

# **MA4K0 Introduction to Uncertainty Quantification**

T. J. Sullivan  
Mathematics Institute  
University of Warwick  
Coventry CV4 7AL UK  
`Tim.Sullivan@warwick.ac.uk`

DRAFT  
Not For General Distribution  
Version 11 (2013-10-16 15:45)

DRAFT

# Contents

Contents . . . . .	i
Preface . . . . .	v
<b>Introduction to Uncertainty Quantification</b>	<b>1</b>
<b>1 Introduction</b>	<b>3</b>
1.1 What is Uncertainty Quantification? . . . . .	3
1.2 Mathematical Prerequisites . . . . .	6
1.3 The Road Not Travelled . . . . .	7
<b>2 Recap of Measure and Probability Theory</b>	<b>9</b>
2.1 Measure and Probability Spaces . . . . .	9
2.2 Random Variables and Stochastic Processes . . . . .	12
2.3 Aside: Interpretations of Probability . . . . .	13
2.4 Lebesgue Integration . . . . .	14
2.5 The Radon–Nikodým Theorem and Densities . . . . .	16
2.6 Product Measures and Independence . . . . .	17
2.7 Gaussian Measures . . . . .	18
Bibliography . . . . .	21
<b>3 Recap of Banach and Hilbert Spaces</b>	<b>23</b>
3.1 Basic Definitions and Properties . . . . .	23
3.2 Dual Spaces and Adjoints . . . . .	26
3.3 Orthogonality and Direct Sums . . . . .	27
3.4 Tensor Products . . . . .	30
Bibliography . . . . .	32
<b>4 Basic Optimization Theory</b>	<b>33</b>
4.1 Optimization Problems and Terminology . . . . .	33
4.2 Unconstrained Global Optimization . . . . .	34
4.3 Constrained Optimization . . . . .	37
4.4 Convex Optimization . . . . .	39
4.5 Linear Programming . . . . .	42
4.6 Least Squares . . . . .	43
Bibliography . . . . .	46
Exercises . . . . .	47

<b>5</b>	<b>Measures of Information and Uncertainty</b>	<b>49</b>
5.1	The Existence of Uncertainty . . . . .	49
5.2	Interval Estimates . . . . .	50
5.3	Variance, Information and Entropy . . . . .	50
5.4	Information Gain . . . . .	53
	Bibliography . . . . .	55
	Exercises . . . . .	55
<b>6</b>	<b>Bayesian Inverse Problems</b>	<b>57</b>
6.1	Inverse Problems and Regularization . . . . .	57
6.2	Bayesian Inversion in Banach Spaces . . . . .	62
6.3	Well-Posedness and Approximation . . . . .	63
	Bibliography . . . . .	67
	Exercises . . . . .	67
<b>7</b>	<b>Filtering and Data Assimilation</b>	<b>71</b>
7.1	State Estimation in Discrete Time . . . . .	72
7.2	Linear Kálmán Filter . . . . .	74
7.3	Extended Kálmán Filter . . . . .	77
7.4	Ensemble Kálmán Filter . . . . .	78
7.5	Eulerian and Lagrangian Data Assimilation . . . . .	80
	Bibliography . . . . .	80
	Exercises . . . . .	81
<b>8</b>	<b>Orthogonal Polynomials</b>	<b>85</b>
8.1	Basic Definitions and Properties . . . . .	85
8.2	Recurrence Relations . . . . .	89
8.3	Roots of Orthogonal Polynomials . . . . .	90
8.4	Polynomial Interpolation . . . . .	92
8.5	Polynomial Approximation . . . . .	93
8.6	Orthogonal Polynomials of Several Variables . . . . .	96
	Bibliography . . . . .	97
	Exercises . . . . .	97
<b>9</b>	<b>Numerical Integration</b>	<b>99</b>
9.1	Quadrature in One Dimension . . . . .	99
9.2	Gaussian Quadrature . . . . .	101
9.3	Clenshaw–Curtis / Fejér Quadrature . . . . .	104
9.4	Quadrature in Multiple Dimensions . . . . .	104
9.5	Monte Carlo Methods . . . . .	105
9.6	Pseudo-Random Methods . . . . .	107
	Bibliography . . . . .	109
	Exercises . . . . .	109
<b>10</b>	<b>Sensitivity Analysis and Model Reduction</b>	<b>111</b>
10.1	Model Reduction for Linear Models . . . . .	111
10.2	Derivatives . . . . .	112
10.3	McDiarmid Diameters . . . . .	112
10.4	ANOVA/HDMR Decompositions . . . . .	115
	Bibliography . . . . .	119

Exercises . . . . .	119
<b>11 Spectral Expansions</b>	<b>121</b>
11.1 Karhunen–Loève Expansions . . . . .	121
11.2 Wiener–Hermite Polynomial Chaos . . . . .	126
11.3 Generalized PC Expansions . . . . .	129
Bibliography . . . . .	133
Exercises . . . . .	133
<b>12 Stochastic Galerkin Methods</b>	<b>135</b>
12.1 Lax–Milgram Theory and Galerkin Projection . . . . .	136
12.2 Stochastic Galerkin Projection . . . . .	140
12.3 Nonlinearities . . . . .	145
Bibliography . . . . .	146
Exercises . . . . .	146
<b>13 Non-Intrusive Spectral Methods</b>	<b>149</b>
13.1 Pseudo-Spectral Methods . . . . .	150
13.2 Stochastic Collocation . . . . .	150
Bibliography . . . . .	153
Exercises . . . . .	153
<b>14 Distributional Uncertainty</b>	<b>155</b>
14.1 Maximum Entropy Distributions . . . . .	155
14.2 Distributional Robustness . . . . .	157
14.3 Functional and Distributional Robustness . . . . .	162
Bibliography . . . . .	166
Exercises . . . . .	166
<b>Bibliography and Index</b>	<b>169</b>
<b>Bibliography</b>	<b>171</b>
<b>Index</b>	<b>179</b>

DRAFT

## Preface

These notes are designed as an introduction to Uncertainty Quantification (UQ) at the fourth year (senior) undergraduate or beginning postgraduate level, and are aimed primarily at students from a mathematical (rather than, say, engineering) background; the mathematical prerequisites are listed in Section 1.2, and the early chapters of the text recapitulate some of this material in more detail. These notes accompany the University of Warwick mathematics module [MA4K0 Introduction to Uncertainty Quantification](#); while the notes are intended to be general, certain contextual remarks and assumptions about prior knowledge will be Warwick-specific, and will be indicated by a large “W” in the margin, like the one to the right.

The aim is 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. There are, of course, even more UQ problems and solution methods in the world that are not covered in these notes, which are intended — with the exception of the preliminary material on measure theory and functional analysis — to comprise approximately 30 hours’ worth of lectures. For any grievous omissions in this regard, I ask for your indulgence, and would be happy to receive suggestions for improvements.

The exercises contain, by deliberate choice, a number of terribly ill-posed or under-specified problems of the messy type often encountered in practice. It is my hope that these exercises will encourage students to grapple with the questions of mathematical modelling that are a necessary precursor to doing applied mathematics outside the tame classroom environment. Theoretical knowledge is important; however, problem solving, which begins with problem formulation, is an equally vital skill that too often goes neglected in undergraduate mathematics courses.

These notes have benefitted, from initial conception to nearly finished product, from discussions with many people. I would like to thank Charlie Elliott, Dave McCormick, Mike McKerns, Michael Ortiz, Houman Owhadi, Clint Scovel, Andrew Stuart, and all the students on the 2013–14 iteration of [MA4K0](#) for their useful comments.

*T.J.S.*

University of Warwick, U.K.  
Wednesday 16<sup>th</sup> October, 2013

DRAFT



# **Introduction to Uncertainty Quantification**

DRAFT

# 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?

Uncertainty Quantification (UQ) 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 1990–1995, risk modelling for catastrophe insurance and re-insurance (i.e. insurance for property owners against risks arising from earthquakes, hurricanes, terrorism, &c., and then insurance for the providers of such insurance) was done on a purely statistical basis. Since that time, catastrophe modellers have started 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.
- Over roughly the same period of time, deterministic engineering models of complex physical processes began to incorporate some element of probability to account for lack of knowledge about important physical parameters, random variability in operating circumstances, or outright uncertainty about what the form of a ‘correct’ model would be. Again the aim is to provide more accurate predictions about systems’ behaviour.

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”; other perspectives on UQ include the study

of numerical error and the stability of algorithms. 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.”* [109, 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*, bearing in mind that they are only approximations of reality, not the ‘truth’ of those pieces of information / assumptions. There is always a risk of ‘Garbage In, Garbage Out’. A mathematician performing a UQ analysis 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). Naturally, a respectable analysis will include a meta-analysis examining the sensitivity of the original analysis to perturbations of the governing assumptions. In the author’s view, this is the proper interpretation of philosophically sound but practically unhelpful assertions like “Verification and validation of numerical models of natural systems is impossible” and “The primary value of models is heuristic.” [74].

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

$$\begin{aligned} -\nabla \cdot (\kappa \nabla u) &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega. \end{aligned}$$

This PDE is a simple but not naïve model for the pressure field  $u$  of some fluid occupying a domain  $\Omega$ , the permeability of which to the fluid is described by a tensor field  $\kappa: \Omega \rightarrow \mathbb{R}^{n \times n}$ ; there is a source term  $f$  and the boundary condition specifies that the pressure vanishes on the boundary of  $\Omega$ . This simple 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 one is naturally very interested in where fluid following the velocity field  $v$  will end up, and how soon.

In a course on PDE theory, you will learn that if the permeability field  $\kappa$  is positive-definite and essentially bounded, then this problem has a unique weak solution  $u$  in the Sobolev space  $H_0^1(\Omega)$  for each forcing term  $f$  in the dual Sobolev space  $H^{-1}(\Omega)$ . One objective of this course 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 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’  $\kappa$  and  $f$  are not known precisely, so our knowledge about  $u =$

$u(\kappa, f)$  is also uncertain in some way. If  $\kappa$  and  $f$  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 &c. — and to do so it is necessary to build up the machinery of probability on function spaces.

Another issue is that often we want to solve the *inverse problem*: we know something about  $f$  and something about  $u$  and want to infer  $\kappa$ . Even a simple 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 fields  $f$ ,  $\kappa$  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 in the literature 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. To a certain extent, the distinction 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, &c. 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 *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,

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

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

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

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

- The *parameter identification* or *inverse problem*. 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.



**A Word of Warning.** In this second decade of the third millennium, there is as yet no elegant unified theory of UQ. 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, MA4K0 has certain prerequisites. If you are just following the course for fun, and attending the lectures merely to stay warm and dry in what is almost sure to be a fine English autumn, then good for you. However, if you actually want to understand what is going on, then it’s better for your own health if you can use your nearest time machine to ensure that you have already taken and understood, in addition to the standard G100/G103 Mathematics core courses, the following non-core courses:

- [ST112 Probability B](#)
- [Either MA359 Measure Theory or ST318 Probability Theory](#)
- [MA3G7 Functional Analysis I](#)

As a crude diagnostic test, read the following sentence:

Given any  $\sigma$ -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  $\mu$ -almost everywhere, is a Hilbert space with respect to the inner product  $\langle f, g \rangle := \int_{\mathcal{X}} \bar{f}g d\mu$ .

If any of the symbols, concepts or terms used or implicit in that sentence give you more than a few moments’ pause, then you should think again before attempting MA4K0.

If, in addition, you have taken the following courses, then certain techniques, examples and remarks will make more sense to you:

- [MA117 Programming for Scientists](#)
- [MA228 Numerical Analysis](#)
- [MA250 Introduction to Partial Differential Equations](#)
- [MA398 Matrix Analysis and Algorithms](#)
- [MA3H0 Numerical Analysis and PDEs](#)

- [ST407 Monte Carlo Methods](#)
- [MA482 Stochastic Analysis](#)
- [MA4A2 Advanced PDEs](#)
- [MA607 Data Assimilation](#)

However, none of these courses are essential. That said, some ability and willingness to implement UQ methods — even in simple settings — in e.g. C/C++, Mathematica, Matlab, or Python<sup>(1.1)</sup> is highly desirable. UQ is a topic best learned in the doing, not through pure theory.

## 1.3 The Road Not Travelled

There are many topics relevant to UQ that are either not covered or discussed only briefly in these notes, 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; optimal control; and machine learning.

---

<sup>(1.1)</sup>The author's current language of choice.

DRAFT



## Chapter 2

# Recap of 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 mathematical treatment of uncertainty. 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 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

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

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

**Definition 2.2.** A *signed measure* (or *charge*) on a measurable space  $(\Theta, \mathcal{F})$  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  $E \in \mathcal{F}$ , the series  $\sum_{n \in \mathbb{N}} \mu(E_n)$  converges absolutely to  $\mu(E)$ . A *measure* is a signed measure that does not take negative values. A *probability measure* is a measure such that  $\mu(\Theta) = 1$ . The triple  $(\Theta, \mathcal{F}, \mu)$  is called a *signed measure space*, *measure space*, or *probability space* as appropriate. The sets of all signed measures, measures, and probability measures on  $(\Theta, \mathcal{F})$  are denoted  $\mathcal{M}_{\pm}(\Theta, \mathcal{F})$ ,  $\mathcal{M}_{+}(\Theta, \mathcal{F})$ , and  $\mathcal{M}_1(\Theta, \mathcal{F})$  respectively.

- Examples 2.3.** 1. The *trivial measure*:  $\tau(E) := 0$  for every  $E \in \mathcal{F}$ .  
 2. The *unit Dirac measure* at  $a \in \mathcal{X}$ :

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

3. *Counting measure*:

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

4. *Lebesgue measure* on  $\mathbb{R}^n$ : the unique measure on  $\mathbb{R}^n$  (equipped with its Borel  $\sigma$ -algebra  $\mathcal{B}(\mathbb{R}^n)$ ) that assigns to every cube its volume. To be more precise, Lebesgue measure is actually defined on the completion of  $\mathcal{B}(\mathbb{R}^n)$ , a larger  $\sigma$ -algebra than  $\mathcal{B}(\mathbb{R}^n)$ .

**Definition 2.4.** Let  $(\mathcal{X}, \mathcal{F}, \mu)$  be a measure space. 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  $\mu$ -null set. If the set of  $x \in \mathcal{X}$  for which some property  $P(x)$  does not hold is  $\mu$ -null, then  $P$  is said to hold  $\mu$ -almost everywhere (or, when  $\mu$  is a probability measure,  $\mu$ -almost surely). If every  $\mu$ -null set is in fact an  $\mathcal{F}$ -measurable set, then the measure space  $(\mathcal{X}, \mathcal{F}, \mu)$  is said to be *complete*.

When the sample space is a topological space, it is usual to use the *Borel  $\sigma$ -algebra* (i.e. the smallest  $\sigma$ -algebra that contains all the open sets); measures on the Borel  $\sigma$ -algebra are called *Borel measures*. Unless noted otherwise, this is the convention followed in these notes.

**Definition 2.5.** The *support* of a measure  $\mu$  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  $\mu$ -measure. Equivalently,  $\text{supp}(\mu)$  is the complement of the union of all open sets of  $\mu$ -measure zero, or the set of all points  $x \in \mathcal{X}$  for which every neighbourhood of  $x$  has strictly positive  $\mu$ -measure.

$\mathcal{M}_1(\mathcal{X})$  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. 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  $\mu$  on the power set of  $\mathcal{X}$  are in bijection with row vectors

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

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  $\mu$  is the  $(n-1)$ -dimensional simplex in  $\mathbb{R}^n$  that is

the convex hull of the  $n$  points

$$\begin{aligned}\delta_1 &= [1 \ 0 \ \cdots \ 0], \\ \delta_2 &= [0 \ 1 \ \cdots \ 0], \\ &\vdots \\ \delta_n &= [0 \ 0 \ \cdots \ 1].\end{aligned}$$

Looking ahead, the expected value of  $f$  under  $\mu$  is exactly the matrix product:

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

The geometry of  $\mathcal{M}_1(\mathcal{X})$  is something that one forgets in favour of the algebraic properties of measures and functions at one's peril, as poetically highlighted by Sir Michael Atiyah [4, 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:

ΓΕΩΜΕΤΡΗΤΟΣ ΜΗΔΕΙΣ ΕΙΣΙΤΩ

In a sense that will be made precise later, 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.

**Definition 2.6.** 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  $(\Theta, \mathcal{F})$  is defined by

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

**Theorem 2.7** (Bayes' rule). *If  $(\Theta, \mathcal{F}, \mu)$  is a probability space and  $A, B \in \mathcal{F}$  have  $\mu(A), \mu(B) > 0$ , then*

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

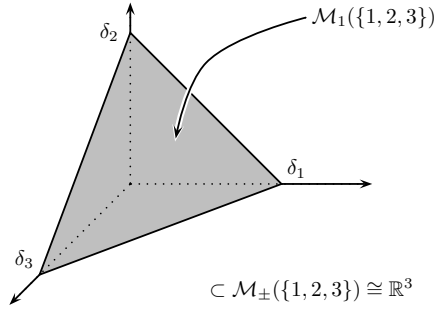


Figure 2.1: The probability simplex  $\mathcal{M}_1(\{1, 2, 3\})$ , drawn as the triangle spanned by the unit Dirac masses  $\delta_i$ ,  $i \in \{1, 2, 3\}$ , in the space of signed measures on  $\{1, 2, 3\}$ .

## 2.2 Random Variables and Stochastic Processes

**Definition 2.8.** Let  $(\mathcal{X}, \mathcal{F})$  and  $(\mathcal{Y}, \mathcal{G})$  be measurable spaces. A function  $f: \mathcal{X} \rightarrow \mathcal{Y}$  generates a  $\sigma$ -algebra on  $\mathcal{X}$  by

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

and  $f$  is called a *measurable function* if  $\sigma(f) \subseteq \mathcal{F}$ . A measurable function whose domain is a probability space is usually called a *random variable*.

**Definition 2.9.** Let  $\Omega$  be any set and let  $(\Theta, \mathcal{F}, \mu)$  be a probability space. A function  $U: \Omega \times \Theta \rightarrow \mathcal{X}$  such that each  $U(\omega, \cdot)$  is a random variable is called an  $\mathcal{X}$ -valued *stochastic process* on  $\Omega$ .

**Definition 2.10.** 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  $\mu$  by  $f$ , by

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

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

**Definition 2.11.** A *filtration* of a  $\sigma$ -algebra  $\mathcal{F}$  is a family  $\mathcal{F}_\bullet = \{\mathcal{F}_i \mid i \in I\}$  of sub- $\sigma$ -algebras of  $\mathcal{F}$ , indexed by an ordered set  $I$ , such that

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

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

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

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  $\sigma$ -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 Aside: 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  $\mu$ .

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  $\theta$  (taking values in a space  $\Theta$ ) in advance of observing data are summarized by a *prior* probability measure  $\pi$  on  $\Theta$ . 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  $\theta$  were the truth. A Bayesian’s belief about  $\theta$  given the prior  $\pi$  and the observed datum  $y$  is the *posterior* probability measure  $\pi(\cdot|y)$  on  $\Theta$ , 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 fancier way that generalizes better to infinite-dimensional  $\Theta$ ,

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

Both the previous two equations are referred to as *Bayes’ rule*, and are at this stage informal applications of the standard Bayes’ rule (Theorem 2.7) for events  $A$  and  $B$  of non-zero probability.

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  $\theta$  and variance 1. We set our frequentist and Bayesian statisticians the challenge of estimating  $\theta$  from the data  $X_1, \dots, X_n$ .

- To the frequentist,  $\theta$  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  $\theta$  with high probability, and hence to give a confidence interval for  $\theta$ , but  $\theta$  itself does not have a distribution.

- To the Bayesian,  $\theta$  is a *random variable*, and its distribution in advance of seeing the data is encoded in a prior  $\pi$ . 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  $\theta$  in view of  $\pi$ ,  $L(y|\theta) \propto e^{-|y-\theta|^2/2}$  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*

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

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 of these notes.

## 2.4 Lebesgue Integration

**Definition 2.12.** Let  $(\mathcal{X}, \mathcal{F}, \mu)$  be a measure space. A function  $f: \mathcal{X} \rightarrow \mathbb{K}$  is called *simple* if  $f = \sum_{i=1}^n \alpha_i \mathbb{1}_{E_i}$  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  $f := \sum_{i=1}^n \alpha_i \mathbb{1}_{E_i}$  is defined to be

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

**Definition 2.13.** 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

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

**Definition 2.14.** 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

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

One of the major attractions of the Lebesgue integral is that, subject to a simple domination condition, pointwise convergence of integrands is enough to ensure convergence of integral values:

**Theorem 2.15** (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  $\mu$ -almost everywhere.

**Definition 2.16.** 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

$$\begin{aligned} C &:= \mathbb{E}_{\mu}[(X - \mathbb{E}_{\mu}[X])(X - \mathbb{E}_{\mu}[X])^*] \in \mathbb{K}^{d \times d} \\ \text{i.e. } C_{ij} &:= \mathbb{E}_{\mu}[(X_i - \mathbb{E}_{\mu}[X_i])(\overline{X_j - \mathbb{E}_{\mu}[X_j]})] \in \mathbb{K} \end{aligned}$$

is the *covariance matrix* of  $X$ .

**Definition 2.17.** 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

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

where

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

for  $1 \leq p < \infty$  and

$$\begin{aligned} \|f\|_{L^\infty(\mu)} &:= \inf \{ \|g\|_\infty \mid f = g: \mathcal{X} \rightarrow \mathbb{K} \text{ } \mu\text{-almost everywhere} \} \\ &= \inf \{ t \geq 0 \mid |f| \leq t \text{ } \mu\text{-almost everywhere} \} \end{aligned}$$

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  $\mu$ -measure zero are identified.

**Theorem 2.18** (Chebyshev's inequality). *Let  $X \in L^p(\Theta, \mu; \mathbb{K})$ ,  $1 \leq p < \infty$ , be a random variable. Then, for all  $t \geq 0$ ,*

$$\mathbb{P}_\mu[|X - \mathbb{E}_\mu[X]| \geq t] \leq t^{-p} \mathbb{E}_\mu[|X|^p]. \quad (2.1)$$

**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. Many interesting 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:

1. 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.

2. 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 [44], and so on.

## 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

$$\nu: E \mapsto \int_E \rho(x) d\mu(x) \quad (2.2)$$



defines a measure  $\nu$  on  $(\mathcal{X}, \mathcal{F})$ . It is natural to ask whether every measure  $\nu$  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  $\rho$  that will make (2.2) hold when  $\mu$  and  $\nu$  are Lebesgue measure and a unit Dirac measure (or vice versa) on  $\mathbb{R}$ .

**Definition 2.19.** Let  $\mu$  and  $\nu$  be measures on a measurable space  $(\mathcal{X}, \mathcal{F})$ . If, for  $E \in \mathcal{F}$ ,  $\nu(E) = 0$  whenever  $\mu(E) = 0$ , then  $\nu$  is said to be *absolutely continuous* with respect to  $\mu$ , denoted  $\nu \ll \mu$ . If  $\nu \ll \mu \ll \nu$ , then  $\mu$  and  $\nu$  are said to be *equivalent*, and this is denoted  $\mu \approx \nu$ . If there exists  $E \in \mathcal{F}$  such that  $\mu(E) = 0$  and  $\nu(\mathcal{X} \setminus E) = 0$ , then  $\mu$  and  $\nu$  are said to be *mutually singular*, denoted  $\mu \perp \nu$ .

**Theorem 2.20 (Radon–Nikodým).** Suppose that  $\mu$  and  $\nu$  are  $\sigma$ -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  $\rho$  with this property are equal  $\mu$ -almost everywhere.

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

## 2.6 Product Measures and Independence

**Definition 2.21.** Let  $(\Theta, \mathcal{F}, \mu)$  be a probability space.

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

**Definition 2.22.** Let  $(\mathcal{X}, \mathcal{F}, \mu)$  and  $(\mathcal{Y}, \mathcal{G}, \nu)$  be  $\sigma$ -finite measure spaces. The *product  $\sigma$ -algebra*  $\mathcal{F} \otimes \mathcal{G}$  is the  $\sigma$ -algebra on  $\mathcal{X} \times \mathcal{Y}$  that is generated by the measurable rectangles, i.e. the smallest  $\sigma$ -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 2.23.** 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  $\sigma$ -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 2.24.** 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  $\mu$  be the (joint) distribution of  $X$ , and  $\mu_i$  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 2.25** (Fubini–Tonelli). Let  $(\mathcal{X}, \mathcal{F}, \mu)$  and  $(\mathcal{Y}, \mathcal{G}, \nu)$  be  $\sigma$ -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:

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

## 2.7 Gaussian Measures

An important class of probability measures and random variables is the class of Gaussian measures (also known as normal distributions) and random variables. Gaussian measures are particularly important because, unlike Lebesgue measure, they are well-defined on infinite-dimensional topological vector spaces; the non-existence of an infinite-dimensional Lebesgue measure is a consequence of the following theorem:

**Theorem 2.26.** Let  $\mathcal{V}$  be an infinite-dimensional, separable Banach space. Then the only Borel measure  $\mu$  on  $\mathcal{V}$  that is locally finite (i.e. every point of  $\mathcal{V}$  has a neighbourhood of finite  $\mu$ -measure) and translation invariant (i.e.  $\mu(x + E) = \mu(E)$  for all  $x \in \mathcal{V}$  and measurable  $E \subseteq \mathcal{V}$ ) is the trivial measure.

Gaussian measures on  $\mathbb{R}^d$  are defined using a Radon–Nikodým derivative with respect to Lebesgue measure:

**Definition 2.27.** 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

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

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

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 2.26, there may not be a Lebesgue measure with respect to which densities can be taken:

**Definition 2.28.** Let  $\mathcal{V}$  be a (locally convex) topological vector space. A Borel measure  $\gamma$  on  $\mathcal{V}$  is said to be a (*non-degenerate*) *Gaussian measure* if, for every  $\ell \in \mathcal{V}'$ , the push-forward measure  $\ell_*\gamma$  is a (non-degenerate) Gaussian measure on  $\mathbb{R}$ . Equivalently,  $\gamma$  is Gaussian if, for every linear map  $T: \mathcal{V} \rightarrow \mathbb{R}^d$ ,  $T_*\gamma = \mathcal{N}(m, C)$  for some  $m \in \mathbb{R}^d$  and some symmetric positive-definite  $C \in \mathbb{R}^{d \times d}$ .

**Definition 2.29.** Let  $\mu$  be a probability measure on a Banach space  $\mathcal{V}$ . An element  $m_\mu \in \mathcal{V}$  is called the *mean* of  $\mu$  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  $\mu$  is said to be *centred*. The *covariance operator* is the symmetric operator  $C_\mu: \mathcal{V}' \times \mathcal{V}' \rightarrow \mathbb{K}$  defined by

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

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

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

The inverse of  $C_\mu$ , if it exists, is called the *precision operator* of  $\mu$ .

**Theorem 2.30 (Vakhania).** *Let  $\mu$  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  $\mu$  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 2.31.** *For a Gaussian measure  $\mu$  on a separable, reflexive Banach space  $\mathcal{V}$ , the following are equivalent:*

1.  $\mu$  is non-degenerate;
2.  $C_\mu: \mathcal{V}' \rightarrow \mathcal{V}$  is one-to-one;
3.  $\overline{C_\mu \mathcal{V}'} = \mathcal{V}$ .

**Example 2.32.** 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,  $\mu$ ) 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

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

**Theorem 2.33.** A centred probability measure  $\mu$  on  $\mathcal{V}$  is a Gaussian measure if and only if its Fourier transform  $\hat{\mu}: \mathcal{V}' \rightarrow \mathbb{C}$  satisfies

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

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

**Theorem 2.34 (Fernique).** Let  $\mu$  be a centered 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,  $\mu$  has moments of all orders: for all  $k \geq 0$ ,

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

**Definition 2.35.** Let  $K: \mathcal{H} \rightarrow \mathcal{H}$  be a linear operator on a separable Hilbert space  $\mathcal{H}$ .

1.  $K$  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  $(\sigma_n)$  such that

$$K = \sum_n \sigma_n \langle v^*, \cdot \rangle u_n,$$

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

2.  $K$  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.
3. 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 2.36.** Let  $\mu$  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, symmetric and of trace class, then there is a Gaussian measure  $\mu$  on  $\mathcal{H}$  such that  $C_{\mu} = K$ .

**Definition 2.37.** Let  $\mu = \mathcal{N}(m_{\mu}, C_{\mu})$  be a Gaussian measure on a Banach space  $\mathcal{V}$ . The *Cameron-Martin space* is the Hilbert space  $\mathcal{H}_{\mu}$  defined equivalently by:

- $\mathcal{H}_{\mu}$  is the completion of

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

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

- $\mathcal{H}_\mu$  is the completion of the range of the covariance operator  $C_\mu: \mathcal{V}' \rightarrow \mathcal{V}$  with respect to this inner product (cf. the closure with respect to the norm in  $\mathcal{V}$  in Theorem 2.30).
- If  $\mathcal{V}$  is Hilbert, then  $\mathcal{H}_\mu$  is the completion of  $\mathcal{R}(C_\mu^{1/2})$  with the inner product  $\langle h, k \rangle_\mu = \langle C_\mu^{-1/2}h, C_\mu^{-1/2}k \rangle_{\mathcal{V}}$ .
- $\mathcal{H}_\mu$  is the set of all  $v \in \mathcal{V}$  such that  $(T_v)_*\mu \approx \mu$ .
- $\mathcal{H}_\mu$  is the intersection of all linear subspaces of  $\mathcal{V}$  that have full  $\mu$ -measure.

Note, however, that if  $\mathcal{H}_\mu$  is infinite-dimensional, then  $\mu(\mathcal{H}_\mu) = 0$ . Furthermore, infinite-dimensional spaces have the rather alarming 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 — and surprisingly simple operations can destroy equivalence.

**Theorem 2.38** (Feldman–Hájek). *Let  $\mu, \nu$  be Gaussian probability measures on a locally convex topological vector space  $\mathcal{V}$ . Then either*

- $\mu$  and  $\nu$  are equivalent, i.e.  $\mu(E) = 0 \iff \nu(E) = 0$ ; or
- $\mu$  and  $\nu$  are mutually singular, i.e. there exists  $E$  such that  $\mu(E) = 0$  and  $\nu(E) = 1$ .

Furthermore, equivalence holds if and only if

1.  $\mathcal{R}(C_\mu^{1/2}) = \mathcal{R}(C_\nu^{1/2}) = E$ ; and
2.  $m_\mu - m_\nu \in E$ ; and
3.  $T := (C_\mu^{-1/2}C_\nu^{1/2})(C_\mu^{-1/2}C_\nu^{1/2})^* - I$  is Hilbert–Schmidt in  $\overline{E}$ .

**Proposition 2.39.** *Let  $\mu$  be a centred Gaussian measure on a separable 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  $\mu$  if and only if  $c \in \{\pm 1\}$ , and  $(D_c)_*\mu$  and  $\mu$  are mutually singular otherwise.*

## Bibliography

At Warwick, this material is mostly covered in [MA359 Measure Theory](#) and [ST318 Probability Theory](#). Gaussian measure theory in infinite-dimensional spaces is covered in [MA482 Stochastic Analysis](#) and [MA612 Probability on Function Spaces and Bayesian Inverse Problems](#). Vakhania’s theorem (Theorem 2.30) on the support of a Gaussian measure can be found in [110]. Fernique’s theorem (Theorem 2.34) on the integrability of Gaussian vectors was proved in [31]. The Feldman–Hájek dichotomy (Theorem 2.38) was proved independently by Feldman [30] and Hájek [36] in 1958.

Gordon’s book [35] is mostly a text on the gauge integral, but its first chapters provide an excellent condensed introduction to measure theory and Lebesgue integration. The book of Capiński & Kopp [18] 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 [10]. The Bochner integral was introduced by Bochner in [12]; recent texts on the topic include those of Diestel & Uhl [24] and Mikusiński [67]. For detailed treatment of the Pettis integral, see Talagrand [102]. Further dis-



W

cussion of the relationship between tensor products and spaces of vector-valued integrable functions can be found in the book of Ryan [85].

Bourbaki [15] 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.

The modern origins of imprecise probability lie in treatises like those of Boole [13] and Keynes [50]; more recent foundations and expositions have been put forward by Walley [114], Kuznetsov [56], Weichselberger [115], and by Dempster [23] and Shafer [87].

## Chapter 3

# Recap of 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 basic and useful constructions like 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 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 of the fields  $\mathbb{R}$  or  $\mathbb{C}$ .

**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

1. *positive semi-definite*: for all  $x \in \mathcal{V}$ ,  $\|x\| \geq 0$ ;
2. *positive definite*: for all  $x \in \mathcal{V}$ ,  $\|x\| = 0$  if and only if  $x = 0$ ;
3. *homogeneous*: for all  $x \in \mathcal{V}$  and  $\alpha \in \mathbb{K}$ ,  $\|\alpha x\| = |\alpha| \|x\|$ ;
4. *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 *semi-norm*. A vector space equipped with a (semi-)norm is called a *(semi-)normed space*.

**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

1. *positive semi-definite*: for all  $x \in \mathcal{V}$ ,  $\langle x, x \rangle \geq 0$ ;
2. *positive definite*: for all  $x \in \mathcal{V}$ ,  $\langle x, x \rangle = 0$  if and only if  $x = 0$ ;
3. *conjugate symmetric*: for all  $x, y \in \mathcal{V}$ ,  $\langle x, y \rangle = \overline{\langle y, x \rangle}$ ;

4. *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.

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*

$$|\langle x, y \rangle| \leq \|x\|^{1/2} \|y\|^{1/2} \quad \text{for all } x, y \in \mathcal{V}. \quad (3.1)$$

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, 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.

**Example 3.3.** 1. For any  $n \in \mathbb{N}$ , the coordinate space  $\mathbb{K}^n$  is an inner product space under the *Euclidean inner product*

$$\langle x, y \rangle := \sum_{i=1}^n \overline{x_i} y_i.$$

In the real case, this is usually known as the *dot product* and denoted  $x \cdot y$ .

2. For any  $m, n \in \mathbb{N}$ , the space  $\mathbb{K}^{m \times n}$  of  $m \times n$  matrices is an inner product space under the *Frobenius inner product*

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

**Definition 3.4.** Let  $(\mathcal{V}, \|\cdot\|)$  be a normed space. 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  $n \geq N$ ,  $\|x_n - x\| < \varepsilon$ . 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  $m, n \geq N$ ,  $\|x_m - x_n\| < \varepsilon$ . 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*.

**Example 3.5.** 1.  $\mathbb{K}^n$  and  $\mathbb{K}^{m \times n}$  are finite-dimensional Hilbert spaces with respect to their usual inner products.



2. The standard example of an infinite-dimensional Hilbert space is the space of *square-summable sequences*,

$$\ell^2(\mathbb{K}) := \left\{ x = (x_n)_{n \in \mathbb{N}} \in \mathbb{K}^{\mathbb{N}} \mid \|x\|_{\ell^2} := \sum_{n \in \mathbb{N}} |x_n|^2 < \infty \right\},$$

is a Hilbert space with respect to the inner product

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

3. Given a measure space  $(\mathcal{X}, \mathcal{F}, \mu)$ , the space  $L^2(\mathcal{X}, \mu; \mathbb{K})$  of (equivalence classes modulo equality  $\mu$ -almost everywhere) of square-integrable functions from  $\mathcal{X}$  to  $\mathbb{K}$  is a Hilbert space with respect to the inner product

$$\langle f, g \rangle_{L^2(\mu)} := \int_{\mathcal{X}} \overline{f(x)} g(x) d\mu(x). \quad (3.5)$$

Note that it is necessary to take the quotient by the equivalence relation of equality  $\mu$ -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:

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

When  $L^2(\mathcal{X}, \mu; \mathbb{K})$  is a separable space, it is isometrically isomorphic to  $\ell^2(\mathbb{K})$  (see Theorem 3.16).

4. 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 a  $\mathbb{K}$ -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 3.6.** 1. For any compact topological space  $\mathcal{X}$ , the space  $\mathcal{C}(\mathcal{X}; \mathbb{K})$  of continuous functions  $f: \mathcal{X} \rightarrow \mathbb{K}$  is a Banach space with respect to the *supremum norm*

$$\|f\|_{\infty} := \sup_{x \in \mathcal{X}} |f(x)|. \quad (3.6)$$

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

2. For  $1 \leq p \leq \infty$ , the spaces  $L^p(\mathcal{X}, \mu; \mathbb{K})$  with their norms

$$\|f\|_{L^p(\mu)} := \left( \int_{\mathcal{X}} |f(x)|^p d\mu(x) \right)^{1/p} \quad (3.7)$$

for  $1 \leq p < \infty$  and

$$\begin{aligned} \|f\|_{L^{\infty}(\mu)} &:= \inf \{ \|g\|_{\infty} \mid f = g: \mathcal{X} \rightarrow \mathbb{K} \text{ } \mu\text{-almost everywhere} \} \\ &= \inf \{ t \geq 0 \mid |f| \leq t \text{ } \mu\text{-almost everywhere} \} \end{aligned} \quad (3.8)$$

are Banach spaces, but only the  $L^2$  spaces are Hilbert spaces.

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 Hilbert spaces. To save space, we use multi-index notation for derivatives:  $\alpha := (\alpha_1, \dots, \alpha_n) \in \mathbb{N}_0^n$  denotes a multi-index, and  $|\alpha| := \alpha_1 + \dots + \alpha_n$ .

**Definition 3.7.** Let  $\Omega \subseteq \mathbb{R}^n$ , let  $\alpha \in \mathbb{N}_0^n$ , and consider  $f: \Omega \rightarrow \mathbb{R}$ . A *weak derivative of order  $\alpha$*  for  $f: \Omega \rightarrow \mathbb{R}$  is a function  $v: \Omega \rightarrow \mathbb{R}$  such that

$$\int_{\Omega} f(x) \frac{\partial^{|\alpha|}}{\partial^{\alpha_1} x_1 \dots \partial^{\alpha_n} x_n} \phi(x) dx = (-1)^{|\alpha|} \int_{\Omega} v(x) \phi(x) dx \quad (3.9)$$

for every  $\phi \in \mathcal{C}_c^\infty(\Omega; \mathbb{R})$ . Such a weak derivative is usually denoted  $D^\alpha f$ , and coincides with the classical (strong) derivative if it exists. The *Sobolev space*  $H^k(\Omega)$  is

$$H^k(\Omega) := \left\{ f \in L^2(\Omega) \mid \begin{array}{l} \text{for all } \alpha \in \mathbb{N}_0^n \text{ with } |\alpha| \leq k, \\ f \text{ has a weak derivative } D^\alpha f \in L^2(\Omega) \end{array} \right\} \quad (3.10)$$

with the inner product

$$\langle u, v \rangle_{H^k} := \sum_{|\alpha| \leq k} \langle D^\alpha u, D^\alpha v \rangle_{L^2}. \quad (3.11)$$

## 3.2 Dual Spaces and Adjoints

**Definition 3.8.** The (*continuous*) *dual space* of a normed space  $\mathcal{V}$  over  $\mathbb{K}$  is the vector space  $\mathcal{V}'$  of all 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  $\ell$ , 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 3.9.** 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. This stands in stark contrast to the duals of even elementary Banach spaces.

**Theorem 3.10** (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.

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 the relation

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

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 and define the adjoint in terms of inner products:

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

Given a matrix  $\{e_i\}_{i \in I}$  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.

### 3.3 Orthogonality and Direct Sums

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

**Definition 3.11.** 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

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

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 3.12.** 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

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

*Proof.* Deferred to Lemma 4.20 and the more general context of convex optimization.  $\square$

The operator  $\Pi_{\mathcal{K}}: \mathcal{H} \rightarrow \mathcal{K}$  is called the *orthogonal projection* onto  $\mathcal{K}$ .

**Theorem 3.13.** Let  $\mathcal{K}$  and  $\mathcal{L}$  be closed subspaces of a Hilbert space  $\mathcal{H}$ . The corresponding orthogonal projection operators

1. are continuous linear operators of norm at most 1;
2. are such that  $\mathbb{I} - \Pi_{\mathcal{K}} = \Pi_{\mathcal{K}^\perp}$ ;

and satisfy, for every  $x \in \mathcal{H}$ ,

3.  $\|x\|^2 = \|\Pi_{\mathcal{K}}x\|^2 + \|(\mathbb{I} - \Pi_{\mathcal{K}})x\|^2$ ;
4.  $\Pi_{\mathcal{K}}x = x \iff x \in \mathcal{K}$ ;
5.  $\Pi_{\mathcal{K}}x = 0 \iff x \in \mathcal{K}^\perp$ .

**Example 3.14** (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  $\sigma$ -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  $\sigma$ -algebra generated by a single event  $E$  of positive  $\mu$ -probability, i.e.

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

The orthogonal projection point of view makes two important properties of conditional expectation intuitively obvious:

1. 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  $\sigma$ -algebra  $\{\emptyset, \Theta\}$ ,

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

2. 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 3.12) 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 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.

Orthogonal direct sums and orthogonal bases are among the most important constructions in Hilbert space theory, and will be very useful in what follows. The prototypical example to bear in mind is 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<sup>(3.1)</sup> that any periodic function  $f$  could be written as

$$\begin{aligned} f(x) &= \sum_{n \in \mathbb{Z}} \hat{f}_n e_n(x), \\ \hat{f}_n &:= \int_{\mathbb{S}^1} f(y) \overline{e_n(y)} dy, \end{aligned}$$

can be seen as one of the historical drivers behind the development of much of analysis. Other important examples are the systems of orthogonal polynomials that will be considered in Chapter 8. Some important results about orthogonal systems are summarized below:

**Lemma 3.15 (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  $\sum_{n \in \mathbb{N}} |\langle x, e_n \rangle|^2$  converges and satisfies*

$$\sum_{n \in \mathbb{N}} |\langle x, e_n \rangle|^2 \leq \|x\|^2. \quad (3.12)$$

**Theorem 3.16 (Parseval identity).** *Let  $\mathcal{H}$  be a Hilbert space, let  $(e_n)_{n \in \mathbb{N}}$  be an orthonormal sequence in  $\mathcal{H}$ , and let  $(\alpha_n)_{n \in \mathbb{N}}$  be a sequence in  $\mathbb{K}$ . Then the series*

<sup>(3.1)</sup>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.

$\sum_{n \in \mathbb{N}} \alpha_n e_n$  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

$$\left\| \sum_{n \in \mathbb{N}} \alpha_n e_n \right\|^2 = \sum_{n \in \mathbb{N}} |\alpha_n|^2. \quad (3.13)$$

**Corollary 3.17.** Let  $\mathcal{H}$  be a Hilbert space and let  $(e_n)_{n \in \mathbb{N}}$  be an orthonormal sequence in  $\mathcal{H}$ . Then, for any  $x \in \mathcal{H}$ , the series  $\sum_{n \in \mathbb{N}} \langle x, e_n \rangle e_n$  converges.

**Theorem 3.18.** Let  $\mathcal{H}$  be a Hilbert space and let  $(e_n)_{n \in \mathbb{N}}$  be an orthonormal sequence in  $\mathcal{H}$ . Then the following are equivalent:

1.  $\{e_n \mid n \in \mathbb{N}\}^\perp = \{0\}$ ;
2.  $\mathcal{H} = \overline{\text{span}\{e_n \mid n \in \mathbb{N}\}}$ ;
3.  $\mathcal{H} = \bigoplus_{n \in \mathbb{N}} \mathbb{K} e_n$  as a direct sum of Hilbert spaces;
4. for all  $x \in \mathcal{H}$ ,  $\|x\|^2 = \sum_{n \in \mathbb{N}} |\langle x, e_n \rangle|^2$ ;
5. for all  $x \in \mathcal{H}$ ,  $x = \sum_{n \in \mathbb{N}} \langle x, e_n \rangle e_n$ .

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

*Proof.* Let  $v := \sum_{n=1}^N v_n e_n$  be any element of  $\text{span}\{e_1, \dots, e_N\}$ . By completeness,  $x = \sum_{n \in \mathbb{N}} \langle x, e_n \rangle e_n$ . Hence,

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

since  $\langle e_n, e_m \rangle = \delta_{nm}$ . □

### 3.4 Tensor Products

Intuitively, the tensor product  $\mathcal{V} \otimes \mathcal{W}$  of two vector spaces  $\mathcal{V}$  and  $\mathcal{W}$  over a common field  $\mathbb{K}$  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  $\mathcal{W}$ . 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 3.20.** 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:

$$F_{\mathcal{V} \times \mathcal{W}} := \left\{ \sum_{i=1}^n \alpha_i e_{(v_i, w_i)} \mid n \in \mathbb{N}, \alpha_i \in \mathbb{K}, (v_i, w_i) \in \mathcal{V} \times \mathcal{W} \right\}$$

The ‘freeness’ of  $F_{\mathcal{V} \times \mathcal{W}}$  is that the elements  $e_{(v,w)}$  are, by definition linearly independent for distinct pairs  $(v, w) \in \mathcal{V} \times \mathcal{W}$ . 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 3.21.** 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 3.22.** 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 3.23.** 1. 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.

2. 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  $L^2(\mathcal{X}, \mu; \mathcal{H})$ .

3. 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 *Bochner spaces* of functions (random variables) taking values in a Banach space  $\mathcal{V}$  that are  $p^{\text{th}}$ -power-integrable in the sense that  $\int_{\mathcal{X}} \|f(x)\|_{\mathcal{V}}^p d\mu(x)$  is finite, and identify this 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’ tensor product of Banach spaces as there is for Hilbert spaces. Readers who are interested in such spaces should consult the book of Ryan [85].

## Bibliography

At Warwick, the theory of Hilbert and Banach spaces is covered in courses such as [MA3G7 Functional Analysis I](#) and [MA3G8 Functional Analysis II](#). Sobolev spaces are covered in [MA4A2 Advanced PDEs](#).

Classic reference texts on elementary functional analysis, including Banach and Hilbert space theory, include the monographs of Reed & Simon [79], Rudin [83], and Rynne & Youngson [86]. Further discussion of the relationship between tensor products and spaces of vector-valued integrable functions can be found in the book of Ryan [85].

Truly intrepid students may wish to consult Bourbaki [14], but the standard warnings about Bourbaki texts apply: the presentation is comprehensive but often forbiddingly austere, and so is perhaps better as a reference text than as a learning tool. Aliprantis & Border’s *Hitchhiker’s Guide* [2] is another encyclopædic text, but is surprisingly readable despite the Bourbakiste order in which material is presented.



## Chapter 4

# Basic 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.

### 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.** The *arg min* or *set of global minimizers* of  $f: \mathcal{X} \rightarrow \mathbb{R} \cup \{\pm\infty\}$  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  $f: \mathcal{X} \rightarrow \mathbb{R} \cup \{\pm\infty\}$  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.** A constraint is said to be

1. *redundant* if it does not change the feasible set, and *relevant* otherwise;
2. *non-binding* if it does not change the extreme value, and *binding* otherwise;
3. *active* if it holds as an equality at the extremizer, and *inactive* otherwise.

## 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, 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 4.3.** Given a differentiable function  $f$  and an initial state  $x_0$ , *Newton's method* for finding a zero of  $f$  is the sequence generated by the iteration

$$x_{n+1} := x_n - (Df(x_n))^{-1} f(x_n).$$

Newton's method is often applied to find critical points of  $f$ , i.e. points where  $Df$  vanishes, in which case the iteration is.

$$x_{n+1} := x_n - (D^2 f(x_n))^{-1} Df(x_n).$$

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  $f$ , in order to illustrate some of the relevant issues. For simplicity, suppose that  $f$  has a unique global minimizer  $\mathbf{x}_{\min}$  and write  $f_{\min}$  for  $f(\mathbf{x}_{\min})$ .

**Algorithm 4.4** (Random sampling). For simplicity, the following algorithm runs for  $n_{\max}$  steps with no convergence checks. The algorithm returns an approximate minimizer  $\mathbf{x}_{\text{best}}$  along with the corresponding value of  $f$ . Suppose that `random()` generates independent samples of  $\mathcal{X}$  from a probability measure  $\mu$  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 4.4 is that it completely neglects local information about  $f$ . Even if the current state  $\mathbf{x}_{\text{best}}$  is very close to the global minimizer  $\mathbf{x}_{\min}$ , the algorithm may continue to sample points  $\mathbf{x}_{\text{new}}$  that are very far away and have  $f(\mathbf{x}_{\text{new}}) \gg f(\mathbf{x}_{\text{best}})$ . It would be preferable to explore a neighbourhood of  $\mathbf{x}_{\text{best}}$  more thoroughly and hence find a better approximation of  $[\mathbf{x}_{\min}, f_{\min}]$ . The next algorithm attempts to rectify this deficiency.

**Algorithm 4.5** (Random walk). As before, this algorithm runs for  $n_{\max}$  steps. The algorithm returns an approximate minimizer  $\mathbf{x}_{\text{best}}$  along with the corresponding value of  $f$ . Suppose that an initial state  $\mathbf{x}_0$  is given, and that `jump()` generates independent samples of  $\mathcal{X}$  from a probability measure  $\mu$  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 4.5 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, 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  $\mathbf{x}$  of the algorithm and the best state so far,  $\mathbf{x}_{\text{best}}$ ; unlike in the previous two algorithms, proposed states  $\mathbf{x}_{\text{new}}$  may be accepted and become  $\mathbf{x}$  even if  $f(\mathbf{x}_{\text{new}}) > f(\mathbf{x}_{\text{best}})$ . 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 4.6** (Simulated annealing). Suppose that an initial state  $\mathbf{x}_0$  is given, and that functions `temperature()`, `neighbour()` and `acceptance_prob()` have been specified. Suppose that `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 4.4, the simulated annealing method can guarantee to find the global minimizer of  $f$  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*:

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

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 4.7 (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  $u_0$  is a constrained extremizer of  $f$  subject to the constraint that  $g(u_0) = 0$ . Suppose also that the Fréchet derivative  $Dg(u_0): \mathcal{X} \rightarrow \mathcal{Y}$  is surjective. Then there exists a Lagrange multiplier  $\lambda_0 \in \mathcal{Y}'$  such that  $(u_0, \lambda_0)$  is an unconstrained critical point of the Lagrangian  $\mathcal{L}$  defined by*

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

i.e.  $Df(u_0) = \lambda_0 \circ Dg(u_0)$  as linear maps from  $\mathcal{X}$  to  $\mathbb{R}$ .

The corresponding result for inequality constraints is the Karush–Kuhn–Tucker theorem:

**Theorem 4.8 (Karush–Kuhn–Tucker).** *Suppose that  $x^* \in \mathbb{R}^n$  is a local minimizer of  $f \in C^1(\mathbb{R}^n; \mathbb{R})$  subject to inequality constraints  $g_i(x) \leq 0$  and equality constraints  $h_j(x) = 0$ , where  $g_i, h_j \in C^1(\mathbb{R}^n; \mathbb{R})$  for  $i = 1, \dots, m$  and  $j = 1, \dots, \ell$ . Then there exist  $\mu \in \mathbb{R}^m$  and  $\lambda \in \mathbb{R}^\ell$  such that*

$$-\nabla f(x^*) = \mu \cdot \nabla g(x^*) + \lambda \cdot \nabla h(x^*),$$

where  $x^*$  is feasible, and  $\mu$  satisfies the dual feasibility criteria  $\mu_i \geq 0$  and the complementary slackness criteria  $\mu_i g_i(x^*) = 0$  for  $i = 1, \dots, m$ .

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  $x^*$ .

**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 `x_new`, accepted states `x`, or even the final solution `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  $\mathbf{x}_{\text{new}}$  is the following:

```

 $\mathbf{x}' = \text{neighbour}(\mathbf{x})$ 
while  $c(\mathbf{x}') > 0$ :
     $\mathbf{x}' = \text{neighbour}(\mathbf{x})$ 
 $\mathbf{x}_{\text{new}} = \mathbf{x}'$ 

```

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  $\mu > 0$ .

