Chapter 2
Measure and Probability Theory

> To be conscious that you are ignorant is a great step to knowledge.
>
> *Sybil*
> Benjamin Disraeli

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

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

2.1 Measure and Probability Spaces

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

T.J. Sullivan, *Introduction to Uncertainty Quantification*, Texts in Applied Mathematics 63, DOI 10.1007/978-3-319-23395-6_2

Definition 2.1. A *measurable space* is a pair $(\mathcal{X}, \mathscr{F})$, where
(a) $\mathcal{X}$ is a set, called the *sample space*; and
(b) $\mathscr{F}$ is a *σ-algebra* on $\mathcal{X}$, i.e. a collection of subsets of $\mathcal{X}$ containing $\varnothing$ and closed under countable applications of the operations of union, intersection and complementation relative to $\mathcal{X}$; elements of $\mathscr{F}$ are called *measurable sets* or *events*.

Example 2.2. (a) On any set $\mathcal{X}$, there is a *trivial σ-algebra* in which the only measurable sets are the empty set $\varnothing$ and the whole space $\mathcal{X}$.
(b) On any set $\mathcal{X}$, there is also the *power set σ-algebra* in which every subset of $\mathcal{X}$ is measurable. It is a fact of life that this σ-algebra contains too many measurable sets to be useful for most applications in analysis and probability.
(c) When $\mathcal{X}$ is a topological — or, better yet, metric or normed — space, it is common to take $\mathscr{F}$ to be the *Borel σ-algebra* $\mathscr{B}(\mathcal{X})$, the smallest σ-algebra on $\mathcal{X}$ so that every open set (and hence also every closed set) is measurable.

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

The triple $(\mathcal{X}, \mathscr{F}, \mu)$ is called a *signed measure space*, *measure space*, or *probability space* as appropriate. The sets of all signed measures, measures, and probability measures on $(\mathcal{X}, \mathscr{F})$ are denoted $\mathcal{M}_{\pm}(\mathcal{X}, \mathscr{F})$, $\mathcal{M}_{+}(\mathcal{X}, \mathscr{F})$, and $\mathcal{M}_{1}(\mathcal{X}, \mathscr{F})$ respectively.

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

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

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

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

(d) *Lebesgue measure* on $\mathbb{R}^n$ is the unique measure on $\mathbb{R}^n$ (equipped with its Borel σ-algebra $\mathscr{B}(\mathbb{R}^n)$, generated by the Euclidean open balls) that assigns to every rectangle its n-dimensional volume in the ordinary sense.

To be more precise, Lebesgue measure is actually defined on the completion $\mathscr{B}_0(\mathbb{R}^n)$ of $\mathscr{B}(\mathbb{R}^n)$, which is a larger σ-algebra than $\mathscr{B}(\mathbb{R}^n)$. The rigorous construction of Lebesgue measure is a non-trivial undertaking.

(e) Signed measures/charges arise naturally in the modelling of distributions with positive and negative values, e.g. $\mu(E) =$ the net electrical charge within some measurable region $E \subseteq \mathbb{R}^3$. They also arise naturally as differences of non-negative measures: see Theorem 2.24 later on.

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

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

(a) If $N \subseteq \mathcal{X}$ is a subset of a measurable set $E \in \mathscr{F}$ such that $\mu(E) = 0$, then N is called a *μ-null set*.

(b) If the set of $x \in \mathcal{X}$ for which some property $P(x)$ does not hold is μ-null, then P is said to hold *μ-almost everywhere* (or, when μ is a probability measure, *μ-almost surely*).

(c) If every μ-null set is in fact an $\mathscr{F}$-measurable set, then the measure space $(\mathcal{X}, \mathscr{F}, \mu)$ is said to be *complete*.

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

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

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

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

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

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

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

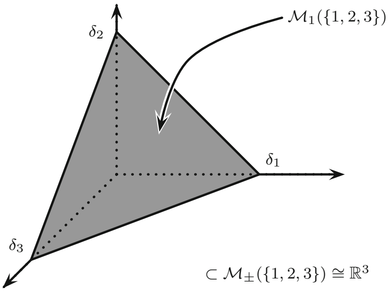
$\mathcal{M}_1(\{1,2,3\})$
δ_2
δ_1
δ_3
$\subset \mathcal{M}_\pm(\{1,2,3\}) \cong \mathbb{R}^3$

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

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

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

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

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

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

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

$$\begin{bmatrix} \mu(\{1\}) & \cdots & \mu(\{n\}) \end{bmatrix}$$

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

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

$$\delta_i = \begin{bmatrix} 0 & \cdots 0 & 1 & 0 & \cdots & 0 \end{bmatrix},$$

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

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

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

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

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

> ΑΓΕΩΜΕΤΡΗΤΟΣ ΜΗΔΕΙΣ ΕΙΣΙΤΩ

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

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

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

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

can be thought of as a polytope of probability measures.

