a given σ -algebra. For instance, if the event $[X \in E]$ is not \mathcal{F} -measurable, then it does not even make sense to ask about the probability $\mathbb{P}_\mu[X \in E]$. For another example, suppose that some stream of observed data is modelled as a stochastic process Y , and it is necessary to make some decision $U(t)$ at each time t . It is common sense to require that the decision stochastic process be \mathcal{F}_\bullet^Y -adapted, since the decision $U(t)$ must be made on the basis of the observations $Y(s)$, $s \leq t$, not on observations from any future time.

2.3 Lebesgue Integration

Integration of a measurable function with respect to a (signed or non-negative) measure is referred to as *Lebesgue integration*. Despite the many technical details that must be checked in the construction of the Lebesgue integral, it remains the integral of choice for most mathematical and probabilistic applications because it extends the simple Riemann integral of functions of a single real variable, can handle worse singularities than the Riemann integral, has better convergence properties, and also naturally captures the notion of an expected value in probability theory. The issue of numerical evaluation of integrals — a vital one in UQ applications — will be addressed separately in Chapter 9.

The construction of the Lebesgue integral is accomplished in three steps: first, the integral is defined for simple functions, which are analogous to step functions from elementary calculus, except that their plateaus are not intervals in \mathbb{R} but measurable events in the sample space.

Definition 2.16. Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space. The *indicator function* \mathbb{I}_E of a set $E \in \mathcal{F}$ is the measurable function defined by

$$\mathbb{I}_E(x) := \begin{cases} 1, & \text{if } x \in E \\ 0, & \text{if } x \notin E. \end{cases}$$

A function $f: \mathcal{X} \rightarrow \mathbb{K}$ is called *simple* if

$$f = \sum_{i=1}^n \alpha_i \mathbb{I}_{E_i}$$

for some scalars $\alpha_1, \dots, \alpha_n \in \mathbb{K}$ and some pairwise disjoint measurable sets $E_1, \dots, E_n \in \mathcal{F}$ with $\mu(E_i)$ finite for $i = 1, \dots, n$. The *Lebesgue integral* of a simple function $f := \sum_{i=1}^n \alpha_i \mathbb{I}_{E_i}$ is defined to be

$$\int_{\mathcal{X}} f \, d\mu := \sum_{i=1}^n \alpha_i \mu(E_i).$$

In the second step, the integral of a non-negative measurable function is defined through approximation from below by the integrals of simple functions:

Definition 2.17. Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space and let $f: \mathcal{X} \rightarrow [0, +\infty]$ be a measurable function. The *Lebesgue integral* of f is defined to be

$$\int_{\mathcal{X}} f \, d\mu := \sup \left\{ \int_{\mathcal{X}} \phi \, d\mu \mid \begin{array}{l} \phi: \mathcal{X} \rightarrow \mathbb{R} \text{ is a simple function, and} \\ 0 \leq \phi(x) \leq f(x) \text{ for } \mu\text{-almost all } x \in \mathcal{X} \end{array} \right\}.$$

Finally, the integral of a real- or complex-valued function is defined through integration of positive and negative real and imaginary parts, with care being taken to avoid the undefined expression ‘ $\infty - \infty$ ’:

Definition 2.18. Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space and let $f: \mathcal{X} \rightarrow \mathbb{R}$ be a measurable function. The *Lebesgue integral* of f is defined to be

$$\int_{\mathcal{X}} f \, d\mu := \int_{\mathcal{X}} f_+ \, d\mu - \int_{\mathcal{X}} f_- \, d\mu$$

provided that at least one of the integrals on the right-hand side is finite. The integral of a complex-valued measurable function $f: \mathcal{X} \rightarrow \mathbb{C}$ is defined to be

$$\int_{\mathcal{X}} f \, d\mu := \int_{\mathcal{X}} (\operatorname{Re} f) \, d\mu + i \int_{\mathcal{X}} (\operatorname{Im} f) \, d\mu.$$

The Lebesgue integral satisfies all the natural requirements for a useful notion of integration: integration is a linear function of the integrand, integrals are additive over disjoint domains of integration, and in the case $\mathcal{X} = \mathbb{R}$ every Riemann-integrable function is Lebesgue integrable. However, one of the chief attractions of the Lebesgue integral over other notions of integration is that, subject to a simple domination condition, pointwise convergence of integrands is enough to ensure convergence of integral values:

Theorem 2.19 (Dominated convergence theorem). *Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space and let $f_n: \mathcal{X} \rightarrow \mathbb{K}$ be a measurable function for each $n \in \mathbb{N}$. If $f: \mathcal{X} \rightarrow \mathbb{K}$ is such that $\lim_{n \rightarrow \infty} f_n(x) = f(x)$ for every $x \in \mathcal{X}$ and there is a measurable function $g: \mathcal{X} \rightarrow [0, \infty]$ such that $\int_{\mathcal{X}} |g| d\mu$ is finite and $|f_n(x)| \leq g(x)$ for all $x \in \mathcal{X}$ and all large enough $n \in \mathbb{N}$, then*

$$\int_{\mathcal{X}} f d\mu = \lim_{n \rightarrow \infty} \int_{\mathcal{X}} f_n d\mu.$$

Furthermore, if the measure space is complete, then the conditions on pointwise convergence and pointwise domination of $f_n(x)$ can be relaxed to hold μ -almost everywhere.

As alluded to earlier, the Lebesgue integral is the standard one in probability theory, and is used to define the mean or expected value of a random variable:

Definition 2.20. When $(\Theta, \mathcal{F}, \mu)$ is a probability space and $X: \Theta \rightarrow \mathbb{K}$ is a random variable, it is conventional to write $\mathbb{E}_{\mu}[X]$ for $\int_{\Theta} X(\theta) d\mu(\theta)$ and to call $\mathbb{E}_{\mu}[X]$ the *expected value* or *expectation* of X . Also,

$$\mathbb{V}_{\mu}[X] := \mathbb{E}_{\mu}[|X - \mathbb{E}_{\mu}[X]|^2] \equiv \mathbb{E}_{\mu}[|X|^2] - |\mathbb{E}_{\mu}[X]|^2$$

is called the *variance* of X . If X is a \mathbb{K}^d -valued random variable, then $\mathbb{E}_{\mu}[X]$, if it exists, is an element of \mathbb{K}^d , and

$$C := \mathbb{E}_{\mu}[(X - \mathbb{E}_{\mu}[X])(X - \mathbb{E}_{\mu}[X])^*] \in \mathbb{K}^{d \times d}$$

i.e. $C_{ij} := \mathbb{E}_{\mu}[(X_i - \mathbb{E}_{\mu}[X_i])(\overline{X_j - \mathbb{E}_{\mu}[X_j]})] \in \mathbb{K}$

is the *covariance matrix* of X .

Spaces of Lebesgue-integrable functions are ubiquitous in analysis and probability theory:

Definition 2.21. Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space. For $1 \leq p \leq \infty$, the L^p space (or *Lebesgue space*) is defined by

$$L^p(\mathcal{X}, \mu; \mathbb{K}) := \{f: \mathcal{X} \rightarrow \mathbb{K} \mid f \text{ is measurable and } \|f\|_{L^p(\mu)} \text{ is finite}\}.$$

For $1 \leq p < \infty$, the norm is defined by the integral expression

$$\|f\|_{L^p(\mu)} := \left(\int_{\mathcal{X}} |f(x)|^p d\mu(x) \right)^{1/p}; \quad (2.1)$$

for $p = \infty$, the norm is defined by the essential supremum (cf. Example 2.7)

$$\begin{aligned} \|f\|_{L^\infty(\mu)} &:= \operatorname{ess\,sup}_{x \in \mathcal{X}} |f(x)| \\ &= \inf \{ \|g\|_{\infty} \mid f = g: \mathcal{X} \rightarrow \mathbb{K} \text{ } \mu\text{-almost everywhere} \} \\ &= \inf \{ t \geq 0 \mid |f| \leq t \text{ } \mu\text{-almost everywhere} \}. \end{aligned} \quad (2.2)$$

To be more precise, $L^p(\mathcal{X}, \mu; \mathbb{K})$ is the set of equivalence classes of such functions, where functions that differ only on a set of μ -measure zero are identified.

When $(\Theta, \mathcal{F}, \mu)$ is a probability space, we have the containments

$$1 \leq p \leq q \leq \infty \implies L^p(\Theta, \mu; \mathbb{R}) \supseteq L^q(\Theta, \mu; \mathbb{R}).$$

Thus, random variables in higher-order Lebesgue spaces are ‘better behaved’ than those in lower-order ones. As a simple example of this slogan, the following inequality shows that the L^p -norm of a random variable X provides control on the probability X deviates strongly from its mean value:

Theorem 2.22 (Chebyshev’s inequality). *Let $X \in L^p(\Theta, \mu; \mathbb{K})$, $1 \leq p < \infty$, be a random variable. Then, for all $t \geq 0$,*

$$\mathbb{P}_\mu[|X - \mathbb{E}_\mu[X]| \geq t] \leq t^{-p} \mathbb{E}_\mu[|X|^p]. \quad (2.3)$$

(The case $p = 1$ is also known as Markov’s inequality.) It is natural to ask if (2.3) is the *best* inequality of this type given the stated assumptions on X , and this is a question that will be addressed in Chapter 14, and specifically Example 14.18.

Integration of Vector-Valued Functions. Lebesgue integration of functions that take values in \mathbb{R}^n can be handled componentwise, as indeed was done above for complex-valued integrands. However, many UQ problems concern random fields, i.e. random variables with values in infinite-dimensional spaces of functions. For definiteness, consider a function f defined on a measure space $(\mathcal{X}, \mathcal{F}, \mu)$ taking values in a Banach space \mathcal{V} . There are two ways to proceed, and they are in general inequivalent:

- (a) The *strong integral* or *Bochner integral* of f is defined by integrating simple \mathcal{V} -valued functions as in the construction of the Lebesgue integral, and then defining

$$\int_{\mathcal{X}} f \, d\mu := \lim_{n \rightarrow \infty} \int_{\mathcal{X}} \phi_n \, d\mu$$

whenever $(\phi_n)_{n \in \mathbb{N}}$ is a sequence of simple functions such that the (scalar-valued) Lebesgue integral $\int_{\mathcal{X}} \|f - \phi_n\| \, d\mu$ converges to 0 as $n \rightarrow \infty$. It transpires that f is Bochner integrable if and only if $\|f\|$ is Lebesgue integrable. The Bochner integral satisfies a version of the Dominated Convergence Theorem, but there are some subtleties concerning the Radon–Nikodým theorem.

- (b) The *weak integral* or *Pettis integral* of f is defined using duality: $\int_{\mathcal{X}} f \, d\mu$ is defined to be an element $v \in \mathcal{V}$ such that

$$\langle \ell | v \rangle = \int_{\mathcal{X}} \langle \ell | f(x) \rangle \, d\mu(x) \quad \text{for all } \ell \in \mathcal{V}'.$$

Since this is a weaker integrability criterion, there are naturally more Pettis-integrable functions than Bochner-integrable ones, but the Pettis integral has deficiencies such as the space of Pettis-integrable functions being incomplete, the existence of a Pettis-integrable function $f: [0, 1] \rightarrow \mathcal{V}$ such that $F(t) := \int_{[0, t]} f(\tau) \, d\tau$ is not differentiable (Kadets, 1994), and so on.

2.4 Decomposition and Total Variation of Signed Measures

If a good mental model for a non-negative measure is a distribution of mass, then a good mental model for a signed measure is a distribution of electrical charge. A natural question to ask is whether every distribution of charge can be decomposed into regions of purely positive and purely negative charge, and hence whether it can be written as the difference of two non-negative distributions, with one supported entirely on the positive set and the other on the negative set. The answer is provided by the Hahn and Jordan decomposition theorems.

Definition 2.23. Two non-negative measures μ and ν on a measurable space $(\mathcal{X}, \mathcal{F})$ are said to be *mutually singular*, denoted $\mu \perp \nu$, if there exists $E \in \mathcal{F}$ such that $\mu(E) = \nu(\mathcal{X} \setminus E) = 0$.

Theorem 2.24 (Hahn–Jordan decomposition). *Let μ be a signed measure on a measurable space $(\mathcal{X}, \mathcal{F})$.*

- (a) *Hahn decomposition: there exist sets $P, N \in \mathcal{F}$ such that $P \cup N = \mathcal{X}$, $P \cap N = \emptyset$, and*

$$\begin{aligned} &\text{for all measurable } E \subseteq P, \quad \mu(E) \geq 0, \\ &\text{for all measurable } E \subseteq N, \quad \mu(E) \leq 0. \end{aligned}$$

This decomposition is essentially unique in the sense that if P' and N' also satisfy these conditions, then every measurable subset of the symmetric differences $P \triangle P'$ and $N \triangle N'$ is of μ -measure zero.

- (b) *Jordan decomposition: there are unique mutually singular non-negative measures μ_+ and μ_- on $(\mathcal{X}, \mathcal{F})$, at least one of which is a finite measure, such that $\mu = \mu_+ - \mu_-$; indeed, for all $E \in \mathcal{F}$,*

$$\begin{aligned} \mu_+(E) &= \mu(E \cap P), \\ \mu_-(E) &= -\mu(E \cap N). \end{aligned}$$

From a probabilistic perspective, the main importance of signed measures and their Hahn and Jordan decompositions is that they provide a useful notion of distance between probability measures:

Definition 2.25. Let μ be a signed measure on a measurable space $(\mathcal{X}, \mathcal{F})$, with Jordan decomposition $\mu = \mu_+ - \mu_-$. The associated *total variation measure* is the non-negative measure $|\mu| := \mu_+ + \mu_-$. The *total variation* of μ is $\|\mu\|_{\text{TV}} := |\mu|(\mathcal{X})$.

Remark 2.26. (a) As the notation $\|\mu\|_{\text{TV}}$ suggests, $\|\cdot\|_{\text{TV}}$ is a norm on the space $\mathcal{M}_{\pm}(\mathcal{X}, \mathcal{F})$ of signed measures on $(\mathcal{X}, \mathcal{F})$.

- (b) The total variation measure can be equivalently defined using measurable partitions:

$$|\mu|(E) = \sup \left\{ \sum_{i=1}^n |\mu(E_i)| \mid \begin{array}{l} n \in \mathbb{N}_0, E_1, \dots, E_n \in \mathcal{F}, \\ \text{and } E = E_1 \cup \dots \cup E_n \end{array} \right\}.$$

- (c) The total variation distance between two probability measures μ and ν (i.e. the total variation norm of their difference) can thus be characterized as

$$d_{\text{TV}}(\mu, \nu) \equiv \|\mu - \nu\|_{\text{TV}} = 2 \sup \{ |\mu(E) - \nu(E)| \mid E \in \mathcal{F} \}, \quad (2.4)$$

i.e. twice the greatest absolute difference in the two probability values that μ and ν assign to any measurable event E .

2.5 The Radon–Nikodým Theorem and Densities

Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space and let $\rho: \mathcal{X} \rightarrow [0, +\infty]$ be a measurable function. The operation

$$\nu: E \mapsto \int_E \rho(x) \, d\mu(x) \quad (2.5)$$

defines a measure ν on $(\mathcal{X}, \mathcal{F})$. It is natural to ask whether every measure ν on $(\mathcal{X}, \mathcal{F})$ can be expressed in this way. A moment's thought reveals that the answer, in general, is no: there is no such function ρ that will make (2.5) hold when μ and ν are Lebesgue measure and a unit Dirac measure (or vice versa) on \mathbb{R} .

Definition 2.27. Let μ and ν be measures on a measurable space $(\mathcal{X}, \mathcal{F})$. If, for $E \in \mathcal{F}$, $\nu(E) = 0$ whenever $\mu(E) = 0$, then ν is said to be *absolutely continuous* with respect to μ , denoted $\nu \ll \mu$. If $\nu \ll \mu \ll \nu$, then μ and ν are said to be *equivalent*, and this is denoted $\mu \approx \nu$.

Definition 2.28. A measure space $(\mathcal{X}, \mathcal{F}, \mu)$ is said to be σ -finite if \mathcal{X} can be expressed as a countable union of \mathcal{F} -measurable sets, each of finite μ -measure.

Theorem 2.29 (Radon–Nikodým). *Suppose that μ and ν are σ -finite measures on a measurable space $(\mathcal{X}, \mathcal{F})$ and that $\nu \ll \mu$. Then there exists a measurable function $\rho: \mathcal{X} \rightarrow [0, \infty]$ such that, for all measurable functions $f: \mathcal{X} \rightarrow \mathbb{R}$ and all $E \in \mathcal{F}$,*

$$\int_E f \, d\nu = \int_E f \rho \, d\mu$$

whenever either integral exists. Furthermore, any two functions ρ with this property are equal μ -almost everywhere.

The function ρ in the Radon–Nikodým theorem is called the *Radon–Nikodým derivative* of ν with respect to μ , and the suggestive notation $\rho = \frac{d\nu}{d\mu}$ is often used. In probability theory, when ν is a probability measure, $\frac{d\nu}{d\mu}$ is called the *probability density function* (PDF) of ν (or any ν -distributed random variable) with respect to μ . Radon–Nikodým derivatives behave very much like the derivatives of elementary calculus:

Theorem 2.30 (Chain rule). *Suppose that μ , ν and π are σ -finite measures on a measurable space $(\mathcal{X}, \mathcal{F})$ and that $\pi \ll \nu \ll \mu$. Then $\pi \ll \mu$ and*

$$\frac{d\pi}{d\mu} = \frac{d\pi}{d\nu} \frac{d\nu}{d\mu} \quad \mu\text{-almost everywhere.}$$

Remark 2.31. The Radon–Nikodým theorem also holds for a signed measure ν and a non-negative measure μ , but in this case the absolute continuity condition is that the total variation measure $|\nu|$ satisfies $|\nu| \ll \mu$, and of course the density ρ is no longer required to be a non-negative function.

2.6 Product Measures and Independence

The previous section considered one way of making new measures from old ones, namely by re-weighting them using a locally integrable density function. By way of contrast, this section considers another way of making new measures from old, namely forming a product measure. Geometrically speaking, the product of two measures is analogous to ‘area’ as the product of two ‘length’ measures. Products of measures also arise naturally in probability theory, since they are the distributions of mutually independent random variables.

Definition 2.32. Let $(\Theta, \mathcal{F}, \mu)$ be a probability space.

- (a) Two measurable sets (events) $E_1, E_2 \in \mathcal{F}$ are said to be *independent* if $\mu(E_1 \cap E_2) = \mu(E_1)\mu(E_2)$.
- (b) Two sub- σ -algebras \mathcal{G}_1 and \mathcal{G}_2 of \mathcal{F} are said to be *independent* if E_1 and E_2 are independent events whenever $E_1 \in \mathcal{G}_1$ and $E_2 \in \mathcal{G}_2$.
- (c) Two measurable functions (random variables) $X: \Theta \rightarrow \mathcal{X}$ and $Y: \Theta \rightarrow \mathcal{Y}$ are said to be *independent* if the σ -algebras generated by X and Y are independent.

Definition 2.33. Let $(\mathcal{X}, \mathcal{F}, \mu)$ and $(\mathcal{Y}, \mathcal{G}, \nu)$ be σ -finite measure spaces. The *product σ -algebra* $\mathcal{F} \otimes \mathcal{G}$ is the σ -algebra on $\mathcal{X} \times \mathcal{Y}$ that is generated by the measurable rectangles, i.e. the smallest σ -algebra for which all the products

$$F \times G, \quad F \in \mathcal{F}, G \in \mathcal{G},$$

are measurable sets. The *product measure* $\mu \otimes \nu: \mathcal{F} \otimes \mathcal{G} \rightarrow [0, +\infty]$ is the measure such that

$$(\mu \otimes \nu)(F \times G) = \mu(F)\nu(G), \quad \text{for all } F \in \mathcal{F}, G \in \mathcal{G}.$$

In the other direction, given a measure on a product space, we can consider the measures induced on the factor spaces:

Definition 2.34. Let $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}, \mu)$ be a measure space and suppose that the factor space \mathcal{X} is equipped with a σ -algebra such that the projections $\Pi_{\mathcal{X}}: (x, y) \mapsto x$ is a measurable function. Then the *marginal measure* $\mu_{\mathcal{X}}$ is the measure on \mathcal{X} defined by

$$\mu_{\mathcal{X}}(E) := ((\Pi_{\mathcal{X}})_* \mu)(E) = \mu(E \times \mathcal{Y}).$$

The marginal measure $\mu_{\mathcal{Y}}$ on \mathcal{Y} is defined similarly.

Theorem 2.35. Let $X = (X_1, X_2)$ be a random variable taking values in a product space $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$. Let μ be the (joint) distribution of X , and μ_i the (marginal) distribution of X_i for $i = 1, 2$. Then X_1 and X_2 are independent random variables if and only if $\mu = \mu_1 \otimes \mu_2$.

The important property of integration with respect to a product measure, and hence taking expected values of independent random variables, is that it can be performed by iterated integration:

Theorem 2.36 (Fubini–Tonelli). Let $(\mathcal{X}, \mathcal{F}, \mu)$ and $(\mathcal{Y}, \mathcal{G}, \nu)$ be σ -finite measure spaces, and let $f: \mathcal{X} \times \mathcal{Y} \rightarrow [0, +\infty]$ be measurable. Then, of the following three integrals, if one exists in $[0, \infty]$, then all three exist and are equal:

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} f(x, y) d\nu(y) d\mu(x), \quad \int_{\mathcal{Y}} \int_{\mathcal{X}} f(x, y) d\mu(x) d\nu(y),$$

$$\text{and } \int_{\mathcal{X} \times \mathcal{Y}} f(x, y) d(\mu \otimes \nu)(x, y).$$

Infinite product measures (or, put another way, infinite sequences of independent random variables) have some interesting extreme properties. Informally, the following result says that any property of a sequence of independent random variables that is independent of any finite subcollection (i.e. depends only on the ‘infinite tail’ of the sequence) must be almost surely true or almost surely false:

Theorem 2.37 (Kolmogorov zero-one law). Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of independent random variables defined over a probability space $(\Theta, \mathcal{F}, \mu)$, and let $\mathcal{F}_n := \sigma(X_n)$. For each $n \in \mathbb{N}$, let $\mathcal{G}_n := \sigma(\bigcup_{k \geq n} \mathcal{F}_k)$, and let

$$\mathcal{T} := \bigcap_{n \in \mathbb{N}} \mathcal{G}_n = \bigcap_{n \in \mathbb{N}} \sigma(X_n, X_{n+1}, \dots) \subseteq \mathcal{F}$$

be the so-called tail σ -algebra. Then, for every $E \in \mathcal{T}$, $\mu(E) \in \{0, 1\}$.

Thus, for example, it is impossible to have a sequence of real-valued random variables $(X_n)_{n \in \mathbb{N}}$ such that $\lim_{n \rightarrow \infty} X_n$ exists with probability $\frac{1}{2}$; either the sequence converges with probability one, or else with probability one it has no limit at all. There are many other zero-one laws in probability and statistics: one that will come up later in the study of Monte Carlo averages is Kesten’s theorem (Theorem 9.17).

2.7 Gaussian Measures

An important class of probability measures and random variables is the class of Gaussians, also known as normal distributions. For many practical problems, especially those that are linear or nearly so, Gaussian measures can serve as appropriate descriptions of uncertainty; even in the nonlinear situation, the Gaussian picture can be an appropriate approximation, though not always. In either case, a significant attraction of Gaussian measures is that many operations on them (e.g. conditioning) can be performed using elementary linear algebra.

On a theoretical level, Gaussian measures are particularly important because, unlike Lebesgue measure, they are well defined on infinite-dimensional spaces, such as function spaces. In \mathbb{R}^d , Lebesgue measure is characterized up to normalization as the unique Borel measure that is simultaneously

- locally finite, i.e. every point of \mathbb{R}^d has an open neighbourhood of finite Lebesgue measure;
- strictly positive, i.e. every open subset of \mathbb{R}^d has strictly positive Lebesgue measure; and
- translation invariant, i.e. $\lambda(x + E) = \lambda(E)$ for all $x \in \mathbb{R}^d$ and measurable $E \subseteq \mathbb{R}^d$.

In addition, Lebesgue measure is σ -finite. However, the following theorem shows that there can be nothing like an infinite-dimensional Lebesgue measure:

Theorem 2.38. *Let μ be a Borel measure on an infinite-dimensional Banach space \mathcal{V} , and, for $v \in \mathcal{V}$, let $T_v: \mathcal{V} \rightarrow \mathcal{V}$ be the translation map $T_v(x) := v + x$.*

- (a) *If μ is locally finite and invariant under all translations, then μ is the trivial (zero) measure.*
- (b) *If μ is σ -finite and quasi-invariant under all translations (i.e. $(T_v)_*\mu$ is equivalent to μ), then μ is the trivial (zero) measure.*

Gaussian measures on \mathbb{R}^d are defined using a Radon–Nikodým derivative with respect to Lebesgue measure. To save space, when P is a self-adjoint and positive-definite matrix or operator on a Hilbert space (see Section 3.3), write

$$\begin{aligned}\langle x, y \rangle_P &:= \langle x, Py \rangle \equiv \langle P^{1/2}x, P^{1/2}y \rangle, \\ \|x\|_P &:= \sqrt{\langle x, x \rangle_P} \equiv \|P^{1/2}x\|\end{aligned}$$

for the new inner product and norm induced by P .

Definition 2.39. Let $m \in \mathbb{R}^d$ and let $C \in \mathbb{R}^{d \times d}$ be symmetric and positive definite. The Gaussian measure with mean m and covariance C is denoted $\mathcal{N}(m, C)$ and defined by

$$\begin{aligned}\mathcal{N}(m, C)(E) &:= \frac{1}{\sqrt{\det C} \sqrt{2\pi}^d} \int_E \exp\left(-\frac{(x-m) \cdot C^{-1}(x-m)}{2}\right) dx \\ &:= \frac{1}{\sqrt{\det C} \sqrt{2\pi}^d} \int_E \exp\left(-\frac{1}{2}\|x-m\|_{C^{-1}}^2\right) dx\end{aligned}$$

for each measurable set $E \subseteq \mathbb{R}^d$. The Gaussian measure $\gamma := \mathcal{N}(0, I)$ is called the *standard Gaussian measure*. A Dirac measure δ_m can be considered as a degenerate Gaussian measure on \mathbb{R} , one with variance equal to zero.