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

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  $C()$  that takes as input a possibly infeasible  $\mathbf{x}'$  and returns some  $\mathbf{x}_{\text{new}} = C(\mathbf{x}')$  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 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  $\alpha_i$  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 4.9.** 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  $t \in (0, 1)$ . 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  $\text{ext}(K)$ . 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 4.10.** 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  $(\pm 1, \pm 1)$ . 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  $\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$ . See Figure 4.1 for further examples.

**Example 4.11.**  $\mathcal{M}_1(\mathcal{X})$  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,  $\mathcal{M}_1(\mathcal{X}) = \overline{\text{co}}(\{\delta_x \mid x \in \mathcal{X}\})$  as a subset of the space  $\mathcal{M}_\pm(\mathcal{X})$  of signed measures on  $\mathcal{X}$ .

The reason that these notes restrict attention to Hausdorff, locally convex topological vector spaces  $\mathcal{X}$  is that it is just too much of a headache to work with spaces for which the following ‘common sense’ results do not hold:

**Theorem 4.12** (Kreĭn–Milman [54]). *Let  $\mathcal{X}$  be a Hausdorff, locally convex topological vector space, and let  $K \subseteq \mathcal{X}$  be compact and convex. Then  $K$  is the closed convex hull of its extreme points.*

**Theorem 4.13** (Choquet–Bishop–de Leeuw [11]). *Let  $\mathcal{X}$  be a Hausdorff, locally convex topological vector space, let  $K \subseteq \mathcal{X}$  be compact and convex, and let  $c \in K$ .*

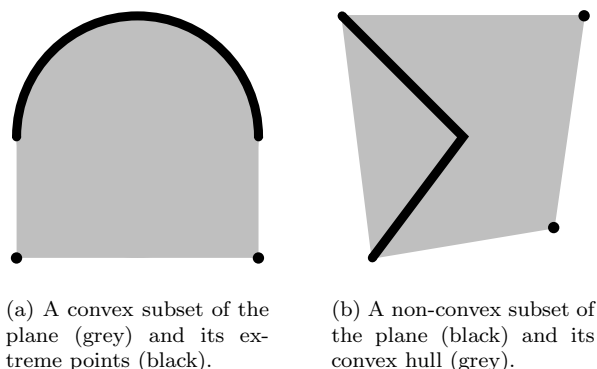


Figure 4.1: Convex sets, extreme points, and convex hulls.

Then there exists a probability measure  $\mu$  supported on  $\text{ext}(K)$  such that, for all affine functions  $f$  on  $K$ ,

$$f(c) = \int_{\text{ext}(K)} f(e) d\mu(e).$$

Informally speaking, the Kreĭn–Milman and Choquet–Bishop–de Leeuw theorems together assure us 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 4.14.** 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  $t \in [0, 1]$ ,

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

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

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

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  $K \times \mathbb{R}$ . Convex functions have many convenient properties with respect to minimization and maximization:

**Theorem 4.15.** Let  $f: K \rightarrow \mathbb{R}$  be a convex function on a compact, convex, non-empty set  $K \subseteq \mathcal{X}$ . Then

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



**Remark.** Note well that Theorem 4.15 does not assert the existence of minimizers, for which simultaneous compactness of  $K$  and lower semicontinuity of  $f$  is required. 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.

*Proof.* 1. 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  $N$ .

2. 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.

3. 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  $K$ .

4. Suppose that  $x$  is a non-extreme point of  $K$  that is also a strict maximizer for  $f$  on  $K$ . Let  $x = x_t = (1-t)x_0 + tx_1$ , where  $x_0, x_1$  are distinct points of  $K$  and  $t \in (0, 1)$ . Since  $x$  is assumed to be a strict maximizer for  $f$  on  $K$ ,  $\max\{f(x_0), f(x_1)\} < f(x_t)$ . Then, since  $f$  is convex,

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

which is a contradiction.  $\square$

**Definition 4.16.** 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 4.17.** 1. 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 4.15, such minimization problems are nicer than general minimization problems, but bona fide convex programs are an even nicer special case.



2. In practice, many problems are not obviously convex programs, but 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, the *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 Chapter 11 of Boyd & Vandenberghe [16].

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  $x_\mu^*$  of  $B(\cdot; \mu)$  lies in  $F$  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  $\mu$  to  $\mu'$ , performs another partial minimization of  $B(\cdot; \mu')$  using Newton's method, and so on in this alternating fashion.

## 4.5 Linear Programming

Theorem 4.15 has the following immediate corollary for the minimization and maximization of affine functions on convex sets:

**Corollary 4.18.** *Let  $\ell: K \rightarrow \mathbb{R}$  be an affine function on a non-empty, compact, convex set  $K \subseteq \mathcal{X}$ . Then*

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

**Definition 4.19.** A *linear program* is an optimization problem of the form

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

where the functions  $f, g_1, \dots, g_m: \mathbb{R}^n \rightarrow \mathbb{R}$  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.

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 worst-case complexity of 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).

**Lemma 4.20.** *Let  $K$  be a closed, convex subset of a Hilbert space  $\mathcal{H}$ . Then, for each  $y \in \mathcal{H}$ , there is a unique element  $\hat{x} \in K$  such that*

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

*Proof.* By Exercise 4.1, the function  $J: \mathcal{X} \rightarrow [0, \infty)$  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  $\mathcal{X}$ , 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,  $\|y - \hat{x}\| = I$ .  $\square$

**Lemma 4.21** (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 4.20, 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  $y \in V$ ,

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

and, in the complex case,

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

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

**Lemma 4.22** (Normal equations). *Let  $A: \mathcal{H} \rightarrow \mathcal{K}$  be a linear operator between Hilbert spaces such that  $\mathcal{R}(A)$  is a closed subspace of  $\mathcal{K}$ . Then, given  $b \in \mathcal{K}$ ,*

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

*the equations on the right-hand side being known as the normal equations.*

*Proof.* Recall that, 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,

$$\begin{aligned}
 & \|Ax - b\|_{\mathcal{K}} \text{ is minimal} \\
 & \iff (Ax - b) \perp Av \text{ for all } v \in \mathcal{H}, \text{ by Lemma 4.21} \\
 & \iff \langle Ax - b, Av \rangle_{\mathcal{K}} = 0 \text{ for all } v \in \mathcal{H} \\
 & \iff \langle A^*Ax - A^*b, v \rangle_{\mathcal{H}} = 0 \text{ for all } v \in \mathcal{H} \\
 & \iff A^*Ax = A^*b \text{ by completeness of } \mathcal{H}. \quad \square
 \end{aligned}$$

**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 operator on  $\mathcal{K}$ .

**Corollary 4.23** (Weighted least squares). *Let  $A: \mathcal{H} \rightarrow \mathcal{K}$  be a linear operator between Hilbert spaces such that  $\mathcal{R}(A)$  is a closed subspace of  $\mathcal{K}$ . Let  $Q: \mathcal{K} \rightarrow \mathcal{K}$  be self-adjoint and positive-definite and let*

$$\langle k, k' \rangle_Q := \langle k, Qk' \rangle_{\mathcal{K}}$$

Then, given  $b \in \mathcal{K}$ ,

$$\hat{x} \in \arg \min_{x \in \mathcal{H}} \|Ax - b\|_Q \iff A^*QA\hat{x} = A^*Qb.$$

*Proof.* Exercise 4.2.  $\square$

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*:

**Corollary 4.24** (Tikhonov-regularized least squares). *Let  $A: \mathcal{H} \rightarrow \mathcal{K}$  be a linear operator between Hilbert spaces such that  $\mathcal{R}(A)$  is a closed subspace of  $\mathcal{K}$ , let  $Q: \mathcal{H} \rightarrow \mathcal{H}$  be self-adjoint and positive-definite, and let  $b \in \mathcal{K}$  and  $x_0 \in \mathcal{H}$  be given. Let*

$$J(x) := \|Ax - b\|^2 + \|x - x_0\|_Q^2.$$

Then

$$\hat{x} \in \arg \min_{x \in \mathcal{H}} J(x) \iff (A^*A + Q)\hat{x} = A^*b + Qx_0.$$

*Proof.* Exercise 4.3.  $\square$

**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  $\theta$ , 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  $\theta$  to minimize the objective function  $J(\theta) := \|r(\theta)\|_2^2$ . Let

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

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  $\delta$  is ‘too large’, in which case it may be appropriate to use a judiciously chosen small positive multiple of  $\delta$ . 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  $\delta$  is essentially the reciprocal of the Tikhonov regularization parameter.

## Bibliography

W Direct and iterative methods for the solution of linear least squares problems are covered in [MA398 Matrix Analysis and Algorithms](#).

The book of Boyd & Vandenberghe [16] 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 [82]. A standard reference on numerical methods for optimization is the book of Nocedal & Wright [72].

For constrained global optimization in the absence of ‘nice’ features, particularly for the UQ methods in Chapter 14, the author recommends the Differential Evolution algorithm [77, 97] within the [Mystic](#) framework.

## Exercises

**Exercise 4.1.** Let  $\|\cdot\|$  be a norm on a vector space  $\mathcal{X}$ , and fix  $y \in \mathcal{X}$ . Show that the function  $J: \mathcal{X} \rightarrow [0, \infty)$  defined by  $J(x) := \|y - x\|^2$  is strictly convex.

**Exercise 4.2.** Let  $A: \mathcal{H} \rightarrow \mathcal{K}$  be a linear operator between Hilbert spaces such that  $\mathcal{R}(A)$  is a closed subspace of  $\mathcal{K}$ . Let  $Q: \mathcal{K} \rightarrow \mathcal{K}$  be self-adjoint and positive-definite and let

$$\langle k, k' \rangle_Q := \langle k, Qk' \rangle_{\mathcal{K}}$$

Show that, given  $b \in \mathcal{K}$ ,

$$\hat{x} \in \arg \min_{x \in \mathcal{H}} \|Ax - b\|_Q \iff A^*QA\hat{x} = A^*Qb.$$

**Exercise 4.3.** Let  $A: \mathcal{H} \rightarrow \mathcal{K}$  be a linear operator between Hilbert spaces such that  $\mathcal{R}(A)$  is a closed subspace of  $\mathcal{K}$ , let  $Q: \mathcal{H} \rightarrow \mathcal{H}$  be self-adjoint and positive-definite, and let  $b \in \mathcal{K}$  and  $x_0 \in \mathcal{H}$  be given. Let

$$J(x) := \|Ax - b\|^2 + \|x - x_0\|_Q^2.$$

Show that

$$\hat{x} \in \arg \min_{x \in \mathcal{H}} J(x) \iff (A^*A + Q)\hat{x} = A^*b + Qx_0.$$

*Hint: Consider the operator from  $\mathcal{H}$  into  $\mathcal{K} \oplus \mathcal{H}$  given in block form as  $[A, Q^{1/2}]^\top$ .*

DRAFT



## 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  $z_2$ .

## 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 commonly-used 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,

$$\begin{aligned} [a, b] + [c, d] &= [a + c, b + d], \\ [a, b] - [c, d] &= [a - d, b - c], \\ [a, b] \cdot [c, d] &= [\min\{a \cdot c, a \cdot d, b \cdot c, b \cdot d\}, \max\{a \cdot c, a \cdot d, b \cdot c, b \cdot d\}], \\ [a, b] / [c, d] &= [\min\{a/c, a/d, b/c, b/d\}, \max\{a/c, a/d, b/c, b/d\}] \text{ when } 0 \notin [c, d]. \end{aligned}$$

The addition and multiplication operations are commutative, associative and sub-distributive:

$$A(B + B) \subseteq AB + AC.$$

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  $f$ .

The distributional robustness approaches covered in Chapter 14 can be seen as an extension of this approach from partially known real numbers to partially known probability measures.

## 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  $\mu$  is a very rich and high-dimensional object; often it is necessary to summarize the degree of uncertainty implicit in  $\mu$  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  $\mu$  has mean  $m$ , is the variance of  $\mu$ , 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  $\mu$  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-definite matrix. Hence, such a  $C(\mu)$  has  $n$  positive real eigenvalues (counted with multiplicity)

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0,$$

with corresponding normalized eigenvectors  $v_1, \dots, v_n \in \mathbb{R}^n$ . The direction  $v_1$  corresponding to the largest eigenvalue  $\lambda_1$  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  $\psi$  is a unit-norm element of a suitable Hilbert space, usually something like  $L^2(\mathbb{R}^n; \mathbb{C})$ . Physical observables like position, momentum &c. 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).  $\square$

**Information and Entropy.** In information theory as pioneered by Claude Shannon, the *information* (or *surprisal*) associated to 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:

base 2  $\leftrightarrow$  bits, base  $e$   $\leftrightarrow$  nats/nits, base 10  $\leftrightarrow$  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  $x$ . So, 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  $\mu$  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  $\mu$  that has Lebesgue density  $\rho$ :

$$I(x) := -\log \rho(x),$$

$$H(\mu) := -\int_{\mathbb{R}^n} \rho(x) \log \rho(x) dx.$$

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  $\mu$ . (Remember that if  $\mu$  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  $\mu$  and  $\nu$  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 5.2. □

## 5.4 Information Gain

Implicit in the definition and entropy (5.3) is the use of 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. In general, if  $\mu$  is a probability measure on a measurable space  $(\mathcal{X}, \mathcal{F})$  with reference measure  $\pi$ , then we would like to define the entropy of  $\mu$  with respect to  $\pi$  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  $\mu$  has a Radon–Nikodým derivative with respect to  $\pi$ . 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  $\mu, \nu$  be  $\sigma$ -finite measures on  $(\mathcal{X}, \mathcal{F})$ . The *Kullback–Leibler divergence* from  $\mu$  to  $\nu$  is defined to be

$$D_{\text{KL}}(\mu \parallel \nu) := \begin{cases} \int_{\mathcal{X}} \frac{d\mu}{d\nu} \log \frac{d\mu}{d\nu} d\nu \equiv \int_{\mathcal{X}} \log \frac{d\mu}{d\nu} d\mu, & \text{if } \mu \ll \nu, \\ +\infty, & \text{otherwise.} \end{cases}$$

While  $D_{\text{KL}}(\cdot \parallel \cdot)$  is non-negative, and vanishes if and only if its arguments are identical, it is neither symmetric nor does it satisfy the triangle inequality. Nevertheless,  $D_{\text{KL}}(\cdot \parallel \cdot)$  generates a topology on  $\mathcal{M}_1(\mathcal{X})$  that is finer than the total variation topology:

**Theorem 5.5 (Pinsker).** For any  $\mu, \nu \in \mathcal{M}_1(\mathcal{X}, \mathcal{F})$ ,

$$d_{\text{TV}}(\mu, \nu) \leq \sqrt{\frac{D_{\text{KL}}(\mu \parallel \nu)}{2}},$$

where the total variation metric is defined by

$$d_{\text{TV}}(\mu, \nu) := \sup\{|\mu(E) - \nu(E)| \mid E \in \mathcal{F}\}. \quad (5.4)$$

**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  $\rho$ . Consider the problem of selecting an optimal experimental design  $\lambda$  for the inference of some parameters / unknowns  $\theta$  from the observed data  $y$  that will result from the experiment  $\lambda$ . If, for each  $\lambda$  and  $\theta$ , we know the conditional distribution  $y|\lambda, \theta$  of the observed data, then the conditional distribution  $y|\lambda$  is obtained by integration with respect to the prior distribution of  $\theta$ :

$$\rho(y|\lambda) = \int \rho(y|\lambda, \theta) \rho(\theta) d\theta.$$

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  $U(y, \lambda)$  to be the Kullback–Leibler divergence  $D_{\text{KL}}(\rho(\cdot|y, \lambda) \parallel \rho(\cdot|\lambda))$  between the prior and posterior distributions on  $\theta$ . An experimental design  $\lambda$  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.

## Bibliography

Comprehensive treatments of interval analysis include the classic monograph of Moore [69] and the more recent text of Jaulin & al. [42].

Information theory was pioneered by Shannon in his seminal 1948 paper [88]. The Kullback–Leibler divergence was introduced by Kullback & Leibler [55], who in fact considered the symmetrized version of the divergence that now bears their names. The book of MacKay [63] provides a thorough introduction to information theory.

## Exercises

**Exercise 5.1.** Prove *Gibbs' inequality* that the Kullback–Leibler divergence is non-negative, i.e.

$$D_{\text{KL}}(\mu\|\nu) := \int_{\mathcal{X}} \log \frac{d\mu}{d\nu} d\mu \geq 0$$

whenever  $\mu, \nu$  are  $\sigma$ -finite measures on  $(\mathcal{X}, \mathcal{F})$  with  $\mu \ll \nu$ . Show also that  $D_{\text{KL}}(\mu\|\nu) = 0$  if and only if  $\mu = \nu$ .

**Exercise 5.2.** Prove Proposition 5.3. That is, suppose that  $\mu$  and  $\nu$  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 5.3.** 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  $D_{\text{KL}}(\mu_0\|\mu_1)$  is

$$\frac{1}{2} \left( \text{tr}(C_1^{-1}C_0) + (m_1 - m_0)^\top C_1^{-1}(m_1 - m_0) - n - \log \left( \frac{\det C_0}{\det C_1} \right) \right).$$

**Exercise 5.4.** Suppose that  $\mu$  and  $\nu$  are equivalent probability measures on  $(\mathcal{X}, \mathcal{F})$  and define

$$d(\mu, \nu) := \text{ess sup}_{x \in \mathcal{X}} \left| \log \frac{d\mu}{d\nu}(x) \right|.$$

Show that this defines a well-defined metric on the measure equivalence class  $\mathcal{E}$  containing  $\mu$  and  $\nu$ . 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 5.5.** Show that Pinsker's inequality (Theorem 5.5) cannot be reversed. That is, show that, for any  $\varepsilon > 0$ , there exist probability measures  $\mu$  and  $\nu$  with  $d_{\text{TV}}(\mu, \nu) \leq \varepsilon$  but  $D_{\text{KL}}(\mu\|\nu) = +\infty$ .

DRAFT



## Chapter 6

# Bayesian Inverse Problems

It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so.

---

MARK TWAIN

This chapter provides a general introduction, at the high level, to the *backward* propagation of uncertainty/information in the solution of *inverse problems*, and specifically a Bayesian probabilistic perspective on such inverse problems. Under the umbrella of inverse problems, we consider parameter estimation and regression. One specific aim is to make clear the connection between regularization and the application of a Bayesian prior. The filtering methods of Chapter 7 fall under the general umbrella of Bayesian approaches to inverse problems, but have an additional emphasis on real-time computational expediency.

### 6.1 Inverse Problems and Regularization

In many applications it is of interest to solve inverse problems, namely to find  $u$ , an input to a mathematical model, given  $y$ , an observation of (some components of, or functions of) the solution of the model. We have an equation of the form

$$y = H(u)$$

where  $\mathcal{X}$ ,  $\mathcal{Y}$  are, say, Banach spaces,  $u \in \mathcal{X}$ ,  $y \in \mathcal{Y}$ , and  $H: \mathcal{X} \rightarrow \mathcal{Y}$  is the *observation operator*. However, inverse problems are typically ill-posed: there may be no solution, the solution may not be unique, or there may be a unique solution that depends sensitively on  $y$ . Indeed, very often we do not actually observe  $H(u)$ , but some noisily corrupted version of it, such as

$$y = H(u) + \eta. \tag{6.1}$$

The inverse problem framework encompasses that problem of *model calibration* (or *parameter estimation*), where a model  $H_\theta$  relating inputs to outputs depends upon some parameters  $\theta \in \Theta$ , e.g., when  $\mathcal{X} = \mathcal{Y} = \Theta$ ,  $H_\theta(u) = \theta u$ .

The problem is, given some observations of inputs  $u_i$  and corresponding outputs  $y_i$ , to find the parameter value  $\theta$  such that

$$y_i = H_\theta(u_i) \quad \text{for each } i.$$

Again, this problem is typically ill-posed.

One approach to the problem of ill-posedness is to seek a least-squares solution: find, for the norm  $\|\cdot\|_{\mathcal{Y}}$  on  $\mathcal{Y}$ ,

$$\arg \min_{u \in \mathcal{X}} \|y - H(u)\|_{\mathcal{Y}}^2.$$

However, this problem, too, can be difficult to solve as it may possess minimizing sequences that do not have a limit in  $\mathcal{X}$ ,<sup>(6.1)</sup> or may possess multiple minima, or may depend sensitively on the observed data  $y$ . Especially in this last case, it may be advantageous to not try to fit the observed data too closely, and instead *regularize* the problem by seeking

$$\arg \min \left\{ \|y - H(u)\|_{\mathcal{Y}}^2 + \|u - \bar{u}\|_E^2 \mid u \in E \subseteq \mathcal{X} \right\}$$

for some Banach space  $E$  embedded in  $\mathcal{X}$  and a chosen  $\bar{u} \in E$ . The standard example of this regularization setup is *Tikhonov regularization*, as in Corollary 4.24: when  $\mathcal{X}$  and  $\mathcal{Y}$  are Hilbert spaces, for suitable self-adjoint positive-definite operators  $Q$  on  $\mathcal{X}$  and  $R$  on  $\mathcal{Y}$ , we seek

$$\arg \min \left\{ \|Q^{-1/2}(y - H(u))\|_{\mathcal{Y}}^2 + \|R^{-1/2}(u - \bar{u})\|_{\mathcal{X}}^2 \mid u \in \mathcal{X} \right\}$$

However, this approach appears to be somewhat ad hoc, especially where the choice of regularization is concerned.

Taking a probabilistic — specifically, Bayesian — viewpoint alleviates these difficulties. If we think of  $u$  and  $y$  as random variables, then (6.1) defines the conditioned random variable  $y|u$ , and we define the ‘solution’ of the inverse problem to be the conditioned random variable  $u|y$ . This allows us to model the noise,  $\eta$ , via its statistical properties, even if we do not know the exact instance of  $\eta$  that corrupted the given data, and it also allows us to specify a priori the form of solutions that we believe to be more likely, thereby enabling us to attach weights to multiple solutions which explain the data. This is the essence of the Bayesian approach to inverse problems.

**Remark 6.1.** In practice the true observation operator is often approximated by some numerical model  $H(\cdot; h)$ , where  $h$  denotes a mesh parameter, or parameter controlling missing physics. In this case (6.1) becomes

$$y = H(u; h) + \varepsilon + \eta,$$

where  $\varepsilon := H(u) - H(u; h)$ . In principle, the observational noise  $\eta$  and the computational error  $\varepsilon$  could be combined into a single term, but keeping them separate is usually more appropriate: unlike  $\eta$ ,  $\varepsilon$  is typically not of mean zero, and is dependent upon  $u$ .

<sup>(6.1)</sup> Take a moment to reconcile the statement “there may exist minimizing sequences that do not have a limit in  $\mathcal{X}$ ” with  $\mathcal{X}$  being a Banach space.

To illustrate the central role that least squares minimization plays in elementary statistical estimation, and hence motivate the more general considerations of the rest of the chapter, consider the following finite-dimensional linear problem. Suppose that we are interested in learning some vector of parameters  $x \in \mathbb{K}^n$ , which gives rise to a vector  $y \in \mathbb{K}^m$  of observations via

$$y = Ax + \eta,$$

where  $A \in \mathbb{K}^{m \times n}$  is a known linear operator (matrix) and  $\eta$  is a Gaussian *noise vector* known to have mean zero and self-adjoint, positive-definite covariance matrix  $\mathbb{E}[\eta\eta^*] = Q \in \mathbb{K}^{m \times m}$ , with  $\eta$  independent of  $x$ . A common approach is to seek an estimate  $\hat{x}$  of  $x$  that is a *linear* function  $Ky$  of the data  $y$ , is *unbiased* in the sense that  $\mathbb{E}[\hat{x}] = x$ , and is the *best* estimate in that it minimizes an appropriate cost function. The following theorem, the Gauss–Markov theorem, states that there is precisely one such estimator, and that it is the best in two very natural senses:

**Theorem 6.2 (Gauss–Markov).** *Suppose that  $A^*Q^{-1}A$  is invertible. Then, among all unbiased linear estimators  $K \in \mathbb{K}^{n \times m}$ , producing an estimate  $\hat{x} = Ky$  of  $x$ , the one that minimizes both the mean-squared error  $\mathbb{E}[\|\hat{x} - x\|^2]$  and the covariance matrix<sup>(6.2)</sup>  $\mathbb{E}[(\hat{x} - x)(\hat{x} - x)^*]$  is*

$$K = (A^*Q^{-1}A)^{-1}A^*Q^{-1},$$

and the resulting estimate  $\hat{x}$  has  $\mathbb{E}[\hat{x}] = x$  and covariance matrix

$$\mathbb{E}[(\hat{x} - x)(\hat{x} - x)^*] = (A^*Q^{-1}A)^{-1}.$$

**Remark 6.3.** Indeed, by Corollary 4.23,  $\hat{x} = (A^*Q^{-1}A)^{-1}A^*Q^{-1}y$  is also the solution to the weighted least squares problem with weight  $Q^{-1}$ , i.e.

$$\hat{x} = \arg \min_{x \in \mathbb{K}^n} J(x), \quad J(x) := \frac{1}{2} \|Ax - y\|_{Q^{-1}}^2.$$

Note that the first and second derivatives (gradient and Hessian) of  $J$  are

$$\nabla J(x) = A^*Q^{-1}Ax - A^*y, \quad D^2J(x) = A^*Q^{-1}A,$$

so the covariance matrix of  $\hat{x}$  is the inverse of the Hessian of  $J$ . These observations will be useful in the construction of the Kálmán filter in Chapter 7.

*Proof of Theorem 6.2.* Note that the first part of this proof is surplus to requirements: we could simply check that  $K := (A^*Q^{-1}A)^{-1}A^*Q^{-1}$  is indeed the minimal linear unbiased estimator, but it is nice to derive the formula for  $K$  from first principles and get some practice at constrained convex optimization into the bargain.

Since  $K$  is required to be unbiased, it follows that  $KA = I$ . Therefore,

$$\|\hat{x} - x\|^2 = \|Ky - x\|^2 = \|K(Ax + \eta) - x\|^2 = \|K\eta\|^2.$$

---

<sup>(6.2)</sup> Here, the minimization is meant in the sense of positive semi-definite matrices: for two matrices  $A$  and  $B$ , we say that  $A \leq B$  if  $B - A$  is a positive semi-definite matrix.

Since  $\|K\eta\|^2 = \eta^* K^* K \eta$  is a scalar and  $\text{tr}(XY) = \text{tr}(YX)$  for any two rectangular matrices  $X$  and  $Y$  of the appropriate sizes,

$$\mathbb{E}[\|\hat{x} - x\|^2] = \mathbb{E}[\eta^* K^* K \eta] = \text{tr}(\mathbb{E}[K \eta \eta^* K^*]) = \text{tr}(K Q K^*).$$

Thus,  $K$  is the solution to the constrained optimization problem

$$K = \arg \min \{ \text{tr}(K Q K^*) \mid K A = I \}.$$

Note that this is a convex optimization problem, since, by Exercise 6.2,  $K \mapsto \sqrt{\text{tr}(K Q K^*)}$  is a norm. Introduce a matrix  $\Lambda \in \mathbb{K}^{n \times n}$  of Lagrange multipliers, so that the minimizer of the constrained problem is the unique critical point of the Lagrangian

$$\begin{aligned} \mathcal{L}(K, \Lambda) &:= \text{tr}(K Q K^*) - \Lambda : (K A - I) \\ &= \text{tr}(K Q K^* - \Lambda^* (K A - I)). \end{aligned}$$

The critical point of the Lagrangian satisfies

$$0 = \nabla_K \mathcal{L}(K, \Lambda) = K Q^* + K Q - \Lambda A^* = 2K Q - \Lambda A^*,$$

since  $Q$  is self-adjoint. Multiplication on the right by  $Q^{-1}A$ , and using the constraint that  $KA = I$ , reveals that  $\Lambda = 2(A^* Q^{-1} A)^{-1}$ , and hence that  $K = (A^* Q^{-1} A)^{-1} A^* Q^{-1}$ .

It is easily verified that  $K$  is an unbiased estimator:

$$\hat{x} = (A^* Q^{-1} A)^{-1} A^* Q^{-1} (Ax + \eta) = x + (A^* Q^{-1} A)^{-1} A^* Q^{-1} \eta$$

and so, taking expectations of both sides,  $\mathbb{E}[\hat{x}] = x$ . Moreover, the covariance of this estimator satisfies

$$\begin{aligned} \mathbb{E}[(\hat{x} - x)(\hat{x} - x)^*] &= (A^* Q^{-1} A)^{-1} A^* Q^{-1} \mathbb{E}[\eta \eta^*] Q^{-1} A (A^* Q^{-1} A)^{-1} \\ &= (A^* Q^{-1} A)^{-1}, \end{aligned}$$

as claimed.

Now suppose that  $L = K + D$  is any linear unbiased estimator; note that  $DA = 0$ . Then the covariance of the estimate  $Ly$  satisfies

$$\begin{aligned} \mathbb{E}[(Ly - x)(Ly - x)^*] &= \mathbb{E}[(K + D)\eta \eta^* (K^* + D^*)] \\ &= (K + D)Q(K^* + D^*) \\ &= KQK^* + DQD^* + KQD^* + (KQD^*)^*. \end{aligned}$$

Since  $DA = 0$ ,

$$KQD^* = (A^* Q^{-1} A)^{-1} A^* Q^{-1} Q D^* = (A^* Q^{-1} A)^{-1} 0 = 0,$$

and so

$$\mathbb{E}[(Ly - x)(Ly - x)^*] = KQK^* + DQD^* \geq KQK^*$$

in the sense of positive semi-definite matrices, as claimed.  $\square$

**Remark 6.4.** In the situation that  $A^*Q^{-1}A$  is not invertible, it is standard to use the estimator

$$K = (A^*Q^{-1}A)^\dagger A^*Q^{-1},$$

where  $B^\dagger$  denotes the *Moore–Penrose pseudo-inverse* of  $B$ , defined equivalently by

$$\begin{aligned} B^\dagger &:= \lim_{\delta \rightarrow 0} (B^*B + \delta I)B^*, \\ B^\dagger &:= \lim_{\delta \rightarrow 0} B^*(BB^* + \delta I)B^*, \text{ or} \\ B^\dagger &:= V\Sigma^+U^*, \end{aligned}$$

where  $B = U\Sigma V^*$  is the singular value decomposition of  $B$ , and  $\Sigma^+$  is the transpose of the matrix obtained from  $\Sigma$  by replacing all the strictly positive singular values by their reciprocals.

**Bayesian Interpretation of Regularization.** The Gauss–Markov estimator is not ideal: for example, because of its characterization as the minimizer of a quadratic cost function, it is sensitive to large outliers in the data, i.e. components of  $y$  that differ greatly from the corresponding component of  $A\hat{x}$ . In such a situation, it may be desirable to not try to fit the observed data  $y$  too closely, and instead *regularize* the problem by seeking  $\hat{x}$ , the minimizer of

$$J(x) := \frac{1}{2}\|Ax - y\|_{Q^{-1}}^2 + \frac{1}{2}\|x - \bar{x}\|_{R^{-1}}^2,$$

for some chosen  $\bar{x} \in \mathbb{K}^n$  and positive-definite *Tikhonov matrix*  $R \in \mathbb{K}^{n \times n}$ . Depending upon the relative sizes of  $Q$  and  $R$ ,  $\hat{x}$  will be influenced more by the data  $y$  and hence lie close to the Gauss–Markov estimator, or be influenced more by the regularization term and hence lie close to  $\bar{x}$ . At first sight this procedure may seem somewhat ad hoc, but it has a natural Bayesian interpretation.

The observation equation

$$y = Ax + \eta$$

in fact defines the conditional distribution  $y|x$  by  $(y - Ax)|x = \eta \sim \mathcal{N}(0, Q)$ . To find the minimizer of  $x \mapsto \frac{1}{2}\|Ax - y\|_{Q^{-1}}^2$ , i.e.  $\hat{x} = Ky$ , amounts to finding the *maximum likelihood estimator* of  $x$  given  $y$ . The Bayesian interpretation of the regularization term is that  $\mathcal{N}(\bar{x}, R)$  is a prior distribution for  $x$ . The resulting posterior distribution for  $x|y$  has Lebesgue density  $\rho(x|y)$  with

$$\begin{aligned} \rho(x|y) &\propto \exp\left(-\frac{1}{2}\|Ax - y\|_{Q^{-1}}^2\right) \exp\left(-\frac{1}{2}\|x - \bar{x}\|_{R^{-1}}^2\right) \\ &= \exp\left(-\frac{1}{2}\|Ax - y\|_{Q^{-1}}^2 - \frac{1}{2}\|x - \bar{x}\|_{R^{-1}}^2\right) \\ &= \exp\left(-\frac{1}{2}\|x - Ky\|_{A^*Q^{-1}A}^2 - \frac{1}{2}\|x - \bar{x}\|_{R^{-1}}^2\right) \\ &= \exp\left(-\frac{1}{2}\|x - (A^*Q^{-1}A + R^{-1})^{-1}(A^*Q^{-1}AKy + R^{-1}\bar{x})\|_{A^*Q^{-1}A + R^{-1}}^2\right) \end{aligned}$$

by the standard result that the product of Gaussian distributions with means  $m_1$  and  $m_2$  and covariances  $C_1$  and  $C_2$  is a Gaussian with covariance  $C_3 =$

$(C_1^{-1} + C_2^{-1})^{-1}$  and mean  $C_3(C_1^{-1}m_1 + C_2^{-1}m_2)$ . The solution to the regularized least squares problem, i.e. minimizing the exponent in the above posterior distribution, is the *maximum a posteriori estimator* of  $x$  given  $y$ . However, the full posterior contains more information than the MAP estimator alone. In particular, the posterior covariance matrix  $(A^*Q^{-1}A + R^{-1})^{-1}$  reveals those components of  $x$  about which we are relatively more or less certain.

## 6.2 Bayesian Inversion in Banach Spaces

This section concerns Bayesian inversion in Banach spaces, and, in particular, establishing the appropriate rigorous statement of Bayes' rule in settings where there is no Lebesgue measure with respect to which we can take densities.

**Example 6.5.** There are many applications in which it is of interest to determine the permeability of subsurface rock, e.g. the prediction of transport of radioactive waste from an underground waste repository, or the optimization of oil recovery from underground fields. The flow velocity  $v$  of a fluid under pressure  $p$  in a medium of permeability  $K$  is given by *Darcy's law*

$$v = -K\nabla p.$$

The pressure field  $p$  within a bounded, open domain  $\Omega \subset \mathbb{R}^d$  is governed by the elliptic PDE

$$\begin{aligned} -\nabla \cdot (K\nabla p) &= 0 && \text{in } \Omega, \\ p &= h && \text{on } \partial\Omega. \end{aligned}$$

For simplicity, take the permeability tensor field  $K$  to be a scalar field  $k$  times the identity tensor; for mathematical and physical reasons, it is important that  $k$  be positive, so write  $k = e^u$ . The objective is to recover  $u$  from, say, observations of the pressure field at known points  $x_1, \dots, x_m \in \Omega$ :

$$y_i = p(x_i) + \eta_i.$$

Note that this fits the general ' $y = H(u) + \eta$ ' setup, with  $H$  being defined implicitly by the solution operator to the elliptic PDE.

In general, let  $u$  be a random variable with (prior) distribution  $\mu_0$  — which we do not at this stage assume to be Gaussian — on a separable Banach space  $\mathcal{X}$ . Suppose that we observe data  $y \in \mathbb{R}^m$  according to (6.1), where  $\eta$  is an  $\mathbb{R}^m$ -valued random variable independent of  $u$  with probability density  $\rho$  with respect to Lebesgue measure. Let  $\Phi(u; y)$  be any function that differs from  $-\log \rho(y - H(u))$  by an additive function of  $y$  alone, so that

$$\frac{\rho(y - H(u))}{\rho(y)} \propto \exp(-\Phi(u; y))$$

with a constant of proportionality independent of  $u$ . An informal application of Bayes' rule suggests that the posterior probability distribution of  $u$  given  $y$ ,  $\mu^y$ , has Radon–Nikodým derivative with respect to the prior,  $\mu_0$ , given by

$$\frac{d\mu^y}{d\mu_0}(u) \propto \exp(-\Phi(u; y)).$$

The next theorem makes this argument rigorous:

**Theorem 6.6** (Generalized Bayes' rule). *Suppose that  $H: \mathcal{X} \rightarrow \mathbb{R}^m$  is continuous, and that  $\eta$  is absolutely continuous with support  $\mathbb{R}^m$ . If  $u \sim \mu_0$ , then  $u|y \sim \mu^y$ , where  $\mu^y \ll \mu_0$  and*

$$\frac{d\mu^y}{d\mu_0}(u) \propto \exp(-\Phi(u; y)).$$

**Lemma 6.7.** *Let  $\mu, \nu$  be probability measures on  $S \times T$ , where  $(S, A)$  and  $(T, B)$  are measurable spaces. Assume that  $\mu \ll \nu$  and that  $\frac{d\mu}{d\nu} = \varphi$ . Assume further that the conditional distribution of  $x|y$  under  $\nu$ , denoted by  $\nu^y(dx)$ , exists. Then distribution of  $x|y$  under  $\mu$ , denoted  $\mu^y(dx)$ , exists and  $\mu^y \ll \nu^y$ , with Radon–Nikodým derivative given by*

$$\frac{d\mu^y}{d\nu^y}(x) = \begin{cases} \frac{\varphi(x, y)}{Z(y)}, & \text{if } Z(y) > 0, \\ 1, & \text{otherwise,} \end{cases}$$

where  $Z(y) := \int_S \varphi(x, y) d\nu^y(x)$ .

*Proof.* See Section 10.2 of [25]. □

*Proof of Theorem 6.6.* Let  $\mathbb{Q}_0(dy) := \rho(y) dy$  on  $\mathbb{R}^m$  and  $\mathbb{Q}(du|y) := \rho(y - H(u)) dy$ , so that, by construction

$$\frac{d\mathbb{Q}}{d\mathbb{Q}_0}(y|u) = C(y) \exp(-\Phi(u; y)).$$

Define measures  $\nu_0$  and  $\nu$  on  $\mathbb{R}^m \times \mathcal{X}$  by

$$\begin{aligned} \nu_0(dy, du) &:= \mathbb{Q}_0(dy) \otimes \mu_0(du), \\ \nu(dy, du) &:= \mathbb{Q}_0(dy|u) \mu_0(du). \end{aligned}$$

Note that  $\nu_0$  is a product measure under which  $u$  and  $y$  are independent, whereas  $\nu$  is not. Since  $H$  is continuous, so is  $\Phi$ ; since  $\mu_0(\mathcal{X}) = 1$ , it follows that  $\Phi$  is  $\mu_0$ -measurable. Therefore,  $\nu$  is well-defined,  $\nu \ll \nu_0$ , and

$$\frac{d\nu}{d\nu_0}(y, u) = C(y) \exp(-\Phi(u; y)).$$

Note that

$$\int_{\mathcal{X}} \exp(-\Phi(u; y)) d\mu_0(u) = C(y) \int_{\mathcal{X}} \rho(y - H(u)) d\mu_0(u) > 0,$$

since  $\rho$  is strictly positive on  $\mathbb{R}^m$  and  $H$  is continuous. Since  $\nu_0(du|y) = \mu_0(du)$ , the result follows from Lemma 6.7. □

**Proposition 6.8.** *If the prior  $\mu_0$  is a Gaussian measure and the potential  $\Phi$  is quadratic in  $u$ , then, for all  $y$ , the posterior  $\mu^y$  is Gaussian.*

*Proof.* Exercise 6.1. □

## 6.3 Well-Posedness and Approximation

This section concerns the well-posedness of the Bayesian inference problem for Gaussian priors on Banach spaces. To save space later on, the following will be taken as our *standard assumptions* on the negative log-likelihood / potential  $\Phi$ :

**Assumptions on  $\Phi$ .** Assume that  $\Phi: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  satisfies:

1. For every  $\varepsilon > 0$  and  $r > 0$ , there exists  $M = M(\varepsilon, r) \in \mathbb{R}$  such that, for all  $u \in \mathcal{X}$  and all  $y \in \mathcal{Y}$  with  $\|y\|_{\mathcal{Y}} < r$ ,

$$\Phi(u; y) \geq M - \varepsilon \|u\|_{\mathcal{X}}^2.$$

2. For every  $r > 0$ , there exists  $K = K(r) > 0$  such that, for all  $u \in \mathcal{X}$  and all  $y \in \mathcal{Y}$  with  $\|u\|_{\mathcal{X}}, \|y\|_{\mathcal{Y}} < r$ ,

$$\Phi(u; y) \leq K.$$

3. For every  $r > 0$ , there exists  $L = L(r) > 0$  such that, for all  $u_1, u_2 \in \mathcal{X}$  and all  $y \in \mathcal{Y}$  with  $\|u_1\|_{\mathcal{X}}, \|u_2\|_{\mathcal{X}}, \|y\|_{\mathcal{Y}} < r$ ,

$$|\Phi(u_1; y) - \Phi(u_2; y)| \leq L \|u_1 - u_2\|_{\mathcal{X}}.$$

4. For every  $\varepsilon > 0$  and  $r > 0$ , there exists  $C = C(\varepsilon, r) > 0$  such that, for all  $u \in \mathcal{X}$  and all  $y_1, y_2 \in \mathcal{Y}$  with  $\|y_1\|_{\mathcal{Y}}, \|y_2\|_{\mathcal{Y}} < r$ ,

$$|\Phi(u; y_1) - \Phi(u; y_2)| \leq \exp(\varepsilon \|u\|_{\mathcal{X}}^2 + C) \|y_1 - y_2\|_{\mathcal{Y}}.$$

**Theorem 6.9.** Let  $\Phi$  satisfy standard assumptions (1), (2), and (3) and assume that  $\mu_0$  is a Gaussian probability measure on  $\mathcal{X}$ . Then, for each  $y \in \mathcal{Y}$ ,  $\mu^y$  given by

$$\begin{aligned} \frac{d\mu^y}{d\mu_0}(u) &= \frac{\exp(-\Phi(u; y))}{Z(y)}, \\ Z(y) &= \int_{\mathcal{X}} \exp(-\Phi(u; y)) d\mu_0(u), \end{aligned}$$

is a well-defined probability measure on  $\mathcal{X}$ .

*Proof.* Assumption (2) implies that  $Z(y)$  is bounded below:

$$Z(y) \geq \int_{\{u \|u\|_{\mathcal{X}} \leq r\}} \exp(-K(r)) d\mu_0(u) = \exp(-K(r)) \mu_0[\|u\|_{\mathcal{X}} \leq r] > 0$$

for  $r > 0$ , since  $\mu_0$  is a strictly positive measure on  $\mathcal{X}$ . Assumption (3) implies that  $\Phi$  is  $\mu_0$ -measurable, and hence that  $\mu^y$  is a well-defined measure. By assumption (1), for  $\|y\|_{\mathcal{Y}} \leq r$  and  $\varepsilon$  sufficiently small,

$$\begin{aligned} Z(y) &= \int_{\mathcal{X}} \exp(-\Phi(u; y)) d\mu_0(u) \\ &\leq \int_{\mathcal{X}} \exp(\varepsilon \|u\|_{\mathcal{X}}^2 - M(\varepsilon, r)) d\mu_0(u) \\ &\leq C \exp(-M(\varepsilon, r)) < \infty, \end{aligned}$$

since  $\mu_0$  is Gaussian and we may choose  $\varepsilon$  small enough that the Fernique theorem (Theorem 2.34) applies. Thus,  $\mu^y$  can indeed be normalized to be a probability measure on  $\mathcal{X}$ .  $\square$



**Definition 6.10.** The *Hellinger distance* between two measures  $\mu$  and  $\nu$  is defined in terms of any reference measure  $\rho$  with respect to which both  $\mu$  and  $\nu$  are absolutely continuous by:

$$d_{\text{Hell}}(\mu, \nu) := \sqrt{\frac{1}{2} \int_{\Theta} \left| \sqrt{\frac{d\mu}{d\rho}}(\theta) - \sqrt{\frac{d\nu}{d\rho}}(\theta) \right|^2 d\rho(\theta)}.$$

Exercises 6.4, 6.5 and 6.6 establish the major properties of the Hellinger metric. A particularly useful property is that closeness in the Hellinger metric implies closeness of expected values of polynomially bounded functions: if  $f: \mathcal{X} \rightarrow E$ , for some Banach space  $E$ , then

$$\|\mathbb{E}_{\mu}[f] - \mathbb{E}_{\nu}[f]\|_E \leq 2\sqrt{\mathbb{E}_{\mu}[\|f\|_E^2] - \mathbb{E}_{\nu}[\|f\|_E^2]} d_{\text{Hell}}(\mu, \nu).$$

The following theorem shows that Bayesian inference with respect to a Gaussian prior measure is well-posed with respect to perturbations of the observed data  $y$ :

**Theorem 6.11.** *Let  $\Phi$  satisfy the standard assumptions (1), (2) and (4), suppose that  $\mu_0$  is a Gaussian probability measure on  $\mathcal{X}$ , and that  $\mu^y \ll \mu_0$  with density given by the generalized Bayes' rule for each  $y \in \mathcal{Y}$ . Then there exists a constant  $C \geq 0$  such that, for all  $y, y' \in \mathcal{Y}$ ,*

$$d_{\text{Hell}}(\mu^y, \mu^{y'}) \leq C\|y - y'\|_{\mathcal{Y}}.$$

*Proof.* As in the proof of Theorem 6.9, standard assumption (2) gives a lower bound on  $Z(y)$ . By standard assumptions (1) and (4) and the Fernique theorem (Theorem 2.34),

$$\begin{aligned} |Z(y) - Z(y')| &\leq \|y - y'\|_{\mathcal{Y}} \int_{\mathcal{X}} \exp(-\varepsilon\|u\|_{\mathcal{X}}^2 - M) \exp(-\varepsilon\|u\|_{\mathcal{X}}^2 + C) d\mu_0(u) \\ &\leq C\|y - y'\|_{\mathcal{Y}}. \end{aligned}$$

By the definition of the Hellinger distance,

$$\begin{aligned} 2d_{\text{Hell}}(\mu^y, \mu^{y'})^2 &= \int_{\mathcal{X}} \left( \frac{1}{\sqrt{Z(y)}} e^{-\Phi(u;y)/2} - \frac{1}{\sqrt{Z(y')}} e^{-\Phi(u;y')/2} \right)^2 d\mu_0(u) \\ &\leq I_1 + I_2 \end{aligned}$$

where

$$\begin{aligned} I_1 &:= \frac{2}{Z(y)} \int_{\mathcal{X}} \left( e^{-\Phi(u;y)/2} - e^{-\Phi(u;y')/2} \right)^2 d\mu_0(u), \\ I_2 &:= 2 \left| \frac{1}{\sqrt{Z(y)}} - \frac{1}{\sqrt{Z(y')}} \right|^2 \int_{\mathcal{X}} e^{-\Phi(u;y')/2} d\mu_0(u). \end{aligned}$$

By standard assumptions (1) and (4) and the Fernique theorem,

$$\begin{aligned} \frac{Z(y)}{2} I_1 &\leq \int_{\mathcal{X}} \frac{1}{4} \exp(\varepsilon\|u\|_{\mathcal{X}}^2 - M) \exp(2\varepsilon\|u\|_{\mathcal{X}}^2 + 2C) \|y - y'\|_{\mathcal{Y}}^2 d\mu_0(u) \\ &\leq C\|y - y'\|_{\mathcal{Y}}^2. \end{aligned}$$

A similar application of standard assumption (1) and the Fernique theorem shows that the integral in  $I_2$  is finite. Also, the lower bound on  $Z(\cdot)$  implies that

$$\left| \frac{1}{\sqrt{Z(y)}} - \frac{1}{\sqrt{Z(y')}} \right|^2 \leq C \max \left\{ \frac{1}{Z(y)^3}, \frac{1}{Z(y')^3} \right\} |Z(y) - Z(y')|^2 \leq C \|y - y'\|_Y^2.$$

Combining these facts yields the desired Lipschitz continuity in the Hellinger metric.  $\square$

Similarly, the next theorem shows that Bayesian inference with respect to a Gaussian prior measure is well-posed with respect to approximation of measures and log-likelihoods. The approximation of  $\Phi$  by some  $\Phi^N$  typically arises through the approximation of  $H$  by some discretized numerical model  $H^N$ .

**Theorem 6.12.** *Suppose that the probabilities  $\mu$  and  $\mu^N$  are the posteriors arising from potentials  $\Phi$  and  $\Phi^N$  and are all absolutely continuous with respect to  $\mu_0$ , and that  $\Phi, \Phi^N$  satisfy the standard assumptions (1) and (2) with constants uniform in  $N$ . Assume also that, for all  $\varepsilon > 0$ , there exists  $K = K(\varepsilon) > 0$  such that*

$$|\Phi(u) - \Phi^N(u; y)| \leq K \exp(\varepsilon \|u\|_{\mathcal{X}}^2) \psi(N), \quad (6.2)$$

where  $\lim_{N \rightarrow \infty} \psi(N) = 0$ . Then there is a constant  $C$ , independent of  $N$ , such that

$$d_{\text{Hell}}(\mu, \mu^N) \leq C \psi(N).$$

*Proof.* Since  $y$  does not appear in this problem,  $y$ -dependence will be suppressed for the duration of this proof. Let  $Z$  and  $Z^N$  denote the appropriate normalization constants, as in the proof of Theorem 6.11. By standard assumption (1), (6.2), and the Fernique theorem,

$$|Z - Z^N| \leq \int_{\mathcal{X}} K \psi(N) \exp(\varepsilon \|u\|_{\mathcal{X}}^2) - M \exp(\varepsilon \|u\|_{\mathcal{X}}^2) d\mu_0 \leq C \psi(N).$$

By the definition of the Hellinger distance,

$$\begin{aligned} 2d_{\text{Hell}}(\mu, \mu^N)^2 &= \int_{\mathcal{X}} \left( \frac{1}{\sqrt{Z}} e^{-\Phi(u)/2} - \frac{1}{\sqrt{Z^N}} e^{-\Phi^N(u)/2} \right)^2 d\mu_0(u) \\ &\leq I_1 + I_2 \end{aligned}$$

where

$$\begin{aligned} I_1 &:= \frac{2}{Z} \int_{\mathcal{X}} \left( e^{-\Phi(u)/2} - e^{-\Phi^N(u)/2} \right)^2 d\mu_0(u), \\ I_2 &:= 2 \left| \frac{1}{\sqrt{Z}} - \frac{1}{\sqrt{Z^N}} \right|^2 \int_{\mathcal{X}} e^{-\Phi^N(u)/2} d\mu_0(u). \end{aligned}$$

By standard assumption (1), (6.2), and the Fernique theorem,

$$\begin{aligned} \frac{Z}{2} I_1 &\leq \int_{\mathcal{X}} K^2 \exp(3\varepsilon \|u\|_{\mathcal{X}}^2) \psi(N)^2 d\mu_0(u) \\ &\leq C \psi(N)^2. \end{aligned}$$

Similarly,

$$\left| \frac{1}{\sqrt{Z}} - \frac{1}{\sqrt{Z^N}} \right|^2 \leq C \max \left\{ \frac{1}{Z^3}, \frac{1}{(Z^N)^3} \right\} |Z - Z^N|^2 \leq C\psi(N)^2.$$

Combining these facts yields the desired bound on  $d_{\text{Hell}}(\mu, \mu^N)$ .  $\square$

**Remark 6.13.** Note well that, regardless of the value of the observed data  $y$ , the Bayesian posterior  $\mu^y$  is absolutely continuous with respect to the prior  $\mu_0$  and, in particular, cannot associate positive posterior probability to any event of prior probability zero. However, the Feldman–Hájek theorem (Theorem 2.38) says that it is very difficult for probability measures on infinite-dimensional spaces to be absolutely continuous with respect to one another. Therefore, the choice of infinite-dimensional prior  $\mu_0$  is a very strong modelling assumption that, if it is ‘wrong’, cannot be ‘corrected’ even by large amounts of data  $y$ . In this sense, it is not reasonable to expect that Bayesian inference on function spaces should be well-posed with respect to apparently small perturbations of the prior  $\mu_0$ , e.g. by a shift of mean that lies outside the Cameron–Martin space, or a change of covariance arising from a non-unit dilation of the space. Nevertheless, the infinite-dimensional perspective is not without genuine fruits: in particular, the well-posedness results (Theorems 6.11 and 6.12) are very important for the stability of finite-dimensional (discretized) Bayesian problems with respect to discretization dimension  $N$ .



## Bibliography

This material is covered in much greater detail in the module [MA612 Probability on Function Spaces and Bayesian Inverse Problems](#).

W

Tikhonov regularization was introduced in [105, 106]. An introduction to the general theory of regularization and its application to inverse problems is the book of Engl & al. [26]. The book of Tarantola [103] also provides a good applied introduction to inverse problems. Kaipio & Somersalo [45] provide a good introduction to the Bayesian approach to inverse problems, especially in the context of differential equations.

This chapter owes a great deal to the papers of Stuart [98] and Cotter & al. [20, 21], which set out the common structure of Bayesian inverse problems on Banach and Hilbert spaces, focussing on Gaussian priors. Stuart’s article stresses the importance of delaying discretization to the last possible moment, much as in PDE theory, lest one carelessly end up with a family of finite-dimensional problems that are individually well-posed but collectively ill-conditioned as the discretization dimension tends to infinity. Extensions to Besov priors, which are constructed using wavelet bases of  $L^2$  and allow for non-smooth local features in the random fields, can be found in the articles of Dashti & al. [22] and Lassas & al. [57].

## Exercises

**Exercise 6.1.** Let  $\mu_0$  be a Gaussian probability measure on a separable Banach space  $\mathcal{X}$  and suppose that the potential  $\Phi(u; y)$  is quadratic in  $u$ . Show that the posterior  $\mu^y$  is also a Gaussian measure on  $\mathcal{X}$ .

**Exercise 6.2.** Let  $Q \in \mathbb{K}^{n \times n}$  be a self-adjoint and positive-definite matrix. Show that

$$\langle A, B \rangle := \operatorname{tr}(A^*QB) \quad \text{for } A, B \in \mathbb{K}^{n \times m}$$

defines an inner product on the space  $\mathbb{K}^{n \times m}$  of  $n \times m$  matrices over  $\mathbb{K}$ , and hence that  $A \mapsto \sqrt{\operatorname{tr}(A^*QA)}$  is a norm on  $\mathbb{K}^{n \times m}$ .

**Exercise 6.3.** Let  $\mu$  and  $\nu$  be probability measures on  $(\Theta, \mathcal{F})$ , both absolutely continuous with respect to a reference measure  $\rho$ . Define the *total variation distance* between  $\mu$  and  $\nu$  by

$$\begin{aligned} d_{\text{TV}}(\mu, \nu) &:= \frac{1}{2} \mathbb{E}_\rho \left[ \left| \frac{d\mu}{d\rho} - \frac{d\nu}{d\rho} \right| \right] \\ &= \frac{1}{2} \int_\Theta \left| \frac{d\mu}{d\rho}(\theta) - \frac{d\nu}{d\rho}(\theta) \right| d\rho(\theta). \end{aligned}$$

Show that  $d_{\text{TV}}$  is a metric on  $\mathcal{M}_1(\Theta, \mathcal{F})$ , that its values do not depend upon the choice of  $\rho$ , that  $\mathcal{M}_1(\Theta, \mathcal{F})$  has diameter at most 1, and that, if  $\nu \ll \mu$ , then

$$d_{\text{TV}}(\mu, \nu) := \frac{1}{2} \mathbb{E}_\mu \left[ \left| 1 - \frac{d\nu}{d\mu} \right| \right] \equiv \frac{1}{2} \int_\Theta \left| 1 - \frac{d\nu}{d\mu}(\theta) \right| d\mu(\theta).$$

Show also that this definition agrees with the definition given in (5.4).

**Exercise 6.4.** Let  $\mu$  and  $\nu$  be probability measures on  $(\Theta, \mathcal{F})$ , both absolutely continuous with respect to a reference measure  $\rho$ . Define the *Hellinger distance* between  $\mu$  and  $\nu$  by

$$\begin{aligned} d_{\text{Hell}}(\mu, \nu) &:= \sqrt{\frac{1}{2} \mathbb{E}_\rho \left[ \left| \sqrt{\frac{d\mu}{d\rho}} - \sqrt{\frac{d\nu}{d\rho}} \right|^2 \right]} \\ &= \sqrt{\frac{1}{2} \int_\Theta \left| \sqrt{\frac{d\mu}{d\rho}}(\theta) - \sqrt{\frac{d\nu}{d\rho}}(\theta) \right|^2 d\rho(\theta)}. \end{aligned}$$

Show that  $d_{\text{Hell}}$  is a metric on  $\mathcal{M}_1(\Theta, \mathcal{F})$ , that its values do not depend upon the choice of  $\rho$ , that  $\mathcal{M}_1(\Theta, \mathcal{F})$  has diameter at most 1, and that, if  $\nu \ll \mu$ , then

$$d_{\text{Hell}}(\mu, \nu) := \sqrt{\frac{1}{2} \mathbb{E}_\mu \left[ \left| 1 - \sqrt{\frac{d\nu}{d\mu}} \right|^2 \right]} \equiv \sqrt{\frac{1}{2} \int_\Theta \left| 1 - \sqrt{\frac{d\nu}{d\mu}}(\theta) \right|^2 d\mu(\theta)}.$$

**Exercise 6.5.** Show that the total variation and Hellinger distances satisfy, for all  $\mu$  and  $\nu$ ,

$$\frac{1}{\sqrt{2}} d_{\text{TV}}(\mu, \nu) \leq d_{\text{Hell}}(\mu, \nu) \leq d_{\text{TV}}(\mu, \nu)^{1/2}.$$

**Exercise 6.6.** Suppose that  $\mu$  and  $\nu$  are probability measures on a Banach space  $\mathcal{X}$ . Show that, if  $E$  is a Banach space and  $f: \mathcal{X} \rightarrow E$  has finite second moment with respect to both  $\mu$  and  $\nu$ , then

$$\|\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]\|_E \leq 2\sqrt{\mathbb{E}_\mu[\|f\|_E^2] - \mathbb{E}_\nu[\|f\|_E^2]} d_{\text{Hell}}(\mu, \nu).$$

Show also that, if  $E$  is Hilbert and  $f$  has finite fourth moment with respect to both  $\mu$  and  $\nu$ , then

$$\|\mathbb{E}_\mu[f \otimes f] - \mathbb{E}_\nu[f \otimes f]\|_E \leq 2\sqrt{\mathbb{E}_\mu[\|f\|_E^4] - \mathbb{E}_\nu[\|f\|_E^4]} d_{\text{Hell}}(\mu, \nu).$$

Hence show that the differences between the means and covariance operators of two measures on  $\mathcal{X}$  are bounded above by the Hellinger distance between the two measures.

**Exercise 6.7.** Let  $\Gamma \in \mathbb{R}^{q \times q}$  be symmetric and positive definite. Suppose that  $H: \mathcal{X} \rightarrow \mathbb{R}^q$  satisfies

1. For every  $\varepsilon > 0$ , there exists  $M \in \mathbb{R}$  such that, for all  $u \in \mathcal{X}$ ,

$$\|H(u)\|_{\Gamma^{-1}} \leq \exp(\varepsilon\|u\|_{\mathcal{X}}^2 + M).$$

2. For every  $r > 0$ , there exists  $K > 0$  such that, for all  $u_1, u_2 \in \mathcal{X}$  with  $\|u_1\|_{\mathcal{X}}, \|u_2\|_{\mathcal{X}} < r$ ,

$$\|H(u_1) - H(u_2)\|_{\Gamma^{-1}} \leq K\|u_1 - u_2\|_{\mathcal{X}}.$$

Show that  $\Phi: \mathcal{X} \times \mathbb{R}^q \rightarrow \mathbb{R}$  defined by

$$\Phi(u; y) := \frac{1}{2} \langle y - H(u), \Gamma^{-1}(y - H(u)) \rangle$$

satisfies the standard assumptions.

**Exercise 6.8.** An exercise in forensic inference:

**NEWSFLASH: THE PRESIDENT HAS BEEN SHOT!**

While being driven through the streets of the capital in his open-topped limousine, President Marx of Freedonia has been shot by a sniper. To make matters worse, the bullet appears to have come from a twenty-storey building, on each floor of which was stationed a single bodyguard of President Marx's security detail who was meant to protect him. None of the suspects have any obvious marks of guilt such as one bullet missing from their magazine, gunpowder burns, failed lie detector tests... not even an evil moustache. The soundproofing inside the building was good enough that none of the security detail can even say whether the shot came from above or below them. You have been called in as an expert quantifier of uncertainty to try to infer from which floor the assassin took the shot, and hence identify the guilty man. You have the following information:

1. At the time of the shot, the President's limousine was 500m from the building, travelling at about 20mph.
2. The bullet entered the President's body at the base of his neck and exited through the centre of the breastbone. The President is an average-sized man.

Apply the techniques you have learned in this chapter to make a recommendation as to which floor the shot came from, and hence which bodyguard is the traitor. Do your own research on rifle muzzle velocities &c., explaining the assumptions that you make along the way.

**Exercise 6.9.** Construct a short sequence of Bayesian inferences in which the posterior appears to predict one thing and then another. Implications for not cutting short analyses and studies. [??] **FINISH ME!!!**

DRAFT

## Chapter 7

# Filtering and Data Assimilation

It is not bigotry to be certain we are right; but it is bigotry to be unable to imagine how we might possibly have gone wrong.

---

*The Catholic Church and Conversion*

G. K. CHESTERTON

Data assimilation is the integration of two information sources:

- a mathematical model of a time-dependent physical system, or a numerical implementation of such a model; and
- a sequence of observations of that system, usually corrupted by some noise.

The objective is to combine these two ingredients to produce a more accurate estimate of the system's true state, and hence more accurate predictions of the system's future state. Very often, data assimilation is synonymous with *filtering*, which incorporates many of the same ideas but arose in the context of signal processing. An additional component of the data assimilation / filtering problem is that one typically wants to achieve it in *real time*: if today is Monday, then a data assimilation scheme that takes until Friday to produce an accurate prediction of Tuesday's weather using Monday's observations is basically useless.

Data assimilation methods are typically Bayesian, in the sense that the current knowledge of the system state can be thought of as a prior, and the incorporation of the dynamics and observations as an update/conditioning step that produces a posterior. Bearing in mind considerations of computational cost and the imperative for *real time* data assimilation, there are two key ideas underlying filtering: the first is to build up knowledge about the posterior sequentially, and hence perhaps more efficiently; the second is to break up the unknown  $u$  and build up knowledge about its constituent parts sequentially, hence reducing the computational dimension of each sampling problem. Thus, the first idea means decomposing the data sequentially, while the second means decomposing the unknown state sequentially.

## 7.1 State Estimation in Discrete Time

In the Kálmán filter, the probability distributions representing the system state and various noise terms are described purely in terms of their mean and covariance, so they are effectively being approximated as Gaussians.

For simplicity, the first description of the Kálmán filter will be of a controlled linear dynamical system that evolves in discrete time steps

$$t_0 < t_1 < \dots < t_k < \dots$$

The state of the system at time  $t_k$  is a vector  $x_k \in \mathbb{K}^n$ , and it evolves from the state  $x_{k-1} \in \mathbb{K}^n$  at time  $t_{k-1}$  according to the linear model

$$x_k = F_k x_{k-1} + G_k u_k + w_k \quad (7.1)$$

where, for each time  $t_k$ ,

- $F_k \in \mathbb{K}^{n \times n}$  is the state transition model, which is applied to the previous state  $x_{k-1} \in \mathbb{K}^n$ ;
- $G_k \in \mathbb{K}^{n \times p}$  is the control-to-input model, which is applied to the control vector  $u_k \in \mathbb{K}^p$ ;
- $w_k \sim \mathcal{N}(0, Q_k)$  is the process noise, a  $\mathbb{K}^n$ -valued centred Gaussian random variable with self-adjoint positive-definite covariance matrix  $Q_k \in \mathbb{K}^{n \times n}$ .

At time  $t_k$  an observation  $y_k \in \mathbb{K}^q$  of the true state  $x_k$  is made according to

$$y_k = H_k x_k + v_k, \quad (7.2)$$

where

- $H_k \in \mathbb{K}^{q \times n}$  is the *observation operator*, which maps the true state space  $\mathbb{K}^n$  into the observable space  $\mathbb{K}^q$
- $v_k \sim \mathcal{N}(0, R_k)$  is the observation noise, a  $\mathbb{K}^q$ -valued centred Gaussian random variable with self-adjoint positive-definite covariance  $Q_k \in \mathbb{K}^{q \times q}$ .

As an initial condition, the state of the system at time  $t_0$  is taken to be  $x_0 = \mu_0 + w_0$  where  $\mu_0 \in \mathbb{K}^n$  is known and  $w_0 \sim \mathcal{N}(0, Q_0)$ . All the noise vectors are assumed to be mutually independent.

As a preliminary to constructing the full Kálmán filter, we consider the problem of estimating states  $x_1, \dots, x_k$  given the corresponding controls  $u_1, \dots, u_k$  and  $m$  known observations  $y_1, \dots, y_m$ , where  $k \geq m$ . In particular, we seek the best linear unbiased estimate of  $x_1, \dots, x_k$ .

First note that (7.1)–(7.2) is equivalent to the single equation

$$b_{k|m} = A_{k|m} z_k + \eta_{k|m}, \quad (7.3)$$

where

$$b_{k|m} := \begin{bmatrix} \mu_0 \\ G_1 u_1 \\ y_1 \\ \vdots \\ G_m u_m \\ y_m \\ G_{m+1} u_{m+1} \\ \vdots \\ G_k u_k \end{bmatrix} \in \mathbb{K}^{n(k+1)+qm}, \quad z_k := \begin{bmatrix} x_0 \\ \vdots \\ x_k \end{bmatrix}, \quad \eta_{k|m} := \begin{bmatrix} -w_0 \\ -w_1 \\ +v_1 \\ \vdots \\ -w_m \\ +v_m \\ -w_{m+1} \\ \vdots \\ -w_k \end{bmatrix}$$



and  $A_{k|m} \in \mathbb{K}^{(n(k+1)+qm) \times n(k+1)}$  is

$$A_{k|m} := \begin{bmatrix} I & 0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ -F_1 & I & 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & H_1 & 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & -F_2 & I & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & H_2 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -F_m & I & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 & H_m & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & -F_{m+1} & I & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & -F_k & I \end{bmatrix}.$$

Note that the noise vector  $\eta_{k|m}$  is  $\mathbb{K}^{n(k+1)+qm}$ -valued and has mean zero and block-diagonal positive-definite precision operator (inverse covariance)  $W_{k|m}$  given in block form by

$$W_{k|m} := \text{diag}(Q_0^{-1}, Q_1^{-1}, R_1^{-1}, \dots, Q_m^{-1}, R_m^{-1}, Q_{m+1}^{-1}, \dots, Q_k^{-1}).$$

By the Gauss–Markov theorem (Theorem 6.2), the best linear unbiased estimate  $\hat{z}_{k|m} = [\hat{x}_{0|m}, \dots, \hat{x}_{k|m}]^*$  of  $z_k$  satisfies

$$\hat{z}_{k|m} \in \arg \min_{z_k \in \mathbb{K}^n} J_{k|m}(z_k), \quad J_{k|m}(z_k) := \frac{1}{2} \|A_{k|m}z_k - b_{k|m}\|_{W_{k|m}}^2, \quad (7.4)$$

and by Lemma 4.22 is the solution of the normal equations

$$A_{k|m}^* W_{k|m} A_{k|m} \hat{z}_{k|m} = A_{k|m}^* W_{k|m} b_{k|m}.$$

By Exercise 7.1, it follows from the assumptions made above that these normal equations have a unique solution

$$\hat{z}_{k|m} = (A_{k|m}^* W_{k|m} A_{k|m})^{-1} A_{k|m}^* W_{k|m} b_{k|m}. \quad (7.5)$$

By Theorem 6.2 and Remark 6.3,  $\mathbb{E}[\hat{z}_{k|m}] = z_k$  and the covariance matrix of the estimate  $\hat{z}_{k|m}$  is  $(A_{k|m}^* W_{k|m} A_{k|m})^{-1}$ ; a Bayesian statistician would say that  $z_k$ , conditioned upon the control and observation data  $b_{k|m}$ , is the Gaussian random variable with distribution  $\mathcal{N}(\hat{z}_{k|m}, (A_{k|m}^* W_{k|m} A_{k|m})^{-1})$ .

Note that, since  $W_{k|m}$  is block diagonal,  $J_{k|m}$  can be written as

$$J_{k|m}(z_k) = \frac{1}{2} \|x_0 - \mu_0\|_{Q_0^{-1}}^2 + \frac{1}{2} \sum_{i=1}^m \|y_i - H_i x_i\|_{R_i^{-1}}^2 + \frac{1}{2} \sum_{i=1}^k \|x_i - F_i x_{i-1} - G_i u_i\|_{Q_i^{-1}}^2.$$

An expansion of this type will prove very useful in derivation of the linear Kálmán filter in the next section.

## 7.2 Linear Kálmán Filter

We now consider the state estimation problem in the common practical situation that  $k = m$ . Why is the state estimate (7.5) not the end of the story? For one thing, there is an issue of immediacy: one does not want to have to wait for observation  $y_{1000}$  to come in before estimating states  $x_1, \dots, x_{999}$  as well as  $x_{1000}$ , in particular because the choice of the control  $u_{k+1}$  typically depends upon the estimate of  $x_k$ ; what one wants is to estimate  $x_k$  upon observing  $y_k$ . However, there is also an issue of computational cost, and hence computation time: the solution of the least squares problem

$$\text{find } \hat{x} = \arg \min_{x \in \mathbb{K}^n} \|Ax - b\|_{\mathbb{K}^m}^2$$

where  $A \in \mathbb{K}^{m \times n}$ , at least by direct methods such as solving the normal equations or QR factorization, requires of the order of  $mn^2$  floating-point operations, as shown in [MA398 Matrix Analysis and Algorithms](#). Hence, calculation of the state estimate  $\hat{z}_k$  by direct solution of (7.5) takes of the order of

$$(n(k+1) + qm)n^2(k+1)^2$$

operations. It is clearly impractical to work with a state estimation scheme with a computational cost that increases cubically with the number of time steps to be considered. The idea of filtering is to break the state estimation problem down into a sequence of estimation problems that can be solved with constant computational cost per time step, as each observation comes in.

The two-step *linear Kálmán filter* (LKF) is a recursive method for constructing the best linear unbiased estimate  $\hat{x}_{k|k}$  (with covariance matrix  $P_{k|k}$ ) of  $x_k$  in terms of the previous state estimate  $\hat{x}_{k-1|k-1}$  and the data  $u_k$  and  $y_k$ . It is called the *two-step* filter because the process of updating the state estimate ( $\hat{x}_{k-1|k-1}, P_{k-1|k-1}$ ) for time  $t_{k-1}$  into the estimate ( $\hat{x}_{k|k}, P_{k|k}$ ) for  $t_k$  is split into two steps (which could, of course, be algebraically unified into a single step):

- the *prediction step* uses the dynamics alone to update ( $\hat{x}_{k-1|k-1}, P_{k-1|k-1}$ ) into ( $\hat{x}_{k|k-1}, P_{k|k-1}$ ), an estimate for the state at time  $t_k$  that does not use the observation  $y_k$ ;
- the *correction step* updates ( $\hat{x}_{k|k-1}, P_{k|k-1}$ ) into ( $\hat{x}_{k|k}, P_{k|k}$ ) using the observation  $y_k$ .

**Initialization.** We begin by initializing the state estimate as

$$(\hat{x}_{0|0}, P_{0|0}) := (\mu_0, Q_0).$$

**Prediction.** Write

$$\mathcal{F}_k := \begin{bmatrix} 0 & \cdots & 0 & F_k \end{bmatrix} \in \mathbb{K}^{n \times nk},$$

where the  $F_k$  block is the block corresponding to  $x_{k-1}$ , so that  $\mathcal{F}_k z_{k-1} = F_k x_{k-1}$ . A key insight here is to write the cost function  $J_{k|k-1}$  recursively as

$$J_{k|k-1}(z_k) = J_{k-1|k-1}(z_{k-1}) + \frac{1}{2} \|x_k - \mathcal{F}_k z_{k-1} - G_k u_k\|_{Q_k^{-1}}^2,$$

in which case the gradient and Hessian of  $J_{k|k-1}$  are given in block form with respect to  $z_k = [z_{k-1}, x_k]^*$  as

$$\nabla J_{k|k-1}(z_k) = \begin{bmatrix} \nabla J_{k-1|k-1}(z_{k-1}) + \mathcal{F}_k^* Q_k^{-1} (\mathcal{F}_k z_{k-1} - x_k + G_k u_k) \\ -Q_k^{-1} (\mathcal{F}_k z_{k-1} - x_k + G_k u_k) \end{bmatrix}$$

and

$$D^2 J_{k|k-1}(z_k) = \begin{bmatrix} D^2 J_{k-1|k-1}(z_{k-1}) + \mathcal{F}_k^* Q_k^{-1} \mathcal{F}_k & -\mathcal{F}_k^* Q_k^{-1} \\ -Q_k^{-1} \mathcal{F}_k & Q_k^{-1} \end{bmatrix},$$

which, by Exercise 7.2, is positive definite. Hence, by a single iteration of Newton's method with any initial condition  $z_k$ , the minimizer  $\hat{z}_{k|k-1}$  of  $J_{k|k-1}(z_k)$  is simply

$$\hat{z}_{k|k-1} = z_k - (D^2 J_{k|k-1}(z_k))^{-1} \nabla J_{k|k-1}(z_k).$$

Note that  $\nabla J_{k-1|k-1}(\hat{z}_{k-1|k-1}) = 0$  and  $\mathcal{F}_k \hat{z}_{k-1|k-1} = F_k \hat{x}_{k-1|k-1}$ , so clever initial guess is to take

$$z_k = \begin{bmatrix} \hat{z}_{k-1|k-1} \\ F_k \hat{x}_{k-1|k-1} + G_k u_k \end{bmatrix}.$$

With this initial guess, the gradient becomes  $\nabla J_{k|k-1}(z_k) = 0$  i.e. the optimal estimate of  $x_k$  given  $y_1, \dots, y_{k-1}$  is the bottom row of  $\hat{z}_{k|k-1}$  and — by Remark 6.3 — the covariance matrix  $P_{k|k-1}$  of this estimate is the bottom-right block of the inverse Hessian  $(D^2 J_{k|k-1}(z_k))^{-1}$ , calculated using the method of Schur complements (Exercise 7.3):

$$\hat{x}_{k|k-1} = F_k \hat{x}_{k-1|k-1} + G_k u_k, \quad (7.6)$$

$$P_{k|k-1} = F_k P_{k-1|k-1} F_k^* + Q_k. \quad (7.7)$$

These two updates comprise the *prediction step* of the Kálmán filter. The calculation of  $\hat{x}_{k|k-1}$  requires  $\Theta(n^2 + np)$  operations, and the calculation of  $P_{k|k-1}$  requires  $O(n^\alpha)$  operations, assuming that matrix-matrix multiplication for  $n \times n$  matrices can be effected in  $O(n^\alpha)$  operations for some  $2 \leq \alpha \leq 3$ .

**Correction.** The next step is a *correction step* (or *update step*) that corrects the prior estimate-covariance pair  $(\hat{x}_{k|k-1}, P_{k|k-1})$  to a posterior estimate-covariance pair  $(\hat{x}_{k|k}, P_{k|k})$  given the observation  $y_k$ . Write

$$\mathcal{H}_k := \begin{bmatrix} 0 & \cdots & 0 & H_k \end{bmatrix} \in \mathbb{K}^{n \times nk},$$

where the  $H_k$  block is the block corresponding to  $x_k$ , so that  $\mathcal{H}_k z_k = H_k x_k$ . Again, the cost function is written recursively:

$$J_{k|k}(z_k) = J_{k|k-1}(z_k) + \frac{1}{2} \|y_k - \mathcal{H}_k z_k\|_{R_k^{-1}}^2.$$

The gradient and Hessian are

$$\begin{aligned} \nabla J_{k|k}(z_k) &= \nabla J_{k|k-1}(z_k) + \mathcal{H}_k^* R_k^{-1} (\mathcal{H}_k z_k - y_k) \\ &= \nabla J_{k|k-1}(z_k) + \mathcal{H}_k^* R_k^{-1} (H_k x_k - y_k) \end{aligned}$$

and

$$D^2 J_{k|k}(z_k) = D^2 J_{k|k-1}(z_k) + \mathcal{H}_k^* R_k^{-1} \mathcal{H}_k.$$

We now take  $z_k = \hat{z}_{k|k-1}$  as a clever initial guess for a single Newton iteration, so that the gradient becomes

$$\nabla J_{k|k}(\hat{z}_{k|k-1}) = \underbrace{\nabla J_{k|k-1}(\hat{z}_{k|k-1})}_{=0} + \mathcal{H}_k^* R_k^{-1} (\mathcal{H}_k \hat{z}_{k|k-1} - y_k).$$

The posterior estimate  $\hat{x}_{k|k}$  is now obtained as the bottom row of the Newton update, i.e.

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} - P_{k|k} H_k^* R_k^{-1} (H_k \hat{x}_{k|k-1} - y_k) \quad (7.8)$$

where the posterior covariance  $P_{k|k}$  is obtained as the bottom-right block of the inverse Hessian  $(D^2 J_{k|k}(z_k))^{-1}$  by Schur complementation:

$$P_{k|k} = (P_{k|k-1}^{-1} + H_k^* R_k^{-1} H_k)^{-1}. \quad (7.9)$$

Determination of the computational costs of these two steps is left as an exercise (Exercise 7.6).

In many presentations of the Kálmán filter, the correction step is phrased in terms of the *Kálmán gain*  $K_k \in \mathbb{K}^{n \times q}$ :

$$K_k := P_{k|k-1} H_k^* (H_k P_{k|k-1} H_k^* + R_k)^{-1}. \quad (7.10)$$

so that

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k (z_k - H_k \hat{x}_{k|k-1}) \quad (7.11)$$

$$P_{k|k} = (I - K_k H_k) P_{k|k-1}. \quad (7.12)$$

It is also common to refer to

$$\tilde{y}_k := z_k - H_k \hat{x}_{k|k-1}$$

as the *innovation residual* and

$$S_k := H_k P_{k|k-1} H_k^* + R_k$$

as the *innovation covariance*, so that  $K_k = P_{k|k-1} H_k^* S_k^{-1}$  and  $\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k \tilde{y}_k$ . It is an exercise in algebra to show that the first presentation of the correction step (7.8)–(7.9) and the Kálmán gain formulation (7.10)–(7.12) are the same.

The Kálmán filter can also be formulated in continuous time, or in a hybrid form with continuous evolution but discrete observations. For example, the hybrid Kálmán filter has the evolution and observation equations

$$\begin{aligned} \dot{x}(t) &= F(t)x(t) + G(t)u(t) + w(t), \\ y_k &= H_k x_k + v_k, \end{aligned}$$

where  $x_k := x(t_k)$ . The prediction equations are that  $\hat{x}_{k|k-1}$  is the solution at time  $t_k$  of the initial value problem

$$\begin{aligned} \dot{\hat{x}}(t) &= F(t)\hat{x}(t) + G(t)u(t), \\ \hat{x}(t_{k-1}) &= \hat{x}_{k-1|k-1}, \end{aligned}$$

and that  $P_{k|k-1}$  is the solution at time  $t_k$  of the initial value problem

$$\begin{aligned}\dot{P}(t) &= F(t)P(t)F(t)^* + Q(t), \\ P(t_{k-1}) &= P_{k-1|k-1}.\end{aligned}$$

The correction equations (in Kálmán gain form) are as before:

$$\begin{aligned}K_k &= P_{k|k-1}H_k^*(H_kP_{k|k-1}H_k^* + R_k)^{-1} \\ \hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k(z_k - H_k\hat{x}_{k|k-1}) \\ P_{k|k} &= (I - K_kH_k)P_{k|k-1}.\end{aligned}$$

The LKF with continuous time evolution and observation is known as the *Kálmán–Bucy filter*. The evolution and observation equations are

$$\begin{aligned}\dot{x}(t) &= F(t)x(t) + G(t)u(t) + w(t), \\ y(t) &= H(t)x(t) + v(t).\end{aligned}$$

Notably, in the Kálmán–Bucy filter, the distinction between prediction and correction does not exist.

$$\begin{aligned}\dot{\hat{x}}(t) &= F(t)\hat{x}(t) + G(t)u(t) + K(t)(y(t) - H(t)\hat{x}(t)), \\ \dot{P}(t) &= F(t)P(t) + P(t)F(t)^* + Q(t) - K(t)R(t)K(t)^*,\end{aligned}$$

where

$$K(t) := P(t)H(t)^*R(t)^{-1}.$$

### 7.3 Extended Kálmán Filter

The *extended Kálmán filter* (EKF) is an extension of the Kálmán filter to nonlinear dynamical systems. In discrete time, the evolution and observation equations are

$$\begin{aligned}x_k &= f_k(x_{k-1}, u_k) + w_k, \\ y_k &= h_k(x_k) + v_k,\end{aligned}$$

where, as before,  $x_k \in \mathbb{K}^n$  are the states,  $u_k \in \mathbb{K}^p$  are the controls,  $y_k \in \mathbb{K}^q$  are the observations,  $f_k: \mathbb{K}^n \times \mathbb{K}^p \rightarrow \mathbb{K}^n$  are the vector fields for the dynamics,  $h_k: \mathbb{K}^n \rightarrow \mathbb{K}^q$  are the observation maps, and the noise processes  $w_k$  and  $v_k$  are uncorrelated with zero mean and positive-definite covariances  $Q_k$  and  $R_k$  respectively.

The classical derivation of the EKF is to approximate the nonlinear evolution–observation equations with a linear system and then use the LKF on that linear system. In contrast to the LKF, the EKF is neither the unbiased minimum mean-squared error estimator nor the minimum variance unbiased estimator of the state; in fact, the EKF is generally biased. However, the EKF is the best linear unbiased estimator of the linearized dynamical system, which can often be a good approximation of the nonlinear system. As a result, how well the local linear dynamics match the nonlinear dynamics determines in large part how well the EKF will perform.

The approximate linearized system is obtained by first-order Taylor expansion of  $f_k$  about the previous estimated state  $\hat{x}_{k-1|k-1}$  and  $h_k$  about  $\hat{x}_{k|k-1}$

$$\begin{aligned} x_k &= f_k(\hat{x}_{k-1|k-1}, u_k) + Df_k(\hat{x}_{k-1|k-1}, u_k)(x_{k-1} - \hat{x}_{k-1|k-1}) + w_k, \\ y_k &= h_k(\hat{x}_{k|k-1}) + Dh_k(\hat{x}_{k|k-1})(x_k - \hat{x}_{k|k-1}) + v_k. \end{aligned}$$

Taking

$$\begin{aligned} F_k &:= Df_k(\hat{x}_{k-1|k-1}, u_k), \\ H_k &:= Dh_k(\hat{x}_{k|k-1}), \\ \tilde{u}_k &:= f_k(\hat{x}_{k-1|k-1}, u_k) - F_k \hat{x}_{k-1|k-1}, \\ z_k &:= h_k(\hat{x}_{k|k-1}) - H_k \hat{x}_{k|k-1}, \end{aligned}$$

the linearized system is

$$\begin{aligned} x_k &= F_k x_{k-1} + \tilde{u}_k + w_k, \\ y_k &= H_k x_k + z_k + v_k. \end{aligned}$$

We now apply the standard LKF to this system, treating  $\tilde{u}_k$  as the controls for the linear system and  $y_k - z_k$  as the observations, to obtain

$$\hat{x}_{k|k-1} = f_k(\hat{x}_{k-1|k-1}, u_k), \quad (7.13)$$

$$P_{k|k-1} = F_k P_{k-1|k-1} F_k^* + Q_k, \quad (7.14)$$

$$P_{k|k} = (P_{k|k-1}^{-1} + H_k^* R_k^{-1} H_k)^{-1}, \quad (7.15)$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} - P_{k|k} H_k^* R_k^{-1} (h_k(\hat{x}_{k|k-1}) - y_k). \quad (7.16)$$

## 7.4 Ensemble Kálmán Filter

The EnKF is a Monte Carlo approximation of the Kalman filter that avoids evolving the covariance matrix of the state vector  $x \in \mathbb{K}^n$ . Instead, the EnKF uses an ensemble of  $N$  states

$$X = [x^{(1)}, \dots, x^{(N)}].$$

The columns of the matrix  $X \in \mathbb{K}^{n \times N}$  are the ensemble members.

**Initialization.** The ensemble is initialized by choosing the columns of  $\hat{X}_{0|0}$  to be  $N$  independent draws from, say,  $\mathcal{N}(\mu_0, Q_0)$ . However, the ensemble members are not generally independent except in the initial ensemble, since every EnKF step ties them together, but all the calculations proceed as if they actually were independent.

**Prediction.** The prediction step of the EnKF is straightforward: each column  $\hat{x}_{k-1|k-1}^{(i)}$  is evolved to  $\hat{x}_{k|k-1}$  using the LKF prediction step (7.6)

$$\hat{x}_{k|k-1}^{(i)} = F_k \hat{x}_{k-1|k-1}^{(i)} + G_k u_k,$$

or the EKF prediction step (7.13)

$$\hat{x}_{k|k-1} = f_k(\hat{x}_{k-1|k-1}, u_k).$$

**Correction.** The correction step for the EnKF uses a trick called *data replication*: the observed data  $y_k = H_k x_k + v_k$  is replicated into an  $m \times N$  matrix

$$D = [d^{(1)}, \dots, d^{(N)}], \quad d^{(i)} := y_k + \eta_i, \quad \eta_i \sim \mathcal{N}(0, R_k).$$

so that each column  $d_i$  consists of the observed data vector  $y_k$  plus an independent random draw from  $\mathcal{N}(0, R_k)$ . If the columns of  $\hat{X}_{k|k-1}$  are a sample from the prior distribution, then the columns of

$$\hat{X}_{k|k-1} + K_k (D - H_k \hat{X}_{k|k-1})$$

form a sample from the posterior probability distribution, in the sense of a Bayesian prior (before data) and posterior (conditioned upon the data). The EnKF approximates this sample by replacing the exact Kálmán gain matrix (7.10)

$$K_k := P_{k|k-1} H_k^* (H_k P_{k|k-1} H_k^* + R_k)^{-1},$$

which involves the covariance matrix  $P_{k|k-1}$ , which is not tracked in the EnKF, by an approximate covariance matrix. The empirical mean and empirical covariance of  $X$  are

$$\langle X \rangle := \frac{1}{N} \sum_{i=1}^N x^{(i)}, \quad \frac{(X - \langle X \rangle)(X - \langle X \rangle)^*}{N - 1}.$$

The Kálmán gain matrix for the EnKF uses  $C_{k|k-1}$  in place of  $P_{k|k-1}$ :

$$\tilde{K}_k := C_{k|k-1} H_k^* (H_k C_{k|k-1} H_k^* + R_k)^{-1}, \quad (7.17)$$

so that the correction step becomes

$$\hat{X}_{k|k} := \hat{X}_{k|k-1} + \tilde{K}_k (D - H_k \hat{X}_{k|k-1}). \quad (7.18)$$

One can also use sampling to dispense with  $R_k$ , and instead use the empirical covariance of the replicated data,

$$\frac{(D - \langle D \rangle)(D - \langle D \rangle)^*}{N - 1}.$$

Note, however, that the empirical covariance matrix is typically rank-deficient (in practical applications, there are usually many more state variables than ensemble members), in which case the matrix inverse in (7.17) may fail to exist; in such situations, a pseudo-inverse may be used.

**Remark 7.1.** Even when the matrices involved are positive-definite, instead of computing the inverse of a matrix and multiplying by it, it is much better (several times cheaper and also more accurate) to compute the Cholesky decomposition of the matrix and treat the multiplication by the inverse as solution of a system of linear equations (cf. [MA398 Matrix Analysis and Algorithms](#)). This is a general point relevant to the implementation of all KF-like methods.

## 7.5 Eulerian and Lagrangian Data Assimilation

Systems involving some kind of fluid flow are an important application area for data assimilation. It is interesting to consider two classes of observations of such systems, namely *Eulerian* and *Lagrangian* observations: Eulerian observations are observations at fixed points in space, whereas Lagrangian observations are observations at points that are carried along by the flow field. For example, a fixed weather station on the ground with a thermometer, a barometer, an anemometer &c. would report Eulerian observations of temperature, pressure and wind speed. By contrast, an unpowered float set adrift on the ocean currents would report Lagrangian data about temperature, salinity &c. at its current (evolving) position.

Consider for example the incompressible Stokes ( $\iota = 0$ ) or Navier–Stokes ( $\iota = 1$ ) equations on the unit square with periodic boundary conditions, thought of as the two-dimensional torus  $\mathbb{T}^2$ , starting at time  $t = 0$ :

$$\begin{aligned} \frac{\partial u}{\partial t} + \iota u \cdot \nabla u &= \nu \Delta u - \nabla p + f && \text{on } \mathbb{T}^2 \times [0, \infty), \\ \nabla \cdot u &= 0 && \text{on } \mathbb{T}^2 \times [0, \infty), \\ u &= u_0 && \text{on } \mathbb{T}^2 \times \{0\}. \end{aligned}$$

Here  $u, f: \mathbb{T}^2 \times [0, \infty) \rightarrow \mathbb{R}^2$  are the velocity field and forcing term respectively,  $p: \mathbb{T}^2 \times [0, \infty) \rightarrow \mathbb{R}$  is the pressure field,  $u_0: \mathbb{T}^2 \rightarrow \mathbb{R}^2$  is the initial value of the velocity field, and  $\nu \geq 0$  is the viscosity of the fluid.

Eulerian observations of this system might take the form of noisy observations  $y_{j,k}$  of the velocity field at fixed points  $z_j \in \mathbb{T}^2$ ,  $j = 1, \dots, J$ , at an increasing sequence of discrete times  $t_k \geq 0$ ,  $k \in \mathbb{N}$ , i.e.

$$y_{j,k} = u(z_j, t_k) + \eta_{j,k}.$$

On the other hand, Lagrangian observations might take the form of noisy observations  $y_{j,k}$  of the locations  $z_j(t_k) \in \mathbb{T}^2$  at time  $t_k$  of  $J$  passive tracers that start at position  $z_{j,0} \in \mathbb{T}^2$  at time  $t = 0$  and are carried along with the flow thereafter, i.e.

$$\begin{aligned} z_j(t) &= z_{j,0} + \int_0^t u(z_j(s), s) \, ds, \\ y_{j,k} &= z_j(t_k) + \eta_{j,k}. \end{aligned}$$

## Bibliography

The description given of the Kálmán filter, particularly in terms of Newton's method applied to the quadratic objective function  $J$ , follows that of Humpherys & al. [41]. The remarks about Eulerian versus Lagrangian data assimilation borrow from §3.6 of Stuart [98].

The original presentation of the Kálmán [46] and Kálmán–Bucy filters [47] was in the context of signal processing, and encountered some initial resistance from the engineering community, as related in the article of Humpherys & al. Filtering is now fully accepted in applications communities and has a sound algorithmic and theoretical base; for a stochastic processes point of view on filtering, see e.g. the books of Jazwinski [43] and Øksendal [73] (§6.1).



## Exercises

**Exercise 7.1.** Verify that the normal equations for the state estimation problem (7.4) have a unique solution.

**Exercise 7.2.** Suppose that  $A \in \mathbb{K}^{n \times n}$  and  $C \in \mathbb{K}^{m \times m}$  are self-adjoint and positive definite,  $B \in \mathbb{K}^{m \times n}$ , and  $D \in \mathbb{K}^{m \times m}$  is self-adjoint and positive semi-definite. Then

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

is self-adjoint and positive-definite.

**Exercise 7.3** (Schur complements). Let

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

be a square matrix with  $A$ ,  $D$ ,  $A - BD^{-1}C$  and  $D - CA^{-1}B$  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

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

**Exercise 7.4.** Schur complementation is often stated in the more restrictive setting of self-adjoint positive-definite matrices, in which it 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. Show:

1.  $C_{11}$  and  $C_{22}$  are necessarily self-adjoint and positive-definite matrices.
2. With the Schur complement defined by  $S := C_{11} - C_{12}C_{22}^{-1}C_{12}^*$ ,  $S$  is self-adjoint and positive definite, and

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

3. The conditional distribution of  $X$  given that  $Y = y$  is Gaussian:

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

Sketch the PDF of a Gaussian random variable in  $\mathbb{R}^2$  to further convince yourself of this result.

**Exercise 7.5** (Sherman–Morrison–Woodbury formula). Let  $A$  and  $D$  be invertible matrices and let  $B$  and  $C$  be such that  $A + BD^{-1}C$  and  $D + CA^{-1}B$  are non-singular. Then

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

**Exercise 7.6.** Determine the asymptotic computational costs of the correction steps (7.8) and (7.9) of the LKF, and hence the asymptotic computational cost per iteration of the LKF.

**Exercise 7.7 (Fading memory).** In the LKF, the current state variable is updated as the latest inputs and measurements become known, but the estimation is based on the least squares solution of all the previous states where all measurements are weighted according to their covariance. One can also use an estimator that discounts the error in older measurements leading to a greater emphasis on recent observations, which is particularly useful in situations where there is some modeling error in the system.

To do this, consider the objective function

$$J_{k|k}^{(\lambda)}(z_k) := \frac{\lambda^k}{2} \|x_0 - \mu_0\|_{Q_0^{-1}}^2 + \frac{1}{2} \sum_{i=1}^k \lambda^{k-i} \|y_i - H_i x_i\|_{R_i^{-1}}^2 \\ + \frac{1}{2} \sum_{i=1}^k \lambda^{k-1} \|x_i - F_i x_{i-1} - G_i u_i\|_{Q_i^{-1}}^2,$$

where the parameter  $\lambda \in [0, 1]$  is called the *forgetting factor*; note that the standard LKF is the case  $\lambda = 1$ , and the objective function increasingly relies upon recent measurements as  $\lambda \rightarrow 0$ . Find a recursive expression for the objective function  $J_{k|k}^{(\lambda)}$  and follow the steps in the derivation of the usual LKF to derive the LKF with fading memory  $\lambda$ .

**Exercise 7.8.** Write the prediction and correction equations (7.13)–(7.16) for the EKF in terms of the Kálmán gain matrix.

**Exercise 7.9.** Suppose that a fuel tank of an aircraft is (when the aircraft is level) the cuboid  $\Omega := [a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$ . Assume that at some time  $t$ , the aircraft is flying such that

- the original upward-pointing  $[0, 0, 1]^*$  vector of the plane, and hence the tank, is  $\nu(t) \in \mathbb{R}^3$ ;
- the fuel in the tank is in static equilibrium;
- fuel probes inside the tank provide noisy measurements of the fuel depth at the four corners of the tank: specifically, if  $[a_i, b_i, z_i]^*$  is the intersection of the fuel surface with the boundary of the tank at the corner  $[a_i, b_i]^*$ , assume that you are told  $\zeta_i = z_i + \mathcal{N}(0, \sigma^2)$ , independently for each fuel probe.

Using this information:

1. Assuming first that  $\sigma^2 = 0$ , calculate the volume of fuel in the tank.
2. Assuming now that  $\sigma^2 > 0$ , estimate the volume of fuel in the tank.
3. Explain how to estimate the volume of fuel remaining in the tank as a function of time. What other information other than the probe measurements might you use, and why?
4. Generalize your results to more general fuel tank and probe geometries, and more general observational noise.

**Exercise 7.10.** Consider the problem of using the LKF to estimate the position and velocity of a projectile given a few noisy measurements of its position. In fact, the LKF not only provides a relatively smooth profile of the projectile's

trajectory as it passes by a radar sensor, but also effectively predicts the point of impact as well as the point of origin — so that troops on the ground can both duck for cover and return fire before the projectile lands.

The state of the projectile is...

DRAFT

## Chapter 8

# Orthogonal Polynomials

Although our intellect always longs for clarity and certainty, our nature often finds uncertainty fascinating.

---

*On War*  
KARL VON CLAUSEWITZ

Orthogonal polynomials are an important example of orthogonal decompositions of Hilbert spaces. They are also of great practical importance: they play a central role in numerical integration using quadrature rules (Chapter 9) and approximation theory; in the context of UQ, they are also a foundational tool in polynomial chaos expansions (Chapter 11).

For the rest of this chapter,  $\mathcal{N} = \mathbb{N}_0$  or  $\{0, 1, \dots, N\}$  for some  $N \in \mathbb{N}_0$ .

### 8.1 Basic Definitions and Properties

Recall that a *polynomial of degree  $n$*  in a single indeterminate  $x$  is an expression of the form

$$p(x) = c_n x^n + c_{n-1} x^{n-1} + \dots + c_1 x + c_0,$$

where the coefficients  $c_i$  are scalars drawn from some field  $\mathbb{K}$ , and  $c_n \neq 0$ . If  $c_n = 1$ , then  $p$  is said to be *monic*. The space of all polynomials in  $x$  is denoted  $\mathbb{K}[x]$ , and the space of polynomials of degree at most  $n$  is denoted  $\mathbb{K}_{\leq n}[x]$ .

**Definition 8.1.** Let  $\mu$  be a non-negative measure on  $\mathbb{R}$ . A family of polynomials  $\mathcal{Q} = \{q_n \mid n \in \mathcal{N}\}$  is called an *orthogonal system of polynomials* if  $q_n$  is of degree  $n$  and

$$\langle q_m, q_n \rangle_{L^2(\mu)} := \int_{\mathbb{R}} q_m(x) q_n(x) d\mu(x) = \gamma_n \delta_{mn} \quad \text{for } m, n \in \mathcal{N}.$$

The constants

$$\gamma_n = \|q_n\|_{L^2(\mu)}^2 = \int_{\mathbb{R}} q_n^2 d\mu$$

are required to be strictly positive and are called the *normalization constants* of the system  $\mathcal{Q}$ . If  $\gamma_n = 1$  for all  $n \in \mathcal{N}$ , then  $\mathcal{Q}$  is an *orthonormal system*.

In other words, a system of orthogonal polynomials is nothing but a collection of orthogonal elements of the Hilbert space  $L^2(\mathbb{R}, \mu)$  that happen to be polynomials. Note that, given  $\mu$ , orthogonal (resp. orthonormal) polynomials for  $\mu$  can be found inductively by using the Gram–Schmidt orthogonalization (resp. orthonormalization) procedure on the monomials  $1, x, x^2, \dots$ .

**Example 8.2.** 1. The *Legendre polynomials*  $\text{Le}_n$  are the orthogonal polynomials for uniform measure on  $[-1, 1]$ :

$$\int_{-1}^1 \text{Le}_m(x) \text{Le}_n(x) dx = \delta_{mn}.$$

2. The *Hermite polynomials*  $\text{He}_n$  are the orthogonal polynomials for standard Gaussian measure  $(2\pi)^{-1/2} e^{-x^2/2} dx$  on the real line:

$$\int_{-\infty}^{\infty} \text{He}_m(x) \text{He}_n(x) \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx = n! \delta_{mn}.$$

The first few Legendre and Hermite polynomials are given in Table 8.1 and illustrated in Figures 8.1 and 8.2.

3. See Table 8.2 for a summary of other classical systems of orthogonal polynomials corresponding to various probability measures on subsets of the real line.



**Remark 8.3.** Many sources, typically physicists' texts, use the weight function  $e^{-x^2} dx$  instead of probabilists' preferred  $(2\pi)^{-1/2} e^{-x^2/2} dx$  or  $e^{-x^2/2} dx$  for the Hermite polynomials. Changing from one normalization to the other is of course not difficult, but special care must be exercised in practice to see which normalization a source is using, especially when relying on third-party software packages: for example, the GAUSSQ Gaussian quadrature package from <http://netlib.org/> uses the  $e^{-x^2} dx$  normalization. To convert integrals with respect to one Gaussian measure to integrals with respect to another (and hence get the right answers for Gauss–Hermite quadrature), use the following change-of-variables formula:

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(x) e^{-x^2/2} dx = \frac{1}{\pi} \int_{\mathbb{R}} f(\sqrt{2}x) e^{-x^2} dx.$$

It follows from this that conversion between the physicists' and probabilists' Gauss–Hermite quadrature formulæ is achieved by

$$w_i^{\text{prob}} = \frac{w_i^{\text{phys}}}{\sqrt{\pi}}, \quad x_i^{\text{prob}} = \sqrt{2} x_i^{\text{phys}}.$$

**Lemma 8.4.** The  $L^2(\mathbb{R}, \mu)$  inner product is positive definite on  $\mathbb{R}_{\leq d}[x]$  if and only if the Hankel determinant  $\det(H_n)$  is strictly positive for  $n = 1, \dots, d+1$ , where

$$H_n := \begin{bmatrix} m_0 & m_1 & \cdots & m_{n-1} \\ m_1 & m_2 & \cdots & m_n \\ \vdots & \vdots & \ddots & \vdots \\ m_{n-1} & m_n & \cdots & m_{2n-2} \end{bmatrix}, \quad m_n := \int_{\mathbb{R}} x^n d\mu(x).$$

Hence, the  $L^2(\mathbb{R}, \mu)$  inner product is positive definite on  $\mathbb{R}[x]$  if and only if  $\det(H_n) > 0$  for all  $n \in \mathbb{N}$ .

$n$	$Le_n$	$He_n$
0	1	1
1	$x$	$x$
2	$\frac{1}{2}(3x^2 - 1)$	$x^2 - 1$
3	$\frac{1}{2}(5x^3 - 3x)$	$x^3 - 3x$
4	$\frac{1}{8}(35x^4 - 30x^2 + 3)$	$x^4 - 6x^2 + 3$
5	$\frac{1}{8}(63x^5 - 70x^3 + 15x)$	$x^5 - 10x^3 + 15x$

Table 8.1: The first few Legendre polynomials  $Le_n$ , which are the orthogonal polynomials for uniform measure  $dx$  on  $[-1, 1]$ , and Hermite polynomials  $He_n$ , which are the orthogonal polynomials for standard Gaussian measure  $(2\pi)^{-1/2}e^{-x^2/2}dx$  on  $\mathbb{R}$ .

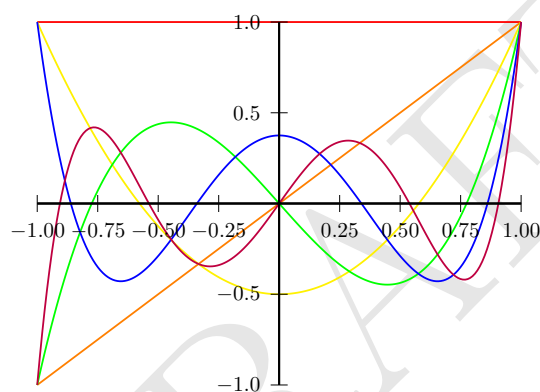


Figure 8.1: The Legendre polynomials  $Le_0$  (red),  $Le_1$  (orange),  $Le_2$  (yellow),  $Le_3$  (green),  $Le_4$  (blue) and  $Le_5$  (purple) on  $[-1, 1]$ .