A non-degenerate Gaussian measure is a strictly positive probability measure on \mathbb{R}^d , i.e. it assigns strictly positive mass to every open subset of \mathbb{R}^d ; however, unlike Lebesgue measure, it is not translation invariant:

Lemma 2.40 (Cameron–Martin formula). *Let $\mu = \mathcal{N}(m, C)$ be a Gaussian measure on \mathbb{R}^d . Then the push-forward $(T_v)_*\mu$ of μ by translation by any $v \in \mathbb{R}^d$, i.e. $\mathcal{N}(m+v, C)$, is equivalent to $\mathcal{N}(m, C)$ and*

$$\frac{d(T_v)_*\mu}{d\mu}(x) = \exp\left(\langle v, x-m \rangle_{C^{-1}} - \frac{1}{2}\|v\|_{C^{-1}}^2\right),$$

i.e., for every integrable function f ,

$$\int_{\mathbb{R}^d} f(x+v) d\mu(x) = \int_{\mathbb{R}^d} f(x) \exp\left(\langle v, x-m \rangle_{C^{-1}} - \frac{1}{2}\|v\|_{C^{-1}}^2\right) d\mu(x).$$

It is easily verified that the push-forward of $\mathcal{N}(m, C)$ by any linear functional $\ell: \mathbb{R}^d \rightarrow \mathbb{R}$ is a Gaussian measure on \mathbb{R} , and this is taken as the defining property of a general Gaussian measure for settings in which, by Theorem 2.38, there may not be a Lebesgue measure with respect to which densities can be taken:

Definition 2.41. A Borel measure μ on a normed vector space \mathcal{V} is said to be a (*non-degenerate*) *Gaussian measure* if, for every continuous linear functional $\ell: \mathcal{V} \rightarrow \mathbb{R}$, the push-forward measure $\ell_*\mu$ is a (non-degenerate) Gaussian measure on \mathbb{R} . Equivalently, μ is Gaussian if, for every linear map $T: \mathcal{V} \rightarrow \mathbb{R}^d$, $T_*\mu = \mathcal{N}(m_T, C_T)$ for some $m_T \in \mathbb{R}^d$ and some symmetric positive-definite $C_T \in \mathbb{R}^{d \times d}$.

Definition 2.42. Let μ be a probability measure on a Banach space \mathcal{V} . An element $m_\mu \in \mathcal{V}$ is called the *mean* of μ if

$$\int_{\mathcal{V}} \langle \ell | x - m_\mu \rangle d\mu(x) = 0 \text{ for all } \ell \in \mathcal{V}',$$

so that $\int_{\mathcal{V}} x d\mu(x) = m_\mu$ in the sense of a Pettis integral. If $m_\mu = 0$, then μ is said to be *centred*. The *covariance operator* is the self-adjoint (i.e. conjugate-symmetric) operator $C_\mu: \mathcal{V}' \times \mathcal{V}' \rightarrow \mathbb{K}$ defined by

$$C_\mu(k, \ell) = \int_{\mathcal{V}} \langle k | x - m_\mu \rangle \overline{\langle \ell | x - m_\mu \rangle} d\mu(x) \text{ for all } k, \ell \in \mathcal{V}'.$$

We often abuse notation and write $C_\mu: \mathcal{V}' \rightarrow \mathcal{V}''$ for the operator defined by

$$\langle C_\mu k | \ell \rangle := C_\mu(k, \ell)$$

In the case that $\mathcal{V} = \mathcal{H}$ is a Hilbert space, it is usual to employ the Riesz representation theorem to identify \mathcal{H} with \mathcal{H}' and \mathcal{H}'' and hence treat C_μ as a linear operator from \mathcal{H} into itself. The inverse of C_μ , if it exists, is called the *precision operator* of μ .

The covariance operator of a Gaussian measure is closely connected to its non-degeneracy:

Theorem 2.43 (Vakhania, 1975). *Let μ be a Gaussian measure on a separable, reflexive Banach space \mathcal{V} with mean $m_\mu \in \mathcal{V}$ and covariance operator $C_\mu: \mathcal{V}' \rightarrow \mathcal{V}$. Then the support of μ is the affine subspace of \mathcal{V} that is the translation by the mean of the closure of the range of the covariance operator, i.e.*

$$\text{supp}(\mu) = m_\mu + \overline{C_\mu \mathcal{V}'}.$$

Corollary 2.44. *For a Gaussian measure μ on a separable, reflexive Banach space \mathcal{V} , the following are equivalent:*

- (a) μ is non-degenerate;
- (b) $C_\mu: \mathcal{V}' \rightarrow \mathcal{V}$ is one-to-one;
- (c) $\overline{C_\mu \mathcal{V}'} = \mathcal{V}$.

Example 2.45. Consider a Gaussian random variable $X = (X_1, X_2) \sim \mu$ taking values in \mathbb{R}^2 . Suppose that the mean and covariance of X (or, equivalently, μ) are, in the usual basis of \mathbb{R}^2 ,

$$m = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad C = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Then $X = (Z, 1)$, where $Z \sim \mathcal{N}(0, 1)$ is a standard Gaussian random variable on \mathbb{R} ; the values of X all lie on the affine line $L := \{(x_1, x_2) \in \mathbb{R}^2 \mid x_2 = 1\}$. Indeed, Vakhania's theorem says that

$$\text{supp}(\mu) = m + \overline{C(\mathbb{R}^2)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \left\{ \begin{bmatrix} x_1 \\ 0 \end{bmatrix} \mid x_1 \in \mathbb{R} \right\} = L.$$

Gaussian measures can also be identified by reference to their Fourier transforms:

Theorem 2.46. *A probability measure μ on \mathcal{V} is a Gaussian measure if and only if its Fourier transform $\hat{\mu}: \mathcal{V}' \rightarrow \mathbb{C}$ satisfies*

$$\hat{\mu}(\ell) := \int_{\mathcal{V}} e^{i\langle \ell | x \rangle} d\mu(x) = \exp\left(i\langle \ell | m \rangle - \frac{Q(\ell)}{2}\right) \quad \text{for all } \ell \in \mathcal{V}'.$$

for some $m \in \mathcal{V}$ and some positive-definite quadratic form Q on \mathcal{V}' . Indeed, m is the mean of μ and $Q(\ell) = C_{\mu}(\ell, \ell)$. Furthermore, if two Gaussian measures μ and ν have the same mean and covariance operator, then $\mu = \nu$.

Not only does a Gaussian measure have a well-defined mean and variance, it in fact has moments of all orders:

Theorem 2.47 (Fernique, 1970). *Let μ be a centred Gaussian measure on a separable Banach space \mathcal{V} . Then there exists $\alpha > 0$ such that*

$$\int_{\mathcal{V}} \exp(\alpha \|x\|^2) d\mu(x) < +\infty.$$

A fortiori, μ has moments of all orders: for all $k \geq 0$,

$$\int_{\mathcal{V}} \|x\|^k d\mu(x) < +\infty.$$

The covariance operator of a Gaussian measure on a Hilbert space \mathcal{H} is a self-adjoint operator from \mathcal{H} into itself. A classification of exactly which self-adjoint operators on \mathcal{H} can be Gaussian covariance operators is provided by the next result, Sazonov's theorem:

Definition 2.48. Let $K: \mathcal{H} \rightarrow \mathcal{H}$ be a linear operator on a separable Hilbert space \mathcal{H} .

- (a) K is said to be *compact* if it has a singular value decomposition, i.e. if there exist finite or countably infinite orthonormal sequences (u_n) and (v_n) in \mathcal{H} and a sequence of non-negative reals (σ_n) such that

$$K = \sum_n \sigma_n \langle v_n, \cdot \rangle u_n,$$

with $\lim_{n \rightarrow \infty} \sigma_n = 0$ if the sequences are infinite.

- (b) K is said to be *trace class* or *nuclear* if $\sum_n \sigma_n$ is finite, and *Hilbert–Schmidt* or *nuclear of order 2* if $\sum_n \sigma_n^2$ is finite.
(c) If K is trace class, then its *trace* is defined to be

$$\text{tr}(K) := \sum_n \langle e_n, K e_n \rangle$$

for any orthonormal basis (e_n) of \mathcal{H} , and (by Lidskiĭ's theorem) this equals the sum of the eigenvalues of K , counted with multiplicity.

Theorem 2.49 (Sazonov, 1958). *Let μ be a centred Gaussian measure on a separable Hilbert space \mathcal{H} . Then $C_{\mu}: \mathcal{H} \rightarrow \mathcal{H}$ is trace class and*

$$\text{tr}(C_{\mu}) = \int_{\mathcal{H}} \|x\|^2 d\mu(x).$$

Conversely, if $K: \mathcal{H} \rightarrow \mathcal{H}$ is positive, self-adjoint and of trace class, then there is a Gaussian measure μ on \mathcal{H} such that $C_{\mu} = K$.

Sazonov's theorem is often stated in terms of the square root $C_\mu^{1/2}$ of C_μ : $C_\mu^{1/2}$ is Hilbert–Schmidt, i.e. has square-summable singular values $(\sigma_n)_{n \in \mathbb{N}}$.

As noted above, even finite-dimensional Gaussian measures are not invariant under translations, and the change-of-measure formula is given by Lemma 2.40. In the infinite-dimensional setting, it is not even true that translation produces a new measure that has a density with respect to the old one. This phenomenon leads to an important object associated with any Gaussian measure, its Cameron–Martin space:

Definition 2.50. Let $\mu = \mathcal{N}(m, C)$ be a Gaussian measure on a Banach space \mathcal{V} . The *Cameron–Martin space* is the Hilbert space \mathcal{H}_μ defined equivalently by:

- \mathcal{H}_μ is the completion of

$$\{h \in \mathcal{V} \mid \text{for some } h^* \in \mathcal{V}', C(h^*, \cdot) = \langle \cdot, h \rangle\}$$


with respect to the inner product $\langle h, k \rangle_\mu := C(h^*, k^*)$.

- \mathcal{H}_μ is the completion of the range of the covariance operator $C: \mathcal{V}' \rightarrow \mathcal{V}$ with respect to this inner product (cf. the closure with respect to the norm in \mathcal{V} in Theorem 2.43).
- If \mathcal{V} is Hilbert, then \mathcal{H}_μ is the completion of $\text{ran } C^{1/2}$ with the inner product $\langle h, k \rangle_{C^{-1}} := \langle C^{-1/2}h, C^{-1/2}k \rangle_{\mathcal{V}}$.
- \mathcal{H}_μ is the set of all $v \in \mathcal{V}$ such that $(T_v)_*\mu \approx \mu$, with

$$\frac{d(T_v)_*\mu}{d\mu}(x) = \exp \left(\langle v, x \rangle_{C^{-1}} - \frac{\|v\|_{C^{-1}}^2}{2} \right)$$

as in Lemma 2.40.

- \mathcal{H}_μ is the intersection of all linear subspaces of \mathcal{V} that have full μ -measure.

By Theorem 2.38, if μ is any probability measure (Gaussian or otherwise) on an infinite-dimensional space \mathcal{V} , then we certainly cannot have $\mathcal{H}_\mu = \mathcal{V}$. In fact, one should think of \mathcal{H}_μ as being a very small subspace of \mathcal{V} : if \mathcal{H}_μ is infinite dimensional, then $\mu(\mathcal{H}_\mu) = 0$. Also, infinite-dimensional spaces have the extreme property that Gaussian measures on such spaces are either equivalent or mutually singular — there is no middle ground in the way that Lebesgue measure on $[0, 1]$ has a density with respect to Lebesgue measure on \mathbb{R} but is not equivalent to it. 

Theorem 2.51 (Feldman–Hájek). *Let μ, ν be Gaussian probability measures on a normed vector space \mathcal{V} . Then either*

- μ and ν are equivalent, i.e. $\mu(E) = 0 \iff \nu(E) = 0$, and hence each has a strictly positive density with respect to the other; or
- μ and ν are mutually singular, i.e. there exists E such that $\mu(E) = 0$ and $\nu(E) = 1$, and so neither μ nor ν can have a density with respect to the other.

Furthermore, equivalence holds if and only if

- (a) $\text{ran } C_\mu^{1/2} = \text{ran } C_\nu^{1/2}$;
- (b) $m_\mu - m_\nu \in \text{ran } C_\mu^{1/2} = \text{ran } C_\nu^{1/2}$; and
- (c) $T := (C_\mu^{-1/2} C_\nu^{1/2})(C_\mu^{-1/2} C_\nu^{1/2})^* - I$ is Hilbert–Schmidt in $\text{ran } C_\mu^{1/2}$.

The Cameron–Martin and Feldman–Hájek theorems show that translation by any vector not in the Cameron–Martin space $\mathcal{H}_\mu \subseteq \mathcal{V}$ produces a new measure that is mutually singular with respect to the old one. It turns out that dilation by a non-unitary constant also destroys equivalence:

Proposition 2.52. *Let μ be a centred Gaussian measure on a separable real Banach space \mathcal{V} such that $\dim \mathcal{H}_\mu = \infty$. For $c \in \mathbb{R}$, let $D_c: \mathcal{V} \rightarrow \mathcal{V}$ be the dilation map $D_c(x) := cx$. Then $(D_c)_*\mu$ is equivalent to μ if and only if $c \in \{\pm 1\}$, and $(D_c)_*\mu$ and μ are mutually singular otherwise.*

Remark 2.53. There is another attractive viewpoint on Gaussian measures on Hilbert spaces, namely that draws from a Gaussian measure $\mathcal{N}(m, C)$ on a Hilbert space are the same as draws from random series of the form

$$m + \sum_{k \in \mathbb{N}} \sqrt{\lambda_k} \xi_k \psi_k,$$

where $\{\psi_k\}_{k \in \mathbb{N}}$ are orthonormal eigenvectors for the covariance operator C , $\{\lambda_k\}_{k \in \mathbb{N}}$ are the corresponding eigenvalues, and $\{\xi_k\}_{k \in \mathbb{N}}$ are independent draws from the standard normal distribution $\mathcal{N}(0, 1)$ on \mathbb{R} . This point of view will be revisited in more detail in Section 11.1 in the context of Karhunen–Loève expansions of Gaussian and Besov measures.

The conditioning properties of Gaussian measures can easily be expressed using an elementary construction from linear algebra, the Schur complement. This result will be very useful in Chapters 6, 7, and 13.

Theorem 2.54 (Conditioning of Gaussian measures). *Let $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ be a direct sum of separable Hilbert spaces. Let $X = (X_1, X_2) \sim \mu$ be an \mathcal{H} -valued Gaussian random variable with mean $m = (m_1, m_2)$ and positive-definite covariance operator C . For $i, j = 1, 2$, let*

$$C_{ij}(k_i, k_j) := \mathbb{E}_\mu \left[\langle k_i, x - m_i \rangle \overline{\langle k_j, x - m_j \rangle} \right] \quad (2.6)$$

for all $k_i \in \mathcal{H}_i$, $k_j \in \mathcal{H}_j$, so that C is decomposed^[2.2] in block form as

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}; \quad (2.7)$$

in particular, the marginal distribution of X_i is $\mathcal{N}(m_i, C_{ii})$, and $C_{21} = C_{12}^*$. Then C_{22} is invertible and, for each $x_2 \in \mathcal{H}_2$, the conditional distribution of X_1 given $X_2 = x_2$ is Gaussian:

$$(X_1 | X_2 = x_2) \sim \mathcal{N}(m_1 + C_{12} C_{22}^{-1} (x_2 - m_2), C_{11} - C_{12} C_{22}^{-1} C_{21}). \quad (2.8)$$

2.8 Interpretations of Probability

It is worth noting that the above discussions are purely mathematical: a probability measure is an abstract algebraic–analytic object with no necessary connection to everyday notions of chance or probability. The question of what *interpretation* of probability to adopt, i.e. what practical meaning to ascribe to probability measures, is a question of philosophy and mathematical modelling. The two main points of view are the *frequentist* and *Bayesian* perspectives. To a frequentist, the probability $\mu(E)$ of an event E is the relative frequency of occurrence of the event E in the limit of infinitely many independent but identical trials; to a Bayesian, $\mu(E)$ is a numerical representation of one’s degree of belief in the truth of a proposition E . The frequentist’s point of view is *objective*; the Bayesian’s is *subjective*; both use the same mathematical machinery of probability measures to describe the properties of the function μ .

Frequentists are careful to distinguish between parts of their analyses that are fixed and deterministic versus those that have a probabilistic character. However, for a Bayesian, *any* uncertainty can be described in terms of a suitable probability measure. In particular, one’s beliefs about some unknown θ (taking values in a space Θ) in advance of observing data are summarized by a *prior* probability measure π on Θ . The other ingredient of a Bayesian analysis is a *likelihood function*, which is up to normalization a conditional probability: given any observed datum y , $L(y|\theta)$ is the likelihood of observing y if the parameter value θ were

^[2.2]Here we are again abusing notation to conflate $C_{ij}: \mathcal{H}_i \oplus \mathcal{H}_j \rightarrow \mathbb{K}$ defined in (2.6) with $C_{ij}: \mathcal{H}_j \rightarrow \mathcal{H}_i$ given by $\langle C_{ij}(k_j), k_i \rangle_{\mathcal{H}_i} = C_{ij}(k_i, k_j)$.

the truth. A Bayesian's belief about θ given the prior π and the observed datum y is the *posterior* probability measure $\pi(\cdot|y)$ on Θ , which is just the conditional probability

$$\pi(\theta|y) = \frac{L(y|\theta)\pi(\theta)}{\mathbb{E}_\pi[L(y|\theta)]} = \frac{L(y|\theta)\pi(\theta)}{\int_\Theta L(y|\theta) d\pi(\theta)}$$

or, written in a way that generalizes better to infinite-dimensional Θ , we have a density / Radon–Nikodým derivative

$$\frac{d\pi(\cdot|y)}{d\pi}(\theta) \propto L(y|\theta).$$

Both the previous two equations are referred to as *Bayes' rule*, and are at this stage informal applications of the standard Bayes' rule (Theorem 2.10) for events A and B of non-zero probability.

Example 2.55. Parameter estimation provides a good example of the philosophical difference between frequentist and subjectivist uses of probability. Suppose that X_1, \dots, X_n are n independent and identically distributed observations of some random variable X , which is distributed according to the normal distribution $\mathcal{N}(\theta, 1)$ of mean θ and variance 1. We set our frequentist and Bayesian statisticians the challenge of estimating θ from the data $d := (X_1, \dots, X_n)$.

- (a) To the frequentist, θ is a well-defined *real number* that happens to be unknown. This number can be estimated using the estimator

$$\hat{\theta}_n := \frac{1}{n} \sum_{i=1}^n X_i,$$

which is a random variable. It makes sense to say that $\hat{\theta}_n$ is close to θ with high probability, and hence to give a confidence interval for θ , but θ itself does not have a distribution.

- (b) To the Bayesian, θ is a *random variable*, and its distribution in advance of seeing the data is encoded in a prior π . Upon seeing the data and conditioning upon it using Bayes' rule, the distribution of the parameter is the posterior distribution $\pi(\theta|d)$. The posterior encodes everything that is known about θ in view of π , $L(y|\theta) \propto e^{-|y-\theta|^2/2}$ and d , although this information may be summarized by a single number such as the *maximum a posteriori estimator*

$$\hat{\theta}^{\text{MAP}} := \arg \max_{\theta \in \mathbb{R}} \pi(\theta|d)$$

or the *maximum likelihood estimator*

$$\hat{\theta}^{\text{MLE}} := \arg \max_{\theta \in \mathbb{R}} L(d|\theta).$$

The Bayesian perspective can be seen as the natural extension of classical Aristotelian bivalent (i.e. true-or-false) logic to propositions of uncertain truth value. This point of view is underwritten by *Cox's theorem* (Cox, 1946, 1961), which asserts that any 'natural' extension of Aristotelian logic to \mathbb{R} -valued truth values is probabilistic, and specifically Bayesian, although the 'naturalness' of the hypotheses has been challenged by e.g. Halpern (1999a,b).

It is also worth noting that there is a significant community that, in addition to being frequentist or Bayesian, asserts that selecting a single probability measure is too precise a description of uncertainty. These 'imprecise probabilists' count such distinguished figures as George Boole and John Maynard Keynes among their ranks, and would prefer to say that $\frac{1}{2} - 2^{-100} \leq \mathbb{P}[\text{heads}] \leq \frac{1}{2} + 2^{-100}$ than commit themselves to the assertion that $\mathbb{P}[\text{heads}] = \frac{1}{2}$; imprecise probabilists would argue that the former assertion can be verified, to a prescribed

level of confidence, in finite time, whereas the latter cannot. Techniques like the use of *lower and upper probabilities* (or *interval probabilities*) are popular in this community, including sophisticated generalizations like Dempster–Shafer theory; one can also consider *feasible sets of probability measures*, which is the approach taken in Chapter 14.

2.9 Bibliography

The book of Gordon (1994) is mostly a text on the gauge integral, but its first chapters provide an excellent condensed introduction to measure theory and Lebesgue integration. Capiński and Kopp (2004) is a clear, readable and self-contained introductory text confined mainly to Lebesgue integration on \mathbb{R} (and later \mathbb{R}^n), including material on L^p spaces and the Radon–Nikodým theorem. Another excellent text on measure and probability theory is the monograph of Billingsley (1995). Readers who prefer to learn mathematics through counterexamples rather than theorems may wish to consult the books of Romano and Siegel (1986) and Stoyanov (1987). The disintegration theorem, alluded to at the end of Section 2.1, can be found in Ambrosio et al. (2008, Section 5.3) and Dellacherie and Meyer (1978, Section III-70).

The Bochner integral was introduced by Bochner (1933); recent texts on the topic include those of Diestel and Uhl (1977) and Mikusiński (1978). For detailed treatment of the Pettis integral, see Talagrand (1984). Further discussion of the relationship between tensor products and spaces of vector-valued integrable functions can be found in the book of Ryan (2002).

Bourbaki (2004) contains a treatment of measure theory from a functional-analytic perspective. The presentation is focussed on Radon measures on locally compact spaces, which is advantageous in terms of regularity but leads to an approach to measurable functions that is cumbersome, particularly from the viewpoint of probability theory. All the standard warnings about Bourbaki texts apply: the presentation is comprehensive but often forbiddingly austere, and so it is perhaps better as a reference text than a learning tool.

Chapters 7 and 8 of the book of Smith (2014) compare and contrast the frequentist and Bayesian perspectives on parameter estimation in the context of UQ. The origins of imprecise probability lie in treatises like those of Boole (1854) and Keynes (1921). More recent foundations and expositions for imprecise probability have been put forward by Walley (1991), Kuznetsov (1991), Weichselberger (2000), and by Dempster (1967) and Shafer (1976).

A general introduction to the theory of Gaussian measures is the book of Bogachev (1998); a complementary viewpoint, in terms of Gaussian stochastic processes, is presented by Rasmussen and Williams (2006).

The non-existence of an infinite-dimensional Lebesgue measure, and related results, can be found in the lectures of Yamasaki (1985, Part B, Chapter 1, Section 5). The Feldman–Hájek dichotomy (Theorem 2.51) was proved independently by Feldman (1958) and Hájek (1958), and can also be found in the book of Da Prato and Zabczyk (1992, Theorem 2.23).

2.10 Exercises

Exercise 2.1. Let X be any \mathbb{C}^n -valued random variable with mean $m \in \mathbb{C}^n$ and covariance matrix

$$C := \mathbb{E}[(X - m)(X - m)^*] \in \mathbb{C}^{n \times n}.$$

- Show that C is conjugate-symmetric and positive semi-definite. For what collection of vectors in \mathbb{C}^n is C the Gram matrix?
- Show that if the support of X is all of \mathbb{C}^n , then C is positive definite. Hint: suppose that C has non-trivial kernel, construct an open half-space H of \mathbb{C}^n such that $X \notin H$ almost surely.

Exercise 2.2. Let X be any random variable taking values in a Hilbert space \mathcal{H} , with mean $m \in \mathcal{H}$ and covariance operator $C: \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{C}$ defined by

$$C(h, k) := \mathbb{E} \left[\langle h, X - m \rangle \overline{\langle k, X - m \rangle} \right]$$

for $h, k \in \mathcal{H}$. Show that C is conjugate-symmetric and positive semi-definite. Show also that if there is no subspace $S \subseteq \mathcal{H}$ with $\dim S \geq 1$ such that $X \perp S$ with probability one, then C is positive definite.

Exercise 2.3. Prove the finite-dimensional Cameron–Martin formula of Lemma 2.40. That is, let $\mu = \mathcal{N}(m, C)$ be a Gaussian measure on \mathbb{R}^d and let $v \in \mathbb{R}^d$, and show that the push-forward of μ by translation by v , namely $\mathcal{N}(m + v, C)$, is equivalent to μ and

$$\frac{d(T_v)_*\mu}{d\mu}(x) = \exp \left(\langle v, x - m \rangle_{C^{-1}} - \frac{1}{2} \|v\|_{C^{-1}}^2 \right),$$

i.e., for every integrable function f ,

$$\int_{\mathbb{R}^d} f(x + v) d\mu(x) = \int_{\mathbb{R}^d} f(x) \exp \left(\langle v, x - m \rangle_{C^{-1}} - \frac{1}{2} \|v\|_{C^{-1}}^2 \right) d\mu(x).$$

Exercise 2.4. Let $T: \mathcal{H} \rightarrow \mathcal{K}$ be a bounded linear map between Hilbert spaces \mathcal{H} and \mathcal{K} , with adjoint $T^*: \mathcal{K} \rightarrow \mathcal{H}$, and let $\mu = \mathcal{N}(m, C)$ be a Gaussian measure on \mathcal{H} . Show that the push-forward measure $T_*\mu$ is a Gaussian measure on \mathcal{K} and that $T_*\mu = \mathcal{N}(Tm, TCT^*)$.

Exercise 2.5. For $i = 1, 2$, let $X_i \sim \mathcal{N}(m_i, C_i)$ independent Gaussian random variables taking values in Hilbert spaces \mathcal{H}_i , and let $T_i: \mathcal{H}_i \rightarrow \mathcal{K}$ be a bounded linear map taking values in another Hilbert space \mathcal{K} , with adjoint $T_i^*: \mathcal{K} \rightarrow \mathcal{H}_i$. Show that $T_1X_1 + T_2X_2$ is a Gaussian random variable in \mathcal{K} with

$$T_1X_1 + T_2X_2 \sim \mathcal{N}(T_1m_1 + T_2m_2, T_1C_1T_1^* + T_2C_2T_2^*).$$

Give an example to show that the independence assumption is necessary.