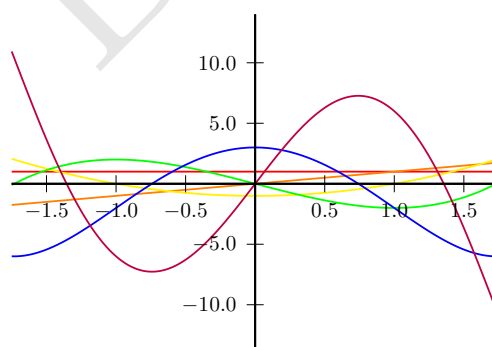


Figure 8.2: The Hermite polynomials  $He_0$  (red),  $He_1$  (orange),  $He_2$  (yellow),  $He_3$  (green),  $He_4$  (blue) and  $He_5$  (purple) on  $\mathbb{R}$ .

	Distribution of $\xi$	Polynomials $q_k(\xi)$	Support
Continuous	Gaussian	Hermite	$\mathbb{R}$
	Gamma	Laguerre	$[0, \infty)$
	Beta	Jacobi	$[a, b]$
	Uniform	Legendre	$[a, b]$
Discrete	Poisson	Charlier	$\mathbb{N}_0$
	Binomial	Krawtchouk	$\{0, 1, \dots, n\}$
	Negative Binomial	Meixner	$\mathbb{N}_0$
	Hypergeometric	Hahn	$\{0, 1, \dots, n\}$

Table 8.2: Families of probability distributions and the corresponding families of orthogonal polynomials.

*Proof.* Let

$$p(x) := c_d x^d + \dots + c_1 x + c_0$$

be any polynomial of degree at most  $d$ . Note that

$$\|p\|_{L^2(\mathbb{R}, \mu)}^2 = \int_{\mathbb{R}} \sum_{k, \ell=0}^d c_k c_\ell x^{k+\ell} d\mu(x) = \sum_{k, \ell=0}^d c_k c_\ell m_{k+\ell},$$

and so  $\|p\|_{L^2(\mathbb{R}, \mu)} > 0$  if and only if  $H_{d+1}$  is positive definite. This, in turn, is equivalent to having  $\det(H_n) > 0$  for  $n = 1, 2, \dots, d+1$ .  $\square$

**Theorem 8.5.** *If the  $L^2(\mathbb{R}, \mu)$  inner product is positive definite on  $\mathbb{R}[x]$ , then there exists an infinite sequence of orthogonal polynomials for  $\mu$ .*

*Proof.* Apply the Gram–Schmidt procedure to the monomials  $x^n$ ,  $n \in \mathbb{N}_0$ . That is, take  $q_0(x) = 1$ , and for  $n \in \mathbb{N}$  recursively define

$$q_n(x) := x^n - \sum_{k=0}^{n-1} \frac{\langle x^k, q_k \rangle}{\langle q_k, q_k \rangle} q_k(x).$$

Since the inner product is positive definite,  $\langle q_k, q_k \rangle > 0$ , and so each  $q_n$  is uniquely defined. By construction, each  $q_n$  is orthogonal to  $q_k$  for  $k < n$ .  $\square$

By Exercise 8.1, the hypothesis of Theorem 8.5 is satisfied if the measure  $\mu$  has infinite support. In the other direction, we have the following:

**Theorem 8.6.** *If the  $L^2(\mathbb{R}, \mu)$  inner product is positive definite on  $\mathbb{K}_{\leq d}[x]$ , but not on  $\mathbb{K}_{\leq n}[x]$  for  $n > d$ , then  $\mu$  admits only  $d+1$  orthogonal polynomials.*

*Proof.* The Gram–Schmidt procedure can be applied so long as the denominators  $\langle q_k, q_k \rangle$  are strictly positive, i.e. for  $k \leq d+1$ . The polynomial  $q_{d+1}$  is orthogonal to  $q_n$  for  $n \leq d$ ; we now show that  $q_{d+1} = 0$ . By assumption, there exists a polynomial  $P$  of degree  $d+1$ , having the same leading coefficient as  $q_{d+1}$ , such that  $\|P\|_{L^2(\mathbb{R}, \mu)} = 0$ . Hence,  $P - q_{d+1}$  has degree  $d$ , so it can be written in the orthogonal basis  $\{q_0, \dots, q_d\}$  as

$$P - q_{d+1} = \sum_{k=0}^d c_k q_k$$



for some coefficients  $c_0, \dots, c_d$ . Hence,

$$0 = \|P\|_{L^2(\mathbb{R}, \mu)}^2 = \|q_{d+1}\|_{L^2(\mathbb{R}, \mu)}^2 + \sum_{k=0}^d c_k^2 \|q_k\|_{L^2(\mathbb{R}, \mu)}^2,$$

which implies, in particular, that  $\|q_{d+1}\|_{L^2(\mathbb{R}, \mu)} = 0$ . Hence, the normalization constant  $\gamma_{d+1} = 0$ , which is not permitted, and so  $q_{d+1}$  is not a member of a sequence of orthogonal polynomials for  $\mu$ .  $\square$

**Theorem 8.7.** *If  $\mu$  has finite moments only of degrees  $0, 1, \dots, r$ , then  $\mu$  admits only a finite system of orthogonal polynomials  $q_0, \dots, q_d$ , where  $d = \lfloor r/2 \rfloor$ .*

*Proof.* Exercise 8.2  $\square$

**Theorem 8.8.** *The coefficients of any system of orthogonal polynomials are determined, up to multiplication by an arbitrary constant for each degree, by the Hankel determinants of the polynomial moments. That is, if*

$$m_n := \int_{\mathbb{R}} x^n d\mu(x)$$

then the  $n^{\text{th}}$  degree orthogonal polynomial  $q_n$  for  $\mu$  is, for some  $c_n \neq 0$ ,

$$\begin{aligned} q_n &= c_n \det \left[ \begin{array}{ccc|c} & & & m_n \\ & & & \vdots \\ & & & m_{2n-1} \\ \hline 1 & \dots & x^{n-1} & x^n \end{array} \right] \\ &= c_n \det \begin{bmatrix} m_0 & m_1 & m_2 & \dots & m_n \\ m_1 & m_2 & m_3 & \dots & m_{n+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{n-1} & m_n & m_{n+2} & \dots & m_{2n-1} \\ 1 & x & x^2 & \dots & x^n \end{bmatrix}. \end{aligned}$$

*Proof.* FINISH ME!!!  $\square$

## 8.2 Recurrence Relations

The Legendre polynomials satisfy the recurrence relation

$$\text{Le}_{n+1}(x) = \frac{2n+1}{n+1} x \text{Le}_n(x) - \frac{n}{n+1} \text{Le}_{n-1}(x).$$

The Hermite polynomials satisfy the recurrence relation

$$\text{He}_{n+1}(x) = x \text{He}_n(x) - n \text{He}_{n-1}(x).$$

In fact, all systems of orthogonal polynomials satisfy a *three-term recurrence relation* of the form

$$q_{n+1}(x) = (A_n x + B_n) q_n(x) - C_n q_{n-1}(x)$$

for some sequences  $(A_n)$ ,  $(B_n)$ ,  $(C_n)$ , with the initial terms  $q_0(x) = 1$  and  $q_{-1}(x) = 0$ . Furthermore, there is a characterization of precisely which sequences  $(A_n)$ ,  $(B_n)$ ,  $(C_n)$  arise from systems of orthogonal polynomials.

**Theorem 8.9 (Favard).** Let  $(A_n)$ ,  $(B_n)$ ,  $(C_n)$  be real sequences and let  $\mathcal{Q} = \{q_n \mid n \in \mathcal{N}\}$  be defined by

$$\begin{aligned} q_{n+1}(x) &= (A_n x + B_n)q_n(x) - C_n q_{n-1}(x), \\ q_0(x) &= 1, \\ q_{-1}(x) &= 0. \end{aligned}$$

Then  $\mathcal{Q}$  is a system of orthogonal polynomials for some measure  $\mu$  if and only if, for all  $n \in \mathcal{N}$ ,

$$A_n \neq 0, \quad C_n \neq 0, \quad C_n A_n A_{n-1} > 0.$$

**Theorem 8.10 (Christoffel–Darboux formula).** The orthonormal polynomials  $\{P_n \mid n \geq 0\}$  for a measure  $\mu$  satisfy

$$\sum_{k=0}^n P_k(y)P_k(x) = \sqrt{\beta_{n+1}} \frac{P_{n+1}(y)P_n(x) - P_n(y)P_{n+1}(x)}{y - x} \quad (8.1)$$

and

$$\sum_{k=0}^n |P_k(x)|^2 = \sqrt{\beta_{n+1}} (P'_{n+1}(x)P_n(x) - P'_n(x)P_{n+1}(x)). \quad (8.2)$$

*Proof.* Multiply the recurrence relation

$$\sqrt{\beta_{k+1}}P_{k+1}(x) = (x - \alpha_k)P_k(x) - \sqrt{\beta_k}P_{k-1}(x)$$

by  $P_k(y)$  on both sides and subtract the corresponding expression with  $x$  and  $y$  interchanged to obtain

$$\begin{aligned} (y - x)P_k(y)P_k(x) &= \sqrt{\beta_{k+1}}(P_{k+1}(y)P_k(x) - P_k(y)P_{k+1}(x)) \\ &\quad - \sqrt{\beta_k}(P_k(y)P_{k-1}(x) - P_{k-1}(y)P_k(x)). \end{aligned}$$

Sum both sides from  $k = 0$  to  $k = n$  and use the telescoping nature of the sum on the right to obtain (8.1). Take the limit as  $y \rightarrow x$  to obtain (8.2).  $\square$

**Corollary 8.11.** The orthonormal polynomials  $\{P_n \mid n \geq 0\}$  for a measure  $\mu$  satisfy

$$P'_{n+1}(x)P_n(x) - P'_n(x)P_{n+1}(x) > 0.$$

### 8.3 Roots of Orthogonal Polynomials

**Definition 8.12.** The *Jacobi matrix* of a measure  $\mu$  is the infinite, symmetric, tridiagonal matrix

$$J_\infty(\mu) := \begin{bmatrix} \alpha_0 & \sqrt{\beta_1} & 0 & \cdots \\ \sqrt{\beta_1} & \alpha_1 & \sqrt{\beta_2} & \ddots \\ 0 & \sqrt{\beta_2} & \alpha_2 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix}$$

where  $\alpha_k$  and  $\beta_k$  are as in **FINISH ME!!!**. The upper-left  $n \times n$  minor of  $J_\infty(\mu)$  is denoted  $J_n(\mu)$ .

**Theorem 8.13.** Let  $P_0, P_1, \dots$  be the orthonormal polynomials for  $\mu$ . The zeros of  $P_n$  are the eigenvalues of  $J_n(\mu)$ , and the eigenvector corresponding to the zero  $z$  is

$$\begin{bmatrix} P_0(z) \\ \vdots \\ P_{n-1}(z) \end{bmatrix}.$$

**Theorem 8.14** (Zeros of orthogonal polynomials). Let  $\mu$  be supported in a non-degenerate interval  $I \subseteq \mathbb{R}$ , and let  $\mathcal{Q} = \{q_n \mid n \in \mathbb{N}_0\}$  be a system of orthogonal polynomials for  $\mu$

1. For each  $n \in \mathbb{N}_0$ ,  $q_n$  has exactly  $n$  distinct real roots  $z_1^{(n)}, \dots, z_n^{(n)} \in I$ .
2. If  $(a, b)$  is an open interval of  $\mu$ -measure zero, then  $(a, b)$  contains at most one root of any orthogonal polynomial  $q_n$  for  $\mu$ .
3. The zeros of  $q_n$  and  $q_{n+1}$  alternate:

$$z_1^{(n+1)} < z_1^{(n)} < z_2^{(n+1)} < \dots < z_n^{(n+1)} < z_n^{(n)} < z_{n+1}^{(n+1)};$$

hence, whenever  $m > n$ , between any two zeros of  $q_n$  there lies a zero of  $q_m$ .

4. If the support of  $\mu$  is the entire interval  $I$ , then the set of all zeros for the system  $\mathcal{Q}$  is dense in  $I$ :

$$I = \overline{\bigcup_{n \in \mathbb{N}} \{z \in \mathbb{R} \mid q_n(z) = 0\}}.$$

*Proof.* 1. First observe that  $\langle q_n, 1 \rangle_{L^2(\mu)} = 0$ , and so  $q_n$  changes sign in  $I$ . Since  $q_n$  is continuous, the Intermediate Value Theorem implies that  $q_n$  has at least one real root  $z_1^{(n)} \in I$ . For  $n > 1$ , there must be another root  $z_2^{(n)} \in I$  of  $q_n$  distinct from  $z_1^{(n)}$ , since if  $q_n$  were to vanish only at  $z_1^{(n)}$ , then  $(x - z_1^{(n)})q_n$  would not change sign in  $I$ , which would contradict the orthogonality relation  $\langle x - z_1^{(n)}, q_n \rangle_{L^2(\mu)} = 0$ . Similarly, if  $n > 2$ , consider  $(x - z_1^{(n)})(x - z_2^{(n)})q_n$  to deduce the existence of yet a third distinct root  $z_3^{(n)} \in I$ . This procedure terminates when all the  $n$  complex roots of  $q_n$  guaranteed by the Fundamental Theorem of Algebra are shown to lie in  $I$ .

2. Suppose that  $(a, b)$  contains two distinct zeros  $z_i^{(n)}$  and  $z_j^{(n)}$  of  $q_n$ . Then

$$\begin{aligned} \left\langle q_n, \prod_{k \neq i, j} (x - z_k^{(n)}) \right\rangle_{L^2(\mu)} &= \int_{\mathbb{R}} q_n(x) \prod_{k \neq i, j} (x - z_k^{(n)}) d\mu(x) \\ &= \int_{\mathbb{R}} \prod_{k \neq i, j} (x - z_k^{(n)})^2 (x - z_i^{(n)}) (x - z_j^{(n)}) d\mu(x) \\ &> 0, \end{aligned}$$

since the integrand is positive outside of  $(a, b)$ . However, this contradicts the orthogonality of  $q_n$  to all polynomials of degree less than  $n$ .

3. As usual, let  $P_n$  be the normalized version of  $q_n$ . Let  $\sigma, \tau$  be consecutive zeros of  $P_n$ , so that  $P'_n(\sigma)P'_n(\tau) < 0$ . Then Corollary 8.11 implies that  $P_{n+1}$

has opposite signs at  $\sigma$  and  $\tau$ , and so the Intermediate Value Theorem implies that  $P_{n+1}$  has at least one zero between  $\sigma$  and  $\tau$ . This observation accounts for  $n-1$  of the  $n+1$  zeros of  $P_{n+1}$ , namely  $z_2^{(n+1)} < \dots < z_n^{(n+1)}$ . There are two further zeros of  $P_{n+1}$ , one to the left of  $z_1^{(n)}$  and one to the right of  $z_n^{(n)}$ . This follows because  $P'_n(z_n^{(n)}) > 0$ , so Corollary 8.11 implies that  $P_{n+1}(z_n^{(n)}) < 0$ . Since  $P_{n+1}(x) \rightarrow +\infty$  as  $x \rightarrow \infty$ , the Intermediate Value Theorem implies the existence of  $z_{n+1}^{(n+1)} > z_n^{(n)}$ . A similar argument establishes the existence of  $z_1^{(n+1)} < z_1^{(n)}$ .

4. **FINISH ME!!!**

□

## 8.4 Polynomial Interpolation

The existence of a unique polynomial  $p(x) = \sum_{i=0}^n c_i x^i$  of degree at most  $n$  that interpolates the values  $y_0, \dots, y_n$  at  $n+1$  distinct points  $x_0, \dots, x_n$  follows from the invertibility of the *Vandermonde matrix*

$$V_n(x_0, \dots, x_n) := \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^{n+1} \\ 1 & x_1 & x_1^2 & \cdots & x_1^{n+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n+1} \end{bmatrix} \quad (8.3)$$

and hence the unique solvability of the system of simultaneous linear equations

$$V_n(x_0, \dots, x_n) \begin{bmatrix} c_0 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} y_0 \\ \vdots \\ y_n \end{bmatrix}. \quad (8.4)$$

There is, however, another way to express the polynomial interpolation problem, the so-called *Lagrange form*, which amounts to a clever choice of basis for  $\mathbb{R}_{\leq n}[x]$  (instead of the usual monomial basis  $\{1, x, x^2, \dots, x^n\}$ ) so that the matrix in (8.4) in the new basis is the identity matrix.

**Definition 8.15** (Lagrange polynomials). Let  $x_0, \dots, x_K \in \mathbb{R}$  be distinct. For  $0 \leq j \leq K$ , the associated *Lagrange basis polynomial*  $\ell_j$  is defined by

$$\ell_j(x) := \prod_{\substack{0 \leq k \leq K \\ k \neq j}} \frac{x - x_k}{x_j - x_k}.$$

Given also arbitrary values  $y_0, \dots, y_K \in \mathbb{R}$ , the associated *Lagrange interpolation polynomial* is

$$L(x) := \sum_{j=0}^K y_j \ell_j(x).$$

**Theorem 8.16.** Given distinct  $x_0, \dots, x_K \in \mathbb{R}$  and any  $y_0, \dots, y_K \in \mathbb{R}$ , the associated Lagrange interpolation polynomial  $L$  is the unique polynomial of degree at most  $K$  such that  $L(x_k) = y_k$  for  $k = 0, \dots, K$ .

*Proof.* Observe that each Lagrange basis polynomial is a polynomial of degree  $K$ , and so  $L$  is a polynomial of degree at most  $K$ . Observe also that  $\ell_j(x_k) = \delta_{jk}$ . Hence,

$$L(x_k) = \sum_{j=0}^K f(x_j) \ell_j(x_k) = \sum_{j=0}^K f(x_j) \delta_{jk} = f(x_k).$$

For uniqueness, consider the basis  $\{\ell_0, \dots, \ell_K\}$  of  $\mathbb{R}_{\leq K}[x]$  and suppose that  $p = \sum_{j=0}^K c_j \ell_j$  is any polynomial that interpolates the values  $\{y_k\}_{k=0}^K$  at the points  $\{x_k\}_{k=0}^K$ . But then, for each  $k = 0, \dots, K$ ,

$$f(x_k) = \sum_{j=0}^K c_j \ell_j(x_k) = \sum_{j=0}^K c_j \delta_{jk} = c_k,$$

and so  $p = L$ , as claimed.  $\square$

**Runge's Phenomenon.** Given the task of choosing nodes  $x_k \in [a, b]$  between which to interpolate functions  $f: [a, b] \rightarrow \mathbb{R}$ , it might seem natural to choose the nodes  $x_k$  to be equally spaced. Runge's phenomenon [84] shows that this is not always a good choice of interpolation scheme. Consider the function  $f: [-1, 1] \rightarrow \mathbb{R}$  defined by

$$f(x) := \frac{1}{1 + 25x^2}, \quad (8.5)$$

and let  $L_n$  be the degree- $n$  (Lagrange) interpolation polynomial for  $f$  on the equally-spaced nodes  $x_k := \frac{2k}{n} - 1$ . As illustrated in Figure 8.1,  $L_n$  oscillates wildly near the endpoints of the interval  $[-1, 1]$ . Even worse, as  $n$  increases, these oscillations do not die down but increase without bound: it can be shown that

$$\lim_{n \rightarrow \infty} \sup_{x \in [-1, 1]} |f(x) - L_n(x)| = \infty.$$

As a consequence, polynomial interpolation and numerical integration using uniformly spaced nodes — as in the Newton–Cotes formula (Definition 9.5) — can in general be very inaccurate. The oscillations near  $\pm 1$  can be controlled by using a non-uniform set of nodes, in particular one that is denser  $\pm 1$  than near 0; the standard example is the set of *Chebyshev nodes* defined by

$$x_k := \cos\left(\frac{2k-1}{2n}\pi\right),$$

i.e. the roots of the Chebyshev polynomials of the first kind  $T_n$ , which are the orthogonal polynomials for the measure  $(1 - x^2)^{-1/2} dx$  on  $[-1, 1]$ .

However, Chebyshev nodes are not a panacea. Indeed, for every predefined set of interpolation nodes there is a continuous function for which the interpolation process on those nodes diverges. For every continuous function there is a set of nodes on which the interpolation process converges. Interpolation on Chebyshev nodes converges uniformly for every absolutely continuous function.

## 8.5 Polynomial Approximation

The following theorem on the uniform approximation (on compact sets) of continuous functions by polynomials should be familiar from elementary real analysis:

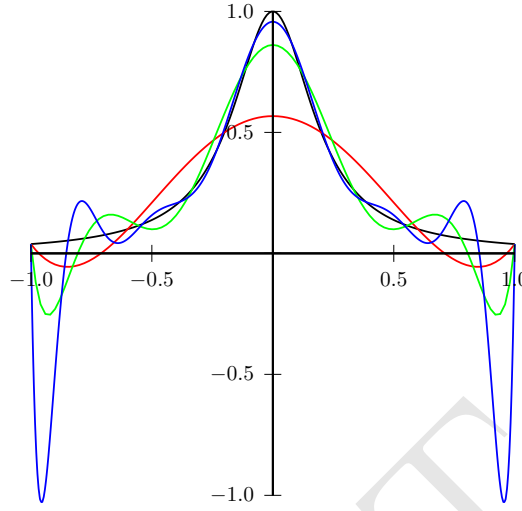


Figure 8.1: Runge's phenomenon. The function  $f(x) := (1 + 25x^2)^{-1}$  in black, and polynomial interpolations of degrees 5 (red), 9 (green), and 13 (blue) on evenly-spaced nodes.

**Theorem 8.17** (Weierstrass). *Let  $[a, b] \subset \mathbb{R}$  be a bounded interval, let  $f: [a, b] \rightarrow \mathbb{R}$  be continuous, and let  $\varepsilon > 0$ . Then there exists a polynomial  $p$  such that*

$$\sup_{a \leq x \leq b} |f(x) - p(x)| < \varepsilon.$$

As a consequence of standard results on orthogonal projection in Hilbert spaces, we have the following:

**Theorem 8.18.** *For any  $f \in L^2(I, \mu)$  and any  $d \in \mathbb{N}_0$ , the orthogonal projection  $\Pi_d f$  of  $f$  onto  $\mathbb{R}_{\leq d}[x]$  is the best degree  $d$  polynomial approximation of  $f$  in the  $L^2(I, \mu)$  norm, i.e.*

$$\Pi_d f = \arg \min_{p(x) \in \mathbb{R}_{\leq d}[x]} \|p - f\|_{L^2(I, \mu)},$$

where, denoting the orthogonal polynomials for  $\mu$  by  $\{q_k \mid k \in \mathbb{N}_0\}$ ,

$$\Pi_d f := \sum_{k=0}^d \frac{\langle f, q_k \rangle_{L^2(\mu)}}{\|q_k\|_{L^2(\mu)}^2} q_k,$$

and the residual is orthogonal to the projection subspace:

$$\langle f - \Pi_d f, p \rangle_{L^2(\mu)} = 0 \quad \text{for all } p(x) \in \mathbb{R}_{\leq d}[x].$$

An important property of polynomial expansions of functions is that the quality of the approximation (i.e. the rate of convergence) improves as the regularity of the function to be approximated increases. This property is referred

to as *spectral convergence* and is easily quantified by using the machinery of Sobolev spaces. Recall that, given a measure  $\mu$  on a subinterval  $I \subseteq \mathbb{R}$

$$\begin{aligned}\langle u, v \rangle_{H^k(\mu)} &:= \sum_{m=0}^k \left\langle \frac{d^m u}{dx^m}, \frac{d^m v}{dx^m} \right\rangle_{L^2(\mu)} = \sum_{m=0}^k \int_I \frac{d^m u}{dx^m} \frac{d^m v}{dx^m} d\mu \\ \|u\|_{H^k(\mu)} &:= \langle u, u \rangle_{H^k(\mu)}^{1/2}.\end{aligned}$$

The Sobolev space  $H^k(\mu)$  consists of all  $L^2(\mu)$  functions that have weak derivatives of all orders up to  $k$  in  $L^2(\mu)$ , and is equipped with the above inner product and norm.

Legendre expansions of Sobolev functions on  $[-1, 1]$  satisfy the following spectral convergence theorem; the analogous results for Hermite expansions of Sobolev functions on  $\mathbb{R}$  and Laguerre expansions of Sobolev functions on  $(0, \infty)$  are Exercise 8.5 and Exercise 8.6 respectively.

**Theorem 8.19** (Spectral convergence of Legendre expansions). *There is a constant  $C \geq 0$  that may depend upon  $k$  but is independent of  $d$  and  $f$  such that, for all  $f \in H^k([-1, 1], dx)$ ,*

$$\|f - \Pi_d f\|_{L^2(dx)} \leq C d^{-k} \|f\|_{H^k(dx)}.$$

*Proof.* Recall that the Legendre polynomials satisfy

$$\mathcal{L} \text{Le}_n = \lambda_n \text{Le}_n,$$

where the differential operator  $\mathcal{L}$  and eigenvalues  $\lambda_n$  are

$$\mathcal{L} = \frac{d}{dx} \left( (1-x^2) \frac{d}{dx} \right) = (1-x^2) \frac{d^2}{dx^2} - 2x \frac{d}{dx}, \quad \lambda_n = -n(n+1)$$

Note that, by the definition of the Sobolev norm and the operator  $\mathcal{L}$ ,  $\|\mathcal{L}f\|_{L^2} \leq C\|f\|_{H^2}$  and hence, for any  $m \in \mathbb{N}$ ,  $\|\mathcal{L}^m f\|_{L^2} \leq C\|f\|_{H^{2m}}$ .

The key ingredient of the proof is integration by parts:

$$\begin{aligned}\langle f, \text{Le}_n \rangle_{L^2} &= \lambda_n^{-1} \int_{-1}^1 (\mathcal{L} \text{Le}_n)(x) f(x) dx \\ &= \lambda_n^{-1} \int_{-1}^1 ((1-x^2) \text{Le}_n''(x) f(x) - 2x \text{Le}_n'(x) f(x)) dx \\ &= -\lambda_n^{-1} \int_{-1}^1 (((1-x^2)f)'(x) \text{Le}_n'(x) + 2x \text{Le}_n'(x) f(x)) dx \\ &= -\lambda_n^{-1} \int_{-1}^1 (1-x^2) f'(x) \text{Le}_n'(x) dx \\ &= \lambda_n^{-1} \int_{-1}^1 ((1-x^2)f')'(x) \text{Le}_n'(x) dx \\ &= \lambda_n^{-1} \langle \mathcal{L}f, \text{Le}_n \rangle_{L^2}.\end{aligned}$$

Hence, for all  $m \in \mathbb{N}_0$  for which  $f$  has  $2m$  weak derivatives,

$$\langle f, \text{Le}_n \rangle = \frac{\langle \mathcal{L}^m f, \text{Le}_n \rangle}{\lambda_n^m}$$

Hence,

$$\begin{aligned}
\|f - \Pi_d f\|^2 &= \sum_{n=d+1}^{\infty} \frac{|\langle f, \text{Le}_n \rangle|^2}{\|\text{Le}_n\|^2} \\
&= \sum_{n=d+1}^{\infty} \frac{|\langle \mathcal{L}^m f, \text{Le}_n \rangle|^2}{\lambda_n^{2m} \|\text{Le}_n\|^2} \\
&\leq \lambda_d^{-2m} \sum_{n=d+1}^{\infty} \frac{|\langle \mathcal{L}^m f, \text{Le}_n \rangle|^2}{\|\text{Le}_n\|^2} \\
&\leq d^{-4m} \|\mathcal{L}^m f\|^2 \\
&\leq C^2 d^{-4m} \|f\|_{H^{2m}}^2.
\end{aligned}$$

Taking  $k = 2m$  and square roots completes the proof.  $\square$

However, in the other direction poor regularity can completely ruin the nice convergence of spectral expansions. The classic example of this is *Gibbs' phenomenon*, in which one tries to approximate the sign function

$$\text{sgn}(x) := \begin{cases} -1, & \text{if } x < 0, \\ 0, & \text{if } x = 0, \\ 1, & \text{if } x > 0, \end{cases}$$

on  $[-1, 1]$  by its expansion with respect to a system of orthogonal polynomials such as the Legendre polynomials  $\text{Le}_n(x)$  or the Fourier polynomials  $e^{\pi n x}$ . **FINISH ME!!!**

## 8.6 Orthogonal Polynomials of Several Variables

For working with polynomials in  $d$  variables, we will use standard multi-index notation. Multi-indices will be denoted by Greek letters  $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ . For  $x = (x_1, \dots, x_d) \in \mathbb{R}^d$  and  $\alpha \in \mathbb{N}_0^d$ , the monomial  $x^\alpha$  is defined by

$$x^\alpha := x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d},$$

and  $|\alpha| := \alpha_1 + \dots + \alpha_d$  is called the *total degree* of  $x^\alpha$ . The total degree of a polynomial  $p$  (i.e. a finite linear combination of such monomials) is denoted  $\deg(p)$  and is the maximum of the total degrees of the summands.

Given a measure  $\mu$  on  $\mathbb{R}^d$ , it is tempting to apply the Gram–Schmidt process with respect to the inner product

$$\langle f, g \rangle_{L^2(\mu)} := \int_{\mathbb{R}^d} f(x)g(x) d\mu(x)$$

to the monomials  $\{x^\alpha \mid \alpha \in \mathbb{N}_0^d\}$  to obtain a system of orthogonal polynomials for the measure  $\mu$ . However, there is an immediate problem, in that orthogonal polynomials of several variables are not unique. In order to apply the Gram–Schmidt process, we need to give a linear order to multi-indices  $\alpha \in \mathbb{N}_0^d$ . There are many choices of well-defined total order (for example, the lexicographic order or the graded lexicographic order); but there is no natural choice and different orders will give different sequences of orthogonal polynomials. Instead of fixing such a total order, we relax Definition 8.1 slightly:



**Definition 8.20.** Let  $\mu$  be a non-negative measure on  $\mathbb{R}^d$ . A family of polynomials  $\mathcal{Q} = \{q_\alpha \mid \alpha \in \mathbb{N}_0^d\}$  is called an *orthogonal system of polynomials* if  $q_\alpha$  is such that

$$\langle q_\alpha, p \rangle_{L^2(\mu)} = 0 \quad \text{for all } p(x) \in \mathbb{R}[x_1, \dots, x_d] \text{ with } \deg(p) < |\alpha|.$$

Hence, in the many-variables case, an orthogonal polynomial of total degree  $n$ , while it is required to be orthogonal to all polynomials of strictly lower total degree, may be non-orthogonal to other polynomials of the same total degree  $n$ . However, the meaning of orthonormality is unchanged: a system of polynomials  $\{P_\alpha \mid \alpha \in \mathbb{N}_0^d\}$  is *orthonormal* if

$$\langle P_\alpha, P_\beta \rangle_{L^2(\mu)} = \delta_{\alpha\beta}.$$

## Bibliography

The classic reference on orthogonal polynomials is the 1939 monograph of Szegő [100]. An excellent more modern reference is the book of Gautschi [33]; some topics covered in that book that are not treated here include. . . **FINISH ME!!!**

Many important properties of orthogonal polynomials, and standard examples, are given in Chapter 22 of Abramowitz & Stegun [1].

## Exercises

**Exercise 8.1.** Show that the  $L^2(\mathbb{R}, \mu)$  inner product is positive definite on the space of polynomials if the measure  $\mu$  has infinite support.

**Exercise 8.2.** Show that if  $\mu$  has finite moments only of degrees  $0, 1, \dots, r$ , then  $\mu$  admits only a finite system of orthogonal polynomials  $q_0, \dots, q_d$ , where  $d = \lfloor r/2 \rfloor$ .

**Exercise 8.3.** Define a Borel measure  $\mu$  on  $\mathbb{R}$  by

$$\frac{d\mu}{dx}(x) = \frac{1}{\pi} \frac{1}{1+x^2}.$$

Show that  $\mu$  is a probability measure, that  $\dim L^2(\mathbb{R}, \mu; \mathbb{R}) = \infty$ , find all orthogonal polynomials for  $\mu$ , and explain your results.

**Exercise 8.4.** Calculate the orthogonal polynomials of Table 8.2 by hand for degree  $p \leq 5$ , and write a numerical program to compute them for higher degree.

**Exercise 8.5 (Spectral convergence of Hermite expansions).** Let  $\gamma = \mathcal{N}(0, 1)$  be standard Gaussian measure on  $\mathbb{R}$ . First establish the integration-by-parts formula

$$\int_{\mathbb{R}} f(x)g'(x) d\gamma(x) = - \int_{\mathbb{R}} (f'(x) - xf(x))g(x) d\gamma(x).$$

Using this, and the fact that the Hermite polynomials satisfy

$$\left( \frac{d^2}{dx^2} - x \frac{d}{dx} \right) \text{He}_n(x) = -n \text{He}_n(x), \quad \text{for } n \in \mathbb{N}_0,$$

mimic the proof of Theorem 8.19 to show that there is a constant  $C \geq 0$  that may depend upon  $k$  but is independent of  $d$  and  $f$  such that, for all  $f \in H^k(\mathbb{R}, \gamma)$ ,  $f$  and its degree  $d$  expansion in the Hermite orthogonal basis of  $L^2(\mathbb{R}, \gamma)$  satisfy

$$\|f - \Pi_d f\|_{L^2(\gamma)} \leq C d^{-k/2} \|f\|_{H^k(\gamma)}.$$

**Exercise 8.6** (Spectral convergence of Laguerre expansions). Let  $d\mu(x) = e^{-x} dx$ , for which the orthogonal polynomials are the Laguerre polynomials  $L_n$ ,  $n \in \mathbb{N}_0$ . Establish an integration-by-parts formula for  $\mu$  and then use this and the fact that  $L_n$  is an eigenfunction for  $x \frac{d^2}{dx^2} + (1-x) \frac{d}{dx}$  with eigenvalue  $-n$  to prove the analogue of Exercise 8.5 but for Laguerre expansions.

## Chapter 9

# Numerical Integration

A turkey is fed for a 1000 days — every day confirms to its statistical department that the human race cares about its welfare “with increased statistical significance”. On the 1001<sup>st</sup> day, the turkey has a surprise.

---

*The Fourth Quadrant: A Map of the  
Limits of Statistics*  
NASSIM TALEB

The topic of this chapter is the numerical (i.e. approximate) evaluation of definite integrals. Such methods of numerical integration will be essential if the expectations the UQ methods of later chapters — with their many expectations — are to be implemented in a practical manner.

The topic of integration has a long history, as one of the twin pillars of calculus, and was historically also known as quadrature. Nowadays, *quadrature* usually refers to a particular method of numerical integration.

### 9.1 Quadrature in One Dimension

This section concerns the numerical integration of a real-valued function  $f$  with respect to a measure  $\mu$  on a sub-interval  $I \subseteq \mathbb{R}$ , and to do so by sampling the function at pre-determined points of  $I$  and taking a suitable weighted average. That is, the aim is to construct an approximation of the form

$$\int_I f(x) d\mu(x) \approx Q(f) := \sum_{k=1}^K w_k f(x_k),$$

with prescribed points  $x_1, \dots, x_K \in I$  called *nodes* and scalars  $w_1, \dots, w_K \in \mathbb{R}$  called *weights*. The approximation  $Q(f)$  is called a *quadrature formula*. The aim is to choose nodes and weights wisely, so that the quality of the approximation  $\int_I f d\mu \approx Q(f)$  is good for a large class of integrands  $f$ . One measure of the quality of the approximation is the following:

**Definition 9.1.** A quadrature formula is said to have *order of accuracy*  $n \in \mathbb{N}_0$  if  $\int_I f \, d\mu = Q(f)$  whenever  $f$  is a polynomial of degree at most  $n$ .

A quadrature formula  $Q(f) = \sum_{k=1}^K w_k f(x_k)$  can be identified with the discrete measure  $\sum_{k=1}^K w_k \delta_{x_k}$ . If some of the weights  $w_k$  are negative, then this measure is a signed measure. This point of view will be particularly useful when considering multi-dimensional quadrature formulas. Regardless of the signature of the weights, the following limitation on the accuracy of quadrature formulas is fundamental:

**Lemma 9.2.** *Let  $I \subseteq \mathbb{R}$  be any interval. Then no quadrature formula with  $n$  distinct nodes in  $I$  can have order of accuracy  $2n$  or greater.*

*Proof.* Let  $\{x_1, \dots, x_n\} \subseteq I$  be any set of  $n$  distinct points, and let  $\{w_1, \dots, w_n\}$  be any set of weights. Let  $f$  be the degree- $2n$  polynomial  $f(x) := \prod_{j=1}^n (x - x_j)^2$ , i.e. the square of the nodal polynomial. Then

$$\int_I f(x) \, dx > 0 = \sum_{j=1}^n w_j f(x_j),$$

since  $f$  vanishes at each node  $x_j$ . Hence, the quadrature formula is not exact for polynomials of degree  $2n$ .  $\square$

The first, simplest, quadrature formulas to consider are those in which the nodes form an equally-spaced discrete set of points in  $[a, b]$ . Many of these quadrature formulas may be familiar from high-school mathematics.

**Definition 9.3 (Midpoint rule).** The *midpoint quadrature formula* has the single node  $x_1 := \frac{b-a}{2}$  and the single weight  $w_1 := \rho(x_1)|b-a|$ . That is, it is the approximation

$$\int_a^b f(x) \rho(x) \, dx \approx I_1(f) := f\left(\frac{b-a}{2}\right) \rho\left(\frac{b-a}{2}\right) |b-a|.$$

Another viewpoint on the midpoint rule is that it is the approximation of the integrand  $f$  by the constant function with value  $f\left(\frac{b-a}{2}\right)$ . The next quadrature formula, on the other hand, amounts to the approximation of  $f$  by the affine function

$$x \mapsto f(a) + \frac{x-a}{b-a}(f(b) - f(a))$$

that equals  $f(a)$  at  $a$  and  $f(b)$  at  $b$ .

**Definition 9.4 (Trapezoidal rule).** The *midpoint quadrature formula* has the nodes  $x_1 := a$  and  $x_2 := b$  and the weights  $w_1 := \rho(a)\frac{|b-a|}{2}$  and  $w_2 := \rho(b)\frac{|b-a|}{2}$ . That is, it is the approximation

$$\int_a^b f(x) \rho(x) \, dx \approx I_2(f) := (f(a)\rho(a) + f(b)\rho(b)) \frac{|b-a|}{2}.$$

Recall the definition of the Lagrange interpolation polynomial  $L$  for a set of nodes and values from Definition 8.15. The midpoint and trapezoidal quadrature formulas amount to approximating  $f$  by a Lagrange interpolation polynomial  $L$  of degree 0 or 1 and hence approximating  $\int_a^b f(x) \, dx$  by  $\int_a^b L(x) \, dx$ . The general such construction is the following:

**Definition 9.5** (Newton–Cotes formula). Consider  $K + 1$  equally-spaced points

$$a = x_0 < x_1 = x_0 + h < x_2 = x_0 + 2h < \cdots < x_K = b,$$

where  $h = \frac{1}{K}$ . The *closed Newton–Cotes quadrature formula* is the quadrature formula that arises from approximating  $f$  by the Lagrange interpolating polynomial  $L$  that runs through the points  $(x_k, f(x_k))_{k=0}^K$ ; the *open Newton–Cotes quadrature formula* is the quadrature formula that arises from approximating  $f$  by the Lagrange interpolating polynomial  $L$  that runs through the points  $(x_k, f(x_k))_{k=1}^{K-1}$ .

**Proposition 9.6.** The weights for the closed Newton–Cotes quadrature formula are given by

$$w_j = \int_a^b \ell_j(x) dx.$$

*Proof.* Let  $L$  be as in the definition of the Newton–Cotes rule. Then

$$\begin{aligned} \int_a^b f(x) dx &\approx \int_a^b L(x) dx \\ &= \int_a^b \sum_{j=0}^K f(x_j) \ell_j(x) dx \\ &= \sum_{j=0}^K f(x_j) \int_a^b \ell_j(x) dx \end{aligned}$$

as claimed. □

The midpoint rule is the open Newton–Cotes quadrature formula on three points; the trapezoidal rule is the closed Newton–Cotes quadrature formula on two points.

The quality of Newton–Cotes quadrature formulas can be very poor, essentially because Runge’s phenomenon can make the quality of the approximation  $f \approx L$  very poor. The weakness of Newton–Cotes quadrature can be seen another way: it is possible for the Newton–Cotes weights to be negative, and any quadrature formula with weights of both signs is prone to errors. For example, the positive-definite function  $f$  that linearly interpolates the values

$$??? \text{ at } x_k$$

has  $Q(f) = 0$ . **FINISH ME!!!**

## 9.2 Gaussian Quadrature

Gaussian quadrature is a powerful method for numerical integration in which both the nodes and the weights are chosen so as to maximize the order of accuracy of the quadrature formula. Remarkably, by the correct choice of  $n$  nodes and weights, the quadrature formula can be made accurate for all polynomials of degree at most  $2n - 1$ . Moreover, the weights in this quadrature formula are all positive, and so the quadrature formula is stable even for high  $n$ .

See [108, Ch. 37].

The objective is to approximate a definite integral

$$\int_a^b f(x)w(x) \, dx,$$

where  $w: [a, b] \rightarrow (0, +\infty)$  is a fixed *weight function*, by a finite sum

$$I_n(f) := \sum_{j=1}^n w_j f(x_j),$$

where the *nodes*  $x_1, \dots, x_n$  and *weights*  $w_1, \dots, w_n$  will be chosen appropriately.

Let  $\{q_0, q_1, \dots\}$  be a system of orthogonal polynomials with respect to the weight function  $w$ . That is,

$$\int_a^b f(x)q_n(x)w(x) \, dx = 0$$

whenever  $f$  is a polynomial of degree at most  $n-1$ . Let the nodes  $x_1, \dots, x_n$  be the zeros of  $q_n$ ; by Theorem 8.14,  $q_n$  has  $n$  distinct roots in  $[a, b]$ . Define the associated weights by

$$w_j := \frac{a_n}{a_{n-1}} \frac{\int_a^b q_{n-1}(x)^2 w(x) \, dx}{q_n'(x_j) q_{n-1}(x_j)},$$

where  $a_k$  is the coefficient of  $x^k$  in  $q_k(x)$ .

Orthogonal polynomials for quadrature formulas can be found in Abramowitz & Stegun [1, §25.4]

1. The nodes actually determine the weights as above.
2. Those weights are positive. DONE
3. Order of accuracy  $2n-1$ . DONE
4. Error estimate [95, Th. 3.6.24]: for  $f \in \mathcal{C}^{2n}$ ,

$$\int_a^b f(x)\rho(x) \, dx - Q_K(f) = \frac{f^{(2n)}(\xi)}{(2n)!} \|p_n\|_\rho^2$$

**Definition 9.7.** The *n-point Gauss–Legendre quadrature formula* is the quadrature formula with nodes  $\{x_1, \dots, x_n\}$  given by the zeros of the Legendre polynomial  $q_{n+1}$ . The nodes are sometimes called *Gauss points*.

**Theorem 9.8.** The *n-point Gauss quadrature formula* has order of accuracy exactly  $2n-1$ , and no quadrature formula has order of accuracy higher than this.

*Proof.* Lemma 9.2 shows that no quadrature formula can have order of accuracy greater than  $2n-1$ .

On the other hand, suppose that  $p$  is any polynomial of degree at most  $2n-1$ . Factor this polynomial as

$$p(x) = g(x)q_{n+1}(x) + r(x),$$

where  $g$  is a polynomial of degree at most  $n-1$ , and the remainder  $r$  is also a polynomial of degree at most  $n-1$ . Since  $q_{n+1}$  is orthogonal to all polynomials

of degree at most  $n$ ,  $\int_a^b g q_{n+1} d\mu = 0$ . However, since  $g(x_j)q_{n+1}(x_j) = 0$  for each node  $x_j$ ,

$$I_n(gq_{n+1}) = \sum_{j=1}^n w_j g(x_j) q_{n+1}(x_j) = 0.$$

Since  $\int_a^b \cdot d\mu$  and  $Q_n(\cdot)$  are both linear operators,

$$\int_a^b f d\mu = \int_a^b r d\mu \text{ and } Q_n(f) = Q_n(r).$$

Since  $r$  is of degree at most  $n-1$ ,  $\int_a^b r d\mu = I_n(r)$ , and so  $\int_a^b f d\mu = Q_n(f)$ , as claimed.  $\square$

**Theorem 9.9.** *The Gauss weights are given by*

$$w_j = \frac{a_n}{a_{n-1}} \frac{\int_a^b q_{n-1}(x)^2 w(x) dx}{q'_n(x_j) q_{n-1}(x_j)},$$

where  $a_k$  is the coefficient of  $x^k$  in  $q_k(x)$ .

*Proof.* Suppose that  $p$  is any polynomial of degree at most  $2n-1$ . Factor this polynomial as

$$p(x) = g(x)q_{n+1}(x) + r(x),$$

where  $g$  is a polynomial of degree at most  $n-1$ , and the remainder  $r$  is also a polynomial of degree at most  $n-1$ . Using Lagrange basis polynomials, write  $r = \sum_{i=1}^n r(x_i)\ell_i$ , so that

$$\int_a^b r d\mu = \sum_{i=1}^n r(x_i) \int_a^b \ell_i d\mu.$$

Since the Gauss quadrature formula is exact for  $r$ , it follows that the Gauss weights satisfy

$$w_i = \int_a^b \ell_i d\mu$$

...

...

...

**FINISH ME!!!**

$\square$

**Theorem 9.10.** *The weights of the Gauss quadrature formula are all positive.*

*Proof.* Fix  $1 \leq i \leq n$  and consider the polynomial

$$p(x) := \prod_{\substack{1 \leq j \leq n \\ j \neq i}} (x - x_j)^2$$

i.e. the square of the nodal polynomial, divided by  $(x - x_i)^2$ . Since the degree of  $p$  is strictly less than  $2n-1$ , the Gauss quadrature formula is exact, and since  $p$  vanishes at every node other than  $x_i$ , it follows that

$$\int_I p d\mu = \sum_{j=1}^n w_j p(x_j) = w_i p(x_i).$$

Since  $\mu$  is a non-negative measure,  $p \geq 0$  everywhere, and  $p(x_i) > 0$ , it follows that  $w_i > 0$ .  $\square$

### 9.3 Clenshaw–Curtis / Fejér Quadrature

Despite its optimal degree of polynomial exactness, Gaussian quadrature has some major drawbacks in practice. One principal drawback is that, by Theorem 8.14, the Gaussian quadrature nodes are never *nested* — that is, if one wishes to increase the accuracy of the numerical integral by passing from using, say,  $n$  nodes to  $2n$  nodes, then none of the first  $n$  nodes will be re-used. If evaluations of the integrand are computationally expensive, then this lack of nesting is a serious concern. Another drawback of Gaussian quadrature on  $n$  nodes is the cost of computing the weights, which is  $O(n^2)$ . By contrast, the Clenshaw–Curtis quadrature rules [19] (although in fact discovered thirty years previously by Fejér [29]) are nested quadrature rules, with accuracy comparable to Gaussian quadrature in many circumstances, and with weights that can be computed with cost  $O(n \log n)$ .

$$f(\cos \theta) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(k\theta)$$

where

$$a_k = \frac{2}{\pi} \int_0^{\pi} f(\cos \theta) \cos(k\theta) d\theta$$

The cosine series expansion of  $f$  is also a Chebyshev polynomial expansion of  $f$ , since by construction  $T_k(\cos \theta) = \cos(k\theta)$ :

$$f(x) = \frac{a_0}{2} T_0(x) + \sum_{k=1}^{\infty} a_k T_k(x)$$

$$\begin{aligned} \int_{-1}^1 f(x) dx &= \int_0^{\pi} f(\cos \theta) \sin \theta d\theta \\ &= a_0 + \sum_{k=1}^{\infty} \frac{2a_{2k}}{1 - (2k)^2} \end{aligned}$$

### 9.4 Quadrature in Multiple Dimensions

Having established quadrature formulæ for integrals with a one-dimensional domain of integration, the next agenda is to produce quadrature formulas for multi-dimensional (i.e. iterated) integrals of the form

$$\int_{\prod_{j=1}^d [a_j, b_j]} f(x) dx = \int_{a_d}^{b_d} \dots \int_{a_1}^{b_1} f(x_1, \dots, x_d) dx_1 \dots dx_d.$$

**Tensor Product Quadrature Formulæ.** The first, obvious, strategy to try is to treat  $d$ -dimensional integration as a succession of  $d$  one-dimensional integrals and apply our favourite one-dimensional quadrature formula  $d$  times. This is the idea underlying *tensor product quadrature formulas*, and it has one major flaw: if the one-dimensional quadrature formula uses  $N$  nodes, then the tensor product rule uses  $N^d$  nodes, which very rapidly leads to an impractically large number of integrand evaluations for even moderately large values of  $N$  and  $d$ . In



general, when the one-dimensional quadrature formula uses  $N$  nodes, the error for an integrand in  $\mathcal{C}^r$  using a tensor product rule is  $O(N^{-r/d})$ .

**Sparse Quadrature Formulæ.** The curse of dimension, which quickly renders tensor product quadrature formulae impractical in high dimension, spurs the consideration of *sparse quadrature formulas*, in which far fewer than  $N^d$  nodes are used, at the cost of some accuracy in the quadrature formula: in practice, we are willing to pay the price of loss of accuracy in order to get any answer at all! One example of a popular sparse quadrature rule is the recursive construction of *Smolyak sparse grids*, which is particularly useful when combined with a nested one-dimensional quadrature rule such as the Clenshaw–Curtis rule.

Assume that we are given, for each  $\ell \in \mathbb{N}$ , we are given a one-dimensional quadrature formula  $Q_\ell^{(1)}$ . The formula for Smolyak quadrature in dimension  $d \in \mathbb{N}$  at level  $\ell \in \mathbb{N}$  is defined in terms of the lower-dimensional quadrature formulae by

$$Q_\ell^{(d)}(f) := \left( \sum_{i=1}^{\ell} \left( Q_i^{(1)} - Q_{i-1}^{(1)} \right) \otimes Q_{\ell-i+1}^{(d-1)} \right) (f)$$

This formula takes a little getting used to, and it helps to first consider the case  $d = 2$  and a few small values of  $\ell$ . First, for  $\ell = 1$ , Smolyak’s rule is the quadrature formula

$$Q_1^{(2)} = Q_1^{(1)} \otimes Q_1^{(1)},$$

i.e. the full tensor product of the one-dimensional quadrature formula  $Q_1^{(1)}$  with itself. For the next level,  $\ell = 2$ , Smolyak’s rule is

$$\begin{aligned} Q_2^{(2)} &= \sum_{i=1}^2 \left( Q_i^{(1)} - Q_{i-1}^{(1)} \right) \otimes Q_{\ell-i+1}^{(1)} \\ &= Q_1^{(1)} \otimes Q_2^{(1)} + \left( Q_2^{(1)} - Q_1^{(1)} \right) \otimes Q_1^{(1)} \\ &= Q_1^{(1)} \otimes Q_2^{(1)} + Q_2^{(1)} \otimes Q_1^{(1)} - Q_1^{(1)} \otimes Q_1^{(1)}. \end{aligned}$$

The “ $-Q_1^{(1)} \otimes Q_1^{(1)}$ ” term is included to avoid double counting. See Figure 9.1 for an illustration of nodes of the Smolyak construction in the case that the one-dimensional quadrature formula  $Q_\ell^{(d)}$  has  $2^\ell - 1$  equally-spaced nodes.

In general, when the one-dimensional quadrature formula at level  $\ell$  uses  $N_\ell$  nodes, the quadrature error for an integrand in  $\mathcal{C}^r$  using Smolyak recursion is  $O(N_\ell^{-r} (\log N_\ell)^{(d-1)(r+1)})$ .

**Remark 9.11.** The right Sobolev space for studying sparse grids, since we need pointwise evaluation, is  $H_{\text{mix}}^1$ , in which functions are weakly differentiable in each coordinate direction.

## 9.5 Monte Carlo Methods

As seen above, tensor product quadrature formulas suffer from the curse of dimensionality: they require exponentially many evaluations of the integrand as a function of the dimension of the integration domain. Sparse grid constructions

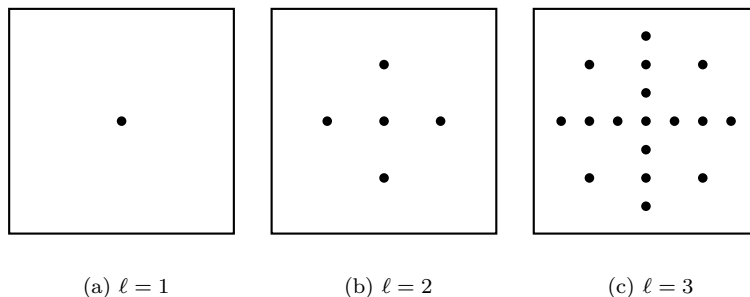


Figure 9.1: Illustration of the nodes of the 2-dimensional Smolyak sparse quadrature formulas  $Q_\ell^{(2)}$  for levels  $\ell = 1, 2, 3$ , in the case that the 1-dimensional quadrature formula  $Q_\ell^{(1)}$  has  $2^\ell - 1$  equally-spaced nodes in the interior of the domain of integration, i.e. is an open Newton–Cotes formula.

only partially alleviate this problem. Remarkably, however, the curse of dimensionality can be entirely circumvented by resorting to random sampling of the integration domain — provided, of course, that it is possible to draw samples from the measure against which the integrand is to be integrated.

Monte Carlo methods are, in essence, an application of the law of large numbers (LLN). Recall that the LLN states that if  $X^{(1)}, X^{(2)}, \dots$  is a sequence of independent samples from a random variable  $X$  with finite expectation  $\mathbb{E}[X]$ , then the sample average

$$\frac{1}{K} \sum_{k=1}^K X^{(k)}$$

converges to  $\mathbb{E}[X]$  as  $K \rightarrow \infty$ . The weak LLN states that the mode of convergence is convergence in probability:

$$\text{for all } \varepsilon > 0, \quad \lim_{K \rightarrow \infty} \mathbb{P} \left[ \left| \frac{1}{K} \sum_{k=1}^K X^{(k)} - \mathbb{E}[X] \right| > \varepsilon \right] = 0;$$

the strong LLN (which is harder to prove than the weak LLN) states that the mode of convergence is actually almost sure:

$$\mathbb{P} \left[ \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=1}^K X^{(k)} = \mathbb{E}[X] \right] = 1.$$

#### ‘Vanilla’ Monte Carlo.

**Theorem 9.12** (Birkhoff–Khinchin ergodic theorem). *Let  $T: \Theta \rightarrow \Theta$  be a measure-preserving map of a probability space  $(\Theta, \mathcal{F}, \mu)$ , and let  $f \in L^1(\Theta, \mu; \mathbb{R})$ . Then, for  $\mu$ -almost every  $\theta \in \Theta$ ,*

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} f(T^k \theta) = \mathbb{E}_\mu[f | \mathcal{G}_T],$$

where  $\mathcal{G}_T$  is the  $\sigma$ -algebra of  $T$ -invariant sets. Hence, if  $T$  is ergodic, then

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} f(T^k \theta) = \mathbb{E}_\mu[f] \quad \mu\text{-a.s.}$$

To obtain an error estimate for Monte Carlo integrals, we simply apply Chebyshev's inequality to  $S_K := \frac{1}{K} \sum_{k=1}^K X^{(k)}$ , which has  $\mathbb{E}[S_K] = \mathbb{E}[X]$  and

$$\mathbb{V}[S_K] = \frac{1}{K^2} \sum_{k=1}^K \mathbb{V}[X] = \frac{\mathbb{V}[X]}{K},$$

to obtain that, for any  $t \geq 0$ ,

$$\mathbb{P}[|S_K - \mathbb{E}[X]| \geq t] \leq \frac{\mathbb{V}[X]}{Kt^2}.$$

That is, for any  $\varepsilon \in (0, 1]$ , with probability at least  $1 - \varepsilon$  with respect to the  $K$  Monte Carlo samples, the Monte Carlo average  $S_K$  lies within  $(\mathbb{V}[X]/K\varepsilon)^{1/2}$  of the true expected value  $\mathbb{E}[X]$ . The fact that the error decays like  $K^{-1/2}$ , i.e. slowly, is a major limitation of ‘vanilla’ Monte Carlo methods; it is undesirable to have to quadruple the number of samples to double the accuracy of the approximate integral.

**CDF Inversion** One obvious criticism of Monte Carlo integration as presented above is the accessibility of the measure of integration  $\mu$ . Even leaving aside the sensitive topic of the generation of truly ‘random’ numbers, it is no easy matter to draw random numbers from an arbitrary probability measure on  $\mathbb{R}$ . The uniform measure on an interval may be said to be easily accessible;  $\rho dx$ , for some positive and integrable function  $\rho$ , is not.

$$F_\nu(x) := \int_{(-\infty, x]} d\nu$$

$$X \sim \text{Unif}([0, 1]) \implies F_\nu^{-1}(X) \sim \nu$$

### Importance Sampling

**Markov Chain Monte Carlo** Markov chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from a probability distribution  $\mu$  based on constructing a Markov chain that has  $\mu$  as its equilibrium distribution. The state of the chain after a large number of steps is then used as a sample of  $\mu$ . The quality of the sample improves as a function of the number of steps. Usually it is not hard to construct a Markov chain with the desired properties; the more difficult problem is to determine how many steps are needed to converge to  $\mu$  within an acceptable error.

## 9.6 Pseudo-Random Methods

This chapter concludes with a very brief survey of numerical integration methods that are in fact based upon deterministic sampling, but in such a way as the sample points ‘might as well be’ random.

Niederreiter [71]

**Definition 9.13.** The *discrepancy*:

$$D_N(P) := \sup_{B \in \mathcal{J}} \left| \frac{\#(P \cap B)}{N} - \lambda^d(B) \right|$$

where  $\mathcal{J}$  is the collection of all products of the form  $\prod_{i=1}^d [a_i, b_i]$ , with  $0 \leq a_i < b_i \leq 1$ . The *star-discrepancy*:

$$D_N^*(P) := \sup_{B \in \mathcal{J}^*} \left| \frac{\#(P \cap B)}{N} - \lambda^d(B) \right|$$

where  $\mathcal{J}^*$  is the collection of all products of the form  $\prod_{i=1}^d [0, b_i]$ , with  $0 \leq b_i < 1$ .

**Lemma 9.14.**

$$D_N^* \leq D_N \leq 2^d D_N^*.$$

**Definition 9.15.** Let  $f: [0, 1]^d \rightarrow \mathbb{R}$ . If  $J \subseteq [0, 1]^d$  is a subrectangle of  $[0, 1]^d$ , i.e. a  $d$ -fold product of subintervals of  $[0, 1]$ , let  $\Delta_J(f)$  be the sum of the values of  $f$  at the  $2^d$  vertices of  $J$ , with alternating signs at nearest-neighbour vertices. The *Vitali variation* of  $f: [0, 1]^d \rightarrow \mathbb{R}$  is defined to be

$$V^{\text{Vit}}(f) := \sup \left\{ \sum_{J \in \Pi} |\Delta_J(f)| \mid \begin{array}{l} \Pi \text{ is a partition of } [0, 1]^d \text{ into finitely} \\ \text{many non-overlapping subrectangles} \end{array} \right\}$$

For  $1 \leq s \leq d$ , the *Hardy–Krause variation* of  $f$  is defined to be

$$V^{\text{HK}}(f) := \sum_F V_F^{\text{Vit}}(f),$$

where the sum runs over all faces  $F$  of  $[0, 1]^d$  having dimension at most  $s$ .

**Theorem 9.16** (Koksma's inequality). *If  $f: [0, 1] \rightarrow \mathbb{R}$  has bounded (total) variation, then, for any  $\{x_1, \dots, x_N\} \subseteq [0, 1]$ ,*

$$\left| \frac{1}{N} \sum_{i=1}^N f(x_i) - \int_{[0,1]} f(x) dx \right| \leq V(f) D_N^*(x_1, \dots, x_N).$$

**Theorem 9.17** (Koksma–Hlawka Inequality). *Let  $f: [0, 1]^d \rightarrow \mathbb{R}$  have bounded Hardy–Krause variation. Then, for any  $\{x_1, \dots, x_N\} \subseteq [0, 1]^N$ ,*

$$\left| \frac{1}{N} \sum_{i=1}^N f(x_i) - \int_{[0,1]^d} f(x) dx \right| \leq V_{[0,1]^d}(f) D_N^*(x_1, \dots, x_N).$$

*Furthermore, this bound is sharp in the sense that, for every  $\{x_1, \dots, x_N\} \subseteq [0, 1]^N$  and every  $\varepsilon > 0$ , there exists  $f: [0, 1]^d \rightarrow \mathbb{R}$  with  $V(f) = 1$  such that*

$$\left| \frac{1}{N} \sum_{i=1}^N f(x_i) - \int_{[0,1]^d} f(x) dx \right| > D_N^*(x_1, \dots, x_N) - \varepsilon.$$

Halton's sequence: [37]

Sobol' sequence: [91]

## Bibliography

W At Warwick, Monte Carlo integration and related topics are covered in the module [ST407 Monte Carlo Methods](#). See also Robert & Casella [81] for a survey of MC methods in statistics.

Orthogonal polynomials for quadrature formulas can be found in Section 25.4 of Abramowitz & Stegun [1]. Gautschi's general monograph [33] on orthogonal polynomials covers applications to Gaussian quadrature in Section 3.1. The article [107] compares the Gaussian and Clenshaw–Curtis quadrature rules and explains their similar accuracy in many circumstances.

Smolyak recursion was introduced in [90].

## Exercises

**Exercise 9.1.** Determine the weights for the open Newton–Cotes quadrature formula. (Cf. Proposition 9.6.)

**Exercise 9.2** (Takahasi–Mori (tanh–sinh) Quadrature [101]). Consider a definite integral over  $[-1, 1]$  of the form  $\int_{-1}^1 f(x) dx$ . Employ a change of variables  $x = \varphi(t) := \tanh(\frac{\pi}{2} \sinh(t))$  to convert this to an integral over the real line. Let  $h > 0$  and  $K \in \mathbb{N}$ , and approximate this integral over  $\mathbb{R}$  using  $2K + 1$  points equally spaced from  $-Kh$  to  $Kh$  to derive a quadrature rule

$$\int_{-1}^1 f(x) dx \approx Q_{h,K}(f) := \sum_{k=-K}^{k=K} w_k f(x_k),$$

where  $x_k := \tanh(\frac{\pi}{2} \sinh(kh))$ ,  
and  $w_k := \frac{\frac{\pi}{2} h \cosh(kh)}{\cosh^2(\frac{\pi}{2} \sinh(kh))}$ .

How are these nodes distributed in  $[-1, 1]$ ? If  $f$  is bounded, then what rate of decay does  $f \circ \varphi$  have? Hence, why is excluding the nodes  $x_k$  with  $|k| > K$  a reasonable approximation?

DRAFT

## Chapter 10

# Sensitivity Analysis and Model Reduction

Le doute n'est pas un état bien agréable,  
mais l'assurance est un état ridicule.

---

VOLTAIRE

The topic of this chapter is *sensitivity analysis*, which may be broadly understood as understanding how  $f(x_1, \dots, x_n)$  depends upon variations not only in the  $x_i$  individually, but also combined or correlated effects among the  $x_i$

### 10.1 Model Reduction for Linear Models

Suppose that the model mapping inputs  $x \in \mathbb{R}^n$  to outputs  $y = f(x) \in \mathbb{R}^m$  is actually a linear map, and so can be represented by a matrix  $A \in \mathbb{R}^{m \times n}$ . There is essentially only one method for the dimensional reduction of such linear models, the *singular value decomposition* (SVD).

**Theorem 10.1** (Singular value decomposition). *For every matrix  $A \in \mathbb{C}^{m \times n}$ , there exist unitary matrices  $U \in \mathbb{C}^{m \times m}$  and  $V \in \mathbb{C}^{n \times n}$  and a diagonal matrix  $\Sigma \in \mathbb{R}_{\geq 0}^{m \times n}$  such that*

$$A = U\Sigma V^*.$$

The columns of  $U$  are called the *left singular vectors* of  $A$ ; the columns of  $V$  are called the *right singular vectors* of  $A$ ; and the diagonal entries of  $\Sigma$  are called the *singular values* of  $A$ . While the singular values are unique, the singular vectors may fail to be. By convention, the singular values and corresponding singular vectors are ordered so that the singular values form a decreasing sequence

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{\min\{m,n\}} \geq 0.$$

Thus, the SVD is a decomposition of  $A$  into a sum of rank-1 operators:

$$A = U\Sigma V^* = \sum_{j=1}^{\min\{m,n\}} \sigma_j u_j \otimes v_j = \sum_{j=1}^{\min\{m,n\}} \sigma_j u_j \langle v_j, \cdot \rangle.$$

The appeal of the SVD is that it is numerically stable, and that it provides optimal low-rank approximation of linear operators: if  $A_k \in \mathbb{R}^{m \times n}$  is defined by

$$A_k := \sum_{j=1}^k \sigma_j u_j \otimes v_j,$$

then  $A_k$  is the optimal rank- $k$  approximation to  $A$  in the sense that

$$\|A - A_k\|_2 = \min \left\{ \|A - X\|_2 \mid \begin{array}{l} X \in \mathbb{R}^{m \times n} \text{ and} \\ \text{rank}(X) \leq k \end{array} \right\},$$

where  $\|\cdot\|_2$  denotes the operator 2-norm on matrices.

Chapter 11 contains an important application of the SVD to the analysis of sample data from random variables, an discrete variant of the Karhunen–Loève expansion. Simply put, when  $A$  is a matrix whose columns are independent samples from some stochastic process (random vector), the SVD of  $A$  is the ideal way to fit a linear structure to those data points. One may consider nonlinear fitting and dimensionality reduction methods in the same way, and this is known as *manifold learning*: see, for instance, the IsoMap algorithm of Tenenbaum & al. [104].

## 10.2 Derivatives

A natural first way to understand the dependence of  $f(x_1, \dots, x_n)$  upon  $x_1, \dots, x_n$  near some nominal point  $x^* = (x_1^*, \dots, x_n^*)$  is to estimate the partial derivatives of  $f$  at  $x^*$ , i.e. to approximate

$$\frac{\partial f}{\partial x_i}(x^*) := \lim_{h \rightarrow 0} \frac{f(x_1^*, \dots, x_i^* + h, \dots, x_n^*) - f(x^*)}{h}.$$

Approximate by e.g.

$$\frac{\partial f}{\partial x_i}(x^*) \approx$$

Ultimately boils down to polynomial approximation  $f \approx p$ ,  $f' \approx p'$ .

## 10.3 McDiarmid Diameters

This section considers an ‘ $L^\infty$ -type’ sensitivity index that measures the sensitivity of a function of  $n$  variables or parameters to variations in those variables/parameters individually.

**Definition 10.2.** The  $i^{\text{th}}$  *McDiarmid subdiameter* of  $f: \prod_{i=1}^n \mathcal{X}_i \rightarrow \mathbb{K}$  is

$$\begin{aligned} \mathcal{D}_i[f] &:= \sup \left\{ |f(x) - f(y)| \mid \begin{array}{l} x, y \in \prod_{i=1}^n \mathcal{X}_i \\ x_j = y_j \text{ for } j \neq i \end{array} \right\} \\ &= \sup \left\{ |f(x_1, \dots, x_i, \dots, x_n) - f(x_1, \dots, x'_i, \dots, x_n)| \mid \begin{array}{l} x_j \in \mathcal{X}_j \text{ for } j = 1, \dots, n \\ x'_i \in \mathcal{X}_i \end{array} \right\}. \end{aligned}$$

The *McDiarmid diameter* of  $f$  is

$$\mathcal{D}[f] := \sqrt{\sum_{i=1}^n \mathcal{D}_i[f]^2}.$$



**Remark 10.3.** Note that although the two definitions of  $\mathcal{D}_i[f]$  given above are obviously mathematically equivalent, they are very different from a computational point of view: the first formulation is ‘obviously’ a constrained optimization problem in  $2n$  variables with  $n - 1$  constraints (i.e. ‘difficult’), whereas the second formulation is ‘obviously’ an unconstrained optimization problem in  $n + 1$  variables (i.e. ‘easy’).

**Lemma 10.4.** For each  $j = 1, \dots, n$ ,  $\mathcal{D}_j[\cdot]$  is a semi-norm on the space of bounded functions  $f: \mathcal{X} \rightarrow \mathbb{K}$ , as is  $\mathcal{D}[\cdot]$ .

*Proof.* Exercise 10.1. □

The McDiarmid subdiameters and diameter are useful not only as sensitivity indices, but also for providing a rigorous upper bound on deviations of a function of independent random variables from its mean value:

**Theorem 10.5** (McDiarmid’s bounded differences inequality). Let  $X = (X_1, \dots, X_n)$  be any random variable with independent components taking values in  $\mathcal{X} = \prod_{i=1}^n \mathcal{X}_i$ , and let  $f: \mathcal{X} \rightarrow \mathbb{R}$  be absolutely integrable with respect to the law of  $X$  and have finite McDiarmid diameter  $\mathcal{D}[f]$ . Then, for any  $t \geq 0$ ,

$$\begin{aligned}\mathbb{P}[f(X) \geq \mathbb{E}[f(X)] + t] &\leq \exp\left(-\frac{2t^2}{\mathcal{D}[f]^2}\right), \\ \mathbb{P}[f(X) \leq \mathbb{E}[f(X)] - t] &\leq \exp\left(-\frac{2t^2}{\mathcal{D}[f]^2}\right), \\ \mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] &\leq 2 \exp\left(-\frac{2t^2}{\mathcal{D}[f]^2}\right).\end{aligned}$$

**Corollary 10.6** (Hoeffding’s inequality). Let  $X = (X_1, \dots, X_n)$  be a random variable with independent components, taking values in the cuboid  $\prod_{i=1}^n [a_i, b_i]$ . Let  $S_n := \frac{1}{n} \sum_{i=1}^n X_i$ . Then, for any  $t \geq 0$ ,

$$\mathbb{P}[S_n - \mathbb{E}[S_n] \geq t] \leq \exp\left(-\frac{2n^2 t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

and similarly for deviations below, and either side, of the mean.

McDiarmid’s and Hoeffding’s inequalities are just two examples of a broad family of inequalities known as *concentration of measure* inequalities. Roughly put, the concentration of measure phenomenon, which was first noticed by Lévy [61], is the fact that a function of a high-dimensional random variable with many independent (or weakly correlated) components has its values overwhelmingly concentrated about the mean (or median). An inequality such as McDiarmid’s provides a rigorous certification criterion: to be sure that  $f(X)$  will deviate above its mean by more than  $t$  with probability no greater than  $\varepsilon \in [0, 1]$ , it suffices to show that

$$\exp\left(-\frac{2t^2}{\mathcal{D}[f]^2}\right) \leq \varepsilon$$

i.e.

$$\mathcal{D}[f] \leq t \sqrt{\frac{2}{\log \varepsilon^{-1}}}.$$

Experimental effort then revolves around determining  $\mathbb{E}[f(X)]$  and  $\mathcal{D}[f]$ ; given those ingredients, the certification criterion is mathematically rigorous. That said, it is unlikely to be the *optimal* rigorous certification criterion, because McDiarmid's inequality is not guaranteed to be sharp. The calculation of optimal probability inequalities is considered in Chapter 14.

To prove McDiarmid's inequality first requires a lemma bounding the moment-generating function of a random variable:

**Lemma 10.7 (Hoeffding's lemma).** *Let  $X$  be a random variable with mean zero taking values in  $[a, b]$ . Then, for  $t \geq 0$ ,*

$$\mathbb{E}[e^{tX}] \leq \exp\left(\frac{t^2(b-a)^2}{8}\right).$$

*Proof.* By the convexity of the exponential function, for each  $x \in [a, b]$ ,

$$e^{tx} \leq \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb}.$$

Therefore, applying the expectation operator,

$$\mathbb{E}[e^{tX}] \leq \frac{b}{b-a}e^{ta} + \frac{a}{b-a}e^{tb} =: e^{\phi(t)}.$$

Observe that  $\phi(0) = 0$ ,  $\phi'(0) = 0$ , and  $\phi''(t) \leq \frac{1}{4}(b-a)^2$ . Hence, since  $\exp$  is an increasing and convex function,

$$\mathbb{E}[e^{tX}] \leq \exp\left(0 + 0t + \frac{(b-a)^2}{4} \frac{t^2}{2}\right) = \exp\left(\frac{t^2(b-a)^2}{8}\right),$$

as claimed.  $\square$

*Proof of McDiarmid's inequality (Theorem 10.5).* The proof uses the properties of conditional expectation outlined in Example 3.14. Let  $\mathcal{F}_i$  be the  $\sigma$ -algebra generated by  $X_1, \dots, X_i$ , and define random variables  $Z_0, \dots, Z_n$  by  $Z_i := \mathbb{E}[f(X)|\mathcal{F}_i]$ . Note that  $Z_0 = \mathbb{E}[f(X)]$  and  $Z_n = f(X)$ . Now consider the conditional increment  $(Z_i - Z_{i-1})|\mathcal{F}_{i-1}$ . First observe that

$$\mathbb{E}[Z_i - Z_{i-1}|\mathcal{F}_{i-1}] = 0,$$

so that the sequence  $(Z_i)_{i \geq 0}$  is a *martingale*. Secondly, observe that

$$L_i \leq Z_i - Z_{i-1}|\mathcal{F}_{i-1} \leq U_i,$$

where

$$L_i := \inf_{\ell} \mathbb{E}[f(X)|\mathcal{F}_{i-1}, X_i = \ell] - \mathbb{E}[f(X)|\mathcal{F}_{i-1}],$$

$$U_i := \sup_u \mathbb{E}[f(X)|\mathcal{F}_{i-1}, X_i = u] - \mathbb{E}[f(X)|\mathcal{F}_{i-1}].$$

Since  $U_i - L_i \leq \mathcal{D}_i[f]$ , Hoeffding's lemma implies that

$$\mathbb{E}\left[e^{s(Z_i - Z_{i-1})} \mid \mathcal{F}_{i-1}\right] \leq e^{s^2 \mathcal{D}_i[f]^2 / 8}.$$

Hence, for any  $s \geq 0$ ,

$$\begin{aligned}
& \mathbb{P}[f(X) - \mathbb{E}[f(X)] \geq t] \\
&= \mathbb{P}[e^{s(f(X) - \mathbb{E}[f(X)])} \geq e^{st}] \\
&\leq e^{-st} \mathbb{E}[e^{s(f(X) - \mathbb{E}[f(X)])}] && \text{(Markov's inequality)} \\
&= e^{-st} \mathbb{E}\left[e^{s \sum_{i=1}^n Z_i - Z_{i-1}}\right] && \text{(telescoping sum)} \\
&= e^{-st} \mathbb{E}\left[\mathbb{E}\left[e^{s \sum_{i=1}^n Z_i - Z_{i-1}} \middle| \mathcal{F}_{n-1}\right]\right] && \text{(tower rule)} \\
&= e^{-st} \mathbb{E}\left[e^{s \sum_{i=1}^{n-1} Z_i - Z_{i-1}} \mathbb{E}\left[e^{s(Z_n - Z_{n-1})} \middle| \mathcal{F}_{n-1}\right]\right] && (Z_0, \dots, Z_{n-1} \text{ are } \mathcal{F}_{n-1}\text{-measurable}) \\
&\leq e^{-st} e^{s^2 \mathcal{D}_n[f]^2 / 8} \mathbb{E}\left[e^{s \sum_{i=1}^{n-1} Z_i - Z_{i-1}}\right]
\end{aligned}$$

by the first part of the proof. Repeating this argument a further  $n - 1$  times shows that

$$\mathbb{P}[f(X) - \mathbb{E}[f(X)] \geq t] \leq \exp\left(-st + \frac{s^2}{8} \mathcal{D}[f]^2\right).$$

The expression on the right-hand side is minimized by  $s = 4t/\mathcal{D}[f]^2$ , which yields the first of McDiarmid's inequalities, and the others follow easily.  $\square$

## 10.4 ANOVA/HDMR Decompositions

The topic of this section is a variance-based decomposition of a function of  $n$  variables that goes by various names such as the *analysis of variance* (ANOVA), the *functional ANOVA*, and the *high-dimensional model representation* (HDMR). As before, let  $(\mathcal{X}_i, \mathcal{F}_i, \mu_i)$  be a probability space for  $i = 1, \dots, n$ , and let  $(\mathcal{X}, \mathcal{F}, \mu)$  be the product space. Write  $\mathcal{N} = \{1, \dots, n\}$ , and consider a ( $\mathcal{F}$ -measurable) function of interest  $f: \mathcal{X} \rightarrow \mathbb{R}$ . Bearing in mind that in practical applications  $n$  may be large ( $10^3$  or more), it is of interest to efficiently identify

- which of the  $x_i$  contribute in the most dominant ways to the variations in  $f(x_1, \dots, x_n)$ ,
- how the effects of multiple  $x_i$  are cooperative or competitive with one another,
- and hence construct a surrogate model for  $f$  that uses a lower-dimensional set of input variables, by using only those that give rise to dominant effects.

The idea is to write  $f(x_1, \dots, x_n)$  as a sum of the form

$$\begin{aligned}
f(x_1, \dots, x_n) &= f_\emptyset + \sum_{i=1}^n f_{\{i\}}(x_i) + \sum_{1 \leq i < j \leq n} f_{\{i,j\}}(x_i, x_j) + \dots \\
&= \sum_{I \subseteq \mathcal{N}} f_I(x_I).
\end{aligned}$$

Experience suggests that 'typical real-world systems'  $f$  exhibit only low-order cooperativity in the effects of the input variables  $x_1, \dots, x_n$ . That is, the terms  $f_I$  with  $|I| \gg 1$  are typically small, and a good approximation of  $f$  is given by, say, a second-order expansion,

$$f(x_1, \dots, x_n) \approx f_\emptyset + \sum_{i=1}^n f_{\{i\}}(x_i) + \sum_{1 \leq i < j \leq n} f_{\{i,j\}}(x_i, x_j).$$

Note, however, that low-order cooperativity does not necessarily imply that there is a small set of significant variables (it is possible that  $f_{\{i\}}$  is large for most  $i \in \{1, \dots, n\}$ ), not does it say anything about the linearity or non-linearity of the input-output relationship. Furthermore, there are many HDMR-type expansions of the form given above; orthogonality criteria can be used to select a particular HDMR representation.

**RS-HDMR / ANOVA.** A long-established decomposition of this type is the *analysis of variance* (ANOVA) or *random sampling HDMR* (RS-HDMR) decomposition. Let

$$f_{\emptyset}(x) := \int_{\mathcal{X}} f \, d\mu,$$

i.e.  $f_{\emptyset}$  is the orthogonal projection of  $f$  onto the one-dimensional space of constant functions, and so it is common to abuse notation and write  $f_{\emptyset} \in \mathbb{R}$ . For  $i = 1, \dots, n$ , let

$$f_{\{i\}}(x) := \int_{\mathcal{X}_1} \dots \int_{\mathcal{X}_{i-1}} \int_{\mathcal{X}_{i+1}} \dots \int_{\mathcal{X}_n} f \, d\mu_1 \dots d\mu_{i-1} d\mu_{i+1} \dots d\mu_n - f_{\emptyset}.$$

Note that  $f_{\{i\}}(x)$  is actually a function of  $x_i$  only, and so it is common to abuse notation and write  $f_{\{i\}}(x_i)$  instead of  $f_{\{i\}}(x)$ ;  $f_{\{i\}}$  is constant with respect to the other  $n - 1$  variables. To take this idea further and capture cooperative effects among two or more  $x_i$ , for  $I \subseteq \mathcal{N} := \{1, \dots, n\}$ , let  $|I|$  denote the cardinality of  $I$  and let  $\sim I$  denote the relative complement  $\mathcal{D} \setminus I$ . For  $I = (i_1, \dots, i_{|I|}) \subseteq \mathcal{N}$  and  $x \in \mathcal{X}$ , define the point  $x_I$  by  $x_I := (x_{i_1}, \dots, x_{i_{|I|}})$ ; similar notation like  $x_{\sim I}$ ,  $\mathcal{X}_I$ ,  $\mu_I$  &c. should hopefully be self-explanatory.

**Definition 10.8.** The *ANOVA decomposition* or *RS-HDMR* of  $f$  is the sum  $f = \sum_{I \subseteq \mathcal{N}} f_I$ , where the functions  $f_I: \mathcal{X} \rightarrow \mathbb{R}$  (or, by abuse of notation,  $f_I: \mathcal{X}_I \rightarrow \mathbb{R}$ ) are defined recursively by

$$\begin{aligned} f_{\emptyset}(x) &:= \int_{\mathcal{X}} f(x) \, d\mu(x) \\ f_I(x_I) &:= \int_{\mathcal{X}_{\sim I}} \left( f(x) - \sum_{J \subsetneq I} f_J(x_J) \right) dx_{\sim I} \\ &= \int_{\mathcal{X}_{\sim I}} f(x) \, dx_{\sim I} - \sum_{J \subsetneq I} f_J(x_J). \end{aligned}$$

**Theorem 10.9 (ANOVA).** Let  $f \in L^2(\mathcal{X}, \mu)$  have variance  $\sigma^2 := \|f - f_{\emptyset}\|_{L^2}^2$ . Then

1. whenever  $i \in I$ ,  $\int_{\mathcal{X}_i} f_I(x) \, d\mu_i(x_i) = 0$ ;
2. whenever  $I \neq J$ ,  $\int f_I(x) f_J(x) \, d\mu(x) = 0$ ;
3.  $\sigma^2 = \sum_{I \subseteq \mathcal{D}} \sigma_I^2$ , where

$$\begin{aligned} \sigma_{\emptyset}^2 &:= 0, \\ \sigma_I^2 &:= \int f_I(x)^2 \, dx. \end{aligned}$$

*Proof.* 1. First suppose that  $I = \{i\}$ . Then

$$\begin{aligned} \int_{\mathcal{X}_i} f_I d\mu_i &= \int_{\mathcal{X}_i} \left( \int_{\mathcal{X}_{\sim I}} f d\mu_{\sim\{i\}} - f_{\emptyset} \right) d\mu_i \\ &= \int_{\mathcal{X}} f d\mu - f_{\emptyset} \\ &= 0. \end{aligned}$$

Now suppose for an induction that  $|I| = k$  and that  $\int_{\mathcal{X}_j} f_J d\mu_j = 0$  for all  $j \in J$  whenever  $|J| < |I|$ . Then

$$\begin{aligned} \int_{\mathcal{X}_i} f_I d\mu_i &= \int_{\mathcal{X}_i} \left( \int_{\mathcal{X}_{\sim I}} f d\mu_{\sim I} - \sum_{\substack{J \subsetneq I \\ i \notin J}} f_J \right) d\mu_i \\ &= \int_{\mathcal{X}_i} \int_{\mathcal{X}_{\sim I}} f d\mu_{\sim I} d\mu_i - \sum_{\substack{J \subsetneq I \\ i \notin J}} \int_{\mathcal{X}_i} f_J d\mu_i \end{aligned}$$

**FINISH ME!!!**

2. **FINISH ME!!!**

3. This follows immediately from the orthogonality relation

$$\int_{\mathcal{X}} f_I(x) f_J(x) d\mu(x) = \langle f_I, f_J \rangle_{L^2(\mu)} = 0 \quad \text{whenever } I \neq J. \quad \square$$

**Cut-HDMR.** In Cut-HDMR, an expansion is performed with respect to a reference point  $\bar{x} \in \mathcal{X}$ :

$$\begin{aligned} f_{\emptyset}(x) &:= f(\bar{x}), \\ f_{\{i\}}(x) &:= f(\bar{x}_1, \dots, \bar{x}_{i-1}, x_i, \bar{x}_{i+1}, \dots, \bar{x}_n) - f_{\emptyset}(x) \\ &\equiv f(x_i, \bar{x}_{\sim\{i\}}) - f_{\emptyset}(x), \\ f_{\{i,j\}}(x) &:= f(\bar{x}_1, \dots, \bar{x}_{i-1}, x_i, \bar{x}_{i+1}, \dots, \bar{x}_{j-1}, x_j, \bar{x}_{j+1}, \dots, \bar{x}_n) - f_{\{i\}}(x) - f_{\{j\}}(x) - f_{\emptyset}(x) \\ &\equiv f(x_{\{i,j\}}, \bar{x}_{\sim\{i,j\}}) - f(x_i, \bar{x}_{\sim\{i\}}) - f(x_j, \bar{x}_{\sim\{j\}}) - f_{\emptyset}(x), \\ f_I(x) &:= f(x_I, \bar{x}_{\sim I}) - \sum_{J \subsetneq I} f_J(x). \end{aligned}$$

Note that a component function  $f_I$  of a Cut-HDMR expansion vanishes at any  $x \in \mathcal{X}$  that has a component in common with  $\bar{x}$ , i.e.

$$f_I(x) = 0 \quad \text{whenever } x_i = \bar{x}_i \text{ for some } i \in I.$$

Hence,

$$f_I(x) f_J(x) = 0 \quad \text{whenever } x_k = \bar{x}_k \text{ for some } k \in I \cup J.$$

Indeed, this orthogonality relation defines the Cut-HDMR expansion.

**HDMR Projectors.** Decomposition of a function  $f$  into an HDMR expansion can be seen as the application of a suitable sequence of orthogonal projection operators, and hence HDMR provides an orthogonal decomposition of  $L^2([0, 1]^n)$ . However, in contrast to orthogonal decompositions such as Fourier and polynomial chaos expansions, in which  $L^2$  is decomposed into an infinite direct sum of finite-dimensional subspaces, the ANOVA/HDMR decomposition is a decomposition of  $L^2$  into a finite direct sum of infinite-dimensional subspaces.

To be more precise, let  $F$  be any vector space of measurable functions from  $\mathcal{X}$  to  $\mathbb{R}$ . The obvious candidate for the projector  $P_\emptyset$  is the orthogonal projection  $P_\emptyset: F \rightarrow F_\emptyset$ , where

$$F_\emptyset := \{f \in F \mid f(x) \equiv a \text{ for some } a \in \mathbb{R} \text{ and all } x \in [0, 1]^n.\}$$

is the space of constant functions. For  $i = 1, \dots, n$ ,  $P_{\{i\}}: F \rightarrow F_{\{i\}}$ , where

$$F_{\{i\}} := \left\{ f \in F \mid f \text{ is independent of } x_j \text{ for } j \neq i \text{ and } \int_0^1 f(x) dx_i = 0 \right\}$$

and, for  $\emptyset \neq I \subseteq \mathcal{N}$ ,  $P_I: F \rightarrow F_I$ , where

$$F_I := \left\{ f \in F \mid f \text{ is independent of } x_j \text{ for } j \notin I \text{ and, for } i \in I, \int_0^1 f(x) dx_i = 0 \right\}.$$

These linear operators  $P_I$  are idempotent, commutative and mutually orthogonal, i.e.

$$P_I P_J f = P_J P_I f = \begin{cases} P_I f, & \text{if } I = J, \\ 0, & \text{if } I \neq J, \end{cases}$$

and form a resolution of the identity

$$\sum_{I \subseteq \mathcal{N}} P_I f = f.$$

Thus, the space of functions  $F$  decomposes as the direct sum  $F = \bigoplus_{I \subseteq \mathcal{N}} F_I$ , and this direct sum is orthogonal when  $F$  is a Hilbert subspace of  $L^2(\mathcal{X}, \mu)$ .

**Sobol' Sensitivity Indices.** The decomposition of the variance given by an HDMR / ANOVA decomposition naturally gives rise to a set of sensitivity indices for ranking the most important input variables and their cooperative effects. An obvious (and naïve) assessment of the relative importance of the variables  $x_I$  is the variance component  $\sigma_I^2$ , or the normalized contribution  $\sigma_I^2/\sigma^2$ . However, this measure neglects the contributions of those  $x_J$  with  $J \subseteq I$ , or those  $x_J$  such that  $J$  has some indices in common with  $I$ . With this in mind, the *Sobol' sensitivity indices* are defined as follows:

**Definition 10.10.** Given an HDMR decomposition of a function  $f$  of  $n$  variables, the *lower and upper Sobol' sensitivity indices* of  $I \subseteq \mathcal{N}$  are, respectively,


$$\underline{\tau}_I^2 := \sum_{J \subseteq I} \sigma_J^2, \text{ and } \bar{\tau}_I^2 := \sum_{J \cap I \neq \emptyset} \sigma_J^2.$$

The *normalized lower and upper Sobol' sensitivity indices* of  $I \subseteq \mathcal{N}$  are, respectively,

$$\underline{s}_I^2 := \underline{\tau}_I^2/\sigma^2, \text{ and } \bar{s}_I^2 := \bar{\tau}_I^2/\sigma^2.$$

Since  $\sum_{I \subseteq \mathcal{N}} \sigma_I^2 = \sigma^2 = \|f - f_\emptyset\|_{L^2}^2$ , it follows immediately that, for each  $I \subseteq \mathcal{N}$ ,

$$0 \leq \underline{s}_I^2 \leq \overline{s}_I^2 \leq 1.$$

Note, however, that while the ANOVA theorem guarantees that  $\sigma^2 = \sum_{I \subseteq \mathcal{D}} \sigma_I^2$ ,  in general Sobol' indices satisfy no such additivity relation:

$$1 \neq \sum_{I \subseteq \mathcal{N}} \underline{s}_I^2 < \sum_{I \subseteq \mathcal{N}} \overline{s}_I^2 \neq 1.$$

## Bibliography

Detailed treatment of the singular value decomposition can be found in any text on (numerical) linear algebra, such as that of Trefethen & Bau [108].

McDiarmid's inequality appears in [65], although the underlying martingale results go back to Hoeffding [39] and Azuma [5]. Ledoux [59] and Ledoux & Talagrand [60] give more general presentations of the concentration-of-measure phenomenon, including geometrical considerations such as isoperimetric inequalities.

In the statistical literature, the analysis of variance (ANOVA) originates with Fisher & Mackenzie [32]. The ANOVA decomposition was generalized by Hoeffding [38] to functions in  $L^2([0, 1]^d)$  for  $d \in \mathbb{N}$ ; for  $d = \infty$ , see Owen [75]. That generalization can easily be applied to  $L^2$  functions on any product domain, and leads to the functional ANOVA of Stone [96]. In the mathematical chemistry literature, the HDMR (with its obvious connections to ANOVA) was popularized by Rabitz & al. [3, 78]. The presentation of ANOVA/HDMR in this chapter draws upon those references and the presentations of Beccacece & Borgonovo [7] and Hooker [40].

Sobol' indices were introduced by Sobol' in [92]. HDMR by Sobol' in [93].

## Exercises

**Exercise 10.1.** Prove Lemma 10.4. That is, show that, for each  $j = 1, \dots, n$ , the McDiarmid subdiameter  $\mathcal{D}_j[\cdot]$  is a semi-norm on the space of bounded functions  $f: \mathcal{X} \rightarrow \mathbb{K}$ , as is the McDiarmid diameter  $\mathcal{D}[\cdot]$ . What are the null-spaces of these semi-norms?

**Exercise 10.2.** Let  $f: [-1, 1]^2 \rightarrow \mathbb{R}$  be a function of two variables. Sketch the vanishing sets of the component functions of  $f$  in a Cut-HDMR expansion through  $\bar{x} = (0, 0)$ . Do the same exercise for  $f: [-1, 1]^3 \rightarrow \mathbb{R}$  and  $\bar{x} = (0, 0, 0)$ , taking particular care with second-order terms like  $f_{\{1,2\}}$ .

DRAFT



# Chapter 11

## Spectral Expansions

The mark of a mature, psychologically healthy mind is indeed the ability to live with uncertainty and ambiguity, but only as much as there really is.

---

JULIAN BAGGINI

This chapter and its sequels consider several *spectral methods* for uncertainty quantification. At their core, these are orthogonal decomposition methods in which a random variable stochastic process (usually the solution of interest) over a probability space  $(\Theta, \mathcal{F}, \mu)$  is expanded with respect to an appropriate orthogonal basis of  $L^2(\Theta, \mu; \mathbb{R})$ . This chapter lays the foundations by considering spectral expansions in general, starting with the *Karhunen–Loève biorthogonal decomposition*, and continuing with orthogonal polynomial bases for  $L^2(\Theta, \mu; \mathbb{R})$  and the resulting *polynomial chaos decompositions*.

### 11.1 Karhunen–Loève Expansions

Fix a compact domain  $\Omega \subseteq \mathbb{R}^d$  (which could be thought of as ‘space’, ‘time’, or a general parameter space) and a probability space  $(\Theta, \mathcal{F}, \mu)$ . The Karhunen–Loève expansion of a stochastic process  $U: \Omega \times \Theta \rightarrow \mathbb{R}$  is a particularly nice spectral decomposition, in that it decomposes  $U$  in a *biorthogonal* fashion, i.e. in terms of components that are both orthogonal over the parameter domain  $\Omega$  and the probability space  $\Theta$ .

To be more precise, consider a stochastic process  $U: \Omega \times \Theta \rightarrow \mathbb{R}$  such that

- for all  $x \in \Omega$ ,  $U(x) \in L^2(\Theta, \mu; \mathbb{R})$ ;
- for all  $x \in \Omega$ ,  $\mathbb{E}_\mu[U(x)] = 0$ ;
- the *covariance function*  $C_U(x, y) := \mathbb{E}_\mu[U(x)U(y)]$  is a well-defined continuous function of  $x, y \in \Omega$ .

**Remark 11.1.** 1. The condition that  $U$  is a zero-mean process is not a serious restriction; if  $U$  is not a zero-mean process, then simply consider  $\tilde{U}$  defined by  $\tilde{U}(x, \theta) := U(x, \theta) - \mathbb{E}_\mu[U(x)]$ .

2. It is common in practice to see the covariance function interpreted as proving some information on the *correlation length* of the process  $U$ . That is,  $C_U(x, y)$  depends only upon  $\|x - y\|$  and, for some function  $g: [0, \infty) \rightarrow [0, \infty)$ ,  $C_U(x, y) = g(\|x - y\|)$ . A typical such  $g$  is  $g(r) = \exp(-r/r_0)$ , and the constant  $r_0$  encodes how similar values of  $U$  at nearby points of  $\Omega$  are expected to be; when the correlation length  $r_0$  is small, the field  $U$  has dissimilar values near to one another, and so is rough; when  $r_0$  is large, the field  $U$  has only similar values near to one another, and so is more smooth.

Define the *covariance operator* of  $U$ , also denoted by  $C_U: L^2(\Omega, dx; \mathbb{R}) \rightarrow L^2(\Omega, dx; \mathbb{R})$  by

$$(C_U f)(x) := \int_{\Omega} C_U(x, y) f(y) dy.$$

Now let  $\{e_n \mid n \in \mathbb{N}\}$  be an orthonormal basis of eigenvectors of  $L^2(\Omega, dx; \mathbb{R})$  with corresponding eigenvalues  $\{\lambda_n \mid n \in \mathbb{N}\}$ , i.e.

$$\int_{\Omega} C_U(x, y) e_n(y) dy = \lambda_n e_n(x)$$

and

$$\int_{\Omega} e_m(x) e_n(x) dx = \delta_{mn}.$$

**Definition 11.2.** Let  $\mathcal{X}$  be a first-countable topological space. A function  $K: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is called a *Mercer kernel* if

1.  $K$  is continuous;
2.  $K$  is symmetric, i.e.  $K(x, x') = K(x', x)$  for all  $x, x' \in \mathcal{X}$ ; and
3.  $K$  is positive semi-definite in the sense that, for all choices of finitely many points  $x_1, \dots, x_n \in \mathcal{X}$ , the *Gram matrix*

$$G := \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & \ddots & \vdots \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{bmatrix}$$

is positive semi-definite, i.e. satisfies  $\xi \cdot G\xi \geq 0$  for all  $\xi \in \mathbb{R}^n$ .

**Theorem 11.3 (Mercer).** Let  $\mathcal{X}$  be a first-countable topological space equipped with a complete Borel measure  $\mu$ . Let  $K: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  be a Mercer kernel. If  $x \mapsto K(x, x)$  lies in  $L^1(\mathcal{X}, \mu; \mathbb{R})$ , then there is an orthonormal basis  $\{e_n\}_{n \in \mathbb{N}}$  of  $L^2(\mathcal{X}, \mu; \mathbb{R})$  consisting of eigenfunctions of the operator

$$f \mapsto \int_{\mathcal{X}} K(\cdot, y) f(y) d\mu(y)$$

with non-negative eigenvalues  $\{\lambda_n\}_{n \in \mathbb{N}}$ . Furthermore, the eigenfunctions corresponding to non-zero eigenvalues are continuous, and

$$K(x, y) = \sum_{n \in \mathbb{N}} \lambda_n e_n(x) e_n(y),$$

and this series converges absolutely, and uniformly over compact subsets of  $\mathcal{X}$ .

**Theorem 11.4** (Karhunen-Loève). *Under the above assumptions on  $U$ , its covariance function and its covariance operator,  $U$  can be written as*

$$U = \sum_{n \in \mathbb{N}} Z_n e_n$$

where the  $\{e_n\}_{n \in \mathbb{N}}$  are orthonormal eigenfunctions of the covariance operator  $C_U$ , the corresponding eigenvalues  $\{\lambda_n\}_{n \in \mathbb{N}}$  are non-negative, the convergence of the series is in  $L^2(\Theta, \mu; \mathbb{R})$  and uniform in  $x \in \Omega$ , with

$$Z_n = \int_{\Omega} U(x) e_n(x) dx.$$

Furthermore, the random variables  $Z_n$  are centred, uncorrelated, and have variance  $\lambda_n$ :

$$\mathbb{E}_{\mu}[Z_n] = 0, \text{ and } \mathbb{E}_{\mu}[Z_m Z_n] = \lambda_n \delta_{mn}.$$

*Proof.* Since it is continuous, the covariance function is a Mercer kernel. Hence, by Mercer's theorem, there is an orthonormal basis  $\{e_n\}_{n \in \mathbb{N}}$  of  $L^2(\Omega, dx; \mathbb{R})$  consisting of eigenfunctions of the covariance operator with non-negative eigenvalues  $\{\lambda_n\}_{n \in \mathbb{N}}$ . In this basis, the covariance function has the representation

$$C_U(x, y) = \sum_{n \in \mathbb{N}} \lambda_n e_n(x) e_n(y).$$

Write the process  $U$  in terms of this basis as

$$U = \sum_{n \in \mathbb{N}} Z_n e_n,$$

where the coefficients  $Z_n$  are random variables given by orthogonal projection:

$$Z_n := \int_{\Omega} U(x) e_n(x) dx.$$

Then

$$\mathbb{E}_{\mu}[Z_n] = \mathbb{E}_{\mu} \left[ \int_{\Omega} U(x) e_n(x) dx \right] = \int_{\Omega} \mathbb{E}[U(x)] e_n(x) dx = 0.$$

and

$$\begin{aligned} \mathbb{E}_{\mu}[Z_m Z_n] &= \mathbb{E}_{\mu} \left[ \int_{\Omega} U(x) e_m(x) dx \int_{\Omega} U(x) e_n(x) dx \right] \\ &= \mathbb{E}_{\mu} \left[ \int_{\Omega} \int_{\Omega} U(x) e_m(x) U(y) e_n(y) dy dx \right] \\ &= \int_{\Omega} \int_{\Omega} \mathbb{E}_{\mu}[U(x) U(y)] e_m(x) e_n(y) dy dx \\ &= \int_{\Omega} \int_{\Omega} C_U(x, y) e_m(x) e_n(y) dy dx \\ &= \int_{\Omega} e_m(x) \int_{\Omega} C_U(x, y) e_n(y) dy dx \\ &= \int_{\Omega} e_m(x) \lambda_n e_n(x) dx \\ &= \lambda_n \delta_{mn}. \end{aligned}$$

Let  $S_N := \sum_{n=1}^N Z_n e_n: \Omega \times \Theta \rightarrow \mathbb{R}$ . Then, for any  $x \in \Omega$ ,

$$\begin{aligned}
& \mathbb{E}_\mu [|U(x) - S_N(x)|^2] \\
&= \mathbb{E}_\mu [U(x)^2] + \mathbb{E}_\mu [S_N(x)^2] - 2\mathbb{E}_\mu [U(x)S_N(x)] \\
&= C_U(x, x) + \mathbb{E}_\mu \left[ \sum_{n=1}^N \sum_{m=1}^N Z_n Z_m e_m(x) e_n(x) \right] - 2\mathbb{E}_\mu \left[ U(x) \sum_{n=1}^N Z_n e_n(x) \right] \\
&= C_U(x, x) + \sum_{n=1}^N \lambda_n e_n(x)^2 - 2\mathbb{E}_\mu \left[ \sum_{n=1}^N \int_{\Omega} U(x) U(y) e_n(y) e_n(x) dy \right] \\
&= C_U(x, x) + \sum_{n=1}^N \lambda_n e_n(x)^2 - 2 \sum_{n=1}^N \int_{\Omega} C_U(x, y) e_n(y) e_n(x) dy \\
&= C_U(x, x) - \sum_{n=1}^N \lambda_n e_n(x)^2 \\
&\rightarrow 0 \text{ as } N \rightarrow \infty, \text{ uniformly in } x, \text{ by Mercer's theorem.} \quad \square
\end{aligned}$$

Among many possible decompositions of a random process, the Karhunen–Loève expansion is optimal in the sense that the mean-square error of any truncation of the expansion after finitely many terms is minimal. However, its utility is limited since the covariance function of the solution process is often not known a priori. Nevertheless, the Karhunen–Loève expansion provides an effective means of representing *input* random processes when their covariance structure is known, and provides a simple method for sampling Gaussian measures on Hilbert spaces, which is a necessary step in the implementation of the methods outlined in Chapter 6.

**Example 11.5.** Suppose that  $C: \mathcal{H} \rightarrow \mathcal{H}$  is a self-adjoint, positive-definite, nuclear operator on a Hilbert space  $\mathcal{H}$  and let  $m \in \mathcal{H}$ . Let  $(\lambda_k, e_k)_{k \in \mathbb{N}}$  be a sequence of orthonormal eigenpairs for  $C$ , ordered by decreasing eigenvalue  $\lambda_k$ . Let  $\xi_1, \xi_2, \dots$  be IID samples from  $\mathcal{N}(0, 1)$ . Then, by the Karhunen–Loève theorem,

$$X := m + \sum_{k=1}^{\infty} \sqrt{\lambda_k} \xi_k e_k$$

is an  $\mathcal{H}$ -valued random variable with distribution  $\mathcal{N}(m, C)$ . Therefore, a finite sum of the form  $m + \sum_{k=1}^K \sqrt{\lambda_k} \xi_k e_k$  for large  $K$  is a reasonable approximation to a draw from  $\mathcal{N}(m, C)$ ; this is the procedure used to generate the sample paths in Figure 11.1.

**Definition 11.6.** A *principal component analysis* of an  $\mathbb{R}^N$ -valued random vector  $U$  is the Karhunen–Loève expansion of  $U$  seen as a stochastic process  $U: \{1, \dots, N\} \times \Omega \rightarrow \mathbb{R}$ . It is also known as the *discrete Karhunen–Loève transform*, the *Hotelling transform*, and the *proper orthogonal decomposition*.