Exercise 2.6. Let \mathcal{H} and \mathcal{K} be Hilbert spaces. Suppose that $A: \mathcal{H} \rightarrow \mathcal{H}$ and $C: \mathcal{K} \rightarrow \mathcal{K}$ are self-adjoint and positive definite, that $B: \mathcal{H} \rightarrow \mathcal{K}$, and that $D: \mathcal{K} \rightarrow \mathcal{K}$ is self-adjoint and positive semi-definite. Show that the operator from $\mathcal{H} \oplus \mathcal{K}$ to itself given in block form by

$$\begin{bmatrix} A + B^*CB & -B^*C \\ -CB & C + D \end{bmatrix}$$

is self-adjoint and positive-definite.

Exercise 2.7 (Inversion lemma). Let \mathcal{H} and \mathcal{K} be Hilbert spaces, and let $A: \mathcal{H} \rightarrow \mathcal{H}$, $B: \mathcal{K} \rightarrow \mathcal{H}$, $C: \mathcal{H} \rightarrow \mathcal{K}$, and $D: \mathcal{K} \rightarrow \mathcal{K}$ be linear maps. Define $M: \mathcal{H} \oplus \mathcal{K} \rightarrow \mathcal{H} \oplus \mathcal{K}$ in block form by

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}.$$

Show that if A , D , $A - BD^{-1}C$ and $D - CA^{-1}B$ are all non-singular, then

$$M^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix}$$

and

$$M^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}.$$

Hence derive the *Woodbury formula*

$$(A + BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1}. \quad (2.9)$$

Exercise 2.8. Exercise 2.7 has a natural interpretation in terms of the conditioning of Gaussian random variables. Let $(X, Y) \sim \mathcal{N}(m, C)$ be jointly Gaussian, where, in block form,

$$m = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \quad C = \begin{bmatrix} C_{11} & C_{12} \\ C_{12}^* & C_{22} \end{bmatrix},$$

and C is self-adjoint and positive definite.

- (a) Show that C_{11} and C_{22} are self-adjoint and positive-definite.
- (b) Show that the Schur complement S defined by $S := C_{11} - C_{12}C_{22}^{-1}C_{12}^*$ is self-adjoint and positive definite, and

$$C^{-1} = \begin{bmatrix} S^{-1} & -S^{-1}C_{12}C_{22}^{-1} \\ -C_{22}^{-1}C_{12}^*S^{-1} & C_{22}^{-1} + C_{22}^{-1}C_{12}^*S^{-1}C_{12}C_{22}^{-1} \end{bmatrix}.$$

- (c) Hence prove Theorem 2.54, that the conditional distribution of X given that $Y = y$ is Gaussian:

$$(X|Y = y) \sim \mathcal{N}(m_1 + C_{12}C_{22}^{-1}(y - m_2), S).$$

Chapter 3

Banach and Hilbert Spaces

Dr. von Neumann, ich möchte gern wissen, was
ist dann eigentlich ein Hilbertscher Raum?

DAVID HILBERT

This chapter covers the necessary concepts from linear functional analysis on Hilbert and Banach spaces: in particular, we review here basic constructions such as orthogonality, direct sums, and tensor products. Like Chapter 2, this chapter is intended as a review of material that should be understood as a prerequisite before proceeding; to an extent, Chapters 2 and 3 are interdependent and so can (and should) be read in parallel with one another.

3.1 Basic Definitions and Properties

In what follows, \mathbb{K} will denote either the real numbers \mathbb{R} or the complex numbers \mathbb{C} , and $|\cdot|$ denotes the absolute value function on \mathbb{K} . All the vector spaces considered in this book will be vector spaces over one of these two fields. In \mathbb{K} , notions of ‘size’ and ‘closeness’ are provided by the absolute value function $|\cdot|$. In a normed vector space, similar notions of ‘size’ and ‘closeness’ are provided by a function called a norm, from which we can build up notions of convergence, continuity, limits, and so on.

Definition 3.1. A *norm* on a vector space \mathcal{V} over \mathbb{K} is a function $\|\cdot\|: \mathcal{V} \rightarrow \mathbb{R}$ that is

- (a) *positive semi-definite*: for all $x \in \mathcal{V}$, $\|x\| \geq 0$;
- (b) *positive definite*: for all $x \in \mathcal{V}$, $\|x\| = 0$ if and only if $x = 0$;
- (c) *positively homogeneous*: for all $x \in \mathcal{V}$ and $\alpha \in \mathbb{K}$, $\|\alpha x\| = |\alpha| \|x\|$; and
- (d) *sublinear*: for all $x, y \in \mathcal{V}$, $\|x + y\| \leq \|x\| + \|y\|$.

If the positive definiteness requirement is omitted, then $\|\cdot\|$ is said to be a *seminorm*. A vector space equipped with a norm (resp. seminorm) is called a *normed space* (resp. *seminormed space*).

In a normed vector space, we can sensibly talk about the ‘size’ or ‘length’ of a single vector, but there is no sensible notion of ‘angle’ between two vectors, and in particular there is no notion of orthogonality. Such notions are provided by an inner product:

Definition 3.2. An *inner product* on a vector space \mathcal{V} over \mathbb{K} is a function $\langle \cdot, \cdot \rangle: \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{K}$ that is

- (a) *positive semi-definite*: for all $x \in \mathcal{V}$, $\langle x, x \rangle \geq 0$;
 - (b) *positive definite*: for all $x \in \mathcal{V}$, $\langle x, x \rangle = 0$ if and only if $x = 0$;
 - (c) *conjugate symmetric*: for all $x, y \in \mathcal{V}$, $\langle x, y \rangle = \overline{\langle y, x \rangle}$; and
-

(d) *sesquilinear*: for all $x, y, z \in \mathcal{V}$ and all $\alpha, \beta \in \mathbb{K}$, $\langle x, \alpha y + \beta z \rangle = \alpha \langle x, y \rangle + \beta \langle x, z \rangle$.

A vector space equipped with an inner product is called an *inner product space*. In the case $\mathbb{K} = \mathbb{R}$, conjugate symmetry becomes symmetry, and sesquilinearity becomes bilinearity.

Many texts have sesquilinear forms be linear in the *first* argument, rather than the second as is done here; this is an entirely cosmetic difference that has no serious consequences, provided that one makes a consistent choice and sticks with it.

It is easily verified that every inner product space is a normed space under the *induced norm*

$$\|x\| := \sqrt{\langle x, x \rangle}.$$

The inner product and norm satisfy the *Cauchy-Schwarz inequality*

$$|\langle x, y \rangle| \leq \|x\| \|y\| \quad \text{for all } x, y \in \mathcal{V}, \quad (3.1)$$

where equality holds in (3.1) if and only if x and y are scalar multiples of one another. Every norm on \mathcal{V} that is induced by an inner product satisfies the *parallelogram identity*

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2 \quad \text{for all } x, y \in \mathcal{V}. \quad (3.2)$$

In the opposite direction, if $\|\cdot\|$ is a norm on \mathcal{V} that satisfies the parallelogram identity (3.2), then the unique inner product $\langle \cdot, \cdot \rangle$ that induces this norm is found by the *polarization identity*

$$\langle x, y \rangle = \frac{\|x + y\|^2 - \|x - y\|^2}{4} \quad (3.3)$$

in the real case, and

$$\langle x, y \rangle = \frac{\|x + y\|^2 - \|x - y\|^2}{4} + i \frac{\|ix - y\|^2 - \|ix + y\|^2}{4} \quad (3.4)$$

in the complex case.

The simplest examples of normed and inner product spaces are the familiar finite-dimensional Euclidean spaces:

Example 3.3. Here are some finite-dimensional examples of norms on \mathbb{K}^n :

- (a) The absolute value function $|\cdot|$ is a norm on \mathbb{K} .
- (b) The most familiar example of a norm is probably the *Euclidean norm* or *2-norm* on \mathbb{K}^n . The Euclidean norm of $v = (v_1, \dots, v_n) \in \mathbb{K}^n$ is given by

$$\|v\|_2 := \sqrt{\sum_{i=1}^n |v_i|^2} = \sqrt{\sum_{i=1}^n |v \cdot e_i|^2}. \quad (3.5)$$

The Euclidean norm is the induced norm for the inner product

$$\langle u, v \rangle := \sum_{i=1}^n \overline{u_i} v_i. \quad (3.6)$$

In the case $\mathbb{K} = \mathbb{R}$ this inner product is commonly called the *dot product* and denoted $u \cdot v$.

- (c) The analogous inner product and norm on $\mathbb{K}^{m \times n}$ of $m \times n$ matrices is the *Frobenius inner product*

$$\langle A, B \rangle \equiv A : B := \sum_{\substack{i=1, \dots, m \\ j=1, \dots, n}} \overline{a_{ij}} b_{ij}.$$

- (d) The *1-norm*, also known as the *Manhattan norm* or *taxicab norm*, on \mathbb{K}^n is defined by

$$\|v\|_1 := \sum_{i=1}^n |v_i|. \quad (3.7)$$

- (e) More generally, for $1 \leq p < \infty$, the *p-norm* on \mathbb{K}^n is defined by

$$\|v\|_p := \left(\sum_{i=1}^n |v_i|^p \right)^{1/p}. \quad (3.8)$$

- (f) Note, however, that the formula in (3.8) does *not* define a norm on \mathbb{K}^n if $p < 1$.
 (g) The analogous norm for $p = \infty$ is the ∞ -norm or *maximum norm* on \mathbb{K}^n :

$$\|v\|_\infty := \max_{i=1,\dots,n} |v_i|. \quad (3.9)$$

There are also many straightforward examples of infinite-dimensional normed spaces. In UQ applications, these spaces often arise as the solution spaces for ordinary or partial differential equations, spaces of random variables, or spaces for sequences of coefficients of expansions of random fields and stochastic processes.

Example 3.4. (a) An obvious norm to define for a sequence $v = (v_n)_{n \in \mathbb{N}}$ is the analogue of the maximum norm. That is, define the *supremum norm* by

$$\|v\|_\infty := \sup_{n \in \mathbb{N}} |v_n|. \quad (3.10)$$

Clearly, if v is not a bounded sequence, then $\|v\|_\infty = \infty$. Since norms are not allowed to take the value ∞ , the supremum norm is only a norm on the space of *bounded sequences*; this space is often denoted ℓ^∞ , or sometimes $\ell^\infty(\mathbb{K})$ if we wish to emphasize the field of scalars, or $\mathcal{B}(\mathbb{N}; \mathbb{K})$ if we want to emphasize that it is a space of bounded functions on some set, in this case \mathbb{N} .

- (b) Similarly, for $1 \leq p < \infty$, the *p-norm* of a sequence is defined by

$$\|v\|_p := \left(\sum_{n \in \mathbb{N}} |v_n|^p \right)^{1/p}. \quad (3.11)$$

The space of sequences for which this norm is finite is the space of *p-summable sequences*, which is often denoted $\ell^p(\mathbb{K})$ or just ℓ^p . The statement from elementary analysis courses that $\sum_{n=1}^{\infty} \frac{1}{n}$ (the harmonic series) diverges but that $\sum_{n=1}^{\infty} \frac{1}{n^2}$ converges is the statement that

$$(1, \frac{1}{2}, \frac{1}{3}, \dots) \in \ell^2 \quad \text{but} \quad (1, \frac{1}{2}, \frac{1}{3}, \dots) \notin \ell^1.$$

- (c) If S is any set, and $\mathcal{B}(S; \mathbb{K})$ denotes the vector space of all bounded \mathbb{K} -valued functions on S , then a norm on $\mathcal{B}(S; \mathbb{K})$ is the *supremum norm* (or *uniform norm*) defined by

$$\|f\|_\infty := \sup_{x \in S} |f(x)|.$$

- (d) Since every continuous function on a closed and bounded interval is bounded, the supremum norm is also a norm on the space $\mathcal{C}^0([0, 1]; \mathbb{R})$ of continuous real-valued functions on the unit interval.

There is a natural norm to use for linear functions between two normed spaces:

Definition 3.5. Given normed spaces \mathcal{V} and \mathcal{W} , the *operator norm* of a linear map $A: \mathcal{V} \rightarrow \mathcal{W}$ is

$$\|A\| := \sup_{0 \neq v \in \mathcal{V}} \frac{\|A(v)\|_{\mathcal{W}}}{\|v\|_{\mathcal{V}}} \equiv \sup_{\substack{v \in \mathcal{V} \\ \|v\|_{\mathcal{V}}=1}} \|A(v)\|_{\mathcal{W}} \equiv \sup_{\substack{v \in \mathcal{V} \\ \|v\|_{\mathcal{V}} \leq 1}} \|A(v)\|_{\mathcal{W}}.$$

If $\|A\|$ is finite, then A is called a *bounded linear operator*. The operator norm of A will also be denoted $\|A\|_{\text{op}}$ or $\|A\|_{\mathcal{V} \rightarrow \mathcal{W}}$. There are many equivalent expressions for this norm: see Exercise 3.1.

Definition 3.6. Two inner product spaces $(\mathcal{V}, \langle \cdot, \cdot \rangle_{\mathcal{V}})$ and $(\mathcal{W}, \langle \cdot, \cdot \rangle_{\mathcal{W}})$ are said to be *isometrically isomorphic* if there is an invertible linear map $T: \mathcal{V} \rightarrow \mathcal{W}$ such that

$$\langle Tu, Tv \rangle_{\mathcal{W}} = \langle u, v \rangle_{\mathcal{V}} \quad \text{for all } u, v \in \mathcal{V}.$$

The two inner product spaces are then ‘the same up to relabelling’. Similarly, two normed spaces are isometrically isomorphic if there is an invertible linear map that preserves the norm.

Finally, normed spaces are examples of topological spaces, in that the norm structure induces a collection of open sets and (as will be revisited in the next section) a notion of convergence:

Definition 3.7. Let \mathcal{V} be a normed space:

- (a) For $x \in \mathcal{V}$ and $r > 0$, the *open ball of radius r centred on x* is

$$\mathbb{B}_r(x) := \{y \in \mathcal{V} \mid \|x - y\| < r\} \quad (3.12)$$

and the *closed ball of radius r centred on x* is

$$\overline{\mathbb{B}}_r(x) := \{y \in \mathcal{V} \mid \|x - y\| \leq r\}. \quad (3.13)$$

- (b) A subset $U \subseteq \mathcal{V}$ is called an *open set* if, for all $x \in U$, there exists $r = r(x) > 0$ such that $\mathbb{B}_r(x) \subseteq U$.
(c) A subset $F \subseteq \mathcal{V}$ is called a *closed set* if $\mathcal{V} \setminus F$ is an open set.

3.2 Banach and Hilbert Spaces

For the purposes of analysis, rather than pure algebra, it is convenient if normed spaces are complete in the same way that \mathbb{R} is complete and \mathbb{Q} is not:

Definition 3.8. Let $(\mathcal{V}, \|\cdot\|)$ be a normed space.

- (a) A sequence $(x_n)_{n \in \mathbb{N}}$ in \mathcal{V} *converges* to $x \in \mathcal{V}$ if, for every $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that, whenever $n \geq N$, $\|x_n - x\| < \varepsilon$.
(b) A sequence $(x_n)_{n \in \mathbb{N}}$ in \mathcal{V} is called *Cauchy* if, for every $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that, whenever $m, n \geq N$, $\|x_m - x_n\| < \varepsilon$.
(c) A *complete space* is one in which each Cauchy sequence in \mathcal{V} converges to some element of \mathcal{V} . Complete normed spaces are called *Banach spaces*, and complete inner product spaces are called *Hilbert spaces*.

It is easily verified that a subset F of a normed space is closed (in the topological sense of being the complement of an open set) if and only if it is closed under the operation of taking limits of sequences (i.e. every convergent sequence in F has its limit also in F), and that closed linear subspaces of Banach (resp. Hilbert) spaces are again Banach (resp. Hilbert) spaces.

Example 3.9. (a) \mathbb{K}^n and $\mathbb{K}^{m \times n}$ are finite-dimensional Hilbert spaces with respect to their usual inner products.

- (b) The standard example of an infinite-dimensional Hilbert space is the space $\ell^2(\mathbb{K})$ of square-summable \mathbb{K} -valued sequences, which is a Hilbert space with respect to the inner product

$$\langle x, y \rangle_{\ell^2} := \sum_{n \in \mathbb{N}} \overline{x_n} y_n.$$

This space is the prototypical example of a separable Hilbert space, i.e. it has a countably infinite dense subset, and hence countably infinite dimension.

- (c) On the other hand, the subspace of ℓ^2 consisting of all sequences with only finitely many non-zero terms is a non-closed subspace of ℓ^2 , and not a Hilbert space. Of course, if the non-zero terms are restricted to lie in a predetermined finite range of indices, say $\{1, \dots, n\}$, then the subspace is an isomorphic copy of the Hilbert space \mathbb{K}^n .
- (d) Given a measure space $(\mathcal{X}, \mathcal{F}, \mu)$, the space $L^2(\mathcal{X}, \mu; \mathbb{K})$ of (equivalence classes modulo equality μ -almost everywhere of) square-integrable functions from \mathcal{X} to \mathbb{K} is a Hilbert space with respect to the inner product

$$\langle f, g \rangle_{L^2(\mu)} := \int_{\mathcal{X}} \overline{f(x)} g(x) d\mu(x). \quad (3.14)$$

Note that it is necessary to take the quotient by the equivalence relation of equality μ -almost everywhere since a function f that vanishes on a set of full measure but is non-zero on a set of zero measure is not the zero function but nonetheless has $\|f\|_{L^2(\mu)} = 0$. When $(\mathcal{X}, \mathcal{F}, \mu)$ is a probability space, elements of $L^2(\mathcal{X}, \mu; \mathbb{K})$ are thought of as random variables of finite variance, and the L^2 inner product is the covariance:

$$\langle X, Y \rangle_{L^2(\mu)} := \mathbb{E}_{\mu}[\overline{XY}] = \text{cov}(X, Y).$$

When $L^2(\mathcal{X}, \mu; \mathbb{K})$ is a separable space, it is isometrically isomorphic to $\ell^2(\mathbb{K})$ (see Theorem 3.24).

- (e) Indeed, Hilbert spaces over a fixed field \mathbb{K} are classified by their dimension: whenever \mathcal{H} and \mathcal{K} are Hilbert spaces of the same dimension over \mathbb{K} , there is an invertible \mathbb{K} -linear map $T: \mathcal{H} \rightarrow \mathcal{K}$ such that $\langle Tx, Ty \rangle_{\mathcal{K}} = \langle x, y \rangle_{\mathcal{H}}$ for all $x, y \in \mathcal{H}$.

Example 3.10. (a) For a compact topological space \mathcal{X} , the space $\mathcal{C}^0(\mathcal{X}; \mathbb{K})$ of continuous functions $f: \mathcal{X} \rightarrow \mathbb{K}$ is a Banach space with respect to the *supremum norm*

$$\|f\|_{\infty} := \sup_{x \in \mathcal{X}} |f(x)|. \quad (3.15)$$

For non-compact \mathcal{X} , the supremum norm is only a bona fide norm if we restrict attention to bounded continuous functions, since otherwise it would take the inadmissible value $+\infty$.

- (b) More generally, if \mathcal{X} is the compact closure of an open subset of a Banach space \mathcal{V} , and $r \in \mathbb{N}_0$, then the space $\mathcal{C}^r(\mathcal{X}; \mathbb{K})$ of all r -times continuously differentiable functions from \mathcal{X} to \mathbb{K} is a Banach space with respect to the norm

$$\|f\|_{\mathcal{C}^r} := \sum_{k=0}^r \|D^k f\|_{\infty}.$$

Here, $Df(x): \mathcal{V} \rightarrow \mathbb{K}$ denotes the first-order *Fréchet derivative* of f at x , the unique bounded linear map such that

$$\lim_{\substack{y \rightarrow x \\ \text{in } \mathcal{X}}} \frac{|f(y) - f(x) - Df(x)(y - x)|}{\|y - x\|} = 0,$$

$D^2f(x) = D(Df)(x): \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{K}$ denotes the second-order Fréchet derivative, etc.

- (c) For $1 \leq p \leq \infty$, the spaces $L^p(\mathcal{X}, \mu; \mathbb{K})$ from Definition 2.21 are Banach spaces, but only the L^2 spaces are Hilbert spaces. As special cases ($\mathcal{X} = \mathbb{N}$, and $\mu =$ counting measure), the sequence spaces ℓ^p are also Banach spaces, and are Hilbert if and only if $p = 2$.

Another family of Banach spaces that arises very often in PDE applications is the family of *Sobolev spaces*. For the sake of brevity, we limit the discussion to those Sobolev spaces that are also Hilbert spaces. To save space, we use multi-index notation for derivatives: for a multi-index $\alpha := (\alpha_1, \dots, \alpha_n) \in \mathbb{N}_0^n$, with $|\alpha| := \alpha_1 + \dots + \alpha_n$,

$$\partial^\alpha u(x) := \frac{\partial^{|\alpha|} u}{\partial^{\alpha_1} x_1 \dots \partial^{\alpha_n} x_n}(x).$$

Sobolev spaces consist of functions^[3.1] that have appropriately integrable *weak* derivatives, as defined by integrating by parts against smooth test functions:

Definition 3.11. Let $\mathcal{X} \subseteq \mathbb{R}^n$, let $\alpha \in \mathbb{N}_0^n$, and consider $u: \mathcal{X} \rightarrow \mathbb{R}$. A *weak derivative* of order α for u is a function $v: \mathcal{X} \rightarrow \mathbb{R}$ such that

$$\int_{\mathcal{X}} u(x) \partial^\alpha \phi(x) \, dx = (-1)^{|\alpha|} \int_{\mathcal{X}} v(x) \phi(x) \, dx \quad (3.16)$$

for every smooth function $\phi: \mathcal{X} \rightarrow \mathbb{R}$ that vanishes outside a compact subset $\text{supp}(\phi) \subseteq \mathcal{X}$. Such a weak derivative is usually denoted $\partial^\alpha u$ as if it were a strong derivative, and indeed coincides with the classical (strong) derivative if the latter exists. For $s \in \mathbb{N}_0$, the *Sobolev space* $H^s(\mathcal{X})$ is

$$H^s(\mathcal{X}) := \left\{ u \in L^2(\mathcal{X}) \mid \begin{array}{l} \text{for all } \alpha \in \mathbb{N}_0^n \text{ with } |\alpha| \leq s, \\ u \text{ has a weak derivative } \partial^\alpha u \in L^2(\mathcal{X}) \end{array} \right\} \quad (3.17)$$

with the inner product

$$\langle u, v \rangle_{H^s} := \sum_{|\alpha| \leq s} \langle \partial^\alpha u, \partial^\alpha v \rangle_{L^2}. \quad (3.18)$$

The following result shows that smoothness in the Sobolev sense implies either a greater degree of integrability or even Hölder continuity. In particular, possibly after modification on sets of Lebesgue measure zero, Sobolev functions in H^s are continuous when $s > n/2$. Thus, such functions can be considered to have well-defined pointwise values.

Theorem 3.12 (Sobolev embedding theorem). *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a Lipschitz domain (i.e. a connected set with non-empty interior, such that $\partial\mathcal{X}$ can always be locally written as the graph of a Lipschitz function of $n - 1$ variables).*

- (a) *If $s < n/2$, then $H^s(\mathcal{X}) \subseteq L^q(\mathcal{X})$, where $\frac{1}{q} = \frac{1}{2} - \frac{s}{n}$, and there is a constant $C = C(s, n, \mathcal{X})$ such that*

$$\|u\|_{L^q(\mathcal{X})} \leq C \|u\|_{H^s(\mathcal{X})} \quad \text{for all } u \in H^s(\mathcal{X}).$$

^[3.1]To be more precise, as with the Lebesgue L^p spaces, Sobolev spaces consist of *equivalence classes* of such functions, with equivalence being equality almost everywhere.

(b) If $s > n/2$, then $H^s(\mathcal{X}) \subseteq \mathcal{C}^{s-\lfloor n/2 \rfloor-1, \gamma}(\mathcal{X})$, where

$$\gamma = \begin{cases} \lfloor n/2 \rfloor + 1 - n/2, & \text{if } n \text{ is odd,} \\ \text{any element of } (0, 1), & \text{if } n \text{ is even,} \end{cases}$$

and there is a constant $C = C(s, n, \gamma, \mathcal{X})$ such that

$$\|u\|_{\mathcal{C}^{s-\lfloor n/2 \rfloor-1, \gamma}(\mathcal{X})} \leq C \|u\|_{H^s(\mathcal{X})} \quad \text{for all } u \in H^s(\mathcal{X}),$$

where the Hölder norm is defined (up to equivalence) by

$$\|u\|_{\mathcal{C}^{k, \gamma}(\mathcal{X})} := \|u\|_{\mathcal{C}^k} + \sup_{\substack{x, y \in \mathcal{X} \\ x \neq y}} \frac{|D^k u(x) - D^k u(y)|}{|x - y|}.$$

3.3 Dual Spaces and Adjoints

Dual Spaces. Many interesting properties of a vector space are encoded in a second vector space whose elements are the linear functions from the first space to its field. When the vector space is a normed space,^[3.2] so that concepts like continuity are defined, it makes sense to study continuous linear functions:

Definition 3.13. The *continuous dual space* of a normed space \mathcal{V} over \mathbb{K} is the vector space \mathcal{V}' of all bounded (equivalently, continuous) linear functionals $\ell: \mathcal{V} \rightarrow \mathbb{K}$. The dual pairing between an element $\ell \in \mathcal{V}'$ and an element $v \in \mathcal{V}$ is denoted $\langle \ell | v \rangle$ or simply $\ell(v)$. For a linear functional ℓ on a seminormed space \mathcal{V} , being continuous is equivalent to being *bounded* in the sense that its *operator norm* (or *dual norm*)

$$\|\ell\|' := \sup_{0 \neq v \in \mathcal{V}} \frac{|\langle \ell | v \rangle|}{\|v\|} \equiv \sup_{\substack{v \in \mathcal{V} \\ \|v\|=1}} |\langle \ell | v \rangle| \equiv \sup_{\substack{v \in \mathcal{V} \\ \|v\| \leq 1}} |\langle \ell | v \rangle|$$

is finite.

Proposition 3.14. For every normed space \mathcal{V} , the dual space \mathcal{V}' is a Banach space with respect to $\|\cdot\|'$.

An important property of Hilbert spaces is that they are naturally *self-dual*: every continuous linear functional on a Hilbert space can be naturally identified with the action of taking the inner product with some element of the space:

Theorem 3.15 (Riesz representation theorem). Let \mathcal{H} be a Hilbert space. For every continuous linear functional $f \in \mathcal{H}'$, there exists $f^\sharp \in \mathcal{H}$ such that $\langle f | x \rangle = \langle f^\sharp, x \rangle$ for all $x \in \mathcal{H}$. Furthermore, the map $f \mapsto f^\sharp$ is an isometric isomorphism between \mathcal{H} and its dual.

The simplicity of the Riesz representation theorem for duals of Hilbert spaces stands in stark contrast to the duals of even elementary Banach spaces, which are identified on a more case-by-case basis:

- For $1 < p < \infty$, $L^p(\mathcal{X}, \mu)$ is isometrically isomorphic to the dual of $L^q(\mathcal{X}, \mu)$, where $\frac{1}{p} + \frac{1}{q} = 1$. This result applies to the sequence space ℓ^p , and indeed to the finite-dimensional Banach spaces \mathbb{R}^n and \mathbb{C}^n with the norm $\|x\|_p := (\sum_{i=1}^n |x_i|^p)^{1/p}$.
- By the Riesz–Markov–Kakutani representation theorem, the dual of the Banach space $\mathcal{C}_c(\mathcal{X})$ of compactly supported continuous functions on a locally compact Hausdorff space \mathcal{X} is isomorphic to the space of regular signed measures on \mathcal{X} .

The second example stands as another piece of motivation for measure theory in general and signed measures in particular. Readers interested in the details of these constructions should refer to a specialist text on functional analysis.

^[3.2]Or even just a topological vector space.

Adjoint Maps. Given a linear map $A: \mathcal{V} \rightarrow \mathcal{W}$ between normed spaces \mathcal{V} and \mathcal{W} , the *adjoint* of A is the linear map $A^*: \mathcal{W}' \rightarrow \mathcal{V}'$ defined by

$$\langle A^* \ell | v \rangle = \langle \ell | Av \rangle \quad \text{for all } v \in \mathcal{V} \text{ and } \ell \in \mathcal{W}'.$$

The following properties of adjoint maps are fundamental:

Proposition 3.16. *Let \mathcal{U} , \mathcal{V} , and \mathcal{W} be normed spaces, let $A, B: \mathcal{V} \rightarrow \mathcal{W}$ and $C: \mathcal{U} \rightarrow \mathcal{V}$ be bounded linear maps, and let α and β be scalars. Then*

- (a) $A^*: \mathcal{W}' \rightarrow \mathcal{V}'$ is bounded, with operator norm $\|A^*\| = \|A\|$;
- (b) $(\alpha A + \beta B)^* = \overline{\alpha} A^* + \overline{\beta} B^*$;
- (c) $(AC)^* = C^* A^*$;
- (d) the kernel and range of A and A^* satisfy

$$\begin{aligned} \ker A^* &= (\text{ran } A)^\perp := \{\ell \in \mathcal{W}' \mid \langle \ell | Av \rangle = 0 \text{ for all } v \in \mathcal{V}\} \\ (\ker A^*)^\perp &= \overline{\text{ran } A}. \end{aligned}$$

When considering a linear map $A: \mathcal{H} \rightarrow \mathcal{K}$ between Hilbert spaces \mathcal{H} and \mathcal{K} , we can appeal to the Riesz representation theorem to identify \mathcal{H}' with \mathcal{H} , \mathcal{K}' with \mathcal{K} , and hence define the adjoint in terms of inner products:

$$\langle A^* k, h \rangle_{\mathcal{H}} = \langle k, Ah \rangle_{\mathcal{K}} \quad \text{for all } h \in \mathcal{H} \text{ and } k \in \mathcal{K}.$$

With this simplification, we can add to Proposition 3.16 the additional properties that $A^{**} = A$ and $\|A^* A\| = \|AA^*\| = \|A\|^2$. Also, in the Hilbert space setting, a linear map $A: \mathcal{H} \rightarrow \mathcal{H}$ is said to be *self-adjoint* if $A = A^*$. A self-adjoint map A is said to be *positive semi-definite* if

$$\inf_{\substack{x \in \mathcal{H} \\ x \neq 0}} \frac{\langle x, Ax \rangle}{\|x\|^2} \geq 0,$$

and *positive definite* if this inequality is strict.

Given a basis $\{e_i\}_{i \in I}$ of \mathcal{H} , the corresponding *dual basis* $\{e_i\}_{i \in I}$ of \mathcal{H} is defined by the relation $\langle e^i, e_j \rangle_{\mathcal{H}} = \delta_{ij}$. The matrix of A with respect to bases $\{e_i\}_{i \in I}$ of \mathcal{H} and $\{f_j\}_{j \in J}$ of \mathcal{K} and the matrix of A^* with respect to the corresponding dual bases are very simply related: the one is the conjugate transpose of the other, and so by abuse of terminology the conjugate transpose of a matrix is often referred to as the adjoint.

Thus, self-adjoint bounded linear maps are the appropriate generalization to Hilbert spaces of symmetric matrices over \mathbb{R} or Hermitian matrices over \mathbb{C} . They are also particularly useful in probability because the covariance operator of an \mathcal{H} -valued random variable is a self-adjoint (and indeed positive semi-definite) bounded linear operator on \mathcal{H} .

3.4 Orthogonality and Direct Sums

Orthogonal decompositions of Hilbert spaces will be fundamental tools in many of the methods considered later on.

Definition 3.17. A subset E of an inner product space \mathcal{V} is said to be *orthogonal* if $\langle x, y \rangle = 0$ for all distinct elements $x, y \in E$; it is said to be *orthonormal* if

$$\langle x, y \rangle = \begin{cases} 1, & \text{if } x = y \in E, \\ 0, & \text{if } x, y \in E \text{ and } x \neq y. \end{cases}$$

Lemma 3.18 (Gram–Schmidt). *Let $(x_n)_{n \in \mathbb{N}}$ be any sequence in an inner product space \mathcal{V} , with the first $d \in \mathbb{N}_0 \cup \{\infty\}$ terms linearly independent. Inductively define $(u_n)_{n \in \mathbb{N}}$ and*

$(e_n)_{n \in \mathbb{N}}$ by

$$u_n := x_n - \sum_{k=1}^{n-1} \frac{\langle x_n, u_k \rangle}{\|u_k\|^2} u_k,$$

$$e_n := \frac{u_n}{\|u_n\|}$$

Then $(u_n)_{n \in \mathbb{N}}$ (resp. $(e_n)_{n \in \mathbb{N}}$) is a sequence of d orthogonal (resp. orthonormal) elements of \mathcal{V} , followed by zeros if $d < \infty$.

Definition 3.19. The *orthogonal complement* E^\perp of a subset E of an inner product space \mathcal{V} is

$$E^\perp := \{y \in \mathcal{V} \mid \text{for all } x \in E, \langle y, x \rangle = 0\}.$$

The orthogonal complement of $E \subseteq \mathcal{V}$ is always a closed linear subspace of \mathcal{V} , and hence if $\mathcal{V} = \mathcal{H}$ is a Hilbert space, then E^\perp is also a Hilbert space in its own right.

Theorem 3.20. Let \mathcal{K} be a closed subspace of a Hilbert space \mathcal{H} . Then, for any $x \in \mathcal{H}$, there is a unique $\Pi_{\mathcal{K}}x \in \mathcal{K}$ that is closest to x in the sense that

$$\|\Pi_{\mathcal{K}}x - x\| = \inf_{y \in \mathcal{K}} \|y - x\|.$$

Furthermore, x can be written uniquely as $x = \Pi_{\mathcal{K}}x + z$, where $z \in \mathcal{K}^\perp$. Hence, \mathcal{H} decomposes as the orthogonal direct sum

$$\mathcal{H} = \mathcal{K} \oplus \mathcal{K}^\perp.$$

Theorem 3.20 can be seen as a special case of closest-point approximation among convex sets: see Lemma 4.25 and Exercise 4.2. The operator $\Pi_{\mathcal{K}}: \mathcal{H} \rightarrow \mathcal{K}$ is called the *orthogonal projection* onto \mathcal{K} .

Theorem 3.21. Let \mathcal{K} be a closed subspace of a Hilbert space \mathcal{H} . The corresponding orthogonal projection operator $\Pi_{\mathcal{K}}$ is

- (a) a continuous linear operator of norm at most 1;
 - (b) with $I - \Pi_{\mathcal{K}} = \Pi_{\mathcal{K}^\perp}$;
- and satisfies, for every $x \in \mathcal{H}$,
- (c) $\|x\|^2 = \|\Pi_{\mathcal{K}}x\|^2 + \|(I - \Pi_{\mathcal{K}})x\|^2$;
 - (d) $\Pi_{\mathcal{K}}x = x \iff x \in \mathcal{K}$;
 - (e) $\Pi_{\mathcal{K}}x = 0 \iff x \in \mathcal{K}^\perp$.

Example 3.22 (Conditional expectation). An important probabilistic application of orthogonal projection is the operation of conditioning a random variable. Let $(\Theta, \mathcal{F}, \mu)$ be a probability space and let $X \in L^2(\Theta, \mathcal{F}, \mu; \mathbb{K})$ be a square-integrable random variable. If $\mathcal{G} \subseteq \mathcal{F}$ is a σ -algebra, then the *conditional expectation* of X with respect to \mathcal{G} , usually denoted $\mathbb{E}[X|\mathcal{G}]$, is the orthogonal projection of X onto the subspace $L^2(\Theta, \mathcal{G}, \mu; \mathbb{K})$. In elementary contexts, \mathcal{G} is usually taken to be the σ -algebra generated by a single event E of positive μ -probability, i.e.

$$\mathcal{G} = \{\emptyset, [X \in E], [X \notin E], \Theta\};$$

or even the trivial σ -algebra $\{\emptyset, \Theta\}$, for which the only measurable functions are the constant functions, and hence the conditional expectation coincides with the usual expectation. The orthogonal projection point of view makes two important properties of conditional expectation intuitively obvious:

- (a) Whenever $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{F}$, $L^2(\Theta, \mathcal{G}_1, \mu; \mathbb{K})$ is a subspace of $L^2(\Theta, \mathcal{G}_2, \mu; \mathbb{K})$ and composition of the orthogonal projections onto these subspace yields the *tower rule* for conditional expectations:

$$\mathbb{E}[X|\mathcal{G}_1] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1],$$

and, in particular, taking \mathcal{G}_1 to be the trivial σ -algebra $\{\emptyset, \Theta\}$,

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}_2]].$$

- (b) Whenever $X, Y \in L^2(\Theta, \mathcal{F}, \mu; \mathbb{K})$ and X is, in fact, \mathcal{G} -measurable,

$$\mathbb{E}[XY|\mathcal{G}] = X\mathbb{E}[Y|\mathcal{G}].$$

Direct Sums. Suppose that \mathcal{V} and \mathcal{W} are vector spaces over a common field \mathbb{K} . The Cartesian product $\mathcal{V} \times \mathcal{W}$ can be given the structure of a vector space over \mathbb{K} by defining the operations componentwise:

$$\begin{aligned} (v, w) + (v', w') &:= (v + v', w + w'), \\ \alpha(v, w) &:= (\alpha v, \alpha w), \end{aligned}$$

for all $v, v' \in \mathcal{V}$, $w, w' \in \mathcal{W}$, and $\alpha \in \mathbb{K}$. The resulting vector space is called the (*algebraic*) *direct sum* of \mathcal{V} and \mathcal{W} and is usually denoted by $\mathcal{V} \oplus \mathcal{W}$, while elements of $\mathcal{V} \oplus \mathcal{W}$ are usually denoted by $v \oplus w$ instead of (v, w) .

If $\{e_i | i \in I\}$ is a basis of \mathcal{V} and $\{e_j | j \in J\}$ is a basis of \mathcal{W} , then $\{e_k | k \in K := I \uplus J\}$ is basis of $\mathcal{V} \oplus \mathcal{W}$. Hence, the dimension of $\mathcal{V} \oplus \mathcal{W}$ over \mathbb{K} is equal to the sum of the dimensions of \mathcal{V} and \mathcal{W} .

When \mathcal{H} and \mathcal{K} are Hilbert spaces, their (algebraic) direct sum $\mathcal{H} \oplus \mathcal{K}$ can be given a Hilbert space structure by defining

$$\langle h \oplus k, h' \oplus k' \rangle_{\mathcal{H} \oplus \mathcal{K}} := \langle h, h' \rangle_{\mathcal{H}} + \langle k, k' \rangle_{\mathcal{K}}$$

for all $h, h' \in \mathcal{H}$ and $k, k' \in \mathcal{K}$. The original spaces \mathcal{H} and \mathcal{K} embed into $\mathcal{H} \oplus \mathcal{K}$ as the subspaces $\mathcal{H} \oplus \{0\}$ and $\{0\} \oplus \mathcal{K}$ respectively, and these two subspaces are mutually orthogonal. For this reason, the orthogonality of the two summands in a Hilbert direct sum is sometimes emphasized by the notation $\mathcal{H} \overset{\perp}{\oplus} \mathcal{K}$. The Hilbert space projection theorem (Theorem 3.20) was the statement that whenever \mathcal{K} is a closed subspace of a Hilbert space \mathcal{H} , $\mathcal{H} = \mathcal{K} \overset{\perp}{\oplus} \mathcal{K}^\perp$.

It is necessary to be a bit more careful in defining the direct sum of countably many Hilbert spaces. Let \mathcal{H}_n be a Hilbert space over \mathbb{K} for each $n \in \mathbb{N}$. Then the Hilbert space direct sum $\mathcal{H} := \bigoplus_{n \in \mathbb{N}} \mathcal{H}_n$ is defined to be

$$\mathcal{H} := \overline{\left\{ x = (x_n)_{n \in \mathbb{N}} \mid \begin{array}{l} x_n \in \mathcal{H}_n \text{ for each } n \in \mathbb{N}, \text{ and} \\ x_n = 0 \text{ for all but finitely many } n \end{array} \right\}},$$

where the completion^[3.3] is taken with respect to the inner product

$$\langle x, y \rangle_{\mathcal{H}} := \sum_{n \in \mathbb{N}} \langle x_n, y_n \rangle_{\mathcal{H}_n},$$

which is always a finite sum when applied to elements of the generating set. This construction ensures that every element x of \mathcal{H} has finite norm $\|x\|_{\mathcal{H}}^2 = \sum_{n \in \mathbb{N}} \|x_n\|_{\mathcal{H}_n}^2$. As before, each of the summands \mathcal{H}_n is a subspace of \mathcal{H} that is orthogonal to all the others.

[3.3] Completions of normed spaces are formed in the same way as the completion of \mathbb{Q} to form \mathbb{R} : the completion is the space of equivalence classes of Cauchy sequences, with sequences whose difference tends to zero in norm being regarded as equivalent.

Orthogonal direct sums and orthogonal bases are among the most important constructions in Hilbert space theory, and will be very useful in what follows. Prototypical examples include the standard ‘Euclidean’ basis of ℓ^2 and the *Fourier basis* $\{e_n \mid n \in \mathbb{Z}\}$ of $L^2(\mathbb{S}^1; \mathbb{C})$, where

$$e_n(x) := \frac{1}{2\pi} \exp(inx).$$

Indeed, Fourier’s claim^[3.4] that any periodic function f could be written as

$$f(x) = \sum_{n \in \mathbb{Z}} \hat{f}_n e_n(x),$$

$$\hat{f}_n := \int_{\mathbb{S}^1} f(y) \overline{e_n(y)} dy,$$

can be seen as one of the historical drivers behind the development of much of analysis. For the purposes of this book’s treatment of UQ, key examples of an orthogonal bases are given by *orthogonal polynomials*, which will be considered at length in Chapter 8.

Some important results about orthogonal systems are summarized below; classically, many of these results arose in the study of Fourier series, but hold for any orthonormal basis of a general Hilbert space.

Lemma 3.23 (Bessel’s inequality). *Let \mathcal{V} be an inner product space and $(e_n)_{n \in \mathbb{N}}$ an orthonormal sequence in \mathcal{V} . Then, for any $x \in \mathcal{V}$, the series $\sum_{n \in \mathbb{N}} |\langle e_n, x \rangle|^2$ converges and satisfies*

$$\sum_{n \in \mathbb{N}} |\langle e_n, x \rangle|^2 \leq \|x\|^2. \quad (3.19)$$

Theorem 3.24 (Parseval identity). *Let $(e_n)_{n \in \mathbb{N}}$ be an orthonormal sequence in a Hilbert space \mathcal{H} , and let $(\alpha_n)_{n \in \mathbb{N}}$ be a sequence in \mathbb{K} . Then the series $\sum_{n \in \mathbb{N}} \alpha_n e_n$ converges in \mathcal{H} if and only if the series $\sum_{n \in \mathbb{N}} |\alpha_n|^2$ converges in \mathbb{R} , in which case*

$$\left\| \sum_{n \in \mathbb{N}} \alpha_n e_n \right\|^2 = \sum_{n \in \mathbb{N}} |\alpha_n|^2. \quad (3.20)$$

Hence, for any $x \in \mathcal{H}$, the series $\sum_{n \in \mathbb{N}} \langle e_n, x \rangle e_n$ converges.

Theorem 3.25. *Let $(e_n)_{n \in \mathbb{N}}$ be an orthonormal sequence in a Hilbert space \mathcal{H} . Then the following are equivalent:*

- (a) $\{e_n \mid n \in \mathbb{N}\}^\perp = \{0\}$;
- (b) $\mathcal{H} = \text{span}\{e_n \mid n \in \mathbb{N}\}$;
- (c) $\mathcal{H} = \bigoplus_{n \in \mathbb{N}} \mathbb{K}e_n$ as a direct sum of Hilbert spaces;
- (d) for all $x \in \mathcal{H}$, $\|x\|^2 = \sum_{n \in \mathbb{N}} |\langle e_n, x \rangle|^2$;
- (e) for all $x \in \mathcal{H}$, $x = \sum_{n \in \mathbb{N}} \langle e_n, x \rangle e_n$.

If one (and hence all) of these conditions holds true, then $(e_n)_{n \in \mathbb{N}}$ is called a *complete orthonormal basis* for \mathcal{H} .

Corollary 3.26. *Let $(e_n)_{n \in \mathbb{N}}$ be a complete orthonormal basis for a Hilbert space \mathcal{H} . For every $x \in \mathcal{H}$, the truncation error $x - \sum_{n=1}^N \langle e_n, x \rangle e_n$ is orthogonal to $\text{span}\{e_1, \dots, e_N\}$.*

Proof. Let $v := \sum_{m=1}^N v_m e_m \in \text{span}\{e_1, \dots, e_N\}$ be arbitrary. By completeness,

$$x = \sum_{n \in \mathbb{N}} \langle e_n, x \rangle e_n.$$

^[3.4]Of course, Fourier did not use the modern notation of Hilbert spaces! Furthermore, if he had, then it would have been ‘obvious’ that his claim could only hold true for L^2 functions and in the L^2 sense, not pointwise for arbitrary functions.

Hence,

$$\begin{aligned}
 \left\langle x - \sum_{n=1}^N \langle e_n, x \rangle e_n, v \right\rangle &= \left\langle \sum_{n>N} \langle e_n, x \rangle e_n, \sum_{m=1}^N v_m e_m \right\rangle \\
 &= \sum_{\substack{n>N \\ m \in \{0, \dots, N\}}} \langle \langle e_n, x \rangle e_n, v_m e_m \rangle \\
 &= \sum_{\substack{n>N \\ m \in \{0, \dots, N\}}} \langle x, e_n \rangle v_m \langle e_n, e_m \rangle \\
 &= 0
 \end{aligned}$$

since $\langle e_n, e_m \rangle = \delta_{nm}$, and $m \neq n$ in the double sum. ■

Remark 3.27. The results cited above (in particular, Theorems 3.20, 3.21, and 3.25, and Corollary 3.26) imply that if we wish to find the closest point of $\text{span}\{e_1, \dots, e_N\}$ to some $x = \sum_{n \in \mathbb{N}} \langle e_n, x \rangle e_n$, then this is a simple matter of series truncation: the optimal approximation is $x \approx x^{(N)} := \sum_{n=1}^N \langle e_n, x \rangle e_n$. Furthermore, this operation is a continuous linear operation as a function of x , and if it is desired to improve the quality of an approximation $x \approx x^{(N)}$ in $\text{span}\{e_1, \dots, e_N\}$ to an approximation in, say, $\text{span}\{e_1, \dots, e_{N+1}\}$, then the improvement is a simple matter of calculating $\langle e_{N+1}, x \rangle$ and adjoining the new term $\langle e_{N+1}, x \rangle e_{N+1}$ to form a new norm-optimal approximation

$$x \approx x^{(N+1)} := \sum_{n=1}^{N+1} \langle e_n, x \rangle e_n = x^{(N)} + \langle e_{N+1}, x \rangle e_{N+1}.$$

However, in Banach spaces (even finite-dimensional ones), closest-point approximation is not as simple as series truncation, and the improvement of approximations is not as simple as adjoining new terms: see Exercise 3.4.

3.5 Tensor Products

The heuristic definition of the tensor product $\mathcal{V} \otimes \mathcal{W}$ of two vector spaces \mathcal{V} and \mathcal{W} over a common field \mathbb{K} is that it is the vector space over \mathbb{K} with basis given by the formal symbols $\{e_i \otimes f_j \mid i \in I, j \in J\}$, where $\{e_i \mid i \in I\}$ is a basis of \mathcal{V} and $\{f_j \mid j \in J\}$ is a basis of \mathcal{W} . Alternatively, we might say that elements of $\mathcal{V} \otimes \mathcal{W}$ are elements of \mathcal{W} with \mathcal{V} -valued rather than \mathbb{K} -valued coefficients (or elements of \mathcal{V} with \mathcal{W} -valued coefficients). However, it is not immediately clear that this definition is independent of the bases chosen for \mathcal{V} and \mathcal{W} . A more thorough definition is as follows.

Definition 3.28. The *free vector space* $F_{\mathcal{V} \times \mathcal{W}}$ on the Cartesian product $\mathcal{V} \times \mathcal{W}$ is defined by taking the vector space in which the elements of $\mathcal{V} \times \mathcal{W}$ are a basis:

$$F_{\mathcal{V} \times \mathcal{W}} := \left\{ \sum_{i=1}^n \alpha_i e_{(v_i, w_i)} \mid \begin{array}{l} n \in \mathbb{N} \text{ and, for } i = 1, \dots, n, \\ \alpha_i \in \mathbb{K}, (v_i, w_i) \in \mathcal{V} \times \mathcal{W} \end{array} \right\}.$$

The ‘freeness’ of $F_{\mathcal{V} \times \mathcal{W}}$ is that the elements $e_{(v, w)}$ are, by definition, linearly independent for distinct pairs $(v, w) \in \mathcal{V} \times \mathcal{W}$; even $e_{(v, 0)}$ and $e_{(-v, 0)}$ are linearly independent. Now define an equivalence relation \sim on $F_{\mathcal{V} \times \mathcal{W}}$ such that

$$\begin{aligned}
 e_{(v+v', w)} &\sim e_{(v, w)} + e_{(v', w)}, \\
 e_{(v, w+w')} &\sim e_{(v, w)} + e_{(v, w')}, \\
 \alpha e_{(v, w)} &\sim e_{(\alpha v, w)} \sim e_{(v, \alpha w)}
 \end{aligned}$$

for arbitrary $v, v' \in \mathcal{V}$, $w, w' \in \mathcal{W}$, and $\alpha \in \mathbb{K}$. Let R be the subspace of $F_{\mathcal{V} \times \mathcal{W}}$ generated by these equivalence relations, i.e. the equivalence class of $e_{(0,0)}$.

Definition 3.29. The (algebraic) tensor product $\mathcal{V} \otimes \mathcal{W}$ is the quotient space

$$\mathcal{V} \otimes \mathcal{W} := \frac{F_{\mathcal{V} \times \mathcal{W}}}{R}.$$

One can easily check that $\mathcal{V} \otimes \mathcal{W}$, as defined in this way, is indeed a vector space over \mathbb{K} . The subspace R of $F_{\mathcal{V} \times \mathcal{W}}$ is mapped to the zero element of $\mathcal{V} \otimes \mathcal{W}$ under the quotient map, and so the above equivalences become equalities in the tensor product space:

$$\begin{aligned} (v + v') \otimes w &= v \otimes w + v' \otimes w, \\ v \otimes (w + w') &= v \otimes w + v \otimes w', \\ \alpha(v \otimes w) &= (\alpha v) \otimes w = v \otimes (\alpha w) \end{aligned}$$

for all $v, v' \in \mathcal{V}$, $w, w' \in \mathcal{W}$, and $\alpha \in \mathbb{K}$.

One can also check that the heuristic definition in terms of bases holds true under the formal definition: if $\{e_i | i \in I\}$ is a basis of \mathcal{V} and $\{f_j | j \in J\}$ is a basis of \mathcal{W} , then $\{e_i \otimes f_j | i \in I, j \in J\}$ is basis of $\mathcal{V} \otimes \mathcal{W}$. Hence, the dimension of the tensor product is the product of dimensions of the original spaces.

Definition 3.30. The Hilbert space tensor product of two Hilbert spaces \mathcal{H} and \mathcal{K} over the same field \mathbb{K} is given by defining an inner product on the algebraic tensor product $\mathcal{H} \otimes \mathcal{K}$ by

$$\langle h \otimes k, h' \otimes k' \rangle_{\mathcal{H} \otimes \mathcal{K}} := \langle h, h' \rangle_{\mathcal{H}} \langle k, k' \rangle_{\mathcal{K}} \quad \text{for all } h, h' \in \mathcal{H} \text{ and } k, k' \in \mathcal{K},$$

extending this definition to all of the algebraic tensor product by sesquilinearity, and defining the Hilbert space tensor product $\mathcal{H} \otimes \mathcal{K}$ to be the completion of the algebraic tensor product with respect to this inner product and its associated norm.

Tensor products of Hilbert spaces arise very naturally when considering spaces of functions of more than one variable, or spaces of functions that take values in other function spaces. A prime example of the second type is a space of stochastic processes.