Principal component analysis is often applied to sample data, and is intimately related to the singular value decomposition:

**Example 11.7.** Let  $X \in \mathbb{R}^{N \times M}$  be a matrix whose columns are  $M$  independent and identically distributed samples from some probability measure on  $\mathbb{R}^N$ ,

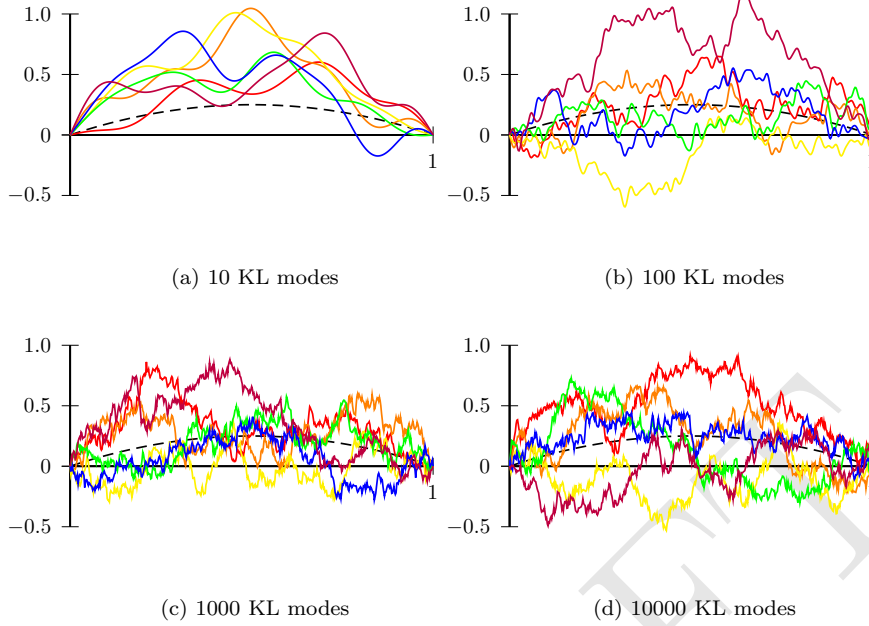


Figure 11.1: Sample paths of the Gaussian distribution on  $H_0^1([0, 1])$  that has mean path  $m(x) = x(1 - x)$  and covariance operator  $(-\frac{d^2}{dx^2})^{-1}$ . Along with the mean path (dashed), six sample paths are shown for Karhunen-Loève expansions using 10, 100, 1000, and 10000 terms.

and assume without loss of generality that the samples have mean zero. The empirical covariance matrix of the samples is

$$\hat{C} := \frac{1}{M^2} X X^\top.$$

The eigenvalues  $\lambda_n$  and eigenfunctions  $e_n$  of the Karhunen-Loève expansion are just the eigenvalues and eigenvectors of this matrix  $\hat{C}$ . Let  $\Lambda \in \mathbb{R}^{N \times N}$  be the diagonal matrix of the eigenvalues  $\lambda_n$  (which are non-negative, and are assumed to be in decreasing order) and  $E \in \mathbb{R}^{N \times N}$  the matrix of corresponding orthonormal eigenvectors, so that  $\hat{C}$  diagonalizes as

$$\hat{C} = E \Lambda E^\top.$$

The principal component transform of the data  $X$  is  $W := E^\top X$ ; this is an orthogonal transformation of  $\mathbb{R}^N$  that transforms  $X$  to a new coordinate system in which the greatest component-wise variance comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

On the other hand, taking the singular value decomposition of the data (normalized by the number of samples) yields

$$\frac{1}{M} X = U \Sigma V^\top,$$

where  $U \in \mathbb{R}^{N \times N}$  and  $V \in \mathbb{R}^{M \times M}$  are orthogonal and  $\Sigma \in \mathbb{R}^{N \times M}$  is diagonal with decreasing non-negative diagonal entries (the singular values of  $\frac{1}{M}X$ ). Then

$$\hat{C} = U\Sigma V^\top (U\Sigma V^\top)^\top = U\Sigma V^\top V\Sigma^\top U^\top = U\Sigma^2 U^\top.$$

from which we see that  $U = E$  and  $\Sigma^2 = \Lambda$ . This is just another instance of the well-known relation that, for any matrix  $A$ , the eigenvalues of  $AA^*$  are the singular values of  $A$  and the right eigenvectors of  $AA^*$  are the left singular vectors of  $A$ ; however, in this context, it also provides an alternative way to compute the principal component transform.

In fact, performing principal component analysis via the singular value decomposition is numerically preferable to forming and then diagonalizing the covariance matrix, since the formation of  $XX^\top$  can cause a disastrous loss of precision; the classic example of this phenomenon is the Lauchli matrix

$$\begin{bmatrix} 1 & \varepsilon & 0 & 0 \\ 1 & 0 & \varepsilon & 0 \\ 1 & 0 & 0 & \varepsilon \end{bmatrix} \quad (0 < \varepsilon \ll 1),$$

for which taking the singular value decomposition is stable, but forming and diagonalizing  $XX^\top$  is unstable.

## 11.2 Wiener–Hermite Polynomial Chaos

The next section will cover polynomial chaos (PC) expansions in greater generality, and this section serves as an introductory prelude. In this, the classical and notationally simplest setting, we consider expansions of a real-valued random variable  $U$  with respect to a single standard Gaussian random variable  $\Xi$ , using appropriate orthogonal polynomials of  $\Xi$ , i.e. the Hermite polynomials. This setting was pioneered by Norbert Wiener, and so it is known as the Wiener–Hermite polynomial chaos.

To this end, let  $\Xi \sim \mathcal{N}(0, 1) =: \gamma$  be a standard Gaussian random variable. The probability density function  $\rho_\Xi: \mathbb{R} \rightarrow \mathbb{R}$  of  $\Xi$  with respect to Lebesgue measure on  $\mathbb{R}$  is

$$\frac{d\gamma}{dx}(\xi) = \rho_\Xi(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\xi^2}{2}\right).$$

Now let  $\text{He}_n(\xi) \in \mathbb{R}[\xi]$ , for  $n \in \mathbb{N}_0$ , be the *Hermite polynomials*, which are a system of orthogonal polynomials for the standard Gaussian measure  $\gamma$ .

**Lemma 11.8.** *If  $p(x) = \sum_{n \in \mathbb{N}_0} p_n x^n$  is any polynomial or convergent power series, then  $\mathbb{E}[p(\Xi)\text{He}_m(\Xi)] = m!p_m$ .*

*Proof.* **TO DO!** □

By the Stone–Weierstrass theorem and the approximability of  $L^2$  functions by continuous ones, the Hermite polynomials form a complete orthogonal basis of the Hilbert space  $L^2(\mathbb{R}, \gamma; \mathbb{R})$  with the inner product

$$\langle U, V \rangle := \mathbb{E}[U(\Xi)V(\Xi)] \equiv \int_{\mathbb{R}} U(\xi)V(\xi)\rho_\Xi(\xi) d\xi.$$

Let us further extend the  $\langle \cdot, \cdot \rangle$  notation for inner products and write  $\langle \cdot \rangle$  for expectation with respect to the distribution  $\gamma$  of  $\Xi$ . So, for example, the orthogonality relation for the Hermite polynomials reads  $\langle \text{He}_m \text{He}_n \rangle = n! \delta_{mn}$ .

**Definition 11.9.** Let  $U \in L^2(\mathbb{R}, \gamma; \mathbb{R})$  be a square-integrable real-valued random variable. The *Wiener–Hermite polynomial chaos expansion* of  $U$  with respect to the standard Gaussian  $\Xi$  is the expansion of  $U$  in the orthogonal basis  $\{\text{He}_n\}_{n \in \mathbb{N}_0}$ , i.e.

$$U = \sum_{n \in \mathbb{N}_0} u_n \text{He}_n(\Xi)$$

with scalar *Wiener–Hermite polynomial chaos coefficients*  $\{u_n\}_{n \in \mathbb{N}_0} \subseteq \mathbb{R}$  given by

$$u_n = \frac{\langle U \text{He}_n \rangle}{\langle \text{He}_n^2 \rangle} = \frac{1}{n! \sqrt{2\pi}} \int_{-\infty}^{\infty} U(x) \text{He}_n(x) e^{-x^2/2} dx.$$

**Remark 11.10.** From the perspective of sampling of random variables, this means that if we wish to draw a sample from the distribution of  $U$ , it is enough to draw a sample  $\xi$  from the standard normal distribution and then evaluate the series  $\sum_{n \in \mathbb{N}_0} u_n \text{He}_n(\xi)$  at that  $\xi$ .

Note that, in particular, since  $\text{He}_0 \equiv 1$ ,

$$\mathbb{E}[U] = \langle \text{He}_0, U \rangle = \sum_{n \in \mathbb{N}_0} u_n \langle \text{He}_0, \text{He}_n \rangle = u_0,$$

so the expected value of  $U$  is simply its 0<sup>th</sup> PC coefficient. Similarly, its variance is a weighted sum of the squares of its PC coefficients:

$$\begin{aligned} \mathbb{V}[U] &= \mathbb{E}[|U - \mathbb{E}[U]|^2] \\ &= \mathbb{E}\left[\left|\sum_{n \in \mathbb{N}} u_n \text{He}_n\right|^2\right] && \text{since } \mathbb{E}[U] = u_0 \\ &= \sum_{m, n \in \mathbb{N}} u_m u_n \langle \text{He}_m, \text{He}_n \rangle \\ &= \sum_{n \in \mathbb{N}} u_n^2 \langle \text{He}_n^2 \rangle && \text{by Hermitian orthogonality.} \end{aligned}$$

**Example 11.11.** Let  $X \sim \mathcal{N}(\mu, \sigma^2)$  be a real-valued Gaussian random variable with mean  $m \in \mathbb{R}$  and variance  $\sigma^2 \geq 0$ . Let  $Y := e^X$ ; since  $\log Y$  is normally distributed, the non-negative-valued random variable  $Y$  is said to be a *log-normal random variable*. As usual, let  $\Xi \sim \mathcal{N}(0, 1)$  be the standard Gaussian random variable; clearly  $X \stackrel{\mathcal{L}}{=} \mu + \sigma \Xi$  and  $Y \stackrel{\mathcal{L}}{=} e^\mu e^{\sigma \Xi}$ . The Wiener–Hermite

expansion of  $Y$  as  $\sum_{k \in \mathbb{N}_0} y_k \text{He}_k(\Xi)$  has coefficients

$$\begin{aligned} y_k &= \frac{\mathbb{E}[e^\mu e^{\sigma \Xi} \text{He}_k(\Xi)]}{\mathbb{E}[\text{He}_k(\Xi)^2]} \\ &= \frac{e^\mu}{k!} \mathbb{E}[e^{\sigma \Xi} \text{He}_k(\Xi)] \\ &= \frac{e^\mu}{k!} \mathbb{E}\left[\frac{\sigma^k \Xi^k}{k!} \text{He}_k(\Xi)\right] && \text{by Hermitian orthogonality} \\ &= \frac{e^\mu e^{\sigma^2/2} \sigma^k}{k!} \end{aligned}$$

i.e.

$$Y = e^{\mu + \sigma^2/2} \sum_{k \in \mathbb{N}_0} \frac{\sigma^k}{k!} \text{He}_k(\Xi).$$

From this expansion it can be seen that  $\mathbb{E}[Y] = e^{\mu + \sigma^2/2}$  and

$$\mathbb{V}[Y] = e^{2\mu + \sigma^2} \sum_{k \in \mathbb{N}_0} \left(\frac{\sigma^k}{k!}\right)^2 \langle \text{He}_k^2 \rangle = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1).$$

Of course, in practice, the series expansion  $U = \sum_{n \in \mathbb{N}_0} u_n \text{He}_n(\Xi)$  must be truncated after finitely many terms, and so it is natural to ask about the quality of the approximation

$$U \approx U^P := \sum_{n=0}^P u_n \text{He}_n(\Xi)$$

Since the Hermite polynomials form a complete orthogonal basis for  $L^2(\mathbb{R}, \gamma; \mathbb{R})$ , the standard results about orthogonal approximations in Hilbert spaces apply. In particular, by Corollary 3.19, the truncation error  $U - U^P$  is orthogonal to the space from which  $U^P$  was chosen, i.e.

$$\text{span}\{\text{He}_0, \text{He}_1, \dots, \text{He}_P\},$$

and tends to zero in mean square.

**Lemma 11.12.** *The truncation error  $U - U^P$  is orthogonal to the subspace*

$$\text{span}\{\text{He}_0, \text{He}_1, \dots, \text{He}_P\}$$

*of  $L^2(\mathbb{R}, \rho_\Xi dx; \mathbb{R})$ . Furthermore,  $\lim_{P \rightarrow \infty} U^P = U$  in  $L^2(\mathbb{R}, \gamma; \mathbb{R})$ .*

*Proof.* Let  $V := \sum_{m=0}^P v_m \text{He}_m$  be any element of the subspace of  $L^2(\mathbb{R}, \gamma; \mathbb{R})$  spanned by the Hermite polynomials of degree at most  $P$ . Then

$$\begin{aligned} \langle U - U^P, V \rangle &= \left\langle \left( \sum_{n > P} u_n \text{He}_n \right) \left( \sum_{m=0}^P v_m \text{He}_m \right) \right\rangle \\ &= \left\langle \sum_{\substack{n > P \\ m \in \{0, \dots, P\}}} u_n v_m \text{He}_n \text{He}_m \right\rangle \\ &= \sum_{\substack{n > P \\ m \in \{0, \dots, P\}}} u_n v_m \langle \text{He}_n \text{He}_m \rangle \\ &= 0. \end{aligned}$$



Hence, by Pythagoras' theorem,

$$\|U\|_{L^2} = \|U^P\|_{L^2(\gamma)} + \|U - U^P\|_{L^2(\gamma)},$$

and hence  $\|U - U^P\|_{L^2(\gamma)} \rightarrow 0$  as  $P \rightarrow \infty$ .  $\square$

## 11.3 Generalized PC Expansions

The ideas of polynomial chaos can be generalized well beyond the setting in which the stochastic germ  $\Xi$  is a standard Gaussian random variable, or even a vector  $\Xi = (\Xi_1, \dots, \Xi_d)$  of mutually orthogonal Gaussian random variables.

Let  $\Xi = (\Xi_1, \dots, \Xi_d)$  be an  $\mathbb{R}^d$ -valued random variable with independent (and hence orthogonal) components. As usual, let  $\mathbb{R}[\xi_1, \dots, \xi_d]$  denote the ring of all polynomials in  $\xi_1, \dots, \xi_d$  with real coefficients, and let  $\mathbb{R}[\xi_1, \dots, \xi_d]_{\leq p}$  denote those polynomials of total degree at most  $p \in \mathbb{N}_0$ . Let  $\Gamma_p \subseteq \mathbb{R}[\xi_1, \dots, \xi_d]_{\leq p}$  be a collection of polynomials that are mutually orthogonal and orthogonal to  $\mathbb{R}[\xi_1, \dots, \xi_d]_{\leq p-1}$ , and let  $\tilde{\Gamma}_p := \text{span } \Gamma_p$ . This yields the orthogonal decomposition

$$L^2(\Theta, \mu; \mathbb{R}) = \bigoplus_{p \in \mathbb{N}_0} \tilde{\Gamma}_p.$$

It is important to note that there is a lack of uniqueness in these basis polynomials whenever  $d \geq 2$ : each choice of ordering of multi-indices  $\alpha \in \mathbb{N}_0^d$  can yield a different orthogonal basis of  $L^2$  when the Gram-Schmidt procedure is applied to the monomials  $\xi^\alpha$ .

Note that (as usual, assuming separability) the  $L^2$  space over the product probability space  $(\Theta, \mathcal{F}, \mu)$  is isomorphic to the Hilbert space tensor product of the  $L^2$  spaces over the marginal probability spaces:

$$L^2(\Theta_1 \times \dots \times \Theta_d, \mu_1 \otimes \dots \otimes \mu_d; \mathbb{R}) = \bigotimes_{i=1}^d L^2(\Theta_i, \mu_i; \mathbb{R}),$$

from which we see that an orthogonal system of multivariate polynomials for  $L^2(\Theta, \mu; \mathbb{R})$  can be found by taking products of univariate orthogonal polynomials for the marginal spaces  $L^2(\Theta_i, \mu_i; \mathbb{R})$ . A *generalized polynomial chaos* (gPC) expansion of a random variable or stochastic process  $U$  is simply the expansion of  $U$  with respect to such a complete orthogonal polynomial basis of  $L^2(\Theta, \mu)$ .

**Example 11.13.** Let  $\Xi = (\Xi_1, \Xi_2)$  be such that  $\Xi_1$  and  $\Xi_2$  are independent (and hence orthogonal) and such that  $\Xi_1$  is a standard Gaussian random variable and  $\Xi_2$  is uniformly distributed on  $[-1, 1]$ . Hence, the univariate orthogonal polynomials for  $\Xi_1$  are the Hermite polynomials  $\text{He}_n$  and the univariate orthogonal polynomials for  $\Xi_2$  are the Legendre polynomials  $\text{Le}_n$ . Then a system of

orthogonal polynomials for  $\Xi$  up to total degree 3 is

$$\begin{aligned}\Gamma_0 &= \{1\}, \\ \Gamma_1 &= \{\text{He}_1(\xi_1), \text{Le}_1(\xi_2)\} \\ &= \{\xi_1, \xi_2\}, \\ \Gamma_2 &= \{\text{He}_2(\xi_1), \text{He}_1(\xi_1)\text{Le}_1(\xi_2), \text{Le}_2(\xi_2)\} \\ &= \{\xi_1^2 - 1, \xi_1\xi_2, \tfrac{1}{2}(3\xi_2^2 - 1)\}, \\ \Gamma_3 &= \{\text{He}_3(\xi_1), \text{He}_2(\xi_1)\text{Le}_1(\xi_2), \text{He}_1(\xi_1)\text{Le}_2(\xi_2), \text{Le}_3(\xi_2)\} \\ &= \{\xi_1^3 - 3\xi_1, \xi_1^2\xi_2 - \xi_2, \tfrac{1}{2}(3\xi_1\xi_2^2 - \xi_1), \tfrac{1}{2}(5\xi_2^3 - 3\xi_2)\}.\end{aligned}$$

Rather than have the orthogonal basis polynomials have two indices, one for the degree  $p$  and one within each set  $\Gamma_p$ , it is useful and conventional to order the basis polynomials using a single index  $k \in \mathbb{N}_0$ . It is common in practice to take  $\Psi_0 = 1$  and to have the polynomial degree be (weakly) increasing with respect to the new index  $k$ . So, to continue Example 11.13, one could take

$$\begin{aligned}\Psi_0(\xi) &= 1, \\ \Psi_1(\xi) &= \xi_1, \\ \Psi_2(\xi) &= \xi_2, \\ \Psi_3(\xi) &= \xi_1^2 - 1, \\ \Psi_4(\xi) &= \xi_1\xi_2, \\ \Psi_5(\xi) &= \tfrac{1}{2}(3\xi_2^2 - 1), \\ \Psi_6(\xi) &= \xi_1^3 - 3\xi_1, \\ \Psi_7(\xi) &= \xi_1^2\xi_2 - \xi_2, \\ \Psi_8(\xi) &= \tfrac{1}{2}(3\xi_1\xi_2^2 - \xi_1), \\ \Psi_9(\xi) &= \tfrac{1}{2}(5\xi_2^3 - 3\xi_2).\end{aligned}$$

**Truncation of gPC Expansions.** Suppose that a gPC expansion  $U = \sum_{k \in \mathbb{N}_0} u_k \Psi_k$  is truncated, i.e. we consider

$$U^P = \sum_{k=0}^P u_k \Psi_k.$$

It is an easy exercise to show that the truncation error  $U - U^P$  is orthogonal to  $\text{span}\{\Psi_0, \dots, \Psi_P\}$ . It is also worth considering how many terms there are in such a truncated gPC expansion. Suppose that the stochastic germ  $\Xi$  has dimension  $d$  (i.e. has  $d$  independent components), and we work only with polynomials of total degree at most  $p$ . The total number of coefficients in the truncated expansion  $U^P$  is

$$P + 1 = \frac{(d + p)!}{d!p!}$$

That is, the total number of gPC coefficients that must be calculated grows combinatorially as a function of the number of input random variables and the degree of polynomial approximation. Such rapid growth limits the usefulness of gPC expansions for practical applications where  $d$  and  $p$  are much greater than, say, 10.

**Remark 11.14.** It is possible to adapt the notion of a gPC expansion to the situation of dependent random variables, but there are some complications. In summary, suppose that  $\Xi = (\Xi_1, \dots, \Xi_d)$  taking values in  $\Theta = \Theta_1 \times \dots \times \Theta_d$  has joint law  $\mu$ , which is not necessarily a product measure. Nevertheless, let  $\mu_i$  denote the marginal law of  $\Xi_i$ , i.e.

$$\mu_i(E_i) := \mu(\Theta_1 \times \dots \times \Theta_{i-1} \times E_i \times \Theta_{i+1} \times \dots \times \Theta_d).$$

To simplify matter further, assume that  $\mu$  (resp.  $\mu_i$ ) has Lebesgue density  $\rho$  (resp.  $\rho_i$ ). Now let  $\phi_p^{(i)}(\xi_i) \in \mathbb{R}[\xi_i]$ ,  $p \in \mathbb{N}_0$ , be univariate orthogonal polynomials for  $\mu_i$ . The *chaos function* associated to a multi-index  $\alpha$  defined to be

$$\Psi_\alpha(\xi) := \sqrt{\frac{\rho_1(\xi_1) \dots \rho_d(\xi_d)}{\rho(\xi)}} \phi_{\alpha_1}^{(1)}(\xi_1) \dots \phi_{\alpha_d}^{(d)}(\xi_d).$$

It can be shown that the family  $\{\Psi_\alpha \mid \alpha \in \mathbb{N}_0^d\}$  is a complete orthonormal basis for  $L^2(\Theta, \mu; \mathbb{R})$ , so we have the usual series expansion  $U = \sum_\alpha u_\alpha \Psi_\alpha$ . Note, however, that with the exception of  $\Psi_0 = 1$ , the functions  $\Psi_\alpha$  are not polynomials. Nevertheless, we still have the usual properties that truncation error is orthogonal to the approximation subspace, and

$$\mathbb{E}_\mu[U] = u_0, \quad \mathbb{V}_\mu[U] = \sum_{\alpha \neq 0} u_\alpha^2 \langle \Psi_\alpha^2 \rangle.$$

**Expansions of Random Variables.** Consider a real-valued random variable  $U$ , which we expand in terms of a stochastic germ  $\xi$ . Let  $U^P$  be a truncated GPC expansion of  $U$ :

$$U^P(\xi) = \sum_{k=0}^P u_k \Psi_k(\xi),$$

where the polynomials  $\Psi_k$  are orthogonal with respect to the law of  $\xi$ , and with the usual convention that  $\Psi_0 = 1$ . A first, easy, observation is that

$$\mathbb{E}[U^P] = \langle \Psi_0, U^P \rangle = \sum_{k=0}^P u_k \langle \Psi_0, \Psi_k \rangle = u_0,$$

so the expected value of  $U^P$  is simply its 0<sup>th</sup> GPC coefficient. Similarly, its variance is a weighted sum of the squares of its GPC coefficients:

$$\begin{aligned} \mathbb{E}[|U^P - \mathbb{E}[U^P]|^2] &= \mathbb{E}\left[\left|\sum_{k=1}^P u_k \Psi_k\right|^2\right] \\ &= \sum_{k,\ell=1}^P u_k u_\ell \langle \Psi_k, \Psi_\ell \rangle \\ &= \sum_{k=1}^P u_k^2 \langle \Psi_k^2 \rangle. \end{aligned}$$

**Expansions of Random Vectors.** Similar remarks can be made for a  $\mathbb{R}^d$ -valued random vector  $U$  having truncated GPC expansion

$$U^P(\xi) = \sum_{k=0}^P u_k \Psi_k(\xi),$$

with coefficients  $u_k = (u_k^1, \dots, u_k^d) \in \mathbb{R}^d$  for each  $k \in \{0, \dots, P\}$ . As before,

$$\mathbb{E}[U^P] = \langle \Psi_0, U^P \rangle = \sum_{k=0}^P u_k \langle \Psi_0, \Psi_k \rangle = u_0 \in \mathbb{R}^d$$

and the covariance matrix  $C \in \mathbb{R}^{d \times d}$  of  $U^P$  is given by

$$C = \sum_{k=1}^P u_k u_k^\top \langle \Psi_k^2 \rangle \text{ i.e. } C_{ij} = \sum_{k=1}^P u_k^i u_k^j \langle \Psi_k^2 \rangle.$$

**Expansions of Stochastic Processes.** Consider now a square-integrable stochastic process  $U: \Omega \times \Theta \rightarrow \mathbb{R}$ ; that is, for each  $x \in \Omega$ ,  $U(x, \cdot) \in L^2(\Theta, \mu)$  is a real-valued random variable, and, for each  $\theta \in \Theta$ ,  $U(\cdot, \theta) \in L^2(\Omega, dx)$  is a scalar field on the domain  $\Omega$ . Recall that

$$L^2(\Theta, \mu; \mathbb{R}) \otimes L^2(\Omega, dx; \mathbb{R}) \cong L^2(\Theta \times \Omega, \mu \otimes dx; \mathbb{R}) \cong L^2(\Theta, \mu; L^2(\Omega, dx)).$$

As usual, take  $\{\Psi_k \mid k \in \mathbb{N}_0\}$  to be an orthogonal polynomial basis of  $L^2(\Theta, \mu; \mathbb{R})$ , ordered (weakly) by total degree, with  $\Psi_0 = 1$ . A GPC expansion of the random field  $U$  is an  $L^2$ -convergent expansion of the form

$$U(x, \xi) = \sum_{k \in \mathbb{N}_0} u_k(x) \Psi_k(\xi),$$

which in practice is truncated to

$$U(x, \xi) \approx U^P(x, \xi) = \sum_{k=0}^P u_k(x) \Psi_k(\xi).$$

The functions  $u_k: \Omega \rightarrow \mathbb{R}$  are called the *stochastic modes* of the process  $U$ . The stochastic mode  $u_0: \Omega \rightarrow \mathbb{R}$  is the *mean field* of  $U$ :

$$\mathbb{E}[U(x)] = \mathbb{E}[U^P(x)] = u_0(x).$$

The variance of the field at  $x \in \Omega$  is

$$\mathbb{V}[U(x)] = \sum_{k=1}^{\infty} u_k^2(x) \langle \Psi_k^2 \rangle \approx \sum_{k=1}^P u_k^2(x) \langle \Psi_k^2 \rangle,$$

whereas, for two points  $x, y \in \Omega$ ,

$$\begin{aligned} \mathbb{E}[U(x)U(y)] &= \left\langle \sum_{k \in \mathbb{N}_0} u_k(x) \Psi_k(\xi) \sum_{\ell \in \mathbb{N}_0} u_\ell(y) \Psi_\ell(\xi) \right\rangle \\ &= \sum_{k \in \mathbb{N}_0} u_k(x) u_k(y) \langle \Psi_k^2 \rangle \\ &\approx \sum_{k=0}^P u_k(x) u_k(y) \langle \Psi_k^2 \rangle, \end{aligned}$$

---

**FINISH ME!!!**

Figure 11.1: ...

---



---

**FINISH ME!!!**

Figure 11.2: ...

---

and so

$$\begin{aligned}
 C_U(x, y) &= \sum_{k \in \mathbb{N}} u_k(x) u_k(y) \langle \Psi_k^2 \rangle \\
 &\approx \sum_{k=1}^P u_k(x) u_k(y) \langle \Psi_k^2 \rangle \\
 &= C_{U^P}(x, y).
 \end{aligned}$$

At least when  $\dim \Omega$  is low, it is very common to see the behaviour of a stochastic field  $U$  (or  $U^P$ ) summarized by plots of the mean field and the variance field, as in Figure 11.1; when  $\dim \Omega = 1$ , a surface or contour plot of the covariance field  $C_U(x, y)$  as in Figure 11.2 can also be informative.

## Bibliography

Spectral expansions in general are covered in Chapter 2 of the monograph of Le Maître & Knio [58], and Chapter 5 of the book of Xiu [117].

The Karhunen–Loève expansion bears the names of Karhunen [48] and Loève [62], but KL-type series expansions of stochastic processes were considered earlier by Kosambi [52]. Lemma 11.12, that the truncation error in a PC expansion is orthogonal to the approximation subspace, is nowadays a simple corollary of standard results in Hilbert spaces, but is an observation that appears to have first first been made in the stochastic context by Cameron & Martin [17]. The application of Wiener–Hermite PC expansions to engineering systems was popularized by Ghanem & Spanos [34]; the extension to gPC and the connection with the Askey scheme is due to Xiu & Karniadakis [118].

The extension of generalized polynomial chaos to arbitrary dependency among the components of the stochastic germ, as in Remark 11.14, is due to Soize & Ghanem [94].

## Exercises

**Exercise 11.1.** Use the Karhunen–Loève expansion to generate samples from a Gaussian random field  $U$  on  $[-1, 1]$  (i.e., for each  $x \in [-1, 1]$ ,  $U(x)$  is a Gaussian random variable) with covariance function

1.  $C_U(x, y) = \exp(-|x - y|^2/a^2)$ ,
2.  $C_U(x, y) = \exp(-|x - y|/a)$ , and
3.  $C_U(x, y) = (1 + |x - y|^2/a^2)^{-1}$

for various values of  $a > 0$ . Plot and comment upon your results, particularly the smoothness of the fields produced.

**Exercise 11.2.** Consider the negative Laplacian operator  $\mathcal{L} := -\frac{d^2}{dx^2}$  acting on real-valued functions on the interval  $[0, 1]$ , with zero boundary conditions. Show that the eigenvalues  $\mu_n$  and eigenfunctions  $e_n$  of  $\mathcal{L}$  are

$$\mu_n = (\pi n)^2, \quad e_n(x) = \sin(\pi n x).$$

Hence show that  $C := \mathcal{L}^{-1}$  has the same eigenfunctions with eigenvalues  $\lambda_n = (\pi n)^{-2}$ . Hence, using the Karhunen–Loève theorem, generate figures similar to Figure 11.1 for your choice of mean field  $m: [0, 1] \rightarrow \mathbb{R}$ .

**Exercise 11.3.** Do the analogue of Exercise 11.3 for the negative Laplacian operator  $\mathcal{L} := -\frac{d^2}{dx^2} - \frac{d^2}{dy^2}$  acting on real-valued functions on the square  $[0, 1]^2$ , again with zero boundary conditions.

**Exercise 11.4.** Show that the eigenvalues  $\lambda_n$  and eigenfunctions  $e_n$  of the exponential covariance function  $C(x, y) = \exp(-|x - y|/a)$  on  $[-b, b]$  are given by

$$\lambda_n = \begin{cases} \frac{2a}{1+a^2w_n^2}, & \text{if } n \in 2\mathbb{Z}, \\ \frac{2a}{1+a^2v_n^2}, & \text{if } n \in 2\mathbb{Z} + 1, \end{cases}$$

$$e_n(x) = \begin{cases} \sin(w_n x) / \sqrt{b - \frac{\sin(2w_n b)}{2w_n}}, & \text{if } n \in 2\mathbb{Z}, \\ \cos(v_n x) / \sqrt{b + \frac{\sin(2v_n b)}{2v_n}}, & \text{if } n \in 2\mathbb{Z} + 1, \end{cases}$$

where  $w_n$  and  $v_n$  solve the transcendental equations

$$\begin{cases} aw_n + \tan(w_n b) = 0, & \text{for } n \in 2\mathbb{Z}, \\ 1 - av_n \tan(v_n b) = 0, & \text{for } n \in 2\mathbb{Z} + 1. \end{cases}$$

Hence, using the Karhunen–Loève theorem, generate sample paths from the Gaussian measure with covariance kernel  $C$  and your choice of mean path.

## Chapter 12

# Stochastic Galerkin Methods

Not to be absolutely certain is, I think,  
one of the essential things in rationality.

---

*Am I an Atheist or an Agnostic?*  
BERTRAND RUSSELL

Unlike non-intrusive approaches, which rely on individual realizations to determine the stochastic model response to random inputs, Galerkin methods use a formalism of *weak solutions*, expressed in terms of inner products, to form systems of governing equations for the solution's PC coefficients, which are generally coupled together. They are in essence the extension to a suitable tensor product Hilbert space of the usual Galerkin formalism that underlies many theoretical and numerical approaches to PDEs. This chapter is devoted to the formulation of Galerkin methods of UQ using PC expansions.

Suppose that the relationship between some input data  $d$  and the output (solution)  $u$  can be expressed formally as

$$\mathcal{M}(u; d) = 0.$$

A good model for this kind of set-up is an elliptic boundary value problem on, say, a bounded, connected domain  $\Omega \subseteq \mathbb{R}^n$  with smooth boundary  $\partial\Omega$ :

$$\begin{aligned} -\nabla \cdot (\kappa(x) \nabla u(x)) &= f(x) & \text{for } x \in \Omega, \\ u(x) &= 0 & \text{for } x \in \partial\Omega. \end{aligned} \tag{12.1}$$

In this case, the input data  $d$  are typically the forcing term  $f: \Omega \rightarrow \mathbb{R}$  and the permeability field  $\kappa: \Omega \rightarrow \mathbb{R}^{n \times n}$ ; in some cases, the domain  $\Omega$  itself might depend upon  $d$ , but this introduces additional complications that will not be considered in this chapter. For a PDE such as this, solutions  $u$  are typically sought in the Sobolev space  $H_0^1(\Omega)$  of  $L^2$  functions that have a weak derivative that itself lies in  $L^2$ , and that vanish on  $\partial\Omega$  in the sense of trace. Moreover, it is usual to seek *weak solutions*, i.e.  $u \in H_0^1(\Omega)$  for which the inner product of (12.1) with any  $v \in H_0^1(\Omega)$  is an equality. That is, integrating by parts, we seek  $u \in H_0^1(\Omega)$  such that

$$\langle \kappa \nabla u, \nabla v \rangle_{L^2(\Omega)} = \langle f, v \rangle_{L^2(\Omega)} \quad \text{for all } v \in H_0^1(\Omega). \tag{12.2}$$

On expressing this problem in a chosen basis of  $H_0^1(\Omega)$ , the column vector  $[u]$  of coefficients of  $u$  in this basis turn out to satisfy a matrix-vector equation (i.e. a system of simultaneous linear equations) of the form  $[a][u] = [b]$  for some matrix  $[a]$  determined by the permeability field  $\kappa$  and a column vector  $[b]$  determined by the forcing term  $f$ .

In this chapter, after reviewing basic Lax–Milgram theory and Galerkin projection for problems like (12.1) / (12.2), we consider the situation in which the input data  $d$  are uncertain and are described as a random variable  $D(\xi)$ . Then the solution is also a random variable  $U(\xi)$  and the model relationship becomes

$$\mathcal{M}(U(\xi); D(\xi)) = 0.$$

Again, this equation is usually interpreted in a weak sense in a suitable Hilbert space of  $H_0^1(\Omega)$ -valued random variables. If  $D$  and  $U$  are expanded in some PC basis, it is natural to ask how the coefficients of  $U$  with respect to this PC basis and a chosen basis of  $H_0^1(\Omega)$  are related to one another. It will turn out that, like in the standard deterministic setting, this problem can be written in the form of a matrix-vector equation  $[A][U] = [B]$  related to, but more complicated than, the deterministic problem  $[a][u] = [b]$ .

## 12.1 Lax–Milgram Theory and Galerkin Projection

Let  $\mathcal{H}$  be a real Hilbert space equipped with a bilinear form  $a: \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$ . Given  $f \in \mathcal{H}^*$  (i.e. a continuous linear functional  $f: \mathcal{H} \rightarrow \mathbb{R}$ ), the associated *weak problem* is:

$$\text{find } u \in \mathcal{H} \text{ such that } a(u, v) = \langle f | v \rangle \text{ for all } v \in \mathcal{V}. \quad (12.3)$$

**Example 12.1.** Let  $\Omega \subseteq \mathbb{R}^n$  be a bounded, connected domain. Let a matrix-valued function  $\kappa: \Omega \rightarrow \mathbb{R}^{n \times n}$  and a scalar-valued function  $f: \Omega \rightarrow \mathbb{R}$  be given, and consider the elliptic problem (12.1). The appropriate bilinear form  $a(\cdot, \cdot)$  is defined by

$$a(u, v) := \langle -\nabla \cdot (\kappa \nabla u), v \rangle_{L^2(\Omega)} = \langle \kappa \nabla u, \nabla v \rangle_{L^2(\Omega)},$$

where the second equality follows from integration by parts when  $u, v$  are smooth functions that vanish on  $\partial\Omega$ ; such functions form a dense subset of the Sobolev space  $H_0^1(\Omega)$ . This short calculation motivates two important developments in the treatment of the PDE (12.1). First, even though the original formulation (12.1) seems to require the solution  $u$  to have two orders of differentiability, the last line of the above calculation makes sense even if  $u$  and  $v$  have only one order of (weak) differentiability, and so we restrict attention to  $H_0^1(\Omega)$ . Second, we declare  $u \in H_0^1(\Omega)$  to be a *weak solution* of (12.1) if the  $L^2(\Omega)$  inner product of (12.1) with any  $v \in H_0^1(\Omega)$  holds as an equality of real numbers, i.e. if

$$-\int_{\Omega} \nabla \cdot (\kappa(x) \nabla u(x)) v(x) dx = \int_{\Omega} f(x) v(x) dx$$

i.e. if

$$a(u, v) = \langle f, v \rangle_{L^2(\Omega)} \quad \text{for all } v \in H_0^1(\Omega).$$



The existence and uniqueness of solutions problems like (12.3), under appropriate conditions on  $a$  (which of course are inherited from appropriate conditions on  $\kappa$ ), is ensured by the Lax–Milgram theorem, which generalizes the Riesz representation theorem that any Hilbert space is isomorphic to its dual space.

**Theorem 12.2** (Lax–Milgram). *Let  $a$  be a bilinear form on a Hilbert space  $\mathcal{H}$  such that*

1. (boundedness) *there exists a constant  $C > 0$  such that, for all  $u, v \in \mathcal{H}$ ,  $|a(u, v)| \leq C\|u\|\|v\|$ ; and*
2. (coercivity) *there exists a constant  $c > 0$  such that, for all  $v \in \mathcal{H}$ ,  $|a(v, v)| \geq c\|v\|^2$ .*

*Then for all  $f \in \mathcal{H}^*$ , there exists a unique  $u \in \mathcal{V}$  such that, for all  $v \in \mathcal{H}$ ,  $a(u, v) = \langle f | v \rangle$ .*

*Proof.* For each  $u \in \mathcal{H}$ ,  $v \mapsto a(u, v)$  is a bounded linear functional on  $\mathcal{H}$ . So, by the Riesz representation theorem, given  $u \in \mathcal{H}$ , there is a unique  $w \in \mathcal{H}$  such that  $\langle w, v \rangle = a(u, v)$ . Define  $Au := w$ . The map  $A: \mathcal{H} \rightarrow \mathcal{H}$  is clearly well-defined. It is also linear: take  $\alpha_1, \alpha_2 \in \mathbb{R}$  and  $u_1, u_2 \in \mathcal{H}$ :

$$\begin{aligned} \langle A(\alpha_1 u_1 + \alpha_2 u_2), v \rangle &= a(\alpha_1 u_1 + \alpha_2 u_2, v) \\ &= \alpha_1 a(u_1, v) + \alpha_2 a(u_2, v) \\ &= \alpha_1 \langle Au_1, v \rangle + \alpha_2 \langle Au_2, v \rangle \\ &= \langle \alpha_1 Au_1 + \alpha_2 Au_2, v \rangle. \end{aligned}$$

$A$  is a bounded map, since

$$\|Au\|^2 = \langle Au, Au \rangle = a(u, Au) \leq C\|u\|\|Au\|,$$

so  $\|Au\| \leq C\|u\|$ . Furthermore,  $A$  is injective since

$$\|Au\|\|u\| \geq \langle Au, u \rangle = a(u, u) \geq c\|u\|^2,$$

so  $Au = 0 \implies u = 0$ .

To see that the range  $\mathcal{R}(A)$  is closed, take a convergent sequence  $(v_n)_{n \in \mathbb{N}}$  in  $\mathcal{R}(A)$  that converges to some  $v \in \mathcal{H}$ . Choose  $u_n \in \mathcal{H}$  such that  $Au_n = v_n$  for each  $n \in \mathbb{N}$ . The sequence  $(Au_n)_{n \in \mathbb{N}}$  is Cauchy, so

$$\begin{aligned} \|Au_n - Au_m\|\|u_n - u_m\| &\geq |\langle Au_n - Au_m, u_n - u_m \rangle| \\ &= |a(u_n - u_m, u_n - u_m)| \\ &\geq c\|u_n - u_m\|^2. \end{aligned}$$

So  $c\|u_n - u_m\| \leq \|v_n - v_m\| \rightarrow 0$ . So  $(u_n)_{n \in \mathbb{N}}$  is Cauchy and converges to some  $u \in \mathcal{H}$ . So  $v_n = Au_n \rightarrow Au = v$  by the continuity (boundedness) of  $A$ , so  $v \in \mathcal{R}(A)$ , and so  $\mathcal{R}(A)$  is closed.

Finally,  $A$  is surjective: for, if not, there is an  $s \in \mathcal{H}$ ,  $s \neq 0$ , such that  $s \perp \mathcal{R}(A)$ . But then

$$c\|s\|^2 \leq a(s, s) = \langle s, As \rangle = 0,$$

so  $s = 0$ , a contradiction.

So, take  $f \in \mathcal{H}^*$ . There is a unique  $w \in \mathcal{H}$  such that  $\langle w, v \rangle = \langle f | v \rangle$  for all  $v \in \mathcal{H}$ . The equation  $Au = w$  has a unique solution since  $A$  is invertible. So  $\langle Au, v \rangle = \langle f | v \rangle$  for all  $v \in \mathcal{H}$ . But  $\langle Au, v \rangle = a(u, v)$ . So there is a unique  $u \in \mathcal{H}$  such that  $a(u, v) = \langle f | v \rangle$ .  $\square$

**Galerkin Projection.** Now consider the problem of finding a good finite-dimensional approximation to  $u$ . Fix a subspace  $\mathcal{V}^{(M)} \subseteq \mathcal{H}$  of dimension  $M$ . The *Galerkin projection* of the weak problem is:

$$\text{find } u^{(M)} \in \mathcal{V}^{(M)} \text{ such that } a(u^{(M)}, v^{(M)}) = \langle f, v^{(M)} \rangle \text{ for all } v^{(M)} \in \mathcal{V}^{(M)}.$$

Note that if the hypotheses of the Lax–Milgram theorem are satisfied on the full space  $\mathcal{H}$ , then they are certainly satisfied on the subspace  $\mathcal{V}^{(M)}$ , thereby ensuring the existence and uniqueness of solutions to the Galerkin problem. Note well, though, that existence of a unique Galerkin solution for each  $M \in \mathbb{N}_0$  does *not* imply the existence of a unique weak solution (nor even multiple weak solutions) to the full problem; for this, one typically needs to show that the Galerkin approximations are uniformly bounded and appeal to a Sobolev embedding theorem to extract a convergent subsequence.

**Example 12.3.** 1. The *Fourier basis*  $\{e_k\}_{k \in \mathbb{Z}}$  of  $L^2(\mathbb{S}^1, dx; \mathbb{C})$  defined by

$$e_k(x) = \frac{1}{\sqrt{2\pi}} \exp(ikx).$$

For Galerkin projection, one can use the finite-dimensional subspace

$$\mathcal{V}^{(2M+1)} := \text{span}\{e_{-M}, \dots, e_{-1}, e_0, e_1, \dots, e_M\}$$

of functions that are band-limited to contain frequencies at most  $M$ . In case of real-valued functions, one can use the functions

$$\begin{aligned} x &\mapsto \cos(kx), & \text{for } k \in \mathbb{N}_0, \\ x &\mapsto \sin(kx), & \text{for } k \in \mathbb{N}. \end{aligned}$$

2. Fix a partition  $a = x_0 < x_1 < \dots < x_M = b$  of a compact interval  $[a, b] \subsetneq \mathbb{R}$  and consider the associated *tent functions* defined by

$$\phi_m(x) := \begin{cases} 0, & \text{if } x \leq a \text{ or } x \leq x_{m-1}; \\ (x - x_{m-1})/(x_m - x_{m-1}), & \text{if } x_{m-1} \leq x \leq x_m; \\ (x_{m+1} - x)/(x_{m+1} - x_m), & \text{if } x_m \leq x \leq x_{m+1}; \\ 0, & \text{if } x \geq b \text{ or } x \geq x_{m+1}. \end{cases}$$

The function  $\phi_m$  takes the value 1 at  $x_m$  and decays linearly to 0 along the two line segments adjacent to  $x_m$ . The  $(M+1)$ -dimensional vector space  $\mathcal{V}^{(M)} := \text{span}\{\phi_0, \dots, \phi_M\}$  consists of all continuous functions on  $[a, b]$  that are piecewise affine on the partition, i.e. have constant derivative on each of the open intervals  $(x_{m-1}, x_m)$ . The space  $\tilde{\mathcal{V}}^{(M)} := \text{span}\{\phi_1, \dots, \phi_{M-1}\}$  consists of the continuous functions that are piecewise affine on the partition and take the value 0 at  $a$  and  $b$ ; hence,  $\tilde{\mathcal{V}}^{(M)}$  is one good choice for a finite-dimensional space to approximate the Sobolev space  $H_0^1([a, b])$ . More generally, one could consider tent functions associated to any simplicial mesh in  $\mathbb{R}^n$ .

An important property of the solution  $u^{(M)}$  of the Galerkin problem, viewed as an approximation to the solution  $u$  of the original problem, is that the error

$u - u^{(M)}$  is  $a$ -orthogonal to the approximation subspace  $\mathcal{V}^{(M)}$ : for any choice of  $v^{(M)} \in \mathcal{V}^{(M)} \subseteq \mathcal{H}$ ,

$$a(u - u^{(M)}, v^{(M)}) = a(u, v^{(M)}) - a(u^{(M)}, v^{(M)}) = \langle f, v^{(M)} \rangle - \langle f, v^{(M)} \rangle = 0.$$



However, note well that  $u^{(M)}$  is generally not the optimal approximation of  $u$  from the subspace  $\mathcal{V}^{(M)}$ , i.e.

$$\|u - u^{(M)}\| \neq \inf \left\{ \|u - v^{(M)}\| \mid v^{(M)} \in \mathcal{V}^{(M)} \right\}.$$

The optimal approximation of  $u$  from  $\mathcal{V}^{(M)}$  is the orthogonal projection of  $u$  onto  $\mathcal{V}^{(M)}$ ; if  $\mathcal{H}$  has an orthonormal basis  $\{e_n\}$  and  $u = \sum_{n \in \mathbb{N}} u^n e_n$ , then the optimal approximation of  $u$  in  $\mathcal{V}^{(M)} = \text{span}\{e_1, \dots, e_M\}$  is  $\sum_{n=1}^M u^n e_n$ , but this is not generally the same as the Galerkin solution  $u^{(M)}$ . However, the next result, Céa's lemma, shows that  $u^{(M)}$  is a quasi-optimal approximation to  $u$  (note that the ratio  $C/c$  is always at least 1):

**Lemma 12.4 (Céa's lemma).** *Let  $a$ ,  $c$  and  $C$  be as in the statement of the Lax–Milgram theorem. Then the weak solution  $u \in \mathcal{H}$  and the Galerkin solution  $u^{(M)} \in \mathcal{V}^{(M)}$  satisfy*

$$\|u - u^{(M)}\| \leq \frac{C}{c} \inf \left\{ \|u - v^{(M)}\| \mid v^{(M)} \in \mathcal{V}^{(M)} \right\}.$$

*Proof.* Exercise 12.2. □

**Matrix Form.** It is helpful to recast the Galerkin problem in matrix form. Let  $\{\phi_1, \dots, \phi_M\}$  be a basis for  $\mathcal{V}^{(M)}$ . Then  $u^{(M)}$  solves the Galerkin problem if and only if

$$a(u^{(M)}, \phi_m) = \langle f, \phi_m \rangle \text{ for } m \in \{1, \dots, M\}.$$

Now expand  $u^{(M)}$  in this basis as  $u^{(M)} = \sum_{m=1}^M u^m \phi_m$  and insert this into the previous equation:

$$a \left( \sum_{m=1}^M u^m \phi_m, \phi_i \right) = \sum_{m=1}^M u^m a(\phi_m, \phi_i) = \langle f, \phi_i \rangle \text{ for } i \in \{1, \dots, M\}.$$

In other words, the vector of coefficients  $[u^{(M)}] = [u^1, \dots, u^M]^\top \in \mathbb{R}^M$  satisfies the matrix equation

$$\underbrace{\begin{bmatrix} a(\phi_1, \phi_1) & \dots & a(\phi_M, \phi_1) \\ \vdots & \ddots & \vdots \\ a(\phi_1, \phi_M) & \dots & a(\phi_M, \phi_M) \end{bmatrix}}_{=: [a]} \underbrace{\begin{bmatrix} u^1 \\ \vdots \\ u^M \end{bmatrix}}_{=: [u^{(M)}]} = \underbrace{\begin{bmatrix} \langle f, \phi_1 \rangle \\ \vdots \\ \langle f, \phi_M \rangle \end{bmatrix}}_{=: [b]}. \quad (12.4)$$

The matrix  $[a] \in \mathbb{R}^{M \times M}$  is the *Gram matrix* of the bilinear form  $a$ , and is of course a symmetric matrix whenever  $a$  is a symmetric bilinear form.

**Remark 12.5.** In practice the matrix-vector equation  $[a][u^{(M)}] = [b]$  is *never* solved by explicitly inverting the Gram matrix  $[a]$  to obtain the coefficients  $u^m$  via  $[u^{(M)}] = [a]^{-1}[b]$ . Indeed, in many situations the Gram matrix is sparse, and so inversion methods that take advantage of that sparsity are used; furthermore, for large systems, the methods used are often iterative rather than direct (e.g. factorization-based).