Example 3.31. (a) Given two measure spaces $(\mathcal{X}, \mathcal{F}, \mu)$ and $(\mathcal{Y}, \mathcal{G}, \nu)$, consider $L^2(\mathcal{X} \times \mathcal{Y}, \mu \otimes \nu; \mathbb{K})$, the space of functions on $\mathcal{X} \times \mathcal{Y}$ that are square integrable with respect to the product measure $\mu \otimes \nu$. If $f \in L^2(\mathcal{X}, \mu; \mathbb{K})$ and $g \in L^2(\mathcal{Y}, \nu; \mathbb{K})$, then we can define a function $h: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{K}$ by $h(x, y) := f(x)g(y)$. The definition of the product measure ensures that $h \in L^2(\mathcal{X} \times \mathcal{Y}, \mu \otimes \nu; \mathbb{K})$, so this procedure defines a bilinear mapping $L^2(\mathcal{X}, \mu; \mathbb{K}) \times L^2(\mathcal{Y}, \nu; \mathbb{K}) \rightarrow L^2(\mathcal{X} \times \mathcal{Y}, \mu \otimes \nu; \mathbb{K})$. It turns out that the span of the range of this bilinear map is dense in $L^2(\mathcal{X} \times \mathcal{Y}, \mu \otimes \nu; \mathbb{K})$ if $L^2(\mathcal{X}, \mu; \mathbb{K})$ and $L^2(\mathcal{Y}, \nu; \mathbb{K})$ are separable. This shows that

$$L^2(\mathcal{X}, \mu; \mathbb{K}) \otimes L^2(\mathcal{Y}, \nu; \mathbb{K}) \cong L^2(\mathcal{X} \times \mathcal{Y}, \mu \otimes \nu; \mathbb{K}),$$

and it also explains why it is necessary to take the completion in the construction of the Hilbert space tensor product.

- (b) Similarly, $L^2(\mathcal{X}, \mu; \mathcal{H})$, the space of functions $f: \mathcal{X} \rightarrow \mathcal{H}$ that are square integrable in the sense that

$$\int_{\mathcal{X}} \|f(x)\|_{\mathcal{H}}^2 d\mu(x) < +\infty,$$

is isomorphic to $L^2(\mathcal{X}, \mu; \mathbb{K}) \otimes \mathcal{H}$ if this space is separable. The isomorphism maps $f \otimes \varphi \in L^2(\mathcal{X}, \mu; \mathbb{K}) \otimes \mathcal{H}$ to the \mathcal{H} -valued function $x \mapsto f(x)\varphi$ in $L^2(\mathcal{X}, \mu; \mathcal{H})$.

- (c) Combining the previous two examples reveals that

$$L^2(\mathcal{X}, \mu; \mathbb{K}) \otimes L^2(\mathcal{Y}, \nu; \mathbb{K}) \cong L^2(\mathcal{X} \times \mathcal{Y}, \mu \otimes \nu; \mathbb{K}) \cong L^2(\mathcal{X}, \mu; L^2(\mathcal{Y}, \nu; \mathbb{K})).$$

Similarly, one can consider a *Bochner space* $L^p(\mathcal{X}, \mu; \mathcal{V})$ of functions (random variables) taking values in a Banach space \mathcal{V} that are p^{th} -power-integrable in the sense that $\int_{\mathcal{X}} \|f(x)\|_{\mathcal{V}}^p d\mu(x)$ is finite, and identify this space with a suitable tensor product $L^p(\mathcal{X}, \mu; \mathbb{R}) \otimes \mathcal{V}$. However, several subtleties arise in doing this, as there is no single ‘natural’ Banach tensor product of Banach spaces as there is for Hilbert spaces.

3.6 Bibliography

Reference texts on elementary functional analysis, including Banach and Hilbert space theory, include the books of Reed and Simon (1972), Rudin (1991), and Rynne and Youngson (2008). The article of Deutsch (1982) gives a good overview of closest-point approximation properties for subspaces of Banach spaces. Further discussion of the relationship between tensor products and spaces of vector-valued integrable functions can be found in the books of Ryan (2002) and Hackbusch (2012); the former is essentially a pure mathematic text, whereas the latter also includes significant treatment of numerical and computational matters. The Sobolev embedding theorem (Theorem 3.12) and its proof can be found in Evans (2010, Section 5.6, Theorem 6).

Intrepid students may wish to consult Bourbaki (1987), but the standard warnings about Bourbaki texts apply: the presentation is comprehensive but often forbiddingly austere, and so it is perhaps better as a reference text than a learning tool. On the other hand, the *Hitchhiker’s Guide* of Aliprantis and Border (2006) is a surprisingly readable encyclopaedic text.

3.7 Exercises

Exercise 3.1 (Formulae for the operator norm). Let $A: \mathcal{V} \rightarrow \mathcal{W}$ be a linear map between normed vector spaces $(\mathcal{V}, \|\cdot\|_{\mathcal{V}})$ and $(\mathcal{W}, \|\cdot\|_{\mathcal{W}})$. Show that the operator norm $\|A\|_{\mathcal{V} \rightarrow \mathcal{W}}$ of A is equivalently defined by any of the following expressions:

$$\begin{aligned} \|A\|_{\mathcal{V} \rightarrow \mathcal{W}} &= \sup_{0 \neq v \in \mathcal{V}} \frac{\|Av\|_{\mathcal{W}}}{\|v\|_{\mathcal{V}}} \\ &= \sup_{\|v\|_{\mathcal{V}}=1} \frac{\|Av\|_{\mathcal{W}}}{\|v\|_{\mathcal{V}}} = \sup_{\|v\|_{\mathcal{V}}=1} \|Av\|_{\mathcal{W}} \\ &= \sup_{0 < \|v\|_{\mathcal{V}} \leq 1} \frac{\|Av\|_{\mathcal{W}}}{\|v\|_{\mathcal{V}}} = \sup_{\|v\|_{\mathcal{V}} \leq 1} \|Av\|_{\mathcal{W}} \\ &= \sup_{0 < \|v\|_{\mathcal{V}} < 1} \frac{\|Av\|_{\mathcal{W}}}{\|v\|_{\mathcal{V}}} = \sup_{\|v\|_{\mathcal{V}} < 1} \|Av\|_{\mathcal{W}}. \end{aligned}$$

Exercise 3.2 (Properties of the operator norm). Suppose that \mathcal{U} , \mathcal{V} , and \mathcal{W} are normed vector spaces, and let $A: \mathcal{U} \rightarrow \mathcal{V}$ and $B: \mathcal{V} \rightarrow \mathcal{W}$ be bounded linear maps. Prove that the operator norm is

(a) *compatible* (or *consistent*) with $\|\cdot\|_{\mathcal{U}}$ and $\|\cdot\|_{\mathcal{V}}$: for all $x \in \mathcal{U}$,

$$\|Au\|_{\mathcal{V}} \leq \|A\|_{\mathcal{U} \rightarrow \mathcal{V}} \|u\|_{\mathcal{U}}.$$

(b) *sub-multiplicative*: $\|B \circ A\|_{\mathcal{U} \rightarrow \mathcal{W}} \leq \|B\|_{\mathcal{V} \rightarrow \mathcal{W}} \|A\|_{\mathcal{U} \rightarrow \mathcal{V}}$.

Exercise 3.3 (Definiteness of the Gram matrix). Let \mathcal{V} be a vector space over \mathbb{K} , equipped with a semi-definite inner product $\langle \cdot, \cdot \rangle$ (i.e. one satisfying all the requirements of Definition 3.2 except possibly positive definiteness). Given vectors $v_1, \dots, v_n \in \mathcal{V}$, the associated *Gram matrix* is

$$G(v_1, \dots, v_n) := \begin{bmatrix} \langle v_1, v_1 \rangle & \cdots & \langle v_1, v_n \rangle \\ \vdots & \ddots & \vdots \\ \langle v_n, v_1 \rangle & \cdots & \langle v_n, v_n \rangle \end{bmatrix}.$$

- (a) Show that, in the case that $\mathcal{V} = \mathbb{K}^n$ with its usual inner product, $G(v_1, \dots, v_n) = V^*V = VV^*$, where V is the matrix with the vectors v_i as its columns, and V^* denotes the conjugate transpose of V .
- (b) Show that $G(v_1, \dots, v_n)$ is a conjugate-symmetric (a.k.a. Hermitian) matrix, and hence is symmetric in the case $\mathbb{K} = \mathbb{R}$.
- (c) Show that $\det G(v_1, \dots, v_n) \geq 0$. Show also that $\det G(v_1, \dots, v_n) = 0$ if v_1, \dots, v_n are linearly dependent, and that this is an ‘if and only if’ if $\langle \cdot, \cdot \rangle$ is positive definite.
- (d) Using the case $n = 2$, prove the Cauchy–Schwarz inequality (3.1).

Exercise 3.4 (Closest-point approximation in Banach spaces). Let $R_\theta: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ denote the linear map that is rotation of the Euclidean plane about the origin through a fixed angle $-\frac{\pi}{4} < \theta < \frac{\pi}{4}$. Define a Banach norm $\|\cdot\|_\theta$ on \mathbb{R}^2 in terms of R_θ and the usual 1-norm by

$$\|(x, y)\|_\theta := \|R_\theta(x, y)\|_1.$$

Find the closest point of the x -axis to the point $(1, 1)$, i.e. find $x' \in \mathbb{R}$ to minimize $\|(x', 0) - (1, 1)\|_\theta$; in particular, show that the closest point is *not* $(1, 0)$. Hint: sketch some norm balls centred on $(1, 1)$.

Exercise 3.5 (Series in normed spaces). Many UQ methods involve series expansions in spaces of deterministic functions and/or random variables, so it is useful to understand when such series converge. Let $(v_n)_{n \in \mathbb{N}}$ be a sequence in a normed space \mathcal{V} . As in \mathbb{R} , we say that the series $\sum_{n \in \mathbb{N}} v_n$ converges to $v \in \mathcal{V}$ if the sequence of partial sums converges to v , i.e. if, for all $\varepsilon > 0$, there exists $N_\varepsilon \in \mathbb{N}$ such that

$$N \geq N_\varepsilon \implies \left\| v - \sum_{n=1}^N v_n \right\| < \varepsilon.$$

- (a) Suppose that $\sum_{n \in \mathbb{N}} v_n$ converges *absolutely* to $v \in \mathcal{V}$, i.e. the series converges and also $\sum_{n \in \mathbb{N}} \|v_n\|$ is finite. Prove the infinite triangle inequality

$$\|v\| \leq \sum_{n \in \mathbb{N}} \|v_n\|.$$

- (b) Suppose that $\sum_{n \in \mathbb{N}} v_n$ converges absolutely to $v \in \mathcal{V}$. Show that $\sum_{n \in \mathbb{N}} v_n$ converges *unconditionally* to $v \in \mathcal{V}$, i.e. $\sum_{n \in \mathbb{N}} v_{\pi(n)}$ converges to $x \in \mathcal{V}$ for every bijection $\pi: \mathbb{N} \rightarrow \mathbb{N}$. Thus, the order of summation ‘does not matter’. (Note that the converse of this result is false: Dvoretzky and Rogers (1950) showed that every infinite-dimensional Banach space contains series that converge unconditionally but not absolutely.)
- (c) Suppose that \mathcal{V} is a Banach space and that $\sum_{n \in \mathbb{N}} \|v_n\|$ is finite. Show that $\sum_{n \in \mathbb{N}} v_n$ converges to some $v \in \mathcal{V}$.



Exercise 3.6 (Weierstrass M -test). Let S be any set, let \mathcal{V} be a Banach space, and, for each $n \in \mathbb{N}$, let $f_n: S \rightarrow \mathcal{V}$. Suppose that M_n is such that

$$\|f_n(x)\| \leq M_n \quad \text{for all } x \in S \text{ and } n \in \mathbb{N},$$

and that $\sum_{n \in \mathbb{N}} M_n$ is finite. Show that the series $\sum_{n \in \mathbb{N}} f_n$ converges *uniformly* on S , i.e. there exists $f: S \rightarrow \mathcal{V}$ such that, for all $\varepsilon > 0$, there exists $N_\varepsilon \in \mathbb{N}$ so that

$$N \geq N_\varepsilon \implies \sup_{x \in S} \left\| f(x) - \sum_{n=1}^N f_n(x) \right\| < \varepsilon.$$

Chapter 4

Optimization Theory

We demand rigidly defined areas of doubt and uncertainty!

The Hitchhiker's Guide to the Galaxy
DOUGLAS ADAMS

This chapter reviews the basic elements of optimization theory and practice, without going into the fine details of numerical implementation. Many UQ problems involve a notion of ‘best fit’, in the sense of minimizing some error function, and so it is helpful to establish some terminology for optimization problems. In particular, many of the optimization problems in this book will fall into the simple settings of linear programming and least squares (quadratic programming), with and without constraints.

4.1 Optimization Problems and Terminology

In an optimization problem, the objective is to find the extreme values (either the minimal value, the maximal value, or both) $f(x)$ of a given function f among all x in a given subset of the domain of f , along with the point or points x that realize those extreme values. The general form of a constrained optimization problem is

$$\begin{aligned} &\text{extremize: } f(x) \\ &\text{with respect to: } x \in \mathcal{X} \\ &\text{subject to: } g_i(x) \in E_i \quad \text{for } i = 1, 2, \dots, \end{aligned}$$

where \mathcal{X} is some set; $f: \mathcal{X} \rightarrow \mathbb{R} \cup \{\pm\infty\}$ is a function called the *objective function*; and, for each i , $g_i: \mathcal{X} \rightarrow \mathcal{Y}_i$ is a function and $E_i \subseteq \mathcal{Y}_i$ some subset. The conditions $\{g_i(x) \in E_i \mid i = 1, 2, \dots\}$ are called *constraints*, and a point $x \in \mathcal{X}$ for which all the constraints are satisfied is called *feasible*; the set of feasible points,

$$\{x \in \mathcal{X} \mid g_i(x) \in E_i \text{ for } i = 1, 2, \dots\},$$

is called the *feasible set*. If there are no constraints, so that the problem is a search over all of \mathcal{X} , then the problem is said to be *unconstrained*. In the case of a minimization problem, the objective function f is also called the *cost function* or *energy*; for maximization problems, the objective function is also called the *utility function*.

From a purely mathematical point of view, the distinction between constrained and unconstrained optimization is artificial: constrained minimization over \mathcal{X} is the same as unconstrained minimization over the feasible set. However, from a practical standpoint,

the difference is huge. Typically, \mathcal{X} is \mathbb{R}^n for some n , or perhaps a simple subset specified using inequalities on one coordinate at a time, such as $[a_1, b_1] \times \cdots \times [a_n, b_n]$; a bona fide non-trivial constraint is one that involves a more complicated function of one coordinate, or two or more coordinates, such as

$$g_1(x) := \cos(x) - \sin(x) > 0$$

or

$$g_2(x_1, x_2, x_3) := x_1 x_2 - x_3 = 0.$$

Definition 4.1. Given $f: \mathcal{X} \rightarrow \mathbb{R} \cup \{\pm\infty\}$, the *arg min* or *set of global minimizers* of f is defined to be

$$\arg \min_{x \in \mathcal{X}} f(x) := \left\{ x \in \mathcal{X} \mid f(x) = \inf_{x' \in \mathcal{X}} f(x') \right\},$$

and the *arg max* or *set of global maximizers* of f is defined to be

$$\arg \max_{x \in \mathcal{X}} f(x) := \left\{ x \in \mathcal{X} \mid f(x) = \sup_{x' \in \mathcal{X}} f(x') \right\}.$$

Definition 4.2. For a given constrained or unconstrained optimization problem, a constraint is said to be

- (a) *redundant* if it does not change the feasible set, and *non-redundant* or *relevant* otherwise;
- (b) *non-binding* if it does not change the extreme value, and *binding* otherwise;
- (c) *active* if it is an inequality constraint that holds as an equality at the extremizer, and *inactive* otherwise.

Example 4.3. Consider $f: \mathbb{R}^2 \rightarrow \mathbb{R}$, $f(x, y) := y$. Suppose that we wish to minimize f over the unbounded w -shaped region

$$W := \{(x, y) \in \mathbb{R}^2 \mid y \geq (x^2 - 1)^2\}.$$

Over W , f takes the minimum value 0 at $(x, y) = (\pm 1, 0)$. Note that the inequality constraint $y \geq (x^2 - 1)^2$ is an active constraint. The additional constraint $y \geq 0$ would be redundant with respect to this feasible set W , and hence also non-binding. The additional constraint $x > 0$ would be non-redundant, but also non-binding, since it excludes the previous minimizer at $(x, y) = (-1, 0)$ but not the one at $(x, y) = (1, 0)$. Similarly, the additional equality constraint $y = (x^2 - 1)^2$ would be non-redundant and non-binding.

The importance of these concepts for UQ lies in the fact that many UQ problems are, in part or in whole, optimization problems: a good example is the calibration of parameters in a model in order to best explain some observed data. Each piece of information about the problem (e.g. a hypothesis about the form of the model, such as a physical law) can be seen as a constraint on that optimization problem. It is easy to imagine that each additional constraint may introduce additional difficulties in computing the parameters of best fit. Therefore, it is natural to want to exclude from consideration those constraints (pieces of information) that are merely complicating the solution process, and not actually determining the optimal parameters, and to have some terminology for describing the various ways in which this can occur.

4.2 Unconstrained Global Optimization

In general, finding a global minimizer of an arbitrary function is *very hard*, especially in high-dimensional settings and without nice features like convexity. Except in very simple settings like linear least squares (Section 4.6), it is necessary to construct an approximate

solution, and to do so iteratively; that is, one computes a sequence $(x_n)_{n \in \mathbb{N}}$ in \mathcal{X} such that x_n converges as $n \rightarrow \infty$ to an extremizer of the objective function within the feasible set. A simple example of a deterministic iterative method for finding the critical points, and hence extrema, of a smooth function is Newton's method:

Definition 4.4. Let \mathcal{X} be a normed vector space. Given a differentiable function $g: \mathcal{X} \rightarrow \mathcal{X}$ and an initial state x_0 , *Newton's method* for finding a zero of g is the sequence generated by the iteration

$$x_{n+1} := x_n - (\text{D}g(x_n))^{-1}g(x_n), \quad (4.1)$$

where $\text{D}g(x_n): \mathcal{X} \rightarrow \mathcal{X}$ is the Fréchet derivative of g at x_n . Newton's method is often applied to find critical points of $f: \mathcal{X} \rightarrow \mathbb{R}$, i.e. points where $\text{D}f$ vanishes, in which case the iteration is.

$$x_{n+1} := x_n - (\text{D}^2f(x_n))^{-1}\text{D}f(x_n). \quad (4.2)$$

(In (4.2), the second derivative (Hessian) $\text{D}^2f(x_n)$ is interpreted as a linear map $\mathcal{X} \rightarrow \mathcal{X}$ rather than a bilinear map $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$.)

Remark 4.5. (a) Newton's method for the determination of critical points of f amounts to local quadratic approximation: we model f about x_n using its Taylor expansion up to second order, and then take as x_{n+1} a critical point of this quadratic approximation. In particular, as shown in Exercise 4.3, Newton's method yields the exact minimizer of f in one iteration when f is in fact a quadratic function.
 (b) We will not dwell at this point on the important practical issue of numerical (and hence approximate) evaluation of derivatives for methods such as Newton iteration. However, this issue will be revisited in Section 10.2 in the context of sensitivity analysis.

For objective functions $f: \mathcal{X} \rightarrow \mathbb{R} \cup \{\pm\infty\}$ that have little to no smoothness, or that have many local extremizers, it is often necessary to resort to random searches of the space \mathcal{X} . For such algorithms, there can only be a probabilistic guarantee of convergence. The rate of convergence and the degree of approximate optimality naturally depend upon features like randomness of the generation of new elements of \mathcal{X} and whether the extremizers of f are difficult to reach, e.g. because they are located in narrow 'valleys'. We now describe three very simple random iterative algorithms for minimization of a prescribed objective function \mathbf{f} , in order to illustrate some of the relevant issues. For simplicity, suppose that \mathbf{f} has a unique global minimizer \mathbf{x}_{\min} and write \mathbf{f}_{\min} for $\mathbf{f}(\mathbf{x}_{\min})$.

Algorithm 4.6 (Random sampling). For simplicity, the following algorithm runs for $\mathbf{n_max}$ steps with no convergence checks. The algorithm returns an approximate minimizer $\mathbf{x_best}$ along with the corresponding value of \mathbf{f} . Suppose that `random()` generates independent samples of \mathcal{X} from a probability measure μ with support \mathcal{X} .

```
f_best = +inf
n = 0
while n < n_max:
    x_new = random()
    f_new = f(x_new)
    if f_new < f_best:
        x_best = x_new
        f_best = f_new
    n = n + 1
return [x_best, f_best]
```

A weakness of Algorithm 4.6 is that it completely neglects local information about \mathbf{f} . Even if the current state $\mathbf{x_best}$ is very close to the global minimizer \mathbf{x}_{\min} , the algorithm may continue to sample points $\mathbf{x_new}$ that are very far away and have $\mathbf{f}(\mathbf{x_new}) \gg \mathbf{f}(\mathbf{x_best})$. It would be preferable to explore a neighbourhood of $\mathbf{x_best}$ more thoroughly

and hence find a better approximation of $[x_{\min}, f_{\min}]$. The next algorithm attempts to rectify this deficiency.

Algorithm 4.7 (Random walk). As before, this algorithm runs for n_{\max} steps. The algorithm returns an approximate minimizer x_{best} along with the corresponding value of f . Suppose that an initial state x_0 is given, and that $\text{jump}()$ generates independent samples of \mathcal{X} from a probability measure μ with support equal to the unit ball of \mathcal{X} .

```
x_best = x0
f_best = f(x_best)
n = 0
while n < n_max:
    x_new = x_best + jump()
    f_new = f(x_new)
    if f_new < f_best:
        x_best = x_new
        f_best = f_new
    n = n + 1
return [x_best, f_best]
```

Algorithm 4.7 also has a weakness: since the state is only ever updated to states with a strictly lower value of f , and only looks for new states within unit distance of the current one, the algorithm is prone to becoming stuck in local minima if they are surrounded by wells that are sufficiently wide, even if they are very shallow. The next algorithm, the *simulated annealing* method of Kirkpatrick et al. (1983), attempts to rectify this problem by allowing the optimizer to make some ‘uphill’ moves, which can be accepted or rejected according to comparison of a uniformly distributed random variable with a user-prescribed acceptance probability function. Therefore, in the simulated annealing algorithm, a distinction is made between the current state x of the algorithm and the best state so far, x_{best} ; unlike in the previous two algorithms, proposed states x_{new} may be accepted and become x even if $f(x_{\text{new}}) > f(x_{\text{best}})$. The idea is to introduce a parameter T , to be thought of as ‘temperature’: the optimizer starts off ‘hot’, and ‘uphill’ moves are likely to be accepted; by the end of the calculation, the optimizer is relatively ‘cold’, and ‘uphill’ moves are unlikely to be accepted.

Algorithm 4.8 (Simulated annealing). Suppose that an initial state x_0 is given. Suppose also that functions $\text{temperature}()$, $\text{neighbour}()$ and $\text{acceptance_prob}()$ have been specified. Suppose that $\text{uniform}()$ generates independent samples from the uniform distribution on $[0, 1]$. Then the simulated annealing algorithm is

```
x = x0
fx = f(x)
x_best = x
f_best = fx
n = 0
while n < n_max:
    T = temperature(n / n_max)
    x_new = neighbour(x)
    f_new = f(x_new)
    if acceptance_prob(fx, f_new, T) > uniform():
        x = x_new
        fx = f_new
    if f_new < f_best:
        x_best = x_new
        f_best = f_new
    n = n + 1
return [x_best, f_best]
```

Like Algorithm 4.6, the simulated annealing method can guarantee to find the global minimizer of \mathbf{f} provided that the `neighbour()` function allows full exploration of the state space and the maximum run time `n_max` is large enough. However, the difficulty lies in coming up with functions `temperature()` and `acceptance_prob()` such that the algorithm finds the global minimizer in reasonable time: simulated annealing calculations can be extremely computationally costly. A commonly used acceptance probability function P is the one from the *Metropolis–Hastings algorithm* (see also Section 9.5):

$$P(e, e', T) = \begin{cases} 1, & \text{if } e' < e, \\ \exp(-(e' - e)/T), & \text{if } e' \geq e. \end{cases}$$

There are, however, many other choices; in particular, it is not necessary to automatically accept downhill moves, and it is permissible to have $P(e, e', T) < 1$ for $e' < e$.

4.3 Constrained Optimization

It is well known that the unconstrained extremizers of smooth enough functions must be critical points, i.e. points where the derivative vanishes. The following theorem, the Lagrange multiplier theorem, states that the constrained minimizers of a smooth enough function, subject to smooth enough equality constraints, are critical points of an appropriately generalized function:

Theorem 4.9 (Lagrange multipliers). *Let \mathcal{X} and \mathcal{Y} be real Banach spaces. Let $U \subseteq \mathcal{X}$ be open and let $f \in C^1(U; \mathbb{R})$. Let $g \in C^1(U; \mathcal{Y})$, and suppose that $x \in U$ is a constrained extremizer of f subject to the constraint that $g(x) = 0$. Suppose also that the Fréchet derivative $Dg(x): \mathcal{X} \rightarrow \mathcal{Y}$ is surjective. Then there exists a Lagrange multiplier $\lambda \in \mathcal{Y}'$ such that (x, λ) is an unconstrained critical point of the Lagrangian \mathcal{L} defined by*

$$U \times \mathcal{Y}' \ni (x, \lambda) \mapsto \mathcal{L}(x, \lambda) := f(x) + \langle \lambda | g(x) \rangle \in \mathbb{R}.$$

i.e. $Df(x) = -\lambda \circ Dg(x)$ as linear maps from \mathcal{X} to \mathbb{R} .

The corresponding result for inequality constraints is the Karush–Kuhn–Tucker theorem, which we state here for a finite system of inequality constraints:

Theorem 4.10 (Karush–Kuhn–Tucker). *Let U be an open subset of a Banach space \mathcal{X} , and let $f \in C^1(U; \mathbb{R})$ and $h \in C^1(U; \mathbb{R}^m)$. Suppose that $x \in U$ is a local minimizer of f subject to the inequality constraints $h_i(x) \leq 0$ for $i = 1, \dots, m$, and suppose that $Dh(x): \mathcal{X} \rightarrow \mathbb{R}^m$ is surjective. Then there exists $\mu = (\mu_1, \dots, \mu_m) \in (\mathbb{R}^m)'$ such that*

$$-Df(x) = \mu \circ Dh(x),$$

where μ satisfies the dual feasibility criteria $\mu_i \geq 0$ and the complementary slackness criteria $\mu_i h_i(x) = 0$ for $i = 1, \dots, m$.

The Lagrange and Karush–Kuhn–Tucker theorems can be combined to incorporate equality constraints g_i and inequality constraints h_j . Strictly speaking, the validity of the Karush–Kuhn–Tucker theorem also depends upon some regularity conditions on the constraints called *constraint qualification conditions*, of which there are many variations that can easily be found in the literature. A very simple one is that if g_i and h_j are affine functions, then no further regularity is needed; another is that the gradients of the active inequality constraints and the gradients of the equality constraints be linearly independent at the optimal point x .

Numerical Implementation of Constraints. In the numerical treatment of constrained optimization problems, there are many ways to implement constraints, not all of which actually *enforce* the constraints in the sense of ensuring that trial states \mathbf{x}_{new} , accepted states \mathbf{x} , or even the final solution \mathbf{x}_{best} are actually members of the feasible set. For definiteness, consider the constrained minimization problem

$$\begin{aligned} & \text{minimize: } f(x) \\ & \text{with respect to: } x \in \mathcal{X} \\ & \text{subject to: } c(x) \leq 0 \end{aligned}$$

for some functions $f, c: \mathcal{X} \rightarrow \mathbb{R} \cup \{\pm\infty\}$. One way of seeing the constraint ' $c(x) \leq 0$ ' is as a Boolean true/false condition: either the inequality is satisfied, or it is not. Supposing that `neighbour(x)` generates new (possibly infeasible) elements of \mathcal{X} given a current state \mathbf{x} , one approach to generating feasible trial states \mathbf{x}_{new} is the following:

```
x' = neighbour(x)
while c(x') > 0:
    x' = neighbour(x)
x_new = x'
```

However, this accept/reject approach is extremely wasteful: if the feasible set is very small, then \mathbf{x}' will 'usually' be rejected, thereby wasting a lot of computational time, and this approach takes no account of how 'nearly feasible' an infeasible \mathbf{x}' might be.

One alternative approach is to use *penalty functions*: instead of considering the constrained problem of minimizing $f(x)$ subject to $c(x) \leq 0$, one can consider the unconstrained problem of minimizing $x \mapsto f(x) + p(x)$, where $p: \mathcal{X} \rightarrow [0, \infty)$ is some function that equals zero on the feasible set and takes larger values the 'more' the constraint inequality $c(x) \leq 0$ is violated, e.g., for $\mu > 0$,

$$p_\mu(x) = \begin{cases} 0, & \text{if } c(x) \leq 0, \\ \exp(c(x)/\mu) - 1, & \text{if } c(x) > 0. \end{cases}$$

The hope is that (a) the minimization of $f + p_\mu$ over all of \mathcal{X} is easy, and (b) as $\mu \rightarrow 0$, minimizers of $f + p_\mu$ converge to minimizers of f on the original feasible set. The penalty function approach is attractive, but the choice of penalty function is rather ad hoc, and issues can easily arise of competition between the penalties corresponding to multiple constraints.

An alternative to the use of penalty functions is to construct *constraining functions* that enforce the constraints exactly. That is, we seek a function $\mathbf{C}()$ that takes as input a possibly infeasible \mathbf{x}' and returns some $\mathbf{x}_{\text{new}} = \mathbf{C}(\mathbf{x}')$ that is guaranteed to satisfy the constraint $c(\mathbf{x}_{\text{new}}) \leq 0$. For example, suppose that $\mathcal{X} = \mathbb{R}^n$ and the feasible set is the Euclidean unit ball, so the constraint is

$$c(x) := \|x\|_2^2 - 1 \leq 0.$$

Then a suitable constraining function could be

$$C(x) := \begin{cases} x, & \text{if } \|x\|_2 \leq 1, \\ x/\|x\|_2, & \text{if } \|x\|_2 > 1. \end{cases}$$

Constraining functions are very attractive because the constraints are treated exactly. However, they must often be designed on a case-by-case basis for each constraint function c , and care must be taken to ensure that multiple constraining functions interact well and do not unduly favour parts of the feasible set over others; for example, the above constraining function C maps the entire infeasible set to the unit sphere, which might be considered undesirable in certain settings, and so a function such as

$$\tilde{C}(x) := \begin{cases} x, & \text{if } \|x\|_2 \leq 1, \\ x/\|x\|_2^2, & \text{if } \|x\|_2 > 1. \end{cases}$$

might be more appropriate. Finally, note that the original accept/reject method of finding feasible states is a constraining function in this sense, albeit a very inefficient one.

4.4 Convex Optimization

The topic of this section is *convex optimization*. As will be seen, convexity is a powerful property that makes optimization problems tractable to a much greater extent than any amount of smoothness (which still permits local minima) or low-dimensionality can do.

In this section, \mathcal{X} will be a normed vector space. (More generally, the properties that are of importance to the discussion hold for any Hausdorff, locally convex topological vector space.) Given two points x_0 and x_1 of \mathcal{X} and $t \in [0, 1]$, x_t will denote the *convex combination*

$$x_t := (1 - t)x_0 + tx_1.$$

More generally, given points x_0, \dots, x_n of a vector space, a sum of the form

$$\alpha_0 x_0 + \dots + \alpha_n x_n$$

is called a *linear combination* if the α_i are any field elements, an *affine combination* if their sum is 1, and a *convex combination* if they are non-negative and sum to 1.

- Definition 4.11.** (a) A subset $K \subseteq \mathcal{X}$ is a *convex set* if, for all $x_0, x_1 \in K$ and $t \in [0, 1]$, $x_t \in K$; it is said to be *strictly convex* if $x_t \in \overset{\circ}{K}$ whenever x_0 and x_1 are distinct points of \bar{K} and $t \in (0, 1)$.
- (b) An *extreme point* of a convex set K is a point of K that cannot be written as a non-trivial convex combination of distinct elements of K ; the set of all extreme points of K is denoted $\text{ext}(K)$.
- (c) The *convex hull* $\text{co}(S)$ (resp. *closed convex hull* $\overline{\text{co}}(S)$) of $S \subseteq \mathcal{X}$ is defined to be the intersection of all convex (resp. closed and convex) subsets of \mathcal{X} that contain S .

- Example 4.12.** (a) The square $[-1, 1]^2$ is a convex subset of \mathbb{R}^2 , but is not strictly convex, and its extreme points are the four vertices $(\pm 1, \pm 1)$.
- (b) The closed unit disc $\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$ is a strictly convex subset of \mathbb{R}^2 , and its extreme points are the points of the unit circle $\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$.
- (c) If $p_0, \dots, p_d \in \mathcal{X}$ are distinct points such that $p_1 - p_0, \dots, p_d - p_0$ are linearly independent, then their (closed) convex hull is called a *d-dimensional simplex*. The points p_0, \dots, p_d are the extreme points of the simplex.
- (d) See Figure 4.1 for further examples.

Example 4.13. $\mathcal{M}_1(\mathcal{X})$ is a convex subset of the space of all (signed) Borel measures on \mathcal{X} . The extremal probability measures are the *zero-one measures*, i.e. those for which, for every measurable set $E \subseteq \mathcal{X}$, $\mu(E) \in \{0, 1\}$. Furthermore, as will be discussed in Chapter 14, if \mathcal{X} is, say, a Polish space, then the zero-one measures (and hence the extremal probability measures) on \mathcal{X} are the Dirac point masses. Indeed, in this situation,

$$\mathcal{M}_1(\mathcal{X}) = \overline{\text{co}}(\{\delta_x \mid x \in \mathcal{X}\}) \subseteq \mathcal{M}_\pm(\mathcal{X}).$$

The principal reason to confine attention to normed spaces^[4.1] \mathcal{X} is that it is highly inconvenient to have to work with spaces for which the following ‘common sense’ results do not hold:

Theorem 4.14 (Kreĭn–Milman). *Let $K \subseteq \mathcal{X}$ be compact and convex. Then K is the closed convex hull of its extreme points.*

^[4.1]Or, more generally, Hausdorff, locally convex, topological vector spaces.

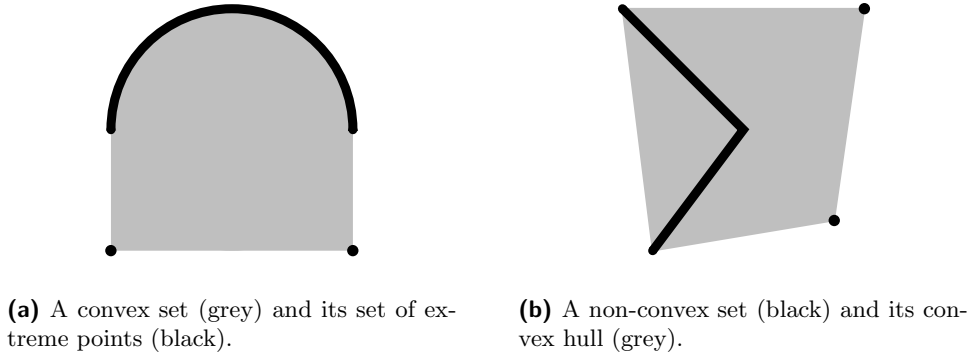


Figure 4.1: Convex sets, extreme points, and convex hulls of some subsets of the plane \mathbb{R}^2 .

Theorem 4.15 (Choquet–Bishop–de Leeuw). *Let $K \subseteq \mathcal{X}$ be compact and convex, and let $c \in K$. Then there exists a probability measure p supported on $\text{ext}(K)$ such that, for all affine functions f on K ,*

$$f(c) = \int_{\text{ext}(K)} f(e) \, dp(e).$$

The point c in Theorem 4.15 is called a *barycentre* of the set K , and the probability measure p is said to *represent* the point c . Informally speaking, the Kreĭn–Milman and Choquet–Bishop–de Leeuw theorems together ensure that a compact, convex subset K of a topologically respectable space is entirely characterized by its set of extreme points in the following sense: every point of K can be obtained as an average of extremal points of K , and, indeed, the value of any affine function at any point of K can be obtained as an average of its values at the extremal points in the same way.

Definition 4.16. Let $K \subseteq \mathcal{X}$ be convex. A function $f: K \rightarrow \mathbb{R} \cup \{\pm\infty\}$ is a *convex function* if, for all $x_0, x_1 \in K$ and $t \in [0, 1]$,

$$f(x_t) \leq (1-t)f(x_0) + tf(x_1), \quad (4.3)$$

and is called a *strictly convex function* if, for all distinct $x_0, x_1 \in K$ and $t \in (0, 1)$,

$$f(x_t) < (1-t)f(x_0) + tf(x_1).$$

The inequality (4.3) defining convexity can be seen as a special case — with $X \sim \mu$ supported on two points x_0 and x_1 — of the following result:

Theorem 4.17 (Jensen). *Let $(\Theta, \mathcal{F}, \mu)$ be a probability space, let $K \subseteq \mathcal{X}$ and $f: K \rightarrow \mathbb{R} \cup \{\pm\infty\}$ be convex, and let $X \in L^1(\Theta, \mu; \mathcal{X})$ take values in K . Then*

$$f(\mathbb{E}_\mu[X]) \leq \mathbb{E}_\mu[f(X)], \quad (4.4)$$

where $\mathbb{E}_\mu[X] \in \mathcal{X}$ is defined by the relation $\langle \ell | \mathbb{E}_\mu[X] \rangle = \mathbb{E}_\mu[\langle \ell | X \rangle]$ for every $\ell \in \mathcal{X}'$. Furthermore, if f is strictly convex, then equality holds in (4.4) if and only if X is μ -almost surely constant.

It is straightforward to see that $f: K \rightarrow \mathbb{R} \cup \{\pm\infty\}$ is convex (resp. strictly convex) if and only if its *epigraph*

$$\text{epi}(f) := \{(x, v) \in K \times \mathbb{R} \mid v \geq f(x)\}$$

is a convex (resp. strictly convex) subset of $K \times \mathbb{R}$. Furthermore, twice-differentiable convex functions are easily characterized in terms of their second derivative (Hessian):

Theorem 4.18. *Let $f: K \rightarrow \mathbb{R}$ be twice continuously differentiable on an open, convex set K . Then f is convex if and only if $D^2f(x)$ is positive semi-definite for all $x \in K$. If $D^2f(x)$ is positive definite for all $x \in K$, then f is strictly convex, though the converse is false.*

Convex functions have many convenient properties with respect to minimization and maximization:

Theorem 4.19. *Let $f: K \rightarrow \mathbb{R}$ be a convex function on a convex set $K \subseteq \mathcal{X}$. Then*

- (a) *any local minimizer of f in K is also a global minimizer;*
- (b) *the set $\arg \min_K f$ of global minimizers of f in K is convex;*
- (c) *if f is strictly convex, then it has at most one global minimizer in K ;*
- (d) *if K is also compact, then f has the same maximum values on K and $\text{ext}(K)$.*

Proof. (a) Suppose that x_0 is a local minimizer of f in K that is not a global minimizer: that is, suppose that x_0 is a minimizer of f in some open neighbourhood N of x_0 , and also that there exists $x_1 \in K \setminus N$ such that $f(x_1) < f(x_0)$. Then, for sufficiently small $t > 0$, $x_t \in N$, but convexity implies that

$$f(x_t) \leq (1-t)f(x_0) + tf(x_1) < (1-t)f(x_0) + tf(x_0) = f(x_0),$$

which contradicts the assumption that x_0 is a minimizer of f in N .

- (b) Suppose that $x_0, x_1 \in K$ are global minimizers of f . Then, for all $t \in [0, 1]$, $x_t \in K$ and

$$f(x_0) \leq f(x_t) \leq (1-t)f(x_0) + tf(x_1) = f(x_0).$$

Hence, $x_t \in \arg \min_K f$, and so $\arg \min_K f$ is convex.

- (c) Suppose that $x_0, x_1 \in K$ are distinct global minimizers of f , and let $t \in (0, 1)$. Then $x_t \in K$ and

$$f(x_0) \leq f(x_t) < (1-t)f(x_0) + tf(x_1) = f(x_0),$$

which is a contradiction. Hence, f has at most one minimizer in K .

- (d) Suppose that $c \in K \setminus \text{ext}(K)$ has $f(c) > \sup_{\text{ext}(K)} f$. By Theorem 4.15, there exists a probability measure p on $\text{ext}(K)$ such that, for all affine functions ℓ on K ,

$$\ell(c) = \int_{\text{ext}(K)} \ell(x) \, dp(x).$$

i.e. $c = \mathbb{E}_{X \sim p}[X]$. Then Jensen's inequality implies that

$$\mathbb{E}_{X \sim p}[f(X)] \geq f(c) > \sup_{\text{ext}(K)} f,$$

which is a contradiction. Hence, since $\sup_K f \geq \sup_{\text{ext}(K)} f$, f must have the same maximum value on $\text{ext}(K)$ as it does on K . ■




Remark 4.20. Note well that Theorem 4.19 does not assert the existence of minimizers, which requires non-emptiness and compactness of K , and lower semicontinuity of f . For example:

- the exponential function on \mathbb{R} is strictly convex, continuous and bounded below by 0 yet has no minimizer;
- the interval $[-1, 1]$ is compact, and the function $f: [-1, 1] \rightarrow \mathbb{R} \cup \{\pm\infty\}$ defined by

$$f(x) := \begin{cases} x, & \text{if } |x| < \frac{1}{2}, \\ +\infty, & \text{if } |x| \geq \frac{1}{2}, \end{cases}$$

is convex, yet f has no minimizer — although $\inf_{x \in [-1, 1]} f(x) = -\frac{1}{2}$, there is no x for which $f(x)$ attains this infimal value.

Definition 4.21. A *convex optimization problem* (or *convex program*) is a minimization problem in which the objective function and all constraints are equalities or inequalities with respect to convex functions.

Remark 4.22. (a) Beware of the common pitfall of saying that a convex program is simply the minimization of a convex function over a convex set. Of course, by Theorem 4.19, such minimization problems are nicer than general minimization problems, but bona fide convex programs are an even nicer special case. 
 (b) In practice, many problems are not obviously convex programs, but can be transformed into convex programs by e.g. a cunning change of variables. Being able to spot the right equivalent problem is a major part of the art of optimization.

It is difficult to overstate the importance of convexity in making optimization problems tractable. Indeed, it has been remarked that lack of convexity is a much greater obstacle to tractability than high dimension. There are many powerful methods for the solution of convex programs, with corresponding standard software libraries such as `cvxopt`. For example, *interior point methods* explore the interior of the feasible set in search of the solution to the convex program, while being kept away from the boundary of the feasible set by a *barrier function*. The discussion that follows is only intended as an outline; for details, see Boyd and Vandenberghe (2004, Chapter 11).

Consider the convex program

$$\begin{aligned} & \text{minimize: } f(x) \\ & \text{with respect to: } x \in \mathbb{R}^n \\ & \text{subject to: } c_i(x) \leq 0 \quad \text{for } i = 1, \dots, m, \end{aligned}$$

where the functions $f, c_1, \dots, c_m: \mathbb{R}^n \rightarrow \mathbb{R}$ are all convex and differentiable. Let F denote the feasible set for this program. Let $0 < \mu \ll 1$ be a small scalar, called the *barrier parameter*, and define the *barrier function* associated to the program by

$$B(x; \mu) := f(x) - \mu \sum_{i=1}^m \log c_i(x).$$

Note that $B(\cdot; \mu)$ is strictly convex for $\mu > 0$, that $B(x; \mu) \rightarrow +\infty$ as $x \rightarrow \partial F$, and that $B(\cdot; 0) = f$; therefore, the unique minimizer x_μ^* of $B(\cdot; \mu)$ lies in $\overset{\circ}{F}$ and (hopefully) converges to the minimizer of the original problem as $\mu \rightarrow 0$. Indeed, using arguments based on convex duality, one can show that

$$f(x_\mu^*) - \inf_{x \in F} f(x) \leq m\mu.$$

The strictly convex problem of minimizing $B(\cdot; \mu)$ can be solved approximately using Newton's method. In fact, however, one settles for a partial minimization of $B(\cdot; \mu)$ using only one or two steps of Newton's method, then decreases μ to μ' , performs another partial minimization of $B(\cdot; \mu')$ using Newton's method, and so on in this alternating fashion.

4.5 Linear Programming

Theorem 4.19 has the following immediate corollary for the minimization and maximization of affine functions on convex sets:

Corollary 4.23. Let $\ell: K \rightarrow \mathbb{R}$ be a continuous affine function on a non-empty, compact, convex set $K \subseteq \mathcal{X}$. Then

$$\text{ext}\{\ell(x) \mid x \in K\} = \text{ext}\{\ell(x) \mid x \in \text{ext}(K)\}.$$

That is, ℓ has the same minimum and maximum values over both K and the set of extreme points of K .

Definition 4.24. A *linear program* is an optimization problem of the form

$$\begin{aligned} &\text{extremize: } f(x) \\ &\text{with respect to: } x \in \mathbb{R}^p \\ &\text{subject to: } g_i(x) \leq 0 \quad \text{for } i = 1, \dots, q, \end{aligned}$$

where the functions $f, g_1, \dots, g_q: \mathbb{R}^p \rightarrow \mathbb{R}$ are all affine functions. Linear programs are often written in the *canonical form*

$$\begin{aligned} &\text{maximize: } c \cdot x \\ &\text{with respect to: } x \in \mathbb{R}^n \\ &\text{subject to: } Ax \leq b \\ &\quad x \geq 0, \end{aligned}$$

where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ are given, and the two inequalities are interpreted componentwise. (Conversion to canonical form, and in particular the introduction of the non-negativity constraint $x \geq 0$, is accomplished by augmenting the original $x \in \mathbb{R}^p$ with additional variables called *slack variables* to form the extended variable $x \in \mathbb{R}^n$.)

Note that the feasible set for a linear program is an intersection of finitely many half-spaces of \mathbb{R}^n , i.e. a *polytope*. This polytope may be empty, in which case the constraints are mutually contradictory and the program is said to be *infeasible*. Also, the polytope may be unbounded in the direction of c , in which case the extreme value of the problem is infinite.

Since linear programs are special cases of convex programs, methods such as interior point methods are applicable to linear programs as well. Such methods approach the optimum point x^* , which is necessarily an extremal element of the feasible polytope, from the interior of the feasible polytope. Historically, however, such methods were preceded by methods such as Dantzig's simplex algorithm, which, sets out to directly explore the set of extreme points in a (hopefully) efficient way. Although the theoretical worst-case complexity of the simplex method as formulated by Dantzig is exponential in n and m , in practice the simplex method is remarkably efficient (polynomial running time) provided that certain precautions are taken to avoid pathologies such as 'stalling'.

4.6 Least Squares

An elementary example of convex programming is unconstrained quadratic minimization, otherwise known as *least squares*. Least squares minimization plays a central role in elementary statistical estimation, as will be demonstrated by the Gauss–Markov theorem (Theorem 6.2). The next three results show that least squares problems have unique solutions, which are given in terms of an orthogonality criterion, which in turn reduces to a system of linear equations, the *normal equations*.

Lemma 4.25. *Let K be a non-empty, closed, convex subset of a Hilbert space \mathcal{H} . Then, for each $y \in \mathcal{H}$, there is a unique element $\hat{x} = \Pi_K y \in K$ such that*

$$\hat{x} \in \arg \min_{x \in K} \|y - x\|.$$

Proof. By Exercise 4.1, the function $J: \mathcal{H} \rightarrow [0, \infty)$ defined by $J(x) := \|y - x\|^2$ is strictly convex, and hence it has at most one minimizer in K . Therefore, it only remains to show that J has at least one minimizer in K . Since J is bounded below (on \mathcal{H} , not just on K), J has a sequence of approximate minimizers: let

$$I := \inf_{x \in K} \|y - x\|^2, \quad I^2 \leq \|y - x_n\|^2 \leq I^2 + \frac{1}{n}.$$

By the parallelogram identity for the Hilbert norm $\|\cdot\|$,

$$\|(y - x_m) + (y - x_n)\|^2 + \|(y - x_m) - (y - x_n)\|^2 = 2\|y - x_m\|^2 + 2\|y - x_n\|^2,$$

and hence

$$\|2y - (x_m + x_n)\|^2 + \|x_n - x_m\|^2 \leq 4I^2 + \frac{2}{n} + \frac{2}{m}.$$

Since K is convex, $\frac{1}{2}(x_m + x_n) \in K$, so the first term on the left-hand side above is bounded below as follows:

$$\|2y - (x_m + x_n)\|^2 = 4 \left\| y - \frac{x_m + x_n}{2} \right\|^2 \geq 4I^2.$$

Hence,

$$\|x_n - x_m\|^2 \leq 4I^2 + \frac{2}{n} + \frac{2}{m} - 4I^2 = \frac{2}{n} + \frac{2}{m},$$

and so the sequence $(x_n)_{n \in \mathbb{N}}$ is Cauchy; since \mathcal{H} is complete and K is closed, this sequence converges to some $\hat{x} \in K$. Since the norm $\|\cdot\|$ is continuous, $\|y - \hat{x}\| = I$. ■

Lemma 4.26 (Orthogonality of the residual). *Let V be a closed subspace of a Hilbert space \mathcal{H} and let $b \in \mathcal{H}$. Then $\hat{x} \in V$ minimizes the distance to b if and only if the residual $\hat{x} - b$ is orthogonal to V , i.e.*

$$\hat{x} = \arg \min_{x \in V} \|x - b\| \iff (\hat{x} - b) \perp V.$$

Proof. Let $J(x) := \frac{1}{2}\|x - b\|^2$, which has the same minimizers as $x \mapsto \|x - b\|$; by Lemma 4.25, such a minimizer exists and is unique. Suppose that $(x - b) \perp V$ and let $y \in V$. Then $y - x \in V$ and so $(y - x) \perp (x - b)$. Hence, by Pythagoras' theorem,

$$\|y - b\|^2 = \|y - x\|^2 + \|x - b\|^2 \geq \|x - b\|^2,$$

and so x minimizes J .

Conversely, suppose that x minimizes J . Then, for every $y \in V$,

$$0 = \left. \frac{\partial}{\partial \lambda} J(x + \lambda y) \right|_{\lambda=0} = \frac{1}{2} (\langle y, x - b \rangle + \langle x - b, y \rangle) = \operatorname{Re} \langle x - b, y \rangle$$

and, in the complex case,

$$0 = \left. \frac{\partial}{\partial \lambda} J(x + \lambda i y) \right|_{\lambda=0} = \frac{1}{2} (-i \langle y, x - b \rangle + i \langle x - b, y \rangle) = -\operatorname{Im} \langle x - b, y \rangle.$$

Hence, $\langle x - b, y \rangle = 0$, and since y was arbitrary, $(x - b) \perp V$. ■

Lemma 4.27 (Normal equations). *Let $A: \mathcal{H} \rightarrow \mathcal{K}$ be a linear operator between Hilbert spaces such that $\operatorname{ran} A \subseteq \mathcal{K}$ is closed. Then, given $b \in \mathcal{K}$,*

$$\hat{x} \in \arg \min_{x \in \mathcal{H}} \|Ax - b\|_{\mathcal{K}} \iff A^* A \hat{x} = A^* b, \quad (4.5)$$

the equations on the right-hand side being known as the normal equations. If, in addition, A is injective, then $A^ A$ is invertible and the least squares problem / normal equations have a unique solution.*

Proof. As a consequence of completeness, the only element of a Hilbert space that is orthogonal to every other element of the space is the zero element. Hence,

$\|Ax - b\|_{\mathcal{K}}$ is minimal

$$\iff (Ax - b) \perp Av \text{ for all } v \in \mathcal{H}$$

by Lemma 4.26

$$\iff \langle Ax - b, Av \rangle_{\mathcal{K}} = 0 \text{ for all } v \in \mathcal{H}$$

$$\iff \langle A^* Ax - A^* b, v \rangle_{\mathcal{H}} = 0 \text{ for all } v \in \mathcal{H}$$

$$\iff A^* Ax = A^* b$$

by completeness of \mathcal{H} ,

and this shows the equivalence (4.5).