**Lax–Milgram Theory for Banach Spaces.** There are many extensions of the now-classical Lax–Milgram lemma beyond the setting of symmetric bilinear forms on Hilbert spaces. For example, the following generalization is due to Kozono & Yanagisawa [53]:

**Theorem 12.6 (Kozono–Yanagisawa).** *Let  $\mathcal{X}$  be a Banach space,  $\mathcal{Y}$  a reflexive Banach space, and  $a: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{C}$  a bilinear form such that*

1. *there is a constant  $M > 0$  such that*

$$|a(x, y)| \leq M \|x\|_{\mathcal{X}} \|y\|_{\mathcal{Y}} \quad \text{for all } x \in \mathcal{X}, y \in \mathcal{Y};$$

2. *the null spaces*

$$\begin{aligned} N_{\mathcal{X}} &:= \{x \in \mathcal{X} \mid a(x, y) = 0 \text{ for all } y \in \mathcal{Y}\} \subseteq \mathcal{X}, \\ N_{\mathcal{Y}} &:= \{y \in \mathcal{Y} \mid a(x, y) = 0 \text{ for all } x \in \mathcal{X}\} \subseteq \mathcal{Y}, \end{aligned}$$

*admit complementary closed subspaces  $R_{\mathcal{X}}$  and  $R_{\mathcal{Y}}$  such that  $\mathcal{X} = N_{\mathcal{X}} \oplus R_{\mathcal{X}}$  and  $\mathcal{Y} = N_{\mathcal{Y}} \oplus R_{\mathcal{Y}}$ ;*

3. *there is a constant  $C > 0$  such that*

$$\begin{aligned} \|x\|_{\mathcal{X}} &\leq C \left( \sup_{y \in \mathcal{Y}} \frac{|a(x, y)|}{\|y\|_{\mathcal{Y}}} + \|P_{\mathcal{X}} x\|_{\mathcal{X}} \right) \quad \text{for all } x \in \mathcal{X}, \\ \|y\|_{\mathcal{Y}} &\leq C \left( \sup_{x \in \mathcal{X}} \frac{|a(x, y)|}{\|x\|_{\mathcal{X}}} + \|P_{\mathcal{Y}} y\|_{\mathcal{Y}} \right) \quad \text{for all } y \in \mathcal{Y}, \end{aligned}$$

*where  $P_{\mathcal{X}}$  (resp.  $P_{\mathcal{Y}}$ ) is the projection of  $\mathcal{X}$  (resp.  $\mathcal{Y}$ ) onto  $N_{\mathcal{X}}$  (resp.  $N_{\mathcal{Y}}$ ) along the direct sum  $\mathcal{X} = N_{\mathcal{X}} \oplus R_{\mathcal{X}}$  (resp.  $\mathcal{Y} = N_{\mathcal{Y}} \oplus R_{\mathcal{Y}}$ ).*

*Then, for every  $f \in \mathcal{Y}'$  such that  $\langle f | y \rangle = 0$  for all  $y \in N_{\mathcal{Y}}$ , there exists  $x \in \mathcal{X}$  such that*

$$a(x, y) = \langle f | y \rangle \quad \text{for all } y \in \mathcal{Y}.$$

*Furthermore, there is a constant  $C$  independent of  $x$  and  $f$  such that  $\|x\|_{\mathcal{X}} \leq C \|f\|_{\mathcal{Y}'}$ .*

## 12.2 Stochastic Galerkin Projection

**Stochastic Lax–Milgram Theory.** The next step is to build appropriate Lax–Milgram theory and Galerkin projection for stochastic problems, for which a good prototype is

$$\begin{aligned} -\nabla \cdot (\kappa(\theta, x) \nabla u(\theta, x)) &= f(\theta, x) & \text{for } x \in \Omega, \\ u(x) &= 0 & \text{for } x \in \partial\Omega, \end{aligned}$$

with  $\theta$  being drawn from some probability space  $(\Theta, \mathcal{F}, \mu)$ . To that end, we introduce a stochastic space  $\mathcal{S}$ , which in the following will be  $L^2(\Theta, \mu; \mathbb{R})$ . We retain also a Hilbert space  $\mathcal{V}$  in which the deterministic solution  $u(\theta)$  is sought for each  $\theta \in \Theta$ ; implicitly,  $\mathcal{V}$  is independent of the problem data, or rather of  $\theta$ . Thus, the space in which the stochastic solution  $U$  is sought is the tensor

product Hilbert space  $\mathcal{H} := \mathcal{V} \otimes \mathcal{S}$ , which is isomorphic to the space  $L^2(\Theta, \mu; \mathcal{V})$  of square-integrable  $\mathcal{V}$ -valued random variables.

In terms of bilinear forms, the setup is that of a bilinear-form-on- $\mathcal{V}$ -valued random variable  $A$  and a  $\mathcal{V}'$ -valued random variable  $F$ . Define a bilinear form  $\alpha$  on  $\mathcal{H}$  by

$$\alpha(X, Y) := \mathbb{E}_\mu[A(X, Y)] \equiv \int_{\Theta} A(\theta)(X(\theta), Y(\theta)) d\mu(\theta)$$

and, similarly, a linear functional  $\beta$  on  $\mathcal{H}$  by

$$\beta(Y) := \mathbb{E}_\mu[\langle F | Y \rangle_{\mathcal{V}}].$$

Clearly, if  $\alpha$  satisfies the boundedness and coercivity assumptions of the Lax–Milgram theorem on  $\mathcal{H}$ , then, for every  $F \in L^2(\Theta, \mu; \mathcal{V}')$ , there is a unique weak solution  $U \in L^2(\Theta, \mu; \mathcal{V})$  satisfying

$$\alpha(U, Y) = \beta(Y) \text{ for all } Y \in L^2(\Theta, \mu; \mathcal{V}).$$

A sufficient, but not necessary, condition for  $\alpha$  to satisfy the hypotheses of the Lax–Milgram theorem on  $\mathcal{H}$  is for  $A(\theta)$  to satisfy those hypotheses uniformly in  $\theta$  on  $\mathcal{V}$ :

**Theorem 12.7** (Stochastic Lax–Milgram theorem). *Let  $(\Theta, \mathcal{F}, \mu)$  be a probability space, and let  $A$  be a random variable on  $\Theta$ , taking values in the space of symmetric bilinear forms on a Hilbert space  $\mathcal{V}$ , and satisfying the hypotheses of the deterministic Lax–Milgram theorem (Theorem 12.2) uniformly with respect to  $\theta \in \Theta$ . Define a bilinear form  $\alpha$  on  $L^2(\Theta, \mu; \mathcal{V})$  by*

$$\alpha(X, Y) := \mathbb{E}_\mu[A(X, Y)].$$

*Then, for every  $F \in L^2(\Theta, \mu; \mathcal{V}')$ , there is a unique  $U \in L^2(\Theta, \mu; \mathcal{V})$  such that*

$$\alpha(U, Y) = \beta(Y) \text{ for all } Y \in L^2(\Theta, \mu; \mathcal{V}).$$

*Proof.* Suppose that  $A(\theta)$  satisfies the boundedness assumption with constant  $C(\theta)$  and the coercivity assumption with constant  $c(\theta)$ . By hypothesis,

$$\begin{aligned} C' &:= \sup_{\theta \in \Theta} C(\theta) \quad \text{and} \\ c' &:= \inf_{\theta \in \Theta} c(\theta) \end{aligned}$$

are both strictly positive and finite. The bilinear form  $\alpha$  satisfies, for all  $X, Y \in \mathcal{H}$ ,

$$\begin{aligned} \alpha(X, Y) &= \mathbb{E}_\mu[A(X, Y)] \\ &\leq \mathbb{E}_\mu[C \|X\|_{\mathcal{V}} \|Y\|_{\mathcal{V}}] \\ &\leq C' \mathbb{E}_\mu[\|X\|_{\mathcal{V}}^2]^{1/2} \mathbb{E}_\mu[\|Y\|_{\mathcal{V}}^2]^{1/2} \\ &= C' \|X\|_{\mathcal{H}} \|Y\|_{\mathcal{H}}, \end{aligned}$$

and

$$\begin{aligned} \alpha(X, X) &= \mathbb{E}_\mu[A(X, X)] \\ &\geq \mathbb{E}_\mu[c \|X\|_{\mathcal{V}}^2] \\ &\geq c' \|X\|_{\mathcal{H}}^2. \end{aligned}$$

Hence, by the deterministic Lax–Milgram theorem applied to the bilinear form  $\alpha$  on the Hilbert space  $\mathcal{H}$ , for every  $F \in L^2(\Theta, \mu; \mathcal{V})$ , there exists a unique  $U \in L^2(\Theta, \mu; \mathcal{V})$  such that

$$\alpha(U, Y) = \beta(Y) \text{ for all } Y \in L^2(\Theta, \mu; \mathcal{V}). \quad \square$$

**Remark 12.8.** Note, however, that uniform boundedness and coercivity of  $A$  are not necessary for  $\alpha$  to be bounded and coercive. For example, the constants  $c(\theta)$  and  $C(\theta)$  may degenerate to 0 or  $\infty$  as  $\theta$  approaches certain points of the sample space  $\Theta$ . Provided that these degeneracies are integrable and yield positive and finite expected values, this will not ruin the boundedness and coercivity of  $\alpha$ . Indeed, there may be an arbitrarily large (but  $\mu$ -measure zero) set of  $\theta$  for which there is no weak solution  $u$  to the deterministic problem

$$A(\theta)(u, v) = \langle F(\theta) | v \rangle \text{ for all } v \in \mathcal{V}.$$

**Stochastic Galerkin Projection.** Let  $\mathcal{V}^{(M)}$  be a finite-dimensional subspace of  $\mathcal{V}$ , with basis  $\phi_1, \dots, \phi_M$ . As indicated above, take the stochastic space  $\mathcal{S}$  to be  $L^2(\Theta, \mu; \mathbb{K})$ , which we assume to be equipped with an orthogonal decomposition such as a PC decomposition. Let  $\mathcal{S}^P$  be a finite-dimensional subspace of  $\mathcal{S}$ , for example the span of the polynomials of degree at most  $P$ . The Galerkin projection of the stochastic problem on  $\mathcal{H}$  is to find  $U = \sum_{m,k} u_k^m \phi_m \otimes \Psi_k \in \mathcal{V}^{(M)} \otimes \mathcal{S}^P$  such that

$$\alpha(U, V) = \beta(V) \text{ for all } V \in L^2(\Theta, \mu; \mathcal{V}).$$

In particular, it suffices to find  $U$  that satisfies this condition for each basis element  $V = \phi_n \otimes \Psi_\ell$  of  $\mathcal{V}^{(M)} \otimes \mathcal{S}^P$ . Recall that  $\phi_n \otimes \Psi_\ell$  is the function  $(\theta, x) \mapsto \phi_n(x) \Psi_\ell(\theta)$ .

As before, the Galerkin problem is equivalent to the matrix-vector equation

$$[\alpha][U] = [\beta]$$

in the basis  $\{\phi_m \otimes \Psi_k \mid m = 1, \dots, M; k = 0, \dots, P\}$  of  $\mathcal{V}^{(M)} \otimes \mathcal{S}^P$ . An obvious question is how the Gram matrix  $[\alpha] \in \mathbb{R}^{M(P+1) \times M(P+1)}$  is related to the Gram matrix of the random bilinear form  $A$ .

...

Suppose that we are already given a linear problem with its deterministic problem discretized and cast in the matrix form

$$[A](\xi)U(\xi) = B(\xi)$$

which, for each fixed  $\xi \in L^2(\Xi, p_\xi)$ , has its solution  $U(\xi) \in \mathcal{V}^{(M)} \cong \mathbb{R}^M$ . The stochastic Galerkin projection for the stochastic solution  $U = \sum_{k=1}^P u_k \Psi_k \in \mathcal{V}^{(M)} \otimes \mathcal{S}^P$  gives

$$\sum_{j=0}^P \langle \Psi_i, [A] \Psi_j \rangle u_j = \langle \Psi_i, B \rangle \text{ for each } i \in \{0, \dots, P\}.$$

This is equivalent to the (large!) block system

$$\begin{bmatrix} [A]_{00} & \cdots & [A]_{0P} \\ \vdots & \ddots & \vdots \\ [A]_{P0} & \cdots & [A]_{PP} \end{bmatrix} \begin{bmatrix} u_0 \\ \vdots \\ u_P \end{bmatrix} = \begin{bmatrix} b_0 \\ \vdots \\ b_P \end{bmatrix}, \quad (12.5)$$

where for  $0 \leq i, j \leq P$ ,

- $[A]_{ij} := \langle \Psi_i, [A] \Psi_j \rangle \in \mathbb{R}^{M \times M}$ , where  $[A] \in \mathbb{R}^{M \times M}$  is the Gram matrix of the random bilinear form  $A$ ;
- $u_i = \sum_{m=1}^M u_i^m \phi_m \in \mathcal{V}^{(M)}$  is the  $i^{\text{th}}$  stochastic mode of the solution  $U$ ;
- and  $b_i := \langle \Psi_i, B \rangle \in \mathbb{R}^M$  is the  $i^{\text{th}}$  stochastic mode of the source term  $B$ .

Note that, in general, the stochastic modes  $u_j$  of the solution  $U$  (and, indeed the coefficients  $u_j^m$  of the stochastic modes in the deterministic basis  $\phi_1, \dots, \phi_M$ ) are all coupled together through the matrix  $[A]$ . This can be a limitation of stochastic Galerkin methods, and will be remarked upon later.

**Example 12.9.** As a special case, suppose that the random data have no impact on the differential operator and affect only the right-hand side  $B$ . In this case  $A(\theta) = a$  for all  $\theta \in \Theta$  and so

$$[A]_{ij} := \langle \Psi_i, [a] \Psi_j \rangle = [a] \langle \Psi_i, \Psi_j \rangle = [a] \delta_{ij} \langle \Psi_i^2 \rangle.$$

Hence, the stochastic Galerkin system, in its matrix form (12.5), becomes

$$\begin{bmatrix} [a] & [0] & \cdots & [0] \\ [0] & [a] \langle \Psi_1^2 \rangle & \ddots & \vdots \\ \vdots & \ddots & \ddots & [0] \\ [0] & \cdots & [0] & [a] \langle \Psi_P^2 \rangle \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_P \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_P \end{bmatrix}.$$

Hence the stochastic modes  $u_j$  decouple and are given by

$$u_j = [a]^{-1} \frac{\langle b, \Psi_j \rangle}{\langle \Psi_j^2 \rangle}.$$

**The Galerkin Tensor.** In contrast to Example 12.9, in which the differential operator is deterministic, we can consider the case in which the matrix  $[\alpha]$  has a (truncated) PC expansion

$$[\alpha] = \sum_{k=0}^P [\alpha]_k \Psi_k$$

with coefficient matrices  $[\alpha]_k \in \mathbb{R}^{M \times M}$  for  $k \geq 0$ . In this case, the blocks  $[\alpha]_{ij}$  are given by

$$[\alpha]_{ij} = \langle \Psi_i, [\alpha] \Psi_j \rangle = \sum_{k=0}^P [\alpha]_k \langle \Psi_i, \Psi_j \Psi_k \rangle.$$

Hence, the Galerkin block system (12.5) is equivalent to

$$\begin{bmatrix} [\bar{\alpha}]_{00} & \cdots & [\bar{\alpha}]_{0P} \\ \vdots & \ddots & \vdots \\ [\bar{\alpha}]_{P0} & \cdots & [\bar{\alpha}]_{PP} \end{bmatrix} \begin{bmatrix} u_0 \\ \vdots \\ u_P \end{bmatrix} = \begin{bmatrix} \bar{b}_0 \\ \vdots \\ \bar{b}_P \end{bmatrix}, \quad (12.6)$$

where

$$\begin{aligned}\bar{b}_i &:= \frac{\langle B, \Psi_i \rangle}{\langle \Psi_i^2 \rangle}, \\ \overline{[\alpha]}_{ij} &:= \sum_{k=0}^P [\alpha]_k C_{kji}, \\ C_{ijk} &:= \frac{\langle \Psi_i \Psi_j \Psi_k \rangle}{\langle \Psi_k \Psi_k \rangle}.\end{aligned}$$

The rank-3 tensor  $C_{ijk}$  is called the *Galerkin tensor*:

- it is symmetric in the first two indices:  $C_{ijk} = C_{jik}$ ;
- this induces symmetry in the problem (12.6):  $\overline{[\alpha]}_{ij} = \overline{[\alpha]}_{ji}$ ;
- and since the  $\Psi_k$  are an orthogonal system, many of the  $(P+1)^3$  entries of  $C_{ijk}$  are zero, leading to sparsity for the matrix problem; for example,

$$\overline{[\alpha]}_{00} = \sum_{k=0}^P [\alpha]_k C_{k00} = [\alpha]_0.$$

Note that the Galerkin tensor  $C_{ijk}$  is determined entirely by the PC system  $\{\Psi_k \mid k \geq 0\}$ , and so while there is a significant computational cost associated to evaluating its entries, this is a one-time cost: the Galerkin tensor can be pre-computed, stored, and then used for many different problems, i.e. many  $A$ s and  $B$ s.

**Example 12.10** (Ordinary differential equations). Consider random variables  $Z, B \in L^2(\Theta, \mu; \mathbb{R})$  and the random linear first-order ordinary differential equation

$$\frac{dU}{dt} = -ZU, \quad U(t) = B,$$

for  $U: [0, T] \times \Theta \rightarrow \mathbb{R}$ . Let  $\{\Psi_k\}_{k \in \mathbb{N}_0}$  be an orthogonal basis for  $L^2(\Theta, \mu; \mathbb{R})$  with the usual convention that  $\Psi_0 = 1$ . Give  $Z, B$  and  $U$  the chaos expansions  $Z = \sum_{k \in \mathbb{N}_0} z_k \Psi_k$ ,  $B = \sum_{k \in \mathbb{N}_0} b_k \Psi_k$  and  $U(t) = \sum_{k \in \mathbb{N}_0} u_k(t) \Psi_k$  respectively. Projecting the evolution equation onto the basis  $\{\Psi_k\}_{k \in \mathbb{N}_0}$  yields

$$\mathbb{E} \left[ \frac{dU}{dt} \Psi_k \right] = -\mathbb{E}[ZU \Psi_k] \text{ for each } k \in \mathbb{N}_0.$$

Inserting the chaos expansions for  $Z$  and  $U$  into this yields, for every  $k \in \mathbb{N}_0$ ,

$$\mathbb{E} \left[ \sum_{i \in \mathbb{N}_0} \dot{u}_i(t) \Psi_i \Psi_k \right] = -\mathbb{E} \left[ \sum_{j \in \mathbb{N}_0} z_j \Psi_j \sum_{i \in \mathbb{N}_0} u_i(t) \Psi_i \Psi_k \right],$$

$$\text{i.e.} \quad \dot{u}_k(t) \langle \Psi_k^2 \rangle = - \sum_{i,j \in \mathbb{N}_0} z_j u_i(t) \mathbb{E}[\Psi_j \Psi_i \Psi_k],$$

$$\text{i.e.} \quad \dot{u}_k(t) = - \sum_{i,j \in \mathbb{N}_0} C_{ijk} z_j u_i(t).$$

The coefficients  $u_k$  are a coupled system of countably many ordinary differential equations. If all the chaos expansions are truncated at order  $P$ , then all



the above summations over  $\mathbb{N}_0$  become summations over  $\{0, \dots, P\}$ , yielding a coupled system of  $P + 1$  ordinary differential equations. In matrix form:

$$\frac{d}{dt} \begin{bmatrix} u_0(t) \\ \vdots \\ u_P(t) \end{bmatrix} = A^\top \begin{bmatrix} u_0 \\ \vdots \\ u_P \end{bmatrix}, \quad \begin{bmatrix} u_0(0) \\ \vdots \\ u_P(0) \end{bmatrix} = \begin{bmatrix} b_0 \\ \vdots \\ b_P \end{bmatrix},$$

where  $A \in \mathbb{R}^{(P+1) \times (P+1)}$  is  $A_{ik} = -\sum_j C_{ijk} z_j$ .

## 12.3 Nonlinearities

Nonlinearities of various types occur throughout practical problems, and their treatment is critical in the context of stochastic Galerkin methods, which require the projection of these nonlinearities onto the finite-dimensional solution spaces. For example, given the approximation

$$U(\xi) \approx U^P(\xi) = \sum_{k=0}^P u_k \Psi_k(\xi)$$

how does one calculate the coefficients of, say,  $U(\xi)^2$  or  $\sqrt{U(\xi)}$ ? The first example,  $U^2$ , is a special case of taking the product of two Galerkin approximations, and can be resolved using the Galerkin tensor  $C_{ijk}$  of the previous section.

**Galerkin Products.** Consider two random variables  $U, V \in L^2(\Theta, \mu; \mathbb{R})$ . The product random variable  $UV$  is again an element of  $L^2(\Theta, \mu; \mathbb{R})$ . The natural question to ask, given expansions  $U = \sum_{k \in \mathbb{N}_0} u_k \Psi_k$  and  $V = \sum_{k \in \mathbb{N}_0} v_k \Psi_k$ , is how to quickly compute the coefficients of  $UV$  in terms of  $\{u_k\}_{k \in \mathbb{N}_0}$  and  $\{v_k\}_{k \in \mathbb{N}_0}$  — particularly if expansions are truncated to finite precision.

**Example 12.11.** Suppose that  $U = \sum_{k=0}^P u_k \Psi_k$  and  $V = \sum_{k=0}^P v_k \Psi_k$ . Then their product  $W := UV$  has the expansion

$$W = \sum_{i,j=0}^P u_i v_j \Psi_i \Psi_j.$$

Note that, while  $W$  is guaranteed to be in  $L^2$ , it is not necessarily in  $\mathcal{S}^P$ . Nevertheless, if we write  $W = \sum_{k \geq 0} w_k \Psi_k$ , it follows that

$$w_k = \frac{\langle W, \Psi_k \rangle}{\langle \Psi_k, \Psi_k \rangle} = \sum_{i,j=0}^P u_i v_j C_{ijk}.$$

The truncation of the expansion  $W = \sum_{k \geq 0} w_k \Psi_k$  to  $k = 0, \dots, P$  is the orthogonal projection of  $W$  onto  $\mathcal{S}^P$  (and hence the  $L^2$ -closest approximation of  $W$  in  $\mathcal{S}^P$ ) and is called the *Galerkin product*, or *pseudo-spectral product*, of  $U$  and  $V$ , denoted  $U * V$ .

The fact that multiplication of two random variables can be handled efficiently, albeit with some truncation error, in terms of their expansions in the

gPC basis and the Galerkin tensor is very useful: it adds to the list of reasons why one might wish to pre-compute and store the Galerkin tensor for use in many problems.

However, the situation of binary products (and hence squares) is...  
Triple and higher products... non-commutativity

**Galerkin Inversion.** Given

$$U = \sum_{k \geq 0} u_k \Psi_k \approx \sum_{k=0}^P u_k \Psi_k$$

we seek a random variable  $V = \sum_{k \geq 0} v_k \Psi_k \approx \sum_{k=0}^P v_k \Psi_k$  such that  $U(\xi)V(\xi) = 1$  for almost every  $\xi$ . The weak interpretation of this desideratum is that  $U * V = \Psi_0$ . Thus we are led to the following matrix-vector equation for the gPC coefficients of  $V$ :

$$\begin{bmatrix} \sum_{k=0}^P C_{k00} u_k & \cdots & \sum_{k=0}^P C_{kP0} u_k \\ \vdots & \ddots & \vdots \\ \sum_{k=0}^P C_{k0P} u_k & \cdots & \sum_{k=0}^P C_{kPP} u_k \end{bmatrix} \begin{bmatrix} v_0 \\ \vdots \\ v_P \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (12.7)$$

Naturally, if  $U(\xi) = 0$  for a positive probability set of  $\xi$ , then  $V(\xi)$  will be undefined for those same  $\xi$ . Furthermore, if  $U \approx 0$  with ‘too large’ probability, then  $V$  may exist a.e. but fail to be in  $L^2$ . Hence, it is not surprising to learn that while (12.7) has a unique solution whenever the matrix on the left-hand is non-singular, the system becomes highly ill-conditioned as the amount of probability mass near  $U = 0$  increases.

**FINISH ME!!!**

## Bibliography

Basic Lax–Milgram theory and Galerkin methods for PDEs can be found in any modern textbook on PDEs, such as those by Evans [27] (see Chapter 6) and Renardy & Rogers [80] (see Chapter 9).

The monograph of Xiu [117] provides a general introduction to spectral methods for uncertainty quantification, including Galerkin methods (Chapter 6), but is light on proofs. The book of Le Maître & Knio [58] covers Galerkin methods in Chapter 4, including an extensive treatment of nonlinearities in Section 4.5.

## Exercises

**Exercise 12.1.** Let  $a$  be a bilinear form satisfying the hypotheses of the Lax–Milgram theorem. Given  $f \in \mathcal{H}^*$ , show that the unique  $u$  such that  $a(u, v) = \langle f | v \rangle$  for all  $v \in \mathcal{H}$  satisfies  $\|u\|_{\mathcal{H}} \leq c^{-1} \|f\|_{\mathcal{H}'}$ .

**Exercise 12.2 (Céa’s lemma).** Let  $a$ ,  $c$  and  $C$  be as in the statement of the Lax–Milgram theorem. Show that the weak solution  $u \in \mathcal{H}$  and the Galerkin

solution  $u^{(M)} \in \mathcal{V}^{(M)}$  satisfy

$$\|u - u^{(M)}\| \leq \frac{C}{c} \inf \left\{ \|u - v^{(M)}\| \mid v^{(M)} \in \mathcal{V}^{(M)} \right\}.$$

**Exercise 12.3.** Consider a partition of the unit interval  $[0, 1]$  into  $N + 1$  equally spaced nodes

$$0 = x_0 < x_1 = h < x_2 = 2h < \cdots < x_N = 1,$$

where  $h = \frac{1}{N} > 0$ . For  $n = 0, \dots, N$ , let

$$\phi_n(x) := \begin{cases} 0, & \text{if } x \leq 0 \text{ or } x \geq x_{n+1}; \\ (x - x_{n-1})/h, & \text{if } x_{n-1} \leq x \leq x_n; \\ (x_{n+1} - x)/h, & \text{if } x_n \leq x \leq x_{n+1}; \\ 0, & \text{if } x \geq 1 \text{ or } x \leq x_{n+1}. \end{cases}$$

What space of functions is spanned by  $\phi_0, \dots, \phi_N$ ? For these functions  $\phi_0, \dots, \phi_N$ , calculate the Gram matrix for the bilinear form

$$a(u, v) := \int_0^1 u'(x)v'(x) \, dx$$

corresponding to the Laplace operator. Determine also the vector components  $\langle f, \phi_n \rangle$  in the Galerkin equation (12.4).

**Exercise 12.4.** Show that, for fixed  $P$ , the Galerkin product satisfies for all  $U, V, W \in \mathcal{S}^P$  and  $\alpha, \beta \in \mathbb{R}$ ,

$$\begin{aligned} U * V &= V * U, \\ (\alpha U) * (\beta V) &= \alpha\beta(U * V), \\ (U + V) * W &= U * W + V * W. \end{aligned}$$

**Exercise 12.5. Galerkin division: WRITE ME!!!**

DRAFT

## Chapter 13

# Non-Intrusive Spectral Methods

[W]hen people thought the Earth was flat, they were wrong. When people thought the Earth was spherical, they were wrong. But if *you* think that thinking the Earth is spherical is *just as wrong* as thinking the Earth is flat, then your view is wronger than both of them put together.

---

*The Relativity of Wrong*  
ISAAC ASIMOV

Chapter 12 considered a spectral approach to UQ, namely Galerkin expansion, that is mathematically very attractive in that it is a natural extension of the Galerkin methods that are commonly used for deterministic PDEs and minimizes the stochastic residual, but has the severe disadvantage that the stochastic modes of the solution are coupled together by a large system such as (12.5). Hence, the Galerkin formalism is not suitable for situations in which deterministic solutions are slow and expensive to obtain, and the deterministic solution method cannot be modified. Many so-called *legacy codes* are not amenable to such *intrusive* methods of UQ.

In contrast, this chapter considers *non-intrusive* spectral methods for UQ. These are characterized by the feature that the solution of the deterministic problem is a ‘black box’ that does not need to be modified for use in the spectral method, beyond being able to be evaluated at any desired point  $\theta$  of the probability space  $(\Theta, \mathcal{F}, \mu)$ .

### 13.1 Pseudo-Spectral Methods

Consider a square-integrable stochastic process  $u: \Theta \rightarrow \mathcal{H}$  taking values in a separable Hilbert space  $\mathcal{H}$ , with a spectral expansion

$$u = \sum_{k \in \mathbb{N}_0} u_k \Psi_k$$

of  $u \in L^2(\Theta, \mu; \mathcal{H}) \cong \mathcal{H} \otimes L^2(\Theta, \mu; \mathbb{R})$  in terms of coefficients (stochastic modes)  $u_k \in \mathcal{H}$  and an orthogonal basis  $\{\Psi_k \mid k \in \mathbb{N}_0\}$  of  $L^2(\Theta, \mu; \mathbb{R})$ . As usual, the stochastic modes are given by

$$\hat{u}_k = \frac{\mathbb{E}_\mu[u \Psi_k]}{\mathbb{E}_\mu[\Psi_k^2]} = \frac{1}{\gamma_k} \int_\Theta u(\theta) \Psi_k(\theta) d\mu(\theta).$$

If the normalization constants  $\gamma_k := \|\Psi_k\|_{L^2(\mu)}^2$  are known ahead of time, then it remains only to approximate the integral with respect to  $\mu$  of the product of  $u$  with each basis function  $\Psi_k$ .

**Deterministic Quadrature.** If the dimension of  $\Theta$  is low and  $u(\theta)$  is relatively smooth as a function of  $\theta$ , then an appealing approach to the estimation of  $\mathbb{E}_\mu[u \Psi_k]$  is deterministic quadrature. For optimal polynomial accuracy, Gaussian quadrature (i.e. nodes at the roots of  $\mu$ -orthogonal polynomials) may be used. In practice, nested quadrature rules such as Clenshaw–Curtis may be preferable since one does not wish to have to discard past solutions of  $u$  upon passing to a more accurate quadrature rule. For multi-dimensional domains of integration  $\Theta$ , sparse quadrature rules may be used to alleviate the curse of dimension.

**Monte Carlo Integration.** If the dimension of  $\Theta$  is high, or  $u(\theta)$  is a non-smooth function of  $\theta$ , then it is tempting to resort to Monte Carlo approximation of  $\mathbb{E}_\mu[u \Psi_k]$ . This approach is also appealing because the calculation of the stochastic modes  $u_k$  can be written as a straightforward (but often large) matrix-matrix multiplication, as in Exercise 13.1. The problem with Monte Carlo methods, as ever, is the slow convergence rate of  $\sim (\text{number of samples})^{-1/2}$ .

**Sources of Error.** In practice, the following sources of error arise when computing pseudo-spectral expansions of this type:

1. *discretization error* comes about through the approximation of  $\mathcal{H}$  by a finite-dimensional subspace, i.e. the approximation of and of the stochastic modes  $u_k$  by a finite sum  $u_k \approx \sum_{i=1}^m u_{k,i} \varphi_i$ , where  $\{\varphi_i \mid i \in \mathbb{N}\}$  is some basis for  $\mathcal{H}$ ;
2. *truncation error* comes about through the truncation of the spectral expansion for  $u$  after finitely many terms, i.e.  $u \approx \sum_{k=0}^K u_k \Psi_k$ ;
3. *quadrature error* comes about through the approximate nature of the numerical integration scheme used to find the stochastic modes.

### 13.2 Stochastic Collocation

Collocation methods for ordinary and partial differential equations are a form of polynomial interpolation. The idea is to find a low-dimensional object —

usually a polynomial — that approximates the true solution to the differential equation by means of *exactly* satisfying the differential equation at a selected set of points, called *collocation points* or *collocation nodes*.

**Example 13.1** (Collocation for an ODE). Consider for example the initial value problem

$$\begin{aligned} \dot{u}(t) &= f(t, u(t)), & \text{for } t \in [a, b] \\ u(a) &= u_a, \end{aligned}$$

to be solved on an interval of time  $[a, b]$ . Choose  $n$  points

$$a \leq t_1 < t_2 < \cdots < t_n \leq b,$$

called *collocation nodes*. Now find a polynomial  $p(t) \in \mathbb{R}_{\leq n}[t]$  so that the ODE

$$\dot{p}(t_k) = f(t_k, p(t_k))$$

is satisfied for  $k = 1, \dots, n$ , as is the initial condition  $p(a) = u_a$ . For example, if  $n = 2$ ,  $t_1 = a$  and  $t_2 = b$ , then the coefficients  $c_2, c_1, c_0 \in \mathbb{R}$  of the polynomial approximation

$$p(t) = \sum_{k=0}^2 c_k (t-a)^k,$$

which has derivative  $\dot{p}(t) = 2c_2(t-a) + c_1$ , are required to satisfy

$$\begin{aligned} \dot{p}(a) &= c_1 = f(a, p(a)) \\ \dot{p}(b) &= 2c_2(b-a) + c_1 = f(b, p(b)) \\ p(a) &= c_0 = u_a \end{aligned}$$

i.e.

$$p(t) = \frac{f(b, p(b)) - f(a, u_a)}{2(b-a)}(t-a)^2 + f(a, u_a)(t-a) + u_a.$$

The above equation implicitly defines the final value  $p(b)$  of the collocation solution. This method is also known as the *trapezoidal rule* for ODEs, since the same solution is obtained by rewriting the differential equation as

$$u(t) = u(a) + \int_a^t f(s, u(s)) \, ds$$

and approximating the integral on the right-hand side by the trapezoidal quadrature rule for integrals.

It should be made clear at the outset that there is nothing stochastic about ‘stochastic collocation’, just as there is nothing chaotic about ‘polynomial chaos’. The meaning of the term ‘stochastic’ in this case is that the collocation principle is being applied across the ‘stochastic space’ (i.e. the probability space) of a stochastic process, rather than the space/time/space-time domain. Consider for example the random PDE

$$\begin{aligned} \mathcal{L}_\theta[u(x, \theta)] &= 0 & \text{for } x \in \Omega, \theta \in \Theta, \\ \mathcal{B}_\theta[u(x, \theta)] &= 0 & \text{for } x \in \partial\Omega, \theta \in \Theta, \end{aligned}$$

where, for  $\mu$ -a.e.  $\theta$  in some probability space  $(\Theta, \mathcal{F}, \mu)$ , the differential operator  $\mathcal{L}_\theta$  and boundary operator  $\mathcal{B}_\theta$  are well-defined and the PDE admits a unique solution  $u(\cdot, \theta): \Omega \rightarrow \mathbb{R}$ . The solution  $u: \Omega \times \Theta \rightarrow \mathbb{R}$  is then a stochastic process. We now let  $\Theta_M = \{\theta^{(1)}, \dots, \theta^{(M)}\} \subseteq \Theta$  be a finite set of prescribed collocation nodes. The collocation problem is to find an approximate solution  $u^{(M)} \approx u$  that satisfies

$$\begin{aligned}\mathcal{L}_{\theta^{(m)}}[u^{(M)}(x, \theta^{(m)})] &= 0 & \text{for } x \in \Omega, \\ \mathcal{B}_{\theta^{(m)}}[u^{(M)}(x, \theta^{(m)})] &= 0 & \text{for } x \in \partial\Omega,\end{aligned}$$

for  $m = 1, \dots, M$ ; there is, however, some flexibility in how to approximate  $u(x, \theta)$  for  $\theta \notin \Theta_M$ .

**Interpolation Approach.** An obvious first approach is to use interpolating polynomials when they are available. This is easiest when the stochastic space  $\Theta$  is one-dimensional.

**Example 13.2.** Consider the initial value problem

$$\frac{d}{dt}u(t, \theta) = -e^\theta u(t, \theta), \quad u(0, \theta) = 1,$$

with  $\theta \sim \mathcal{N}(0, 1)$ . Take as the collocation nodes  $\theta^{(1)}, \dots, \theta^{(M)} \in \mathbb{R}$  the  $M$  roots of the Hermite polynomial  $\text{He}_M$  of degree  $M$ . The collocation solution  $u^{(m)}$  is given at the collocation nodes  $\theta^{(m)}$  by

$$\frac{d}{dt}u^{(m)}(t, \theta^{(m)}) = -e^{\theta^{(m)}} u^{(m)}(t, \theta^{(m)}), \quad u^{(m)}(0, \theta^{(m)}) = 1,$$

i.e.

$$u(t, \theta^{(m)}) = \exp(-e^{\theta^{(m)}} t)$$

Away from the collocation nodes,  $u^{(M)}$  is defined by polynomial interpolation: for each  $t$ ,  $u^{(M)}(t, \theta)$  is a polynomial in  $\theta$  of degree at most  $M$  with prescribed values at the collocation nodes. Writing this interpolation in Lagrange form yields

$$\begin{aligned}u^{(M)}(t, \theta) &= \sum_{m=1}^M u^{(m)}(t, \theta^{(m)}) \ell_m(\theta) \\ &= \sum_{m=1}^M \exp(-e^{\theta^{(m)}} t) \prod_{\substack{1 \leq k \leq M \\ k \neq m}} \frac{\theta - \theta^{(k)}}{\theta^{(m)} - \theta^{(k)}}.\end{aligned}$$

**Tensor Product Collocation.**

**Sparse Grid Collocation.** Sparse grid interpolation using Smolyak–Chebyshev nodes [6]

**Stochastic Collocation on Unstructured Grids.** Stochastic collocation methods for arbitrary unstructured sets of nodes is a notably tricky subject, essentially because it boils down to polynomial interpolation through an unstructured set of nodes, which, as we have seen (**WHERE???**), is generally impossible.



## Bibliography

The monograph of Xiu [117] provides a general introduction to spectral methods for uncertainty quantification, including collocation methods, but is light on proofs. The recent paper of Narayan & Xiu [70] presents a method for stochastic collocation on arbitrary sets of nodes using the framework of least orthogonal interpolation.

Non-intrusive methods for UQ, including NISP and stochastic collocation, are covered in Chapter 3 of Le Maître & Knio [58].

## Exercises

**Exercise 13.1.** Let  $u = (u_1, \dots, u_M) : \Theta \rightarrow \mathbb{R}^M$  be a square-integrable random vector defined over a probability space  $(\Theta, \mathcal{F}, \mu)$ , and let  $\{\Psi_k \mid k \in \mathbb{N}_0\}$ , with normalization constants  $\gamma_k := \|\Psi_k\|_{L^2(\mu)}^2$ , be an orthogonal basis for  $L^2(\Theta, \mu; \mathbb{R})$ . Suppose that  $N$  independent samples  $\{(\theta^{(n)}, u(\theta^{(n)})) \mid n = 1, \dots, N\}$  with  $\theta^{(n)} \sim \mu$ , are given, and it is desired to use these  $N$  Monte Carlo samples to form a truncated pseudo-spectral expansion

$$u \approx u^{(N)} := \sum_{k=0}^K u_k^{(N)} \Psi_k$$

of  $u$ , where the approximate stochastic modes are obtained using Monte Carlo integration. Write down the defining equation for the  $m^{\text{th}}$  component of the  $k^{\text{th}}$  approximate stochastic mode,  $u_{k,m}^{(N)}$ , and hence show that the approximate stochastic modes solve the matrix equation

$$\begin{bmatrix} u_{1,1}^{(N)} & \dots & u_{K,1}^{(N)} \\ \vdots & \ddots & \vdots \\ u_{1,M}^{(N)} & \dots & u_{K,M}^{(N)} \end{bmatrix} = \Gamma^{-1} D P^\top,$$

where

$$\begin{aligned} \Gamma &:= \text{diag}(\gamma_0, \dots, \gamma_K), \\ D &:= \begin{bmatrix} u_1(\theta^{(1)}) & \dots & u_1(\theta^{(N)}) \\ \vdots & \ddots & \vdots \\ u_M(\theta^{(1)}) & \dots & u_M(\theta^{(N)}) \end{bmatrix}, \\ P &:= \begin{bmatrix} \Psi_0(\theta^{(1)}) & \dots & \Psi_1(\theta^{(N)}) \\ \vdots & \ddots & \vdots \\ \Psi_K(\theta^{(1)}) & \dots & \Psi_K(\theta^{(N)}) \end{bmatrix}. \end{aligned}$$

**Exercise 13.2.** What is the analogue of the result of Exercise 13.1 when the integrals are approximated using a quadrature rule, rather than using Monte Carlo?

DRAFT

## Chapter 14

# Distributional Uncertainty

Technology, in common with many other activities, tends toward avoidance of risks by investors. Uncertainty is ruled out if possible. [P]eople generally prefer the predictable. Few recognize how destructive this can be, how it imposes severe limits on variability and thus makes whole populations fatally vulnerable to the shocking ways our universe can throw the dice.

---

*Heretics of Dune*  
FRANK HERBERT

In the previous chapters, it has been assumed that an exact model is available for the probabilistic components of a system, i.e. that all probability distributions involved are known and can be sampled. In practice, however, such assumptions about probability distributions are always wrong to some degree: the distributions used in practice may only be simple approximations of more complicated real ones, or there may be significant uncertainty about what the real distributions actually are. The same is true of uncertainty about the correct form of the forward physical model.

### 14.1 Maximum Entropy Distributions

**Principle of Maximum Entropy.** If all one knows about a probability measure  $\mu$  is that it lies in some set  $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$ , then one should take  $\mu$  to be the element  $\mu^{\text{ME}} \in \mathcal{A}$  of maximum entropy.

((Heuristic justifications... Wallis–Jaynes derivation?))

FINISH ME!!!

**Example 14.1** (Unconstrained maximum entropy distributions). If  $\mathcal{X} = \{1, \dots, m\}$  and  $p \in \mathbb{R}_{>0}^m$  is a probability measure on  $\mathcal{X}$ , then the entropy of  $p$  is

$$H(p) := - \sum_{i=1}^m p_i \log p_i. \quad (14.1)$$

The only constraints on  $p$  are the natural ones that  $p_i \geq 0$  and that  $S(p) := \sum_{i=1}^m p_i = 1$ . Temporarily neglect the inequality constraints and use the method of Lagrange multipliers to find the extrema of  $H(p)$  among all  $p \in \mathbb{R}^m$  with  $S(p) = 1$ ; such  $p$  must satisfy, for some  $\lambda \in \mathbb{R}$ ,

$$0 = \nabla H(p) - \lambda \nabla S(p) = - \begin{bmatrix} 1 + \log p_1 + \lambda \\ \vdots \\ 1 + \log p_m + \lambda \end{bmatrix}$$

It is clear that any solution to this equation must have  $p_1 = \cdots = p_m$ , for if  $p_i$  and  $p_j$  differ, then at most one of  $1 + \log p_i + \lambda$  and  $1 + \log p_j + \lambda$  can equal 0 for the same value of  $\lambda$ . Therefore, since  $S(p) = 1$ , it follows that the unique extremizer of  $H(p)$  among  $\{p \in \mathbb{R}^m \mid S(p) = 1\}$  is  $p_1 = \cdots = p_m = \frac{1}{m}$ . The inequality constraints that were neglected initially are satisfied, and are not active constraints, so it follows that the uniform probability measure on  $\mathcal{X}$  is the unique maximum entropy distribution on  $\mathcal{X}$ .

A similar argument using the calculus of variations shows that the unique maximum entropy probability distribution on an interval  $[a, b] \subseteq \mathbb{R}$  is the uniform distribution  $\frac{1}{|b-a|} dx$ .

**Example 14.2** (Constrained maximum entropy distributions). Consider the set of all probability measures  $\mu$  on  $\mathbb{R}$  that have mean  $m$  and variance  $s^2$ ; what is the maximum entropy distribution in this set? Consider probability measures  $\mu$  that are absolutely continuous with respect to Lebesgue measure, having density  $\rho$ . Then the aim is to find  $\mu$  to maximize

$$H(\mu) = - \int_{\mathbb{R}} \rho(x) \log \rho(x) dx,$$

subject to the constraints that  $\rho \geq 0$ ,  $\int_{\mathbb{R}} \rho(x) dx = 1$ ,  $\int_{\mathbb{R}} x \rho(x) dx = m$  and  $\int_{\mathbb{R}} (x-m)^2 \rho(x) dx = s^2$ . Introduce Lagrange multipliers  $c = (c_0, c_1, c_2)$  and the Lagrangian

$$F_c(\rho) := - \int_{\mathbb{R}} \rho(x) \log \rho(x) dx + c_0 \int_{\mathbb{R}} \rho(x) dx + c_1 \int_{\mathbb{R}} x \rho(x) dx + c_2 \int_{\mathbb{R}} (x-m)^2 \rho(x) dx.$$

Consider a perturbation  $\rho + t\sigma$ ; if  $\rho$  is indeed a critical point of  $F_c$ , then, regardless of  $\sigma$ , it must be true that

$$\left. \frac{d}{dt} F_c(\rho + t\sigma) \right|_{t=0} = 0.$$

This derivative is given by

$$\left. \frac{d}{dt} F_c(\rho + t\sigma) \right|_{t=0} = \int_{\mathbb{R}} \sigma(x) [-\log \rho(x) + c_0 + c_1 x + c_2 (x-m)^2] dx - \int_{\mathbb{R}} \rho(x) dx.$$

Since  $\int_{\mathbb{R}} \rho(x) dx = 1$  and it is required that

$$0 = \int_{\mathbb{R}} [-\log \rho(x) + c_0 - 1 + c_1 x + c_2 (x-m)^2] \sigma(x) dx$$

for every  $\sigma$ , the expression in the brackets must vanish, i.e.

$$\rho(x) = \exp(-c_0 + 1 - c_1x - c_2(x - m)^2).$$

Since  $\rho(x)$  is the exponential of a quadratic form in  $x$ ,  $\mu$  must be a Gaussian of some mean and variance, which, by hypothesis, are  $m$  and  $s^2$  respectively, i.e.

$$\begin{aligned} c_0 &= 1 - \log(1/\sqrt{2\pi s^2}), \\ c_1 &= 0, \\ c_2 &= \frac{1}{2}s^2. \end{aligned}$$

**Discrete Entropy and Convex Programming.** In discrete settings, the entropy of a probability measure  $p \in \mathcal{M}_1(\{1, \dots, m\})$  with respect to the uniform measure as defined in (14.1) is a strictly convex function of  $p \in \mathbb{R}_{>0}^m$ . Therefore, when  $p$  is constrained by a family of convex constraints, finding the maximum entropy distribution is a convex program:

$$\begin{aligned} &\text{minimize: } \sum_{i=1}^m p_i \log p_i \\ &\text{with respect to: } p \in \mathbb{R}^m \\ &\text{subject to: } p \geq 0 \\ &\quad p \cdot \mathbf{1} = 1 \\ &\quad \varphi_i(p) \leq 0 \quad \text{for } i = 1, \dots, n, \end{aligned}$$

for given convex functions  $\varphi_1, \dots, \varphi_n: \mathbb{R}^m \rightarrow \mathbb{R}$ . This is useful because an explicit formula for the maximum entropy distribution, such as in Example 14.2, is rarely available. Therefore, the possibility of efficiently computing the maximum entropy distribution, as in this convex programming situation, is very attractive.

However, it must be remembered that despite the various justifications for the use of the MaxEnt principle, it remains a selection mechanism that in some sense returns a ‘typical’ or ‘representative’ distribution from a given class; what if one is more interested in ‘atypical’ behaviour? This is the topic of the next section.

## 14.2 Distributional Robustness

Suppose that we are interested in the value  $Q(\mu^\dagger)$  of some *quantity of interest* that is a functional of a partially-known probability measure  $\mu^\dagger$  on a space  $\mathcal{X}$ . Very often,  $Q(\mu^\dagger)$  arises as the expected value with respect to  $\mu^\dagger$  of some function  $q: \mathcal{X} \rightarrow \mathbb{R}$ , so the objective is to determine

$$Q(\mu^\dagger) \equiv \mathbb{E}_{X \sim \mu^\dagger} [q(X)].$$

Now suppose that  $\mu^\dagger$  is known only to lie in some subset  $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$ . In the absence of any further information about which  $\mu \in \mathcal{A}$  are more or less likely to be  $\mu^\dagger$ , and particular if the consequences of planning based on an inaccurate estimate of  $Q(\mu^\dagger)$  are very high, it makes sense to adopt a posture of ‘healthy

conservatism' and compute bounds on  $Q(\mu^\dagger)$  that are as tight as justified by the information that  $\mu^\dagger \in \mathcal{A}$ , but no tighter, i.e. to find

$$\underline{Q}(\mathcal{A}) := \inf_{\mu \in \mathcal{A}} Q(\mu) \text{ and } \overline{Q}(\mathcal{A}) := \sup_{\mu \in \mathcal{A}} Q(\mu).$$

When  $Q(\mu)$  is the expected value with respect to  $\mu$  of some function  $q: \mathcal{X} \rightarrow \mathbb{R}$ , the objective is to determine

$$\underline{Q}(\mathcal{A}) := \inf_{\mu \in \mathcal{A}} \mathbb{E}_\mu[q] \text{ and } \overline{Q}(\mathcal{A}) := \sup_{\mu \in \mathcal{A}} \mathbb{E}_\mu[q].$$

The inequality

$$\underline{Q}(\mathcal{A}) \leq Q(\mu^\dagger) \leq \overline{Q}(\mathcal{A})$$

is, by construction, the sharpest possible bound on  $Q(\mu^\dagger)$  given only information that  $\mu^\dagger \in \mathcal{A}$ . The obvious question is, can  $\underline{Q}(\mathcal{A})$  and  $\overline{Q}(\mathcal{A})$  be computed?

**Finite Sample Spaces.** Suppose that the sample space  $\mathcal{X} = \{1, \dots, K\}$  is a finite set equipped with the discrete topology. Then the space of measurable functions  $f: \mathcal{X} \rightarrow \mathbb{R}$  is isomorphic to  $\mathbb{R}^K$  and the space of probability measures  $\mu$  on  $\mathcal{X}$  is isomorphic to the unit simplex in  $\mathbb{R}^K$ . If the available information on  $\mu^\dagger$  is that it lies in the set

$$\mathcal{A} := \{\mu \in \mathcal{M}_1(\mathcal{X}) \mid \mathbb{E}_\mu[\varphi_n] \leq c_n \text{ for } n = 1, \dots, N\}$$

for known measurable functions  $\varphi_1, \dots, \varphi_N: \mathcal{X} \rightarrow \mathbb{R}$  and values  $c_1, \dots, c_N \in \mathbb{R}$ , then the problem of finding the extreme values of  $\mathbb{E}_\mu[q]$  among  $\mu \in \mathcal{A}$  reduces to linear programming:

$$\begin{aligned} &\text{extremize: } p \cdot q \\ &\text{with respect to: } p \in \mathbb{R}^K \\ &\text{subject to: } p \geq 0 \\ &\quad p \cdot 1 = 1 \\ &\quad p \cdot \varphi_n \leq c_n \text{ for } n = 1, \dots, N. \end{aligned}$$

Note that the feasible set  $\mathcal{A}$  for this problem is a convex subset of  $\mathbb{R}^K$ ; indeed,  $\mathcal{A}$  is a *polytope*, i.e. the intersection of finitely many closed half-spaces of  $\mathbb{R}^K$ . Furthermore, as a closed subset of the probability simplex in  $\mathbb{R}^K$ ,  $\mathcal{A}$  is compact. Therefore, by Corollary 4.18, the extreme values of this problem are certain to be found in the extremal set  $\text{ext}(\mathcal{A})$ . This insight can be exploited to great effect in the study of distributional robustness problems for general sample spaces  $\mathcal{X}$ .

Remarkably, when the feasible set  $\mathcal{A}$  of probability measures is sufficiently like a polytope, it is not necessary to consider finite sample spaces. What would appear to be an intractable optimization problem over an infinite-dimensional set of measures is in fact equivalent to a tractable finite-dimensional problem. Thus, the aim of this section is to find a finite-dimensional subset  $\mathcal{A}_\Delta$  of  $\mathcal{A}$  with the property that

$$\text{ext}_{\mu \in \mathcal{A}} Q(\mu) = \text{ext}_{\mu \in \mathcal{A}_\Delta} Q(\mu).$$

To perform this reduction, it is necessary to restrict attention to probability measures, topological spaces, and functionals that are sufficiently well-behaved.

**Extreme Points of Moment Classes.** The first step in this reduction is to classify the extremal measures in sets of probability measures that are prescribed by inequality or equality constraints on the expected value of finitely many arbitrary measurable test functions, so-called *moment classes*. Since, in finite time, we can only verify — even approximately, numerically — the truth of finitely many inequalities, such moment classes are appealing feasible sets from an epistemological point of view because they conform to Karl Popper’s dictum that “Our knowledge can only be finite, while our ignorance must necessarily be infinite.”

**Definition 14.3.** A Borel measure  $\mu$  on a topological space  $\mathcal{X}$  is called *inner regular* if, for every Borel-measurable set  $E \subseteq \mathcal{X}$ ,

$$\mu(E) = \sup\{\mu(K) \mid K \subseteq E \text{ and } K \text{ is compact}\}.$$

A *pseudo-Radon space* is a topological space on which every Borel probability measure is inner regular. A *Radon space* is a separable, metrizable, pseudo-Radon space.