By Proposition 3.16(d), $\ker A^* = (\text{ran } A)^\perp$. Therefore, the restriction of A^* to the range of A is injective. Hence, if A itself is injective, then it follows that A^*A is injective. Again by Proposition 3.16(d), $(\text{ran } A^*)^\perp = \ker A = \{0\}$, and since \mathcal{H} is complete, this implies that A^* is surjective. Since A is surjective onto its range, it follows that A^*A is surjective, and hence bijective and invertible. ■

Weighting and Regularization. It is common in practice that one does not want to minimize the \mathcal{K} -norm directly, but perhaps some re-weighted version of the \mathcal{K} -norm. This re-weighting is accomplished by a self-adjoint and positive definite^[4.2] operator $Q: \mathcal{K} \rightarrow \mathcal{K}$: we define a new inner product and norm on \mathcal{K} by

$$\begin{aligned}\langle k, k' \rangle_Q &:= \langle k, Qk' \rangle_{\mathcal{K}}, \\ \|k\|_Q &:= \langle k, k \rangle_Q^{1/2}.\end{aligned}$$

It is a standard fact that the self-adjoint operator Q possesses an operator square root, i.e. a self-adjoint $Q^{1/2}: \mathcal{K} \rightarrow \mathcal{K}$ such that $Q^{1/2}Q^{1/2} = Q$; for reasons of symmetry, it is common to express the inner product and norm induced by Q using this square root:

$$\begin{aligned}\langle k, k' \rangle_Q &= \langle Q^{1/2}k, Q^{1/2}k' \rangle_{\mathcal{K}}, \\ \|k\|_Q &= \|Q^{1/2}k\|_{\mathcal{K}}.\end{aligned}$$

We then consider the problem, given $b \in \mathcal{K}$, of finding $x \in \mathcal{H}$ to minimize

$$\frac{1}{2}\|Ax - b\|_Q^2 \equiv \frac{1}{2}\|Q^{1/2}(Ax - b)\|_{\mathcal{K}}^2.$$

Another situation that arises frequently in practice is that the normal equations do not have a unique solution (e.g. because A^*A is not invertible) and so it is necessary to select one by some means, or that one has some prior belief that ‘the right solution’ should be close to some initial guess x_0 . A technique that accomplishes both of these aims is *Tikhonov regularization* (known in the statistics literature as *ridge regression*). In this situation, we minimize the following sum of two quadratic functionals:

$$\frac{1}{2}\|Ax - b\|_{\mathcal{K}}^2 + \frac{1}{2}\|x - x_0\|_R^2,$$

where $R: \mathcal{H} \rightarrow \mathcal{H}$ is self-adjoint and positive definite, and $x_0 \in \mathcal{H}$.

These two modifications to ordinary least squares, weighting and regularization, can be combined. The normal equations for weighted and regularized least squares are easily derived from Lemma 4.27:

Theorem 4.28 (Normal equations for weighted and Tikhonov-regularized least squares). *Let \mathcal{H} and \mathcal{K} be Hilbert spaces, let $A: \mathcal{H} \rightarrow \mathcal{K}$ have closed range, let Q and R be self-adjoint and positive definite on \mathcal{K} and \mathcal{H} respectively, and let $b \in \mathcal{K}$, $x_0 \in \mathcal{H}$. Let*

$$J(x) := \frac{1}{2}\|Ax - b\|_Q^2 + \frac{1}{2}\|x - x_0\|_R^2.$$

Then

$$\hat{x} \in \arg \min_{x \in \mathcal{H}} J(x) \iff (A^*QA + R)\hat{x} = A^*Qb + Rx_0.$$

Proof. Exercise 4.4. ■

It is also interesting to consider regularizations that do not come from a Hilbert norm, but instead from some other function. As will be elaborated upon in Chapter 6, there is a strong connection between regularized optimization problems and inverse problems, and the choice of regularization in some sense describes the practitioner’s ‘prior beliefs’ about the structure of the solution.

^[4.2]If Q is not positive definite, but merely positive semi-definite and self-adjoint, then existence of solutions to the associated least squares problems still holds, but uniqueness can fail.

Nonlinear Least Squares and Gauss–Newton Iteration. It often occurs in practice that one wishes to find a vector of parameters $\theta \in \mathbb{R}^p$ such that a function $\mathbb{R}^k \ni x \mapsto f(x; \theta) \in \mathbb{R}^\ell$ best fits a collection of data points $\{(x_i, y_i) \in \mathbb{R}^k \times \mathbb{R}^\ell \mid i = 1, \dots, m\}$. For each candidate parameter vector θ , define the *residual vector*

$$r(\theta) := \begin{bmatrix} r_1(\theta) \\ \vdots \\ r_m(\theta) \end{bmatrix} = \begin{bmatrix} y_1 - f(x_1; \theta) \\ \vdots \\ y_m - f(x_m; \theta) \end{bmatrix} \in \mathbb{R}^m.$$

The aim is to find θ to minimize the objective function $J(\theta) := \|r(\theta)\|_2^2$. Let

$$A := \left[\begin{array}{ccc} \frac{\partial r_1(\theta)}{\partial \theta^1} & \cdots & \frac{\partial r_1(\theta)}{\partial \theta^p} \\ \vdots & \ddots & \vdots \\ \frac{\partial r_m(\theta)}{\partial \theta^1} & \cdots & \frac{\partial r_m(\theta)}{\partial \theta^p} \end{array} \right] \bigg|_{\theta=\theta_n} \in \mathbb{R}^{m \times p}$$

be the Jacobian matrix of the residual vector, and note that $A = -DF(\theta_n)$, where

$$F(\theta) := \begin{bmatrix} f(x_1; \theta) \\ \vdots \\ f(x_m; \theta) \end{bmatrix} \in \mathbb{R}^m.$$

Consider the first-order Taylor approximation

$$r(\theta) \approx r(\theta_n) + A(r(\theta) - r(\theta_n)).$$

Thus, to approximately minimize $\|r(\theta)\|_2$, we find $\delta := r(\theta) - r(\theta_n)$ that makes the right-hand side of the approximation equal to zero. This is an ordinary linear least squares problem, the solution of which is given by the normal equations as

$$\delta = (A^* A)^{-1} A^* r(\theta_n).$$

Thus, we obtain the *Gauss–Newton iteration* for a sequence $(\theta_n)_{n \in \mathbb{N}}$ of approximate minimizers of J :

$$\begin{aligned} \theta_{n+1} &:= \theta_n - (A^* A)^{-1} A^* r(\theta_n) \\ &= \theta_n + ((DF(\theta_n))^* (DF(\theta_n)))^{-1} (DF(\theta_n))^* r(\theta_n). \end{aligned}$$

In general, the Gauss–Newton iteration is not guaranteed to converge to the exact solution, particularly if δ is ‘too large’, in which case it may be appropriate to use a judiciously chosen small positive multiple of δ . The use of Tikhonov regularization in this context is known as the *Levenberg–Marquardt algorithm* or *trust region* method, and the small multiplier applied to δ is essentially the reciprocal of the Tikhonov regularization parameter.

4.7 Bibliography

The book of Boyd and Vandenberghe (2004) is an excellent reference on the theory and practice of convex optimization, as is the associated software library *cvxopt*. The classic reference for convex analysis in general is the monograph of Rockafellar (1997); a more recent text is that of Krantz (2015). A good short reference on Choquet theory is the book of Phelps (2001); in particular, Theorems 4.14 and 4.15 are due to Krein and Milman (1940) and Bishop and de Leeuw (1959) respectively. A standard reference on numerical methods for optimization is the book of Nocedal and Wright (2006). The Banach space version of the Lagrange multiplier theorem, Theorem 4.9, can be found in Zeidler (1995,

Section 4.14). Theorem 4.10 originates with Karush (1939) and Kuhn and Tucker (1951); see, e.g., Gould and Tolle (1975) for discussion of the infinite-dimensional version.

For constrained global optimization in the absence of ‘nice’ features, particularly for the UQ methods in Chapter 14, variations upon the genetic evolution approach, e.g. the differential evolution algorithm (Price et al., 2005; Storn and Price, 1997), have proved up to the task of producing robust results, if not always quick ones. There is no ‘one size fits all’ approach to constrained global optimization: it is basically impossible to be quick, robust, and general all at the same time.

In practice, it is very useful to work using an optimization framework that provides easy interfaces to many optimization methods, with easy interchange among strategies for population generation, enforcement of constraints, termination criteria, and so on: see, for example, the DAKOTA (Adams et al., 2014) and Mystic (McKerns et al., 2009, 2011) projects.

4.8 Exercises

Exercise 4.1. Let $\|\cdot\|$ be a norm on a vector space \mathcal{V} , and fix $\bar{x} \in \mathcal{V}$. Show that the function $J: \mathcal{V} \rightarrow [0, \infty)$ defined by $J(x) := \|x - \bar{x}\|$ is convex, and that $J(x) := \frac{1}{2}\|x - \bar{x}\|^2$ is strictly convex if the norm is induced by an inner product. Give an example of a norm for which $J(x) := \frac{1}{2}\|x - \bar{x}\|^2$ is not strictly convex.

Exercise 4.2. Let K be a non-empty, closed, convex subset of a Hilbert space \mathcal{H} . Lemma 4.25 shows that there is a well-defined function $\Pi_K: \mathcal{H} \rightarrow K$ that assigns to each $y \in \mathcal{H}$ the unique $\Pi_K y \in K$ that is closest to y with respect to the norm on \mathcal{H} .

(a) Prove the variational inequality that $x = \Pi_K y$ if and only if $x \in K$ and

$$\langle x, z - x \rangle \geq \langle y, z - x \rangle \quad \text{for all } z \in K.$$

(b) Prove that Π_K is non-expansive, i.e.

$$\|\Pi_K y_1 - \Pi_K y_2\| \leq \|y_1 - y_2\| \quad \text{for all } y_1, y_2 \in \mathcal{H},$$

and hence a continuous function.

Exercise 4.3. Let $A: \mathcal{H} \rightarrow \mathcal{K}$ be a linear operator between Hilbert spaces such that $\text{ran } A$ is a closed subspace of \mathcal{K} , let $Q: \mathcal{K} \rightarrow \mathcal{K}$ be self-adjoint and positive-definite, and let $b \in \mathcal{K}$. Let

$$J(x) := \frac{1}{2}\|Ax - b\|_Q^2$$

Calculate the gradient and Hessian (second derivative) of J . Hence show that, regardless of the initial condition $x_0 \in \mathcal{H}$, Newton’s method finds the minimum of J in one step.

Exercise 4.4. Prove Theorem 4.28. Hint: Consider the operator from \mathcal{H} into $\mathcal{K} \oplus \mathcal{H}$ given by

$$x \mapsto \begin{bmatrix} Q^{1/2}Ax \\ R^{1/2}x \end{bmatrix}.$$

Chapter 5

Measures of Information and Uncertainty

As we know, there are known knowns. There are things we know we know. We also know there are known unknowns. That is to say we know there are some things we do not know. But there are also unknown unknowns, the ones we don't know we don't know.

DONALD RUMSFELD

This chapter briefly summarizes some basic numerical measures of uncertainty, from interval bounds to information-theoretic quantities such as (Shannon) information and entropy. This discussion then naturally leads to consideration of distances (and distance-like functions) between probability measures.

5.1 The Existence of Uncertainty

At a very fundamental level, the first level in understanding the uncertainties affecting some system is to identify the sources of uncertainty. Sometimes, this can be a challenging task because there may be so much lack of knowledge about, e.g. the relevant physical mechanisms, that one does not even know what a *list* of the important parameters would be, let alone what uncertainty one has about their *values*. The presence of such so-called *unknown unknowns* is of major concern in high-impact settings like risk assessment.

One way of assessing the presence of unknown unknowns is that if one subscribes to a deterministic view of the universe in which reality maps inputs $x \in \mathcal{X}$ to outputs $y = f(x) \in \mathcal{Y}$ by a well-defined single-valued function $f: \mathcal{X} \rightarrow \mathcal{Y}$, then unknown unknowns are additional variables $z \in \mathcal{Z}$ whose existence one infers from contradictory observations like

$$f(x) = y_1 \quad \text{and} \quad f(x) = y_2 \neq y_1.$$

Unknown unknowns explain away this contradiction by asserting the existence of a space \mathcal{Z} containing distinct elements z_1 and z_2 , that in fact f is a function $f: \mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{Y}$, and that the observations were actually

$$f(x, z_1) = y_1 \quad \text{and} \quad f(x, z_2) = y_2.$$

Of course, this viewpoint does nothing to actually identify the relevant space \mathcal{Z} nor the values z_1 and z_2 .

A related issue is that of *model form uncertainty*, i.e. an epistemic lack of knowledge about which of a number of competing models for some system of interest is ‘correct’. Usually, the choice to be made is a qualitative one. For example, should one model some observed data using a linear or a non-linear statistical regression model? Or, should one model a fluid flow through a pipe using a high-fidelity computational fluid dynamics model in three spatial dimensions, or using a coarse model that treats the pipe as one-dimensional? This apparently qualitative choice can be rendered into a quantitative form by placing a Bayesian prior on the discrete index set of the models, conditioning upon observed data, and examining the resulting posterior. However, it is important to not misinterpret the resulting posterior probabilities of the models: we do not claim that the more probable model is ‘correct’, only that it has relatively better explanatory power compared to the other models in the model class.

5.2 Interval Estimates

Sometimes, nothing more can be said about some unknown quantity than a range of possible values, with none more or less probable than any other. In the case of an unknown real number x , such information may boil down to an interval such as $[a, b]$ in which x is known to lie. This is, of course, a very basic form of uncertainty, and one may simply summarize the degree of uncertainty by the length of the interval.

Interval Arithmetic. As well as summarizing the degree of uncertainty by the length of the interval estimate, it is often of interest to manipulate the interval estimates themselves as if they were numbers. One method of manipulating interval estimates of real quantities is *interval arithmetic*. Each of the basic arithmetic operations $*$ $\in \{+, -, \cdot, /\}$ is extended to intervals $A, B \subseteq \mathbb{R}$ by

$$A * B := \{x \in \mathbb{R} \mid x = a * b \text{ for some } a \in A, b \in B\}.$$

Hence,

$$\begin{aligned} [a, b] + [c, d] &= [a + c, b + d], \\ [a, b] - [c, d] &= [a - d, b - c], \\ [a, b] \cdot [c, d] &= [\min\{a \cdot c, a \cdot d, b \cdot c, b \cdot d\}, \max\{a \cdot c, a \cdot d, b \cdot c, b \cdot d\}], \\ [a, b]/[c, d] &= [\min\{a/c, a/d, b/c, b/d\}, \max\{a/c, a/d, b/c, b/d\}], \end{aligned}$$

where the expression for $[a, b]/[c, d]$ is defined only when $0 \notin [c, d]$. The addition and multiplication operations are commutative, associative and sub-distributive:

$$A(B + C) \subseteq AB + AC.$$

These ideas can be extended to elementary functions without too much difficulty: monotone functions are straightforward, and the Intermediate Value Theorem ensures that the continuous image of an interval is again an interval. However, for general functions f , it is not straightforward to compute (the convex hull of) the image of f .

Interval analysis corresponds to a worst-case propagation of uncertainty: the interval estimate on the output f is the greatest lower bound and least upper bound compatible with the interval estimates on the input of f . However, in practical settings, one shortcoming of interval analysis is that it can yield interval bounds on output quantities of interest that are too pessimistic (i.e. too wide) to be useful: there is no scope in the interval arithmetic paradigm to consider how likely or unlikely it would be for the various inputs to ‘conspire’ in a highly correlated way to produce the most extreme output values. (The heuristic idea that a function of many independent or weakly correlated random variables is unlikely to stray

far from its mean or median value is known as the *concentration of measure* phenomenon, and will be discussed in Chapter 10.) In order to produce more refined interval estimates, one will need further information, usually probabilistic in nature, on possible correlations among inputs.

‘Intervals’ of Probability Measures. The distributional robustness approaches covered in Chapter 14 — as well as other theories of imprecise probability, e.g. Dempster–Shafer theory — can be seen as an extension of the interval arithmetic approach from partially known real numbers to partially known probability measures. As hybrid interval-probabilistic approaches, they are one way to resolve the ‘overly pessimistic’ shortcomings of classical interval arithmetic as discussed in the previous paragraph. These ideas will be revisited in Chapter 14.

5.3 Variance, Information and Entropy

Suppose that one adopts a subjectivist (e.g. Bayesian) interpretation of probability, so that one’s knowledge about some system of interest with possible values in \mathcal{X} is summarized by a probability measure $\mu \in \mathcal{M}_1(\mathcal{X})$. The probability measure μ is a very rich and high-dimensional object; often it is necessary to summarize the degree of uncertainty implicit in μ with a few numbers — perhaps even just one number.

Variance. One obvious summary statistic, when \mathcal{X} is (a subset of) a normed vector space and μ has mean m , is the variance of μ , i.e.

$$\mathbb{V}(\mu) := \int_{\mathcal{X}} \|x - m\|^2 d\mu(x) \equiv \mathbb{E}_{X \sim \mu} [\|X - m\|^2].$$

If $\mathbb{V}(\mu)$ is small (resp. large), then we are relatively certain (resp. uncertain) that $X \sim \mu$ is in fact quite close to m . A more refined variance-based measure of informativeness is the covariance operator

$$C(\mu) := \mathbb{E}_{X \sim \mu} [(X - m) \otimes (X - m)].$$

A distribution μ for which the operator norm of $C(\mu)$ is large may be said to be a relatively uninformative distribution. Note that when $\mathcal{X} = \mathbb{R}^n$, $C(\mu)$ is an $n \times n$ symmetric positive-semi-definite matrix. Hence, such a $C(\mu)$ has n positive real eigenvalues (counted with multiplicity)

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0,$$

with corresponding normalized eigenvectors $v_1, \dots, v_n \in \mathbb{R}^n$. The direction v_1 corresponding to the largest eigenvalue λ_1 is the direction in which the uncertainty about the random vector X is greatest; correspondingly, the direction v_n is the direction in which the uncertainty about the random vector X is least.

A beautiful and classical result concerning the variance of *two* quantities of interest is the *uncertainty principle* from quantum mechanics. In this setting, the probability distribution is written as $p = |\psi|^2$, where ψ is a unit-norm element of a suitable Hilbert space, usually one such as $L^2(\mathbb{R}^n; \mathbb{C})$. Physical observables like position, momentum etc. act as self-adjoint operators on this Hilbert space; e.g. the position operator Q is

$$(Q\psi)(x) := x\psi(x),$$

so that the expected position is

$$\langle \psi, Q\psi \rangle = \int_{\mathbb{R}^n} \overline{\psi(x)} x \psi(x) dx = \int_{\mathbb{R}^n} x |\psi(x)|^2 dx.$$

In general, for a fixed unit-norm element $\psi \in \mathcal{H}$, the expected value $\langle A \rangle$ and variance $\mathbb{V}(A) \equiv \sigma_A^2$ of a self-adjoint operator $A: \mathcal{H} \rightarrow \mathcal{H}$ are defined by

$$\begin{aligned}\langle A \rangle &:= \langle \psi, A\psi \rangle, \\ \sigma_A^2 &:= \langle (A - \langle A \rangle)^2 \rangle.\end{aligned}$$

The following inequality provides a fundamental lower bound on the product of the variances of any two observables A and B in terms of their commutator $[A, B] := AB - BA$ and their anti-commutator $\{A, B\} := AB + BA$. When this lower bound is positive, the two variances cannot both be close to zero, so simultaneous high-precision measurements of A and B are impossible.

Theorem 5.1 (Uncertainty principle: Schrödinger's inequality). *Let A, B be self-adjoint operators on a Hilbert space \mathcal{H} , and let $\psi \in \mathcal{H}$ have unit norm. Then*

$$\sigma_A^2 \sigma_B^2 \geq \left| \frac{\langle \{A, B\} \rangle - 2\langle A \rangle \langle B \rangle}{2} \right|^2 + \left| \frac{\langle [A, B] \rangle}{2} \right|^2 \quad (5.1)$$

and, a fortiori, $\sigma_A \sigma_B \geq \frac{1}{2} |\langle [A, B] \rangle|$.

Proof. Let $f := (A - \langle A \rangle)\psi$ and $g := (B - \langle B \rangle)\psi$, so that

$$\begin{aligned}\sigma_A^2 &= \langle f, f \rangle = \|f\|^2, \\ \sigma_B^2 &= \langle g, g \rangle = \|g\|^2.\end{aligned}$$

Therefore, by the Cauchy–Schwarz inequality (3.1),

$$\sigma_A^2 \sigma_B^2 = \|f\|^2 \|g\|^2 \geq |\langle f, g \rangle|^2.$$

Now write the right-hand side of this inequality as

$$\begin{aligned}|\langle f, g \rangle|^2 &= (\operatorname{Re}(\langle f, g \rangle))^2 + (\operatorname{Im}(\langle f, g \rangle))^2 \\ &= \left(\frac{\langle f, g \rangle + \langle g, f \rangle}{2} \right)^2 + \left(\frac{\langle f, g \rangle - \langle g, f \rangle}{2i} \right)^2.\end{aligned}$$

Using the self-adjointness of A and B ,

$$\begin{aligned}\langle f, g \rangle &= \langle (A - \langle A \rangle)\psi, (B - \langle B \rangle)\psi \rangle \\ &= \langle AB \rangle - \langle A \rangle \langle B \rangle - \langle A \rangle \langle B \rangle + \langle A \rangle \langle B \rangle \\ &= \langle AB \rangle - \langle A \rangle \langle B \rangle;\end{aligned}$$

similarly, $\langle g, f \rangle = \langle BA \rangle - \langle A \rangle \langle B \rangle$. Hence,

$$\begin{aligned}\langle f, g \rangle - \langle g, f \rangle &= \langle [A, B] \rangle, \\ \langle f, g \rangle + \langle g, f \rangle &= \langle \{A, B\} \rangle - 2\langle A \rangle \langle B \rangle,\end{aligned}$$

which yields (5.1). ■

An alternative measure of information content, not based on variances, is the information-theoretic notion of entropy:

Information and Entropy. In information theory as pioneered by Claude Shannon, the *information* (or *surprisal*) associated with a possible outcome x of a random variable $X \sim \mu$ taking values in a finite set \mathcal{X} is defined to be

$$I(x) := -\log \mathbb{P}_{X \sim \mu}[X = x] \equiv -\log \mu(x). \quad (5.2)$$

Information has units according to the base of the logarithm used:

$$\text{base } 2 \leftrightarrow \text{bits}, \quad \text{base } e \leftrightarrow \text{nats/nits}, \quad \text{base } 10 \leftrightarrow \text{bans/dits/hartleys}.$$

The negative sign in (5.2) makes $I(x)$ non-negative, and logarithms are used because one seeks a quantity $I(\cdot)$ that represents in an additive way the ‘surprise value’ of observing x . For example, if x has half the probability of y , then one is ‘twice as surprised’ to see the outcome $X = x$ instead of $X = y$, and so $I(x) = I(y) + \log 2$. The *entropy* of the measure μ is the expected information:

$$H(\mu) := \mathbb{E}_{X \sim \mu}[I(X)] \equiv - \sum_{x \in \mathcal{X}} \mu(x) \log \mu(x). \quad (5.3)$$

(We follow the convention that $0 \log 0 := \lim_{p \rightarrow 0} p \log p = 0$.) These definitions are readily extended to a random variable X taking values in \mathbb{R}^n and distributed according to a probability measure μ that has Lebesgue density ρ :

$$I(x) := -\log \rho(x),$$

$$H(\mu) := - \int_{\mathbb{R}^n} \rho(x) \log \rho(x) \, dx.$$

Since entropy measures the average information content of the possible values of $X \sim \mu$, entropy is often interpreted as a measure of the uncertainty implicit in μ . (Remember that if μ is very ‘spread out’ and describes a lot of uncertainty about X , then observing a particular value of X carries a lot of ‘surprise value’ and hence a lot of information.)

Example 5.2. Consider a Bernoulli random variable X taking values in $x_1, x_2 \in \mathcal{X}$ with probabilities $p, 1 - p \in [0, 1]$ respectively. This random variable has entropy

$$-p \log p - (1 - p) \log(1 - p).$$

If X is certain to equal x_1 , then $p = 1$, and the entropy is 0; similarly, if X is certain to equal x_2 , then $p = 0$, and the entropy is again 0; these two distributions carry zero information and have minimal entropy. On the other hand, if $p = \frac{1}{2}$, in which case X is uniformly distributed on \mathcal{X} , then the entropy is $\log 2$; indeed, this is the maximum possible entropy for a Bernoulli random variable. This example is often interpreted as saying that when interrogating someone with questions that demand “yes” or “no” answers, one gains maximum information by asking questions that have an equal probability of being answered “yes” versus “no”.

Proposition 5.3. *Let μ and ν be probability measures on discrete sets or \mathbb{R}^n . Then the product measure $\mu \otimes \nu$ satisfies*

$$H(\mu \otimes \nu) = H(\mu) + H(\nu).$$

That is, the entropy of a random vector with independent components is the sum of the entropies of the component random variables.

Proof. Exercise 5.1. ■

5.4 Information Gain, Distances, and Divergences

The definition of entropy in (5.3) implicitly uses a uniform measure (counting measure on a finite set, or Lebesgue measure on \mathbb{R}^n) as a reference measure. Upon reflection, there is no need to privilege uniform measure with being the unique reference measure; indeed, in some settings, such as infinite-dimensional Banach spaces, there is no such thing as a uniform

measure (cf. Theorem 2.38). In general, if μ is a probability measure on a measurable space $(\mathcal{X}, \mathcal{F})$ with reference measure π , then we would like to define the entropy of μ with respect to π by an expression like

$$H(\mu|\pi) = - \int_{\mathbb{R}} \frac{d\mu}{d\pi}(x) \log \frac{d\mu}{d\pi}(x) d\pi(x)$$

whenever μ has a Radon–Nikodým derivative with respect to π . The negative of this functional is a distance-like function on the set of probability measures on $(\mathcal{X}, \mathcal{F})$:

Definition 5.4. Let μ, ν be σ -finite measures on $(\mathcal{X}, \mathcal{F})$. The *Kullback–Leibler divergence* from μ to ν is defined to be

$$D_{\text{KL}}(\mu\|\nu) := \int_{\mathcal{X}} \frac{d\mu}{d\nu} \log \frac{d\mu}{d\nu} d\nu \equiv \int_{\mathcal{X}} \log \frac{d\mu}{d\nu} d\mu$$

if $\mu \ll \nu$ and this integral is finite, and $+\infty$ otherwise.