**Examples 14.4.** 1. Lebesgue measure on Euclidean space  $\mathbb{R}^n$  (restricted to the Borel  $\sigma$ -algebra  $\mathcal{B}(\mathbb{R}^n)$ , if pedantry is the order of the day) is an inner regular measure. Similarly, Gaussian measure is an inner regular probability measure on  $\mathbb{R}^n$ . *Proof.* See [MA359 Measure Theory](#).  
 2. However, Lebesgue/Gaussian measures on  $\mathbb{R}$  equipped with the topology of one-sided convergence are not inner regular measures. *Proof.* See Exercise [14.1](#).  
 3. Every Polish space (i.e. every separable and completely metrizable topological space) is a pseudo-Radon space.

Compare the following definition of a barycentre (a centre of mass) for a set of probability measures with the conclusion of the Choquet–Bishop–de Leeuw theorem Theorem [4.13](#):

**Definition 14.5.** A *barycentre* for a set  $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$  is a probability measure  $\mu \in \mathcal{M}_1(\mathcal{X})$  such that there exists  $p \in \mathcal{M}_1(\text{ext}(\mathcal{A}))$  such that

$$\mu(B) = \int_{\text{ext}(\mathcal{A})} \nu(B) dp(\nu) \quad \text{for all measurable } B \subseteq \mathcal{X}. \quad (14.2)$$

**Definition 14.6.** A *Riesz space* (or *vector lattice*) is a vector space  $\mathcal{V}$  together with a partial order  $\leq$  that is

1. (translation invariant) for all  $x, y, z \in \mathcal{V}$ ,  $x \leq y \implies x + z \leq y + z$ ;
2. (positively homogeneous) for all  $x, y \in \mathcal{V}$  and scalars  $\alpha \geq 0$ ,  $x \leq y \implies \alpha x \leq \alpha y$ ;
3. (lattice structure) for all  $x, y \in \mathcal{V}$ , there exists a supremum element  $x \vee y$  that is a least upper bound for  $x$  and  $y$  in the order  $x \leq y$ .

**Definition 14.7.** A subset  $S$  of a vector space  $\mathcal{V}$  is a *Choquet simplex* if the cone

$$C := \{(tx, t) \in \mathcal{V} \times \mathbb{R} \mid t \in \mathbb{R}, t \geq 0, x \in \mathcal{V}\}$$

is such that  $C - C$  is a Riesz space when  $C$  is taken to be the non-negative cone.

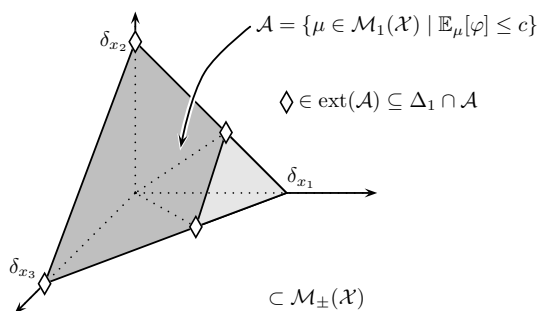


Figure 14.1: Heuristic justification of Winkler's classification of extreme points of moment sets (Theorem 14.8).

The definition of a Choquet simplex extends the usual finite-dimensional definition: a finite-dimensional compact Choquet simplex is a simplex in the ordinary sense of being the closed convex hull of finitely many points.

With these definitions, the extreme points of moment sets of probability measures can be described by the following theorem due to Winkler. The proof is rather technical, and can be found in [116]. The important point is that when  $\mathcal{X}$  is a pseudo-Radon space, Winkler's theorem applies with  $P = \mathcal{M}_1(\mathcal{X})$ , and so the extreme measures in moment classes will simply be finite convex combinations of Dirac measures. Figures like Figure 14.1 should make this an intuitively plausible claim.

**Theorem 14.8 (Winkler).** *Let  $(\mathcal{X}, \mathcal{F})$  be a measurable space and let  $P \subseteq \mathcal{M}_1(\mathcal{F})$  be a Choquet simplex such that  $\text{ext}(P)$  consists of Dirac measures. Fix measurable functions  $\varphi_1, \dots, \varphi_n: \mathcal{X} \rightarrow \mathbb{R}$  and  $c_1, \dots, c_n \in \mathbb{R}$  and let*

$$\mathcal{A} := \left\{ \mu \in P \mid \begin{array}{l} \text{for } i = 1, \dots, n, \\ \varphi_i \in L^1(\mu) \text{ and } \mathbb{E}_\mu[\varphi_i] \leq c_i \end{array} \right\}.$$

*Then  $\mathcal{A}$  is convex and its extremal set satisfies*

$$\text{ext}(\mathcal{A}) \subseteq \mathcal{A}_\Delta := \left\{ \mu \in \mathcal{A} \mid \begin{array}{l} \mu = \sum_{i=1}^m \alpha_i \delta_{x_i}, \\ 1 \leq m \leq n+1, \text{ and} \\ \text{the vectors } (\varphi_1(x_i), \dots, \varphi_n(x_i), 1)_{i=1}^m \\ \text{are linearly independent} \end{array} \right\};$$

*Furthermore, if all the moment conditions defining  $\mathcal{A}$  are given by equalities, then  $\text{ext}(\mathcal{A}) = \mathcal{A}_\Delta$ .*

**Optimization of Measure Affine Functionals.** Having understood the extreme points of moment classes, the next step is to show that the optimization of suitably nice functionals on such classes can be exactly reduced to optimization over the extremal measures in the class.



**Definition 14.9.** For  $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$ , a function  $F: \mathcal{A} \rightarrow \mathbb{R} \cup \{\pm\infty\}$  is said to be *measure affine* if, for all  $\mu \in \mathcal{A}$  and  $p \in \mathcal{M}_1(\text{ext}(\mathcal{A}))$  for which (14.2) holds,  $F$  is  $p$ -integrable with

$$F(\mu) = \int_{\text{ext}(\mathcal{A})} F(\nu) \, dp(\nu). \quad (14.3)$$

As always, the reader should check that the terminology ‘measure affine’ is a sensible choice by verifying that when  $\mathcal{X} = \{1, \dots, K\}$  is a finite sample space, the restriction of any affine function  $F: \mathbb{R}^K \cong \mathcal{M}_\pm(\mathcal{X}) \rightarrow \mathbb{R}$  to a subset  $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$  is a measure affine function in the sense of Definition 14.9.

An important and simple example of a measure affine functional is an evaluation functional, i.e. the integration of a fixed measurable function  $q$ :

**Proposition 14.10.** *If  $q$  is bounded either below or above, then  $\nu \mapsto \mathbb{E}_\nu[q]$  is a measure affine map.*

*Proof.* Exercise 14.2. □

In summary, we now have the following:

**Theorem 14.11.** *Let  $\mathcal{X}$  be a pseudo-Radon space and let  $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$  be a moment class of the form*

$$\mathcal{A} := \{\mu \in \mathcal{M}_1(\mathcal{X}) \mid \mathbb{E}_\mu[\varphi_j] \leq 0 \text{ for } j = 1, \dots, N\}$$

*for prescribed measurable functions  $\varphi_j: \mathcal{X} \rightarrow \mathbb{R}$ . Then the extreme points of  $\mathcal{A}$  are given by*

$$\begin{aligned} \text{ext}(\mathcal{A}) &\subseteq \mathcal{A} \cap \Delta_N(\mathcal{X}) \\ &= \left\{ \mu \in \mathcal{M}_1(\mathcal{A}) \left| \begin{array}{l} \text{for some } \alpha_0, \dots, \alpha_N \in [0, 1], x_0, \dots, x_N \in \mathcal{X}, \\ \mu = \sum_{i=0}^N \alpha_i \delta_{x_i} \\ \sum_{i=0}^N \alpha_i = 1, \\ \text{and } \sum_{i=0}^N \alpha_i \varphi_j(x_i) \leq 0 \text{ for } j = 1, \dots, N \end{array} \right. \right\}. \end{aligned}$$


*Hence, if  $q$  is bounded either below or above, then  $\underline{Q}(\mathcal{A}) = \underline{Q}(\mathcal{A}_\Delta)$  and  $\overline{Q}(\mathcal{A}) = \overline{Q}(\mathcal{A}_\Delta)$ .*

*Proof.* The classification of  $\text{ext}(\mathcal{A})$  is given by Winkler’s theorem (Theorem 14.8). Since  $q$  is bounded on at least one side, Proposition 14.10 implies that  $\mu \mapsto F(\mu) := \mathbb{E}_\mu[q]$  is measure affine. Let  $\mu \in \mathcal{A}$  be arbitrary and choose a probability measure  $p \in \mathcal{M}_1(\text{ext}(\mathcal{A}))$  with barycentre  $\mu$ . Then, it follows from the barycentric formula (14.3) that

$$F(\mu) \leq \sup\{F(\nu) \mid \nu \in \text{ext}(\mathcal{A})\}.$$

This proves the claim for maximization; minimization is similar. □

The kinds of constraints on measures (or, if you prefer, random variables) that can be considered in Theorem 14.11 include values for or bounds on functions of one or more of those random variables, e.g. the mean of  $X_1$ , the variance of  $X_2$ , the covariance of  $X_3$  and  $X_4$ . However, one type of information that is not of this type is that  $X_5$  and  $X_6$  are independent random variables, i.e. that their joint law is a product measure. The problem here is that sets of product

measures can fail to be convex (see Exercise 14.3), so the reduction to extreme points cannot be applied directly. Fortunately, a cunning application of Fubini's theorem resolves this difficulty. Note well, though, that unlike Theorem 14.11, Theorem 14.12 does *not* say that  $\mathcal{A}_\Delta = \text{ext}(\mathcal{A})$ ; it only says that the optimization problem has the same extreme values over  $\mathcal{A}_\Delta$  and  $\mathcal{A}$ . 

**Theorem 14.12.** *Let  $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$  be a moment class of the form*

$$\mathcal{A} := \left\{ \mu = \bigotimes_{k=1}^K \mu_k \in \bigotimes_{k=1}^K \mathcal{M}_1(\mathcal{X}_k) \left| \begin{array}{l} \mathbb{E}_\mu[\varphi_j] \leq 0 \text{ for } j = 1, \dots, N, \\ \mathbb{E}_{\mu_1}[\varphi_{1j}] \leq 0 \text{ for } j = 1, \dots, N_1, \\ \vdots \\ \mathbb{E}_{\mu_K}[\varphi_{Kj}] \leq 0 \text{ for } j = 1, \dots, N_K \end{array} \right. \right\}$$

for prescribed measurable functions  $\varphi_j: \mathcal{X} \rightarrow \mathbb{R}$  and  $\varphi_{kj}: \mathcal{X}_k \rightarrow \mathbb{R}$ . Let

$$\mathcal{A}_\Delta := \{\mu \in \mathcal{A} \mid \mu_k \in \Delta_{N+N_k}(\mathcal{X}_k)\}.$$

Then, if  $q$  is bounded either above or below,  $\underline{Q}(\mathcal{A}) = \underline{Q}(\mathcal{A}_\Delta)$  and  $\overline{Q}(\mathcal{A}) = \overline{Q}(\mathcal{A}_\Delta)$ .

*Proof.* Let  $\varepsilon > 0$  and let  $\mu^* \in \mathcal{A}$  be  $\frac{\varepsilon}{K+1}$ -suboptimal for the maximization of  $\mu \mapsto \mathbb{E}_\mu[q]$  over  $\mu \in \mathcal{A}$ , i.e.

$$\mathbb{E}_{\mu^*}[q] \geq \sup_{\mu \in \mathcal{A}} \mathbb{E}_\mu[q] - \frac{\varepsilon}{K+1}.$$

By Fubini's theorem,

$$\mathbb{E}_{\mu_1^* \otimes \dots \otimes \mu_K^*}[q] = \mathbb{E}_{\mu_1^*}[\mathbb{E}_{\mu_2^* \otimes \dots \otimes \mu_K^*}[q]]$$

By the same arguments used in the proof of Theorem 14.11,  $\mu_1^*$  can be replaced by some probability measure  $\nu_1 \in \mathcal{M}_1(\mathcal{X}_1)$  with support on at most  $N + N_1$  points, such that  $\nu_1 \otimes \mu_2^* \otimes \dots \otimes \mu_K^* \in \mathcal{A}$ , and

$$\mathbb{E}_{\nu_1}[\mathbb{E}_{\mu_2^* \otimes \dots \otimes \mu_K^*}[q]] \geq \mathbb{E}_{\mu_1^*}[\mathbb{E}_{\mu_2^* \otimes \dots \otimes \mu_K^*}[q]] - \frac{\varepsilon}{K+1} \geq \sup_{\mu \in \mathcal{A}} \mathbb{E}_\mu[q] - \frac{2\varepsilon}{K+1}.$$

Repeating this argument a further  $K-1$  times yields  $\nu = \bigotimes_{k=1}^K \nu_k \in \mathcal{A}_\Delta$  such that

$$\mathbb{E}_\nu[q] \geq \sup_{\mu \in \mathcal{A}} \mathbb{E}_\mu[q] - \varepsilon.$$

Since  $\varepsilon > 0$  was arbitrary, it follows that

$$\sup_{\mu \in \mathcal{A}_\Delta} \mathbb{E}_\mu[q] = \sup_{\mu \in \mathcal{A}} \mathbb{E}_\mu[q].$$

The proof for the infimum is similar. □

## 14.3 Functional and Distributional Robustness

In addition to epistemic uncertainty about probability measures, applications often feature epistemic uncertainty about the functions involved. For example, if the system of interest is in reality some function  $g^\dagger$  from a space  $\mathcal{X}$  of inputs to

another space  $\mathcal{Y}$  of outputs, it may only be known that  $g^\dagger$  lies in some subset  $\mathcal{G}$  of the set of all (measurable) functions from  $\mathcal{X}$  to  $\mathcal{Y}$ ; furthermore, our information about  $g^\dagger$  and our information about  $\mu^\dagger$  may be coupled in some way, e.g. by knowledge of  $\mathbb{E}_{X \sim \mu^\dagger}[g^\dagger(X)]$ . Therefore, we now consider admissible sets of the form

$$\mathcal{A} \subseteq \left\{ (g, \mu) \mid \begin{array}{l} g: \mathcal{X} \rightarrow \mathcal{Y} \text{ is measurable} \\ \text{and } \mu \in \mathcal{M}_1(\mathcal{X}) \end{array} \right\},$$

quantities of interest of the form  $Q(g, \mu) = \mathbb{E}_{X \sim \mu}[q(X, g(X))]$  for some measurable function  $q: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ , and seek the extreme values

$$\underline{Q}(\mathcal{A}) := \inf_{(g, \mu) \in \mathcal{A}} \mathbb{E}_{X \sim \mu}[q(X, g(X))] \text{ and } \overline{Q}(\mathcal{A}) := \sup_{(g, \mu) \in \mathcal{A}} \mathbb{E}_{X \sim \mu}[q(X, g(X))].$$

Obviously, if for each  $g: \mathcal{X} \rightarrow \mathcal{Y}$  the set of  $\mu \in \mathcal{M}_1(\mathcal{X})$  such that  $(g, \mu) \in \mathcal{A}$  is a moment class of the form considered in Theorem 14.12, then

$$\begin{aligned} \text{ext}_{(g, \mu) \in \mathcal{A}} \mathbb{E}_{X \sim \mu}[q(X, g(X))] &= \text{ext}_{\substack{(g, \mu) \in \mathcal{A} \\ \mu \in \bigotimes_{k=1}^K \Delta_{N+N_k}(\mathcal{X}_k)}} \mathbb{E}_{X \sim \mu}[q(X, g(X))]. \end{aligned}$$

In principle, though, although the search over  $\mu$  is finite-dimensional for each  $g$ , the search over  $g$  is still infinite-dimensional. However, the passage to discrete measures often enables us to finite-dimensionalize the search over  $g$ , since, in some sense, only the values of  $g$  on the finite set  $\text{supp}(\mu)$  ‘matter’ in computing  $\mathbb{E}_{X \sim \mu}[q(X, g(X))]$ .

The idea is quite simple: instead of optimizing with respect to  $g \in \mathcal{G}$ , we optimize with respect to the finite-dimensional vector  $y = g|_{\text{supp}(\mu)}$ . However, this reduction step requires some care:

1. Some ‘functions’ do not have their values defined pointwise, e.g. ‘functions’ in Lebesgue and Sobolev spaces, which are actually equivalence classes of functions modulo equality almost everywhere. If isolated points have measure zero, then it makes no sense to restrict such ‘functions’ to a finite set  $\text{supp}(\mu)$ . These difficulties are circumvented by insisting that  $\mathcal{G}$  be a space of functions with pointwise-defined values.
2. In the other direction, it is sometimes difficult to verify whether a vector  $y$  indeed arises as the restriction to  $\text{supp}(\mu)$  of some  $g \in \mathcal{G}$ ; we need functions that can be extended from  $\text{supp}(\mu)$  to all of  $\mathcal{X}$ . Suitable extension properties are ensured if we restrict attention to smooth enough functions between the right kinds of spaces.

**Theorem 14.13 (Minty’s extension theorem).** *Let  $(\mathcal{X}, d)$  be a metric space, let  $\mathcal{H}$  be a Hilbert space, let  $E \subseteq \mathcal{H}$ , and suppose that  $f: E \rightarrow \mathcal{H}$  satisfies*

$$\|f(x) - f(y)\|_{\mathcal{H}} \leq d(x, y)^\alpha \quad \text{for all } x, y \in E \quad (14.4)$$

*with Hölder constant  $0 < \alpha \leq 1$ . Then there exists  $F: \mathcal{X} \rightarrow \mathcal{H}$  such that  $F|_E = f$  and (14.4) holds for all  $x, y \in \mathcal{X}$  if either  $\alpha \leq \frac{1}{2}$  or if  $\mathcal{X}$  is an inner product space with metric given by  $d(x, y) = k^{1/\alpha} \|x - y\|$  for some  $k > 0$ . Furthermore, the extension can be performed so that  $F(\mathcal{X}) \subseteq \overline{\text{co}}(f(E))$ , and hence without increasing the Hölder norm*

$$\sup_x \|f(x)\|_{\mathcal{H}} + \sup_{x \neq y} \frac{\|f(x) - f(y)\|_{\mathcal{H}}}{d(x, y)^\alpha}.$$

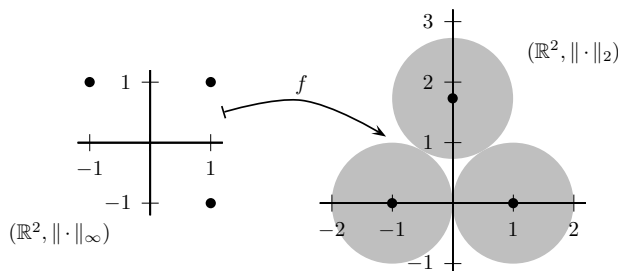


Figure 14.1: The function  $f$  that takes the three points on the left (equipped with  $\|\cdot\|_\infty$ ) to the three points on the right (equipped with  $\|\cdot\|_2$ ) has Lipschitz constant 1, but has no 1-Lipschitz extension  $F$  to  $(0,0)$ , let alone the whole plane  $\mathbb{R}^2$ , since  $F((0,0))$  would have to lie in the (empty) intersection of the three grey discs. Cf. Example 14.14.

Special cases of Minty's theorem include the Kirszbraun–Valentine theorem (which assures that Lipschitz functions between Hilbert spaces can be extended without increasing the Lipschitz constant) and McShane's theorem (which assures that scalar-valued continuous functions on metric spaces can be extended without increasing a prescribed convex modulus of continuity). However, the extensibility property fails for Lipschitz functions between Banach spaces, even finite-dimensional ones:

**Example 14.14.** Let  $E \subseteq \mathbb{R}^2$  be given by  $E := \{(1, -1), (-1, 1), (1, 1)\}$  and define  $f: E \rightarrow \mathbb{R}^2$  by

$$f((1, -1)) := (1, 0), \quad f((-1, 1)) := (-1, 0), \quad \text{and} \quad f((1, 1)) := (0, \sqrt{3}).$$

Suppose that we wish to extend this  $f$  to  $F: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ , where  $E$  and the domain copy of  $\mathbb{R}^2$  are given the metric arising from the maximum norm  $\|\cdot\|_\infty$  and the range copy of  $\mathbb{R}^2$  is given the metric arising from the Euclidean norm  $\|\cdot\|_2$ . Then, for all distinct  $x, y \in E$ ,

$$\|x - y\|_\infty = 2 = \|f(x) - f(y)\|_2,$$

so  $f$  has Lipschitz constant 1 on  $E$ . What value should  $F$  take at the origin,  $(0,0)$ , if it is to have Lipschitz constant at most 1? Since  $(0,0)$  lies at  $\|\cdot\|_\infty$ -distance 1 from all three points of  $E$ ,  $F((0,0))$  must lie within  $\|\cdot\|_2$ -distance 1 of all three points of  $f(E)$ . However, there is no such point of  $\mathbb{R}^2$  within distance 1 of all three points of  $f(E)$ , and hence any extension of  $f$  to  $F: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  must have  $\text{Lip}(F) > 1$ . See Figure 14.1.

**Theorem 14.15.** Let  $\mathcal{G}$  be a collection of measurable functions from  $\mathcal{X}$  to  $\mathcal{Y}$  such that, for every finite subset  $E \subseteq \mathcal{X}$  and  $g: E \rightarrow \mathcal{Y}$ , it is possible to determine whether or not  $g$  can be extended to an element of  $\mathcal{G}$ . Let  $\mathcal{A} \subseteq \mathcal{G} \times \mathcal{M}_1(\mathcal{X})$  be such that, for each  $g \in \mathcal{G}$ , the set of  $\mu \in \mathcal{M}_1(\mathcal{X})$  such that  $(g, \mu) \in \mathcal{A}$  is a

moment class of the form considered in Theorem 14.12. Let

$$\mathcal{A}_\Delta := \left\{ (y, \mu) \left| \begin{array}{l} \mu \in \bigotimes_{k=1}^K \Delta_{N+N_k}(\mathcal{X}_k), \\ y \text{ is the restriction to } \text{supp}(\mu) \text{ of some } g \in \mathcal{G}, \\ \text{and } (g, \mu) \in \mathcal{A} \end{array} \right. \right\}.$$

Then, if  $q$  is bounded either above or below,  $\underline{Q}(\mathcal{A}) = \underline{Q}(\mathcal{A}_\Delta)$  and  $\overline{Q}(\mathcal{A}) = \overline{Q}(\mathcal{A}_\Delta)$ .

*Proof.* Exercise 14.6.  $\square$

**Example 14.16.** Suppose that  $g^\dagger: [-1, 1] \rightarrow \mathbb{R}$  is known to have Lipschitz constant  $\text{Lip}(g^\dagger) \leq L$ . Suppose also that the inputs of  $g^\dagger$  are distributed according to  $\mu^\dagger \in \mathcal{M}_1([-1, 1])$ , and it is known that

$$\mathbb{E}_{X \sim \mu^\dagger}[X] = 0 \quad \text{and} \quad \mathbb{E}_{X \sim \mu^\dagger}[g^\dagger(X)] \geq m > 0.$$

Hence, the corresponding feasible set is

$$\mathcal{A} = \left\{ (g, \mu) \left| \begin{array}{l} g: [-1, 1] \rightarrow \mathbb{R} \text{ has Lipschitz constant } \leq L, \\ \mu \in \mathcal{M}_1([-1, 1]), \mathbb{E}_{X \sim \mu}[X] = 0, \text{ and } \mathbb{E}_{X \sim \mu}[g(X)] \geq m \end{array} \right. \right\}.$$

Suppose that our quantity of interest is the probability of output values below 0, i.e.  $q(x, y) = \mathbb{1}[y \leq 0]$ . Then Theorem 14.15 ensures that the extreme values of

$$Q(g, \mu) = \mathbb{E}_{X \sim \mu}[\mathbb{1}[g(X) \leq 0]] = \mathbb{P}_{X \sim \mu}[g(X) \leq 0]$$

are the solutions of

$$\begin{aligned} & \text{extremize: } \sum_{i=0}^2 w_i \mathbb{1}[y_i \leq 0] \\ & \text{with respect to: } w_0, w_1, w_2 \geq 0 \\ & \quad x_0, x_1, x_2 \in [-1, 1] \\ & \quad y_0, y_1, y_2 \in \mathbb{R} \\ & \text{subject to: } \sum_{i=0}^2 w_i = 1 \\ & \quad \sum_{i=0}^2 w_i x_i = 0 \\ & \quad \sum_{i=0}^2 w_i y_i \geq m \\ & \quad |y_i - y_j| \leq L|x_i - x_j| \text{ for } i, j \in \{0, 1, 2\}. \end{aligned}$$

**Example 14.17 (McDiarmid).** The following admissible set corresponds to the assumptions of McDiarmid's inequality (Theorem 10.5):

$$\mathcal{A} = \left\{ (g, \mu) \left| \begin{array}{l} g: \mathcal{X} \rightarrow \mathbb{R} \text{ has } \mathcal{D}_k[g] \leq D_k, \\ \mu = \bigotimes_{k=1}^K \mu_k \in \mathcal{M}_1(\mathcal{X}), \\ \text{and } \mathbb{E}_{X \sim \mu}[g(X)] = m \end{array} \right. \right\}.$$

McDiarmid's inequality was the upper bound

$$\overline{Q}(\mathcal{A}) := \sup_{(g, \mu) \in \mathcal{A}} \mathbb{P}_\mu[g(X) \leq 0] \leq \exp \left( -\frac{2 \max\{0, m\}^2}{\sum_{k=1}^K D_k^2} \right).$$

Perhaps not surprisingly given its general form, McDiarmid's inequality is not the *least* upper bound on  $\mathbb{P}_\mu[g(X) \leq 0]$ ; the actual least upper bound can be calculated using the reduction theorems.

**FINISH ME!!!**

## Bibliography

Berger [8] makes the case for distributional robustness, with respect to priors and likelihoods, in Bayesian inference. Smith [89] provides theory and several practical examples for generalized Chebyshev inequalities in decision analysis. Boyd & Vandenberghe [16, Sec. 7.2] cover some aspects of distributional robustness under the heading of nonparametric distribution estimation, in the case in which it is a convex problem. Convex optimization approaches to distributional robustness and optimal probability inequalities are also considered by Bertsimas & Popescu [9]. There is also an extensive literature on the related topic of majorization, for which see the book of Marshall & al. [64].

The classification of the extreme points of moment sets and the consequences for the optimization of measure affine functionals are due to von Weizsäcker & Winkler [112, 113] and Winkler [116]. Karr [49] proved similar results under additional topological and continuity assumptions. Theorem 14.12 and the Lipschitz version of Theorem 14.15 can be found in Owhadi & al. [76] and Sullivan & al. [99] respectively. Extension Theorem 14.13 is due to Minty [68], and generalizes earlier results by McShane [66], Kirszbraun [51] and Valentine [111]. Example 14.14 is taken from the example given on p. 202 of Federer [28].

## Exercises

**Exercise 14.1.** Consider the topology  $\mathcal{T}$  on  $\mathbb{R}$  generated by the basis of open sets  $[a, b)$ , where  $a, b \in \mathbb{R}$ .

1. Show that this topology generates the same  $\sigma$ -algebra on  $\mathbb{R}$  as the usual Euclidean topology does. Hence, show that Gaussian measure is a well-defined probability measure on the Borel  $\sigma$ -algebra of  $(\mathbb{R}, \mathcal{T})$ .
2. Show that every compact subset of  $(\mathbb{R}, \mathcal{T})$  is a countable set.
3. Conclude that Gaussian measure on  $(\mathbb{R}, \mathcal{T})$  is not inner regular and that  $(\mathbb{R}, \mathcal{T})$  is not a pseudo-Radon space.

**Exercise 14.2.** Suppose that  $\mathcal{A}$  is a moment class of probability measures on  $\mathcal{X}$  and that  $q: \mathcal{X} \rightarrow \mathbb{R} \cup \{\pm\infty\}$  is bounded either below or above. Show that  $\mu \mapsto \mathbb{E}_\mu[q]$  is a measure affine map. *Hint: verify the assertion for the case in which  $q$  is the indicator function of a measurable set; extend it to bounded measurable functions using the Monotone Class Theorem; for non-negative  $\mu$ -integrable functions  $q$ , use monotone convergence to verify the barycentric formula.*

**Exercise 14.3.** Let  $\lambda$  denote uniform measure on the unit interval  $[0, 1] \subseteq \mathbb{R}$ . Show that the line segment in  $\mathcal{M}_1([0, 1]^2)$  joining the measures  $\lambda \otimes \delta_0$  and  $\delta_0 \otimes \lambda$  contains measures that are not product measures. Hence show that a set  $\mathcal{A}$  of product probability measures like that considered in Theorem 14.12 is typically not convex.

**Exercise 14.4.** Calculate by hand, as a function of  $t \in \mathbb{R}$ ,  $D \geq 0$  and  $m \in \mathbb{R}$ ,

$$\sup_{\mu \in \mathcal{A}} \mathbb{P}_{X \sim \mu}[X \leq t],$$

where

$$\mathcal{A} := \left\{ \mu \in \mathcal{M}_1(\mathbb{R}) \mid \begin{array}{l} \mathbb{E}_{X \sim \mu}[X] \geq m, \text{ and} \\ \text{diam}(\text{supp}(\mu)) \leq D \end{array} \right\}.$$

**Exercise 14.5.** Calculate by hand, as a function of  $t \in \mathbb{R}$ ,  $s \geq 0$  and  $m \in \mathbb{R}$ ,

$$\sup_{\mu \in \mathcal{A}} \mathbb{P}_{X \sim \mu}[X - m \geq st],$$

and

$$\sup_{\mu \in \mathcal{A}} \mathbb{P}_{X \sim \mu}[|X - m| \geq st],$$

where

$$\mathcal{A} := \left\{ \mu \in \mathcal{M}_1(\mathbb{R}) \mid \begin{array}{l} \mathbb{E}_{X \sim \mu}[X] \leq m, \text{ and} \\ \mathbb{E}_{X \sim \mu}[|X - m|^2] \leq s^2 \end{array} \right\}.$$

**Exercise 14.6.** Prove Theorem 14.15.

**Exercise 14.7.** Calculate by hand, as a function of  $t \in \mathbb{R}$ ,  $m \in \mathbb{R}$ ,  $z \in [0, 1]$  and  $v \in \mathbb{R}$ ,

$$\sup_{(g, \mu) \in \mathcal{A}} \mathbb{P}_{X \sim \mu}[g(X) \leq t],$$

where

$$\mathcal{A} := \left\{ (g, \mu) \mid \begin{array}{l} g: [0, 1] \rightarrow \mathbb{R} \text{ has Lipschitz constant } 1, \\ \mu \in \mathcal{M}_1([0, 1]), \mathbb{E}_{X \sim \mu}[g(X)] \geq m, \\ \text{and } g(z) = v \end{array} \right\}.$$

Numerically verify your calculations.

DRAFT



## **Bibliography and Index**

DRAFT

# Bibliography

- [1] M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications Inc., New York, 1992. Reprint of the 1972 edition.
- [2] C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, Berlin, third edition, 2006.
- [3] Ö. F. Alış and H. Rabitz. Efficient implementation of high dimensional model representations. *J. Math. Chem.*, 29(2):127–142, 2001.
- [4] M. Atiyah. *Collected Works. Vol. 6*. Oxford Science Publications. The Clarendon Press Oxford University Press, New York, 2004.
- [5] K. Azuma. Weighted sums of certain dependent random variables. *Tōhoku Math. J. (2)*, 19:357–367, 1967.
- [6] V. Barthelmann, E. Novak, and K. Ritter. High dimensional polynomial interpolation on sparse grids. *Adv. Comput. Math.*, 12(4):273–288, 2000.
- [7] F. Beccacece and E. Borgonovo. Functional ANOVA, ultramodularity and monotonicity: applications in multiattribute utility theory. *European J. Oper. Res.*, 210(2):326–335, 2011.
- [8] J. O. Berger. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994. With comments and a rejoinder by the author.
- [9] D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: a convex optimization approach. *SIAM J. Optim.*, 15(3):780–804 (electronic), 2005.
- [10] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, third edition, 1995. A Wiley-Interscience Publication.
- [11] E. Bishop and K. de Leeuw. The representations of linear functionals by measures on sets of extreme points. *Ann. Inst. Fourier. Grenoble*, 9:305–331, 1959.
- [12] S. Bochner. Integration von Funktionen, deren Werte die Elemente eines Vectorraumes sind. *Fund. Math.*, 20:262–276, 1933.
- [13] G. Boole. *An Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities*. Walton and Maberley, London, 1854.

- [14] N. Bourbaki. *Topological Vector Spaces. Chapters 1–5*. Elements of Mathematics (Berlin). Springer-Verlag, Berlin, 1987. Translated from the French by H. G. Eggleston and S. Madan.
- [15] N. Bourbaki. *Integration. I. Chapters 1–6*. Elements of Mathematics (Berlin). Springer-Verlag, Berlin, 2004. Translated from the 1959, 1965 and 1967 French originals by Sterling K. Berberian.
- [16] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.
- [17] R. H. Cameron and W. T. Martin. The orthogonal development of non-linear functionals in series of Fourier–Hermite functionals. *Ann. of Math. (2)*, 48:385–392, 1947.
- [18] M. Capiński and E. Kopp. *Measure, Integral and Probability*. Springer Undergraduate Mathematics Series. Springer-Verlag London Ltd., London, second edition, 2004.
- [19] C. W. Clenshaw and A. R. Curtis. A method for numerical integration on an automatic computer. *Numer. Math.*, 2:197–205, 1960.
- [20] S. L. Cotter, M. Dashti, J. C. Robinson, and A. M. Stuart. Bayesian inverse problems for functions and applications to fluid mechanics. *Inverse Problems*, 25(11):115008, 43, 2009.
- [21] S. L. Cotter, M. Dashti, and A. M. Stuart. Approximation of Bayesian inverse problems for PDEs. *SIAM J. Numer. Anal.*, 48(1):322–345, 2010.
- [22] M. Dashti, S. Harris, and A. Stuart. Besov priors for Bayesian inverse problems. *Inverse Probl. Imaging*, 6(2):183–200, 2012.
- [23] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.*, 38:325–339, 1967.
- [24] J. Diestel and J. J. Uhl, Jr. *Vector Measures*. Number 15 in Mathematical Surveys. American Mathematical Society, Providence, R.I., 1977.
- [25] R. M. Dudley. *Real Analysis and Probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.
- [26] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.
- [27] L. C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.
- [28] H. Federer. *Geometric Measure Theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York, 1969.

- [29] L. Fejér. On the infinite sequences arising in the theories of harmonic analysis, of interpolation, and of mechanical quadratures. *Bull. Amer. Math. Soc.*, 39(8):521–534, 1933.
- [30] J. Feldman. Equivalence and perpendicularity of Gaussian processes. *Pacific J. Math.*, 8:699–708, 1958.
- [31] X. Fernique. Intégrabilité des vecteurs gaussiens. *C. R. Acad. Sci. Paris Sér. A-B*, 270:A1698–A1699, 1970.
- [32] R. A. Fisher and W. A. Mackenzie. The manurial response of different potato varieties. *J. Agric. Sci.*, 13:311–320, 1923.
- [33] W. Gautschi. *Orthogonal Polynomials: Computation and Approximation*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2004. Oxford Science Publications.
- [34] R. G. Ghanem and P. D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer-Verlag, New York, 1991.
- [35] R. A. Gordon. *The Integrals of Lebesgue, Denjoy, Perron, and Henstock*, volume 4 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1994.
- [36] J. Hájek. On a property of normal distribution of any stochastic process. *Czechoslovak Math. J.*, 8 (83):610–618, 1958.
- [37] J. H. Halton. On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numer. Math.*, 2:84–90, 1960.
- [38] W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statistics*, 19:293–325, 1948.
- [39] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.
- [40] G. Hooker. Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *J. Comput. Graph. Statist.*, 16(3):709–732, 2007.
- [41] J. Humpherys, P. Redd, and J. West. A Fresh Look at the Kalman Filter. *SIAM Rev.*, 54(4):801–823, 2012.
- [42] L. Jaulin, M. Kieffer, O. Didrit, and É. Walter. *Applied Interval Analysis: With Examples in Parameter and State Estimation, Robust Control and Robotics*. Springer-Verlag London Ltd., London, 2001.
- [43] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*, volume 64 of *Mathematics in Science and Engineering*. Academic Press, New York, 1970.
- [44] V. M. Kadets. Non-differentiable indefinite Pettis integrals. *Quaestiones Math.*, 17(2):137–139, 1994.

- [45] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*, volume 160 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2005.
- [46] R. E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME Ser. D. J. Basic Engrg.*, 82:35–45, 1960.
- [47] R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Trans. ASME Ser. D. J. Basic Engrg.*, 83:95–108, 1961.
- [48] K. Karhunen. Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys.*, 1947(37):79, 1947.
- [49] Alan F. Karr. Extreme points of certain sets of probability measures, with applications. *Math. Oper. Res.*, 8(1):74–85, 1983.
- [50] J. M. Keynes. *A Treatise on Probability*. Macmillan and Co., London, 1921.
- [51] M. D. Kirszbraun. Über die zusammenziehende und Lipschitzsche Transformationen. *Fund. Math.*, 22:77–108, 1934.
- [52] D. D. Kosambi. Statistics in function space. *J. Indian Math. Soc. (N.S.)*, 7:76–88, 1943.
- [53] H. Kozono and T. Yanagisawa. Generalized Lax–Milgram theorem in Banach spaces and its application to the elliptic system of boundary value problems. *Manuscripta Math.*, 141(3-4):637–662, 2013.
- [54] M. Krein and D. Milman. On extreme points of regular convex sets. *Studia Math.*, 9:133–138, 1940.
- [55] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statistics*, 22:79–86, 1951.
- [56] V. P. Kuznetsov. *Intervalnye statisticheskie modeli*. “Radio i Svyaz”, Moscow, 1991.
- [57] Matti Lassas, Eero Saksman, and Samuli Siltanen. Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Probl. Imaging*, 3(1):87–122, 2009.
- [58] O. P. Le Maître and O. M. Knio. *Spectral Methods for Uncertainty Quantification: With applications to computational fluid dynamics*. Scientific Computation. Springer, New York, 2010.
- [59] M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.
- [60] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. Isoperimetry and Processes, Reprint of the 1991 edition.

- [61] P. Lévy. *Problèmes Concrets d'Analyse Fonctionnelle. Avec un Complément sur les Fonctionnelles Analytiques par F. Pellegrino.* Gauthier-Villars, Paris, 1951. 2d ed.
- [62] M. Loève. *Probability Theory. II.* Springer-Verlag, New York, fourth edition, 1978. Graduate Texts in Mathematics, Vol. 46.
- [63] D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms.* Cambridge University Press, New York, 2003.
- [64] A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: Theory of Majorization and its Applications.* Springer Series in Statistics. Springer, New York, second edition, 2011.
- [65] C. McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Math. Soc. Lecture Note Ser.*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.
- [66] E. J. McShane. Extension of range of functions. *Bull. Amer. Math. Soc.*, 40(12):837–842, 1934.
- [67] J. Mikusiński. *The Bochner Integral.* Birkhäuser Verlag, Basel, 1978. Lehrbücher und Monographien aus dem Gebiete der exakten Wissenschaften, Mathematische Reihe, Band 55.
- [68] G. J. Minty. On the extension of Lipschitz, Lipschitz–Hölder continuous, and monotone functions. *Bull. Amer. Math. Soc.*, 76:334–339, 1970.
- [69] R. E. Moore. *Interval Analysis.* Prentice-Hall Inc., Englewood Cliffs, N.J., 1966.
- [70] A. Narayan and D. Xiu. Stochastic collocation methods on unstructured grids in high dimensions via interpolation. *SIAM J. Sci. Comput.*, 34(3):A1729–A1752, 2012.
- [71] H. Niederreiter. Low-discrepancy and low-dispersion sequences. *J. Number Theory*, 30(1):51–70, 1988.
- [72] J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.
- [73] B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications.* Universitext. Springer-Verlag, Berlin, sixth edition, 2003.
- [74] N. Oreskes, K. Shrader-Frechette, and K. Belitz. Verification, validation, and confirmation of numerical models in the earth sciences. *Science*.
- [75] A. B. Owen. Latin supercube sampling for very high dimensional simulations. *ACM Trans. Mod. Comp. Sim.*, 8(2):71–102, 1998.
- [76] H. Owhadi, C. Scovel, T. J. Sullivan, M. McKerns, and M. Ortiz. Optimal Uncertainty Quantification. *SIAM Rev.*, 55(2):271–345, 2013.

- [77] K. V. Price, R. M. Storn, and J. A. Lampinen. *Differential Evolution: A Practical Approach to Global Optimization*. Natural Computing Series. Springer-Verlag, Berlin, 2005.
- [78] H. Rabitz and Ö. F. Alış. General foundations of high-dimensional model representations. *J. Math. Chem.*, 25(2-3):197–233, 1999.
- [79] M. Reed and B. Simon. *Methods of Modern Mathematical Physics. I. Functional Analysis*. Academic Press, New York, 1972.
- [80] M. Renardy and R. C. Rogers. *An Introduction to Partial Differential Equations*, volume 13 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2004.
- [81] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004.
- [82] R. T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.
- [83] W. Rudin. *Functional Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill Inc., New York, second edition, 1991.
- [84] C. Runge. Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten. *Zeitschrift für Mathematik und Physik*, 46:224–243, 1901.
- [85] R. A. Ryan. *Introduction to Tensor Products of Banach Spaces*. Springer Monographs in Mathematics. Springer-Verlag London Ltd., London, 2002.
- [86] B. P. Rynne and M. A. Youngson. *Linear Functional Analysis*. Springer Undergraduate Mathematics Series. Springer-Verlag London Ltd., London, second edition, 2008.
- [87] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J., 1976.
- [88] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656, 1948.
- [89] J. E. Smith. Generalized Chebychev inequalities: theory and applications in decision analysis. *Oper. Res.*, 43(5):807–825, 1995.
- [90] S. A. Smolyak. Quadrature and interpolation formulae on tensor products of certain function classes. *Dokl. Akad. Nauk SSSR*, 148:1042–1045, 1963.
- [91] I. M. Sobol'. Uniformly distributed sequences with an additional property of uniformity. *Ž. Vychisl. Mat. i Mat. Fiz.*, 16(5):1332–1337, 1375, 1976.
- [92] I. M. Sobol'. Estimation of the sensitivity of nonlinear mathematical models. *Mat. Model.*, 2(1):112–118, 1990.
- [93] I. M. Sobol'. Sensitivity estimates for nonlinear mathematical models. *Math. Modeling Comput. Experiment*, 1(4):407–414 (1995), 1993.



- [94] C. Soize and R. Ghanem. Physical systems with random uncertainties: chaos representations with arbitrary probability measure. *SIAM J. Sci. Comput.*, 26(2):395–410 (electronic), 2004.
- [95] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*, volume 12 of *Texts in Applied Mathematics*. Springer-Verlag, New York, third edition, 2002. Translated from the German by R. Bartels, W. Gautschi and C. Witzgall.
- [96] C. J. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.*, 22(1):118–184, 1994.
- [97] R. Storn and K. Price. Differential evolution — a simple and efficient heuristic for global optimization over continuous spaces. *J. Global Optim.*, 11(4):341–359, 1997.
- [98] A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451–559, 2010.
- [99] T. J. Sullivan, M. McKerns, D. Meyer, F. Theil, H. Owhadi, and M. Ortiz. Optimal uncertainty quantification for legacy data observations of Lipschitz functions. *Math. Model. Numer. Anal.*, 47(6):1657–1689, 2013.
- [100] G. Szegő. *Orthogonal Polynomials*. American Mathematical Society, Providence, R.I., fourth edition, 1975. American Mathematical Society, Colloquium Publications, Vol. XXIII.
- [101] H. Takahasi and M. Mori. Double exponential formulas for numerical integration. *Publ. Res. Inst. Math. Sci.*, 9:721–741, 1973/74.
- [102] M. Talagrand. Pettis integral and measure theory. *Mem. Amer. Math. Soc.*, 51(307):ix+224, 1984.
- [103] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005.
- [104] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
- [105] A. N. Tikhonov. On the stability of inverse problems. *C. R. (Doklady) Acad. Sci. URSS (N.S.)*, 39:176–179, 1943.
- [106] A. N. Tikhonov. On the solution of incorrectly put problems and the regularisation method. In *Outlines Joint Sympos. Partial Differential Equations (Novosibirsk, 1963)*, pages 261–265. Acad. Sci. USSR Siberian Branch, Moscow, 1963.
- [107] L. N. Trefethen. Is Gauss quadrature better than Clenshaw–Curtis? *SIAM Rev.*, 50(1):67–87, 2008.
- [108] L. N. Trefethen and D. Bau, III. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

- [109] U.S. Department of Energy. *Scientific Grand Challenges for National Security: The Role of Computing at the Extreme Scale*. 2009.
- [110] N. N. Vakhania. The topological support of Gaussian measure in Banach space. *Nagoya Math. J.*, 57:59–63, 1975.
- [111] F. A. Valentine. A Lipschitz condition preserving extension for a vector function. *Amer. J. Math.*, 67(1):83–93, 1945.
- [112] H. von Weizsäcker and G. Winkler. Integral representation in the set of solutions of a generalized moment problem. *Math. Ann.*, 246(1):23–32, 1979/80.
- [113] H. von Weizsäcker and G. Winkler. Noncompact extremal integral representations: some probabilistic aspects. In *Functional analysis: surveys and recent results, II (Proc. Second Conf. Functional Anal., Univ. Paderborn, Paderborn, 1979)*, volume 68 of *Notas Mat.*, pages 115–148. North-Holland, Amsterdam, 1980.
- [114] P. Walley. *Statistical Reasoning with Imprecise Probabilities*, volume 42 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London, 1991.
- [115] K. Weichselberger. The theory of interval-probability as a unifying concept for uncertainty. *Internat. J. Approx. Reason.*, 24(2-3):149–170, 2000.
- [116] G. Winkler. Extreme points of moment sets. *Math. Oper. Res.*, 13(4):581–587, 1988.
- [117] D. Xiu. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press, Princeton, NJ, 2010.
- [118] D. Xiu and G. E. Karniadakis. The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM J. Sci. Comput.*, 24(2):619–644 (electronic), 2002.

# Index

- absolute continuity, 17
- affine combination, 39
- almost everywhere, 10
- ANOVA, 116
- $\arg \max$ , 34
- $\arg \min$ , 34
  
- Banach space, 24
- barycentre, 159
- Bayes' rule, 11, 63
- Bessel's inequality, 29
- Birkhoff–Khinchin ergodic theorem, 106
- Bochner integral, 16, 32
- bounded differences inequality, 113
  
- Céa's lemma, 139, 146
- Cameron–Martin space, 20
- Cauchy–Schwarz inequality, 24, 52
- Chebyshev nodes, 93
- Chebyshev's inequality, 16
- Choquet simplex, 159
- Choquet–Bishop–de Leeuw theorem, 39
- Christoffel–Darboux formula, 90
- Clenshaw–Curtis quadrature, 104
- collocation method
  - for ODEs, 151
  - stochastic, 151
- complete measure space, 10
- concentration of measure, 113
- conditional expectation, 28, 114
- conditional probability, 11
- constraining function, 38
- convex combination, 39
- convex function, 40
- convex hull, 39
- convex optimization problem, 41
- convex set, 39
- counting measure, 10
- covariance
  - matrix, 15
  - operator, 19, 51
  
- Cut-HDMR, 117
  
- Dirac measure, 10
- direct sum, 28
- dominated convergence theorem, 15
- dual space, 26
  
- ensemble Kálmán filter, 78
- entropy, 52, 155
- equivalent measures, 17
- Eulerian observations, 80
- expectation, 15
- expected value, 15
- extended Kálmán filter, 77
- extreme point, 39, 160
  
- Favard's theorem, 90
- Fejér quadrature, 104
- Feldman–Hájek theorem, 21, 67
- Fernique's theorem, 20
- filtration, 12
- Fubini–Tonelli theorem, 18
  
- Galerkin product, 145
- Galerkin projection
  - deterministic, 137
  - stochastic, 140
- Galerkin tensor, 144
- Gauss–Markov theorem, 59
- Gauss–Newton iteration, 46
- Gaussian measure, 18, 19
- Gibbs' inequality, 55
- Gibbs' phenomenon, 96
- Gram matrix, 122, 139
  
- Hankel determinant, 86, 89
- HDMR
  - projectors, 118
- Hellinger distance, 65, 68
- Hermite polynomials, 86, 126
- Hilbert space, 24
- Hoeffding's inequality, 113

- Hoeffding's lemma, 114
- Hotelling transform, 124
- independence, 17
- information, 52
- inner product, 23
- inner product space, 24
- inner regular measure, 159
- integral
  - Bochner, 16, 32
  - Lebesgue, 14
  - Pettis, 16, 19
  - strong, 16
  - weak, 16
- interior point method, 42
- interval arithmetic, 50
- Kálmán filter, 74
  - ensemble, 78
  - extended, 77
  - linear, 74
- Karhunen–Loève theorem, 123
  - sampling Gaussian measures, 124
- Karush–Kuhn–Tucker conditions, 37
- Koksma's inequality, 108
- Koksma–Hlawka inequality, 108
- Kozono–Yanagisawa theorem, 140
- Kreĭn–Milman theorem, 39
- Kullback–Leibler divergence, 54
- Lagrange multipliers, 37
- Lagrange polynomials, 92
- Lagrangian observations, 80
- law of large numbers, 106
- Lax–Milgram theorem
  - deterministic, 137
  - stochastic, 141
- Lebesgue  $L^p$  space, 15
- Lebesgue integral, 14
- Lebesgue measure, 10, 18
- Legendre polynomials, 86
- linear Kálmán filter, 74
- linear program, 43
- marginal, 18
- maximum entropy
  - principle of, 155
- McDiarmid diameter, 112
- McDiarmid subdiameter, 112
- McDiarmid's inequality, 113
- measurable function, 12
- measurable space, 9
- measure, 9
- measure affine function, 160
- Mercer kernel, 122
- Mercer's theorem, 122
- midpoint rule, 100
- Minty's extension theorem, 163
- Moore–Penrose pseudo-inverse, 61
- mutually singular measures, 17
- Newton's method, 34
- Newton–Cotes formula, 101
- norm, 23
- normal equations, 44
- normed space, 23
- null set, 10
- orthogonal complement, 27
- orthogonal polynomials, 88
- orthogonal projection, 27
- orthogonal set, 27
- orthonormal set, 27
- parallelogram identity, 24
- Parseval identity, 29
- penalty function, 38
- Pettis integral, 16, 19
- polarization identity, 24
- precision operator, 19
- principal component analysis, 124
- probability density function, 17
- probability measure, 9
- product measure, 17
- push-forward measure, 12
- quadrature formula, 99
- Radon space, 159
- Radon–Nikodým theorem, 17
- random variable, 12
- Riesz representation theorem, 26
- Riesz space, 159
- RS-HDMR, 116
- Runge's phenomenon, 93
- Schrödinger's inequality, 52
- Schur complements, 81
  - and conditioning of Gaussians, 81
- semi-norm, 23

Sherman–Morrison–Woodbury formula, 81

signed measure, 9

simulated annealing, 36

singular value decomposition, 111, 125

Sobol’ indices, 118

Sobolev space, 26, 95

stochastic collocation method, 151

stochastic process, 12

strong integral, 16

support, 10

surprisal, 52

Takahasi–Mori quadrature, 109

tanh–sinh quadrature, 109

tensor product, 30

Tikhonov regularization, 45, 58

total variation distance, 54, 68

trapezoidal rule, 100

trivial measure, 10

uncertainty principle, 52

Vandermonde matrix, 92

variance, 15

vector lattice, 159

weak integral, 16

Wiener–Hermite PC expansion, 127

Winkler’s theorem, 160

zero-one measure, 39