While $D_{\text{KL}}(\cdot\|\cdot)$ is non-negative, and vanishes if and only if its arguments are identical (see Exercise 5.3), it is neither symmetric nor does it satisfy the triangle inequality. Nevertheless, it can be used to define a topology on $\mathcal{M}_+(\mathcal{X})$ or $\mathcal{M}_1(\mathcal{X})$ by taking as a basis of open sets for the topology the ‘balls’

$$U(\mu, \varepsilon) := \{\nu \mid D_{\text{KL}}(\mu\|\nu) < \varepsilon\}$$

for arbitrary μ and $\varepsilon > 0$. The following result and Exercise 5.6 show that $D_{\text{KL}}(\cdot\|\cdot)$ generates a topology on $\mathcal{M}_1(\mathcal{X})$ that is strictly finer/stronger than that generated by the total variation distance (2.4):

Theorem 5.5 (Pinsker, 1964). *For any $\mu, \nu \in \mathcal{M}_1(\mathcal{X}, \mathcal{F})$,*

$$d_{\text{TV}}(\mu, \nu) \leq \sqrt{2D_{\text{KL}}(\mu\|\nu)}.$$

Proof. Consider a Hahn decomposition (Theorem 2.24) of $(\mathcal{X}, \mathcal{F})$ with respect to the signed measure $\mu - \nu$: let A_0 and A_1 be disjoint measurable sets with union \mathcal{X} such that every measurable subset of A_0 (resp. A_1) has non-negative (resp. non-positive) measure under $\mu - \nu$. Let $\mathcal{A} := \{A_0, A_1\}$. Then the induced measures $\mu_{\mathcal{A}}$ and $\nu_{\mathcal{A}}$ on $\{0, 1\}$ satisfy

$$\begin{aligned} d_{\text{TV}}(\mu, \nu) &= \|\mu - \nu\|_{\text{TV}} \\ &= (\mu - \nu)(A_0) - (\mu - \nu)(A_1) \\ &= (\mu_{\mathcal{A}}(0) - \nu_{\mathcal{A}}(0)) - (\mu_{\mathcal{A}}(1) - \nu_{\mathcal{A}}(1)) \\ &= d_{\text{TV}}(\mu_{\mathcal{A}}, \nu_{\mathcal{A}}). \end{aligned}$$

By the partition inequality (Exercise 5.5), $D_{\text{KL}}(\mu\|\nu) \geq D_{\text{KL}}(\mu_{\mathcal{A}}\|\nu_{\mathcal{A}})$, so it suffices to prove Pinsker’s inequality in the case that \mathcal{X} has only two elements and \mathcal{F} is the power set of \mathcal{X} .

To that end, let $\mathcal{X} := \{0, 1\}$, and let

$$\begin{aligned} \mu &= p\delta_0 + (1-p)\delta_1, \\ \nu &= q\delta_0 + (1-q)\delta_1. \end{aligned}$$

Consider, for fixed $c \in \mathbb{R}$ and $p \in [0, 1]$,

$$g(q) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} - 4c(p-q)^2.$$

Note that $g(p) = 0$ and that, for $q \in (0, 1)$,

$$\begin{aligned} \frac{\partial}{\partial q} g(q) &= -\frac{p}{q} + \frac{1-p}{1-q} + 8c(p-q) \\ &= (q-p) \left(\frac{1}{q(1-q)} - 8c \right). \end{aligned}$$

Since, for all $q \in [0, 1]$, $0 \leq q(1 - q) \leq \frac{1}{4}$, it follows that for any $c \leq \frac{1}{2}$, $g(q)$ attains its minimum at $q = p$. Thus, for $c \leq \frac{1}{2}$,

$$\begin{aligned} g(q) &= D_{\text{KL}}(\mu \| \nu) - c(|p - q| + |(1 - p) - (1 - q)|)^2 \\ &= D_{\text{KL}}(\mu \| \nu) - c(d_{\text{TV}}(\mu, \nu))^2 \\ &\geq 0. \end{aligned}$$

Setting $c = \frac{1}{2}$ yields Pinsker's inequality. ■

One practical use of information-theoretic quantities such as the Kullback–Leibler divergence is to design experiments that will, if run, yield a maximal gain in the Shannon information about the system of interest:

Example 5.6 (Bayesian experimental design). Suppose that a Bayesian point of view is adopted, and for simplicity that all the random variables of interest are finite-dimensional with Lebesgue densities $\rho(\cdot)$. Consider the problem of selecting an optimal experimental design λ for the inference of some parameters / unknowns θ from the observed data y that will result from the experiment λ . If, for each λ and θ , we know the conditional distribution $y|\lambda, \theta$ of the observed data y given λ and θ , then the conditional distribution $y|\lambda$ is obtained by integration with respect to the prior distribution of θ :

$$\rho(y|\lambda) = \int \rho(y|\lambda, \theta) \rho(\theta) d\theta.$$

Let $U(y, \lambda)$ be a real-valued measure of the *utility* of the posterior distribution

$$\rho(\theta|y, \lambda) = \frac{\rho(y|\theta, \lambda) \rho(\theta)}{\rho(y|\lambda)}.$$

For example, one could take the utility function $U(y, \lambda)$ to be the Kullback–Leibler divergence $D_{\text{KL}}(\rho(\cdot|y, \lambda) \| \rho(\cdot|\lambda))$ between the prior and posterior distributions on θ . An experimental design λ that maximizes

$$U(\lambda) := \int U(y, \lambda) \rho(y|\lambda) dy$$

is one that is optimal in the sense of maximizing the expected gain in Shannon information.

In general, the optimization problem of finding a maximally informative experimental design is highly non-trivial, especially in the case of computationally intensive likelihood functions. See, e.g., Chaloner and Verdinelli (1995) for a survey of this large field of study.

Divergences and Other Distances. The total variation distance and Kullback–Leibler divergence are special cases of a more general class of distance-like functions between pairs of probability measures:

Definition 5.7. Let μ and ν be σ -finite measures on a common measurable space $(\mathcal{X}, \mathcal{F})$, and let $f: [0, \infty] \rightarrow \mathbb{R} \cup \{+\infty\}$ be any convex function such that $f(1) = 0$. The *f-divergence* from μ to ν is defined to be

$$D_f(\mu \| \nu) := \int_{\mathcal{X}} f\left(\frac{d\mu}{d\nu}\right) d\nu \tag{5.4}$$

if $\mu \ll \nu$ and this integral is finite, and $+\infty$ otherwise. Equivalently, in terms of any reference measure ρ with respect to which both μ and ν are absolutely continuous (such as $\mu + \nu$),

$$D_f(\mu \| \nu) := \int_{\mathcal{X}} f\left(\frac{d\mu}{d\rho} \bigg/ \frac{d\nu}{d\rho}\right) \frac{d\nu}{d\rho} d\rho. \tag{5.5}$$

It is good practice to check that the alternative definition (5.5) is, in fact, independent of the reference measure used:

Lemma 5.8. *Suppose that μ and ν are absolutely continuous with respect to both ρ_1 and ρ_2 . Then ρ_1 and ρ_2 are equivalent measures except for sets of $(\mu + \nu)$ -measure zero, and (5.5) defines the same value with $\rho = \rho_1$ as it does with $\rho = \rho_2$.*

Proof. Suppose that ρ_1 and ρ_2 are inequivalent. Then, without loss of generality, there exists a measurable set E such that $\rho_1(E) = 0$ but $\rho_2(E) > 0$. Therefore, since $\mu \ll \rho_1$ and $\nu \ll \rho_1$, it follows that $\mu(E) = \nu(E) = 0$. Thus, although ρ_1 and ρ_2 may be inequivalent for arbitrary measurable sets, they are equivalent for sets of positive $(\mu + \nu)$ -measure.

Now let E be a set of full measure under ν , so that $\frac{d\rho_2}{d\rho_1}$ exists and is nowhere zero in E . Then the chain rule for Radon–Nikodým derivatives (Theorem 2.30) yields

$$\begin{aligned}
 & \int_{\mathcal{X}} f \left(\frac{d\mu}{d\rho_1} \middle/ \frac{d\nu}{d\rho_1} \right) \frac{d\nu}{d\rho_1} d\rho_1 \\
 &= \int_E f \left(\frac{d\mu}{d\rho_1} \middle/ \frac{d\nu}{d\rho_1} \right) d\nu && \text{since } \nu(\mathcal{X} \setminus E) = 0 \\
 &= \int_E f \left(\left(\frac{d\mu}{d\rho_2} \frac{d\rho_2}{d\rho_1} \right) \middle/ \left(\frac{d\nu}{d\rho_2} \frac{d\rho_2}{d\rho_1} \right) \right) d\nu && \text{by Theorem 2.30} \\
 &= \int_E f \left(\frac{d\mu}{d\rho_2} \middle/ \frac{d\nu}{d\rho_2} \right) d\nu \\
 &= \int_{\mathcal{X}} f \left(\frac{d\mu}{d\rho_2} \middle/ \frac{d\nu}{d\rho_2} \right) \frac{d\nu}{d\rho_2} d\rho_2. \quad \blacksquare
 \end{aligned}$$

Jensen's inequality and the conditions on f immediately imply that f -divergences of probability measures are non-negative:

$$D_f(\mu \parallel \nu) = \int_{\mathcal{X}} f \left(\frac{d\mu}{d\nu} \right) d\nu \geq f \left(\int_{\mathcal{X}} \frac{d\mu}{d\nu} d\nu \right) = f(1) = 0.$$

For strictly convex f , equality holds if and only if $\mu = \nu$, and for the Kullback–Leibler distance this is known as *Gibbs' inequality* (Exercise 5.3).

- Example 5.9.** (a) The total variation distance defined in (2.4) is the f -divergence with $f(t) := |t - 1|$; this can be seen most directly from formulation (5.5). As already discussed, d_{TV} is a metric on the space of probability measures on $(\mathcal{X}, \mathcal{F})$, and indeed it is a norm on the space of signed measures on $(\mathcal{X}, \mathcal{F})$. Under the total variation distance, $\mathcal{M}_1(\mathcal{X})$ has diameter at most 2.
- (b) The Kullback–Leibler divergence is the f -divergence with $f(t) := t \log t$. This does not define a metric, since in general it is neither symmetric nor does it satisfy the triangle inequality.
- (c) The *Hellinger distance* is the square root of the f -divergence with $f(t) := |\sqrt{t} - 1|^2$, i.e.

$$\begin{aligned}
 d_H(\mu, \nu)^2 &= \int_{\mathcal{X}} \left| \sqrt{\frac{d\mu}{d\nu}} - 1 \right|^2 d\nu \\
 &= \int_{\mathcal{X}} \left| \sqrt{\frac{d\mu}{d\rho}} - \sqrt{\frac{d\nu}{d\rho}} \right|^2 d\rho
 \end{aligned}$$


for any reference measure ρ , and is a bona fide metric.

The total variation and Kullback–Leibler distances and their associated topologies are related by Pinsker's inequality (Theorem 5.5); the corresponding result for the total variation and Hellinger distances and their topologies is *Kraft's inequality* (see Steerneman (1983) for generalizations to signed and product measures):

Theorem 5.10 (Kraft, 1955). *Let μ, ν be probability measures on (Θ, \mathcal{F}) . Then*

$$d_H(\mu, \nu)^2 \leq d_{TV}(\mu, \nu) \leq 2d_H(\mu, \nu). \quad (5.6)$$

Hence, the total variation metric and Hellinger metric induce the same topology on $\mathcal{M}_1(\Theta)$.

Remark 5.11. It also is common in the literature to see the total variation distance defined as the f -divergence with $f(t) := \frac{1}{2}|t-1|$ and the Hellinger distance defined as the square root of the f -divergence with $f(t) := \frac{1}{2}|\sqrt{t}-1|^2$. In this case, Kraft's inequality (5.6) becomes 

$$d_H(\mu, \nu)^2 \leq d_{TV}(\mu, \nu) \leq \sqrt{2} d_H(\mu, \nu). \quad (5.7)$$

A useful property of the Hellinger distance is that it provides a Lipschitz-continuous bound on how the expectation of a random variable changes when changing measure from one measure to another. This property will be useful in the results of Chapter 6 on the well-posedness of Bayesian inverse problems.

Proposition 5.12. *Let $(\mathcal{V}, \|\cdot\|)$ be a Banach space, and suppose that $f: \mathcal{X} \rightarrow \mathcal{V}$ has finite second moment with respect to $\mu, \nu \in \mathcal{M}_1(\mathcal{X})$. Then*

$$\|\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]\| \leq 2 \sqrt{\mathbb{E}_\mu[\|f\|^2] + \mathbb{E}_\nu[\|f\|^2]} d_H(\mu, \nu).$$

Proof. Exercise 5.7. ■

There are also useful measures of distance between probability measures that make use of the metric space structure of the sample space, if it has one. The following metric, the Lévy–Prohorov distance, is particularly important in analysis because it corresponds to the often-used topology of weak convergence of probability measures:

Definition 5.13. The *Lévy–Prohorov distance* between probability measures μ and ν on a metric space (\mathcal{X}, d) is defined by

$$d_{LP}(\mu, \nu) := \inf \left\{ \varepsilon \geq 0 \mid \begin{array}{l} \mu(A) \leq \nu(A^\varepsilon) + \varepsilon \text{ and} \\ \nu(A) \leq \mu(A^\varepsilon) + \varepsilon \text{ for all measurable } A \subseteq \mathcal{X} \end{array} \right\},$$

where A^ε denotes the open ε -neighbourhood of A in the metric d , i.e.

$$A^\varepsilon := \bigcup_{a \in A} \mathbb{B}_\varepsilon(a) = \{x \in \mathcal{X} \mid d(a, x) < \varepsilon \text{ for some } a \in A\}.$$

It can be shown that this defines a metric on the space of probability measures on \mathcal{X} . The Lévy–Prohorov metric d_{LP} on $\mathcal{M}_1(\mathcal{X})$ inherits many of the properties of the original metric d on \mathcal{X} : if (\mathcal{X}, d) is separable, then so too is $(\mathcal{M}_1(\mathcal{X}), d_{LP})$; and if (\mathcal{X}, d) is complete, then so too is $(\mathcal{M}_1(\mathcal{X}), d_{LP})$. By (h) below, the Lévy–Prohorov metric metrizes the topology of weak convergence of probability measures, which by (d) below is essentially the topology of convergence of bounded and continuous statistics:

Theorem 5.14 (Portmanteau theorem for weak convergence). *Let $(\mu_n)_{n \in \mathbb{N}}$ be a sequence of probability measures on a topological space \mathcal{X} , and let $\mu \in \mathcal{M}_1(\mathcal{X})$. Then the following are equivalent, and determine the weak convergence of μ_n to μ :*

- (a) $\limsup_{n \rightarrow \infty} \mu_n(F) \leq \mu(F)$ for all closed $F \subseteq \mathcal{X}$;
- (b) $\liminf_{n \rightarrow \infty} \mu_n(U) \geq \mu(U)$ for all open $U \subseteq \mathcal{X}$;
- (c) $\lim_{n \rightarrow \infty} \mu_n(A) = \mu(A)$ for all $A \subseteq \mathcal{X}$ with $\mu(\partial A) = 0$;
- (d) $\lim_{n \rightarrow \infty} \mathbb{E}_{\mu_n}[f] = \mathbb{E}_\mu[f]$ for every bounded and continuous $f: \mathcal{X} \rightarrow \mathbb{R}$;
- (e) $\lim_{n \rightarrow \infty} \mathbb{E}_{\mu_n}[f] = \mathbb{E}_\mu[f]$ for every bounded and Lipschitz $f: \mathcal{X} \rightarrow \mathbb{R}$;
- (f) $\limsup_{n \rightarrow \infty} \mathbb{E}_{\mu_n}[f] \leq \mathbb{E}_\mu[f]$ for every $f: \mathcal{X} \rightarrow \mathbb{R}$ that is upper semi-continuous and bounded above;

- (g) $\liminf_{n \rightarrow \infty} \mathbb{E}_{\mu_n}[f] \geq \mathbb{E}_{\mu}[f]$ for every $f: \mathcal{X} \rightarrow \mathbb{R}$ that is lower semi-continuous and bounded below;
 (h) when \mathcal{X} is metrized by a metric d , $\lim_{n \rightarrow \infty} d_{\text{LP}}(\mu_n, \mu) = 0$.

Some further examples of distances between probability measures are included in the exercises at the end of the chapter, and the bibliography gives references for more comprehensive surveys.

5.5 Bibliography

The book of Ayyub and Klir (2006) provides a wide-ranging discussion of many notions of uncertainty and their description, elicitation, propagation, and visualization, all with a practical eye on applications to engineering and the sciences.

Wasserman (2000) gives a survey of Bayesian model selection and model averaging. Comprehensive treatments of interval analysis include the classic monograph of Moore (1966) and the more recent text of Jaulin et al. (2001). Hansen and Walster (2004) also provide a modern introduction to interval analysis, with an eye on applications to optimization.

The books of Cover and Thomas (2006) and MacKay (2003) provide a thorough introduction to information theory, which was pioneered by Shannon in his seminal 1948 paper (Shannon, 1948). See also Jaynes (2003) for a unified perspective on information theory, inference, and logic in the sciences.

The Kullback–Leibler divergence was introduced by Kullback and Leibler (1951), who in fact considered the symmetrized version of the divergence that now bears their names. The more general theory of f -divergences was introduced and studied independently by Csiszár (1963), Morimoto (1963), and Ali and Silvey (1966). Lindley (1956) was an early proponent of what would now be called Bayesian experimental design; see Chaloner and Verdinelli (1995) for a comprehensive review of this large field.

Weak convergence of probability measures was introduced by Aleksandrov (1940, 1941, 1943). Theorem 5.14, the portmanteau theorem for weak convergence, can be found in many texts on probability theory, e.g. that of Billingsley (1995, Section 2).

The Wasserstein metric (also known as the Kantorovich or Rubinstein metric, or earth-mover's distance) of Exercise 5.11 plays a central role in the theory of optimal transportation; for comprehensive treatments of this topic, see the books of Villani (2003, 2009), and also Ambrosio et al. (2008, Chapter 6). Gibbs and Su (2002) give a short self-contained survey of many distances between probability measures, and the relationships among them. Deza and Deza (2014, Chapter 14) give a more extensive treatment of distances between probability measures, in the context of a wide-ranging discussion of distances of all kinds.

5.6 Exercises

Exercise 5.1. Prove Proposition 5.3. That is, suppose that μ and ν are probability measures on discrete sets or \mathbb{R}^n , and show that the product measure $\mu \otimes \nu$ satisfies

$$H(\mu \otimes \nu) = H(\mu) + H(\nu).$$

That is, the entropy of a random vector with independent components is the sum of the entropies of the component random variables.

Exercise 5.2. Let $\mu_0 = \mathcal{N}(m_0, C_0)$ and $\mu_1 = \mathcal{N}(m_1, C_1)$ be non-degenerate Gaussian measures on \mathbb{R}^n . Show that

$$D_{\text{KL}}(\mu_0 \| \mu_1) = \frac{1}{2} \left(\log \frac{\det C_1}{\det C_0} - n + \text{tr}(C_1^{-1} C_0) + \|m_0 - m_1\|_{C_1^{-1}}^2 \right).$$

Hint: use the fact that, when $X \sim \mathcal{N}(m, C)$ is an \mathbb{R}^n -valued Gaussian random vector and $A \in \mathbb{R}^{n \times n}$ is symmetric,

$$\mathbb{E}[X \cdot AX] = \text{tr}(AC) + m \cdot Am.$$

Exercise 5.3. Let μ and ν be probability measures on a measurable space $(\mathcal{X}, \mathcal{F})$. Prove *Gibbs' inequality* that $D_{\text{KL}}(\mu\|\nu) \geq 0$, with equality if and only if $\mu = \nu$.

Exercise 5.4. Let f satisfy the requirements for $D_f(\cdot\|\cdot)$ to be a divergence.

- (a) Show that the function $(x, y) \mapsto yf(x/y)$ is a convex function from $(0, \infty) \times (0, \infty)$ to $\mathbb{R} \cup \{+\infty\}$.
- (b) Hence show that $D_f(\cdot\|\cdot)$ is jointly convex in its two arguments, i.e. for all probability measures μ_0, μ_1, ν_0 , and ν_1 and $t \in [0, 1]$,

$$D_f((1-t)\mu_0 + t\mu_1\|(1-t)\nu_0 + t\nu_1) \leq (1-t)D_f(\mu_0\|\nu_0) + tD_f(\mu_1\|\nu_1).$$

Exercise 5.5. The following result is a useful one that frequently allows statements about f -divergences to be reduced to the case of a finite or countable sample space. Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a probability space, and let $f: [0, \infty] \rightarrow [0, \infty]$ be convex. Given a partition $\mathcal{A} = \{A_n \mid n \in \mathbb{N}\}$ of \mathcal{X} into countably many pairwise disjoint measurable sets, define a probability measure $\mu_{\mathcal{A}}$ on \mathbb{N} by $\mu_{\mathcal{A}}(n) := \mu(A_n)$.

- (a) Suppose that $\mu(A_n) > 0$ and that $\mu \ll \nu$. Show that, for each $n \in \mathbb{N}$,

$$\frac{1}{\nu(A_n)} \int_{A_n} f\left(\frac{d\mu}{d\nu}\right) d\nu \geq f\left(\frac{\mu(A_n)}{\nu(A_n)}\right).$$

- (b) Hence prove the following result, known as the *partition inequality*: for any two probability measures μ and ν on \mathcal{X} with $\mu \ll \nu$,

$$D_f(\mu\|\nu) \geq D_f(\mu_{\mathcal{A}}\|\nu_{\mathcal{A}}).$$

Show also that, for strictly convex f , equality holds if and only if $\mu(A_n) = \nu(A_n)$ for each n .

Exercise 5.6. Show that Pinsker's inequality (Theorem 5.5) cannot be reversed. In particular, give an example of a measurable space $(\mathcal{X}, \mathcal{F})$ such that, for any $\varepsilon > 0$, there exist probability measures μ and ν on $(\mathcal{X}, \mathcal{F})$ with $d_{\text{TV}}(\mu, \nu) \leq \varepsilon$ but $D_{\text{KL}}(\mu\|\nu) = +\infty$. Hint: consider a 'small' perturbation to the CDF of a probability measure on \mathbb{R} .

Exercise 5.7. Prove Proposition 5.12. That is, let $(\mathcal{V}, \|\cdot\|)$ be a Banach space, and suppose that $f: \mathcal{X} \rightarrow \mathcal{V}$ has finite second moment with respect to $\mu, \nu \in \mathcal{M}_1(\mathcal{X})$. Then

$$\|\mathbb{E}_{\mu}[f] - \mathbb{E}_{\nu}[f]\| \leq 2 \sqrt{\mathbb{E}_{\mu}[\|f\|^2] + \mathbb{E}_{\nu}[\|f\|^2]} d_{\text{H}}(\mu, \nu).$$

Exercise 5.8. Suppose that μ and ν are equivalent probability measures on $(\mathcal{X}, \mathcal{F})$ and define

$$d(\mu, \nu) := \text{ess sup}_{x \in \mathcal{X}} \left| \log \frac{d\mu}{d\nu}(x) \right|.$$

(See Example 2.7 for the definition of the essential supremum.) Show that this defines a well-defined metric on the measure equivalence class \mathcal{E} containing μ and ν . In particular, show that neither the choice of function used as the Radon–Nikodým derivative $\frac{d\mu}{d\nu}$, nor the choice of measure in \mathcal{E} with respect to which the essential supremum is taken, affect the value of $d(\mu, \nu)$.

Exercise 5.9. For a probability measure μ on \mathbb{R} , let $F_{\mu}: \mathbb{R} \rightarrow [0, 1]$ be the cumulative distribution function (CDF) defined by

$$F_{\mu}(x) := \mu((-\infty, x]).$$

Show that the Lévy–Prohorov distance between probability measures $\mu, \nu \in \mathcal{M}_1(\mathbb{R})$ reduces to the *Lévy distance*, defined in terms of their CDFs F_μ, F_ν by

$$d_L(\mu, \nu) := \inf \{ \varepsilon > 0 \mid F_\mu(x - \varepsilon) - \varepsilon \leq F_\nu(x) \leq F_\mu(x + \varepsilon) + \varepsilon \}.$$

Convince yourself that this distance can be visualized as the side length of the largest square with sides parallel to the coordinate axes that can be placed between the graphs of F_μ and F_ν .

Exercise 5.10. Let (\mathcal{X}, d) be a metric space, equipped with its Borel σ -algebra. The *Lukaszyk–Karmowski distance* between probability measures μ and ν is defined by

$$d_{LK}(\mu, \nu) := \int_{\mathcal{X}} \int_{\mathcal{X}} d(x, x') \, d\mu(x) d\nu(x').$$

Show that this satisfies all the requirements to be a metric on the space of probability measures on \mathcal{X} *except* for the requirement that $d_{LK}(\mu, \mu) = 0$. Hint: suppose that $\mu = \mathcal{N}(m, \sigma^2)$ on \mathbb{R} , and show that $d_{LK}(\mu, \mu) = \frac{2\sigma}{\pi}$.

Exercise 5.11. Let (\mathcal{X}, d) be a metric space, equipped with its Borel σ -algebra. The *Wasserstein distance* between probability measures μ and ν is defined by

$$d_W(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} d(x, x') \, d\gamma(x, x'),$$

where the infimum is taken over the set $\Gamma(\mu, \nu)$ of all measures γ on $\mathcal{X} \times \mathcal{X}$ such that the push-forward of γ onto the first (resp. second) copy of \mathcal{X} is μ (resp. ν). Show that this defines a metric on the space of probability measures on \mathcal{X} , bounded above by the Lukaszyk–Karmowski distance, i.e.

$$d_W(\mu, \nu) \leq d_{LK}(\mu, \nu).$$

Verify also that the *p-Wasserstein distance*

$$d_{W,p}(\mu, \nu) := \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} d(x, x')^p \, d\gamma(x, x') \right)^{1/p},$$

where $p \geq 1$, is a metric. Metrics of this type, and in particular the case $p = 1$, are sometimes known as the *earth-mover's distance* or *optimal transportation distance*, since the minimization over $\gamma \in \Gamma(\mu, \nu)$ can be seen as finding the optimal way of moving/rearranging the pile of earth μ into the pile ν .