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

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

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

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

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

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

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

2.2 Random Variables and Stochastic Processes

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Measurability and adaptedness are important properties of stochastic processes, and loosely correspond to certain questions being 'answerable' or 'decidable' with respect to the information contained in a given σ-algebra. For instance, if the event $[X \in E]$ is not $\mathscr{F}$-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 $\mathscr{F}_\bullet^Y$-adapted, since the decision $U(t)$ must be made on the basis of the observations $Y(s)$, $s \leq t$, not on observations from any future time.

2.3 Lebesgue Integration

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

is the *covariance matrix* of X.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

whenever $(\phi_n)_{n \in \mathbb{N}}$ is a sequence of simple functions such that the (scalar-valued) Lebesgue integral $\int_{\mathcal{X}} \|f - \phi_n\| \, \mathrm{d}\mu$ converges to 0 as $n \to \infty$.

It transpires that f is Bochner integrable if and only if $\|f\|$ is Lebesgue integrable. The Bochner integral satisfies a version of the Dominated Convergence Theorem, but there are some subtleties concerning the Radon–Nikodým theorem.

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

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

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 $f \colon [0,1] \to \mathcal{V}$ such that $F(t) := \int_{[0,t]} f(\tau) \, \mathrm{d}\tau$ is not differentiable (Kadets, 1994), and so on.

2.4 Decomposition and Total Variation of Signed Measures

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$d_{\mathrm{TV}}(\mu, \nu) \equiv \|\mu - \nu\|_{\mathrm{TV}} = 2 \sup\bigl\{ |\mu(E) - \nu(E)| \bigm| E \in \mathscr{F} \bigr\}, \tag{2.4}$$

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

2.5 The Radon–Nikodým Theorem and Densities

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

$$\nu\colon E \mapsto \int_E \rho(x) \,\mathrm{d}\mu(x) \tag{2.5}$$

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

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

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

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

$$\int_E f \, \mathrm{d}\nu = \int_E f \rho \, \mathrm{d}\mu$$

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

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

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

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

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

2.6 Product Measures and Independence

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

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

(a) Two measurable sets (events) $E_1, E_2 \in \mathscr{F}$ are said to be *independent* if $\mu(E_1 \cap E_2) = \mu(E_1)\mu(E_2)$.

(b) Two sub-σ-algebras $\mathscr{G}_1$ and $\mathscr{G}_2$ of $\mathscr{F}$ are said to be *independent* if E_1 and E_2 are independent events whenever $E_1 \in \mathscr{G}_1$ and $E_2 \in \mathscr{G}_2$.

(c) Two measurable functions (random variables) $X \colon \Theta \to \mathcal{X}$ and $Y \colon \Theta \to \mathcal{Y}$ are said to be *independent* if the σ-algebras generated by X and Y are independent.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.7 Gaussian Measures

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

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

- locally finite, i.e. every point of $\mathbb{R}^d$ has an open neighbourhood of finite Lebesgue measure;

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

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

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

(a) If μ is locally finite and invariant under all translations, then μ is the trivial (zero) measure.

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

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

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

for the new inner product and norm induced by P.

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

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

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

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

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

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

i.e., for every integrable function f,

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

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

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

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

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

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

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

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

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

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

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

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

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

Corollary 2.44. *For a Gaussian measure μ on a separable, reflexive Banach space $\mathcal{V}$, the following are equivalent:*
(a) μ is non-degenerate;
(b) $C_\mu \colon \mathcal{V}' \to \mathcal{V}$ is one-to-one;
(c) $\overline{C_\mu \mathcal{V}'} = \mathcal{V}$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(c) If K is trace class, then its *trace* is defined to be

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

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

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

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

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

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

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

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

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

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

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

- $\mathcal{H}_\mu$ is the completion of the range of the covariance operator $C\colon \mathcal{V}' \to \mathcal{V}$ with respect to this inner product (cf. the closure with respect to the norm in $\mathcal{V}$ in Theorem 2.43).
- If $\mathcal{V}$ is Hilbert, then $\mathcal{H}_\mu$ is the completion of $\operatorname{ran} C^{1/2}$ with the inner product $\langle h, k \rangle_{C^{-1}} := \langle C^{-1/2}h, C^{-1/2}k \rangle_{\mathcal{V}}$.
- $\mathcal{H}_\mu$ is the set of all $v \in \mathcal{V}$ such that $(T_v)_*\mu \approx \mu$, with
$$\frac{\mathrm{d}(T_v)_*\mu}{\mathrm{d}\mu}(x) = \exp\left(\langle v, x \rangle_{C^{-1}} - \frac{\|v\|_{C^{-1}}^2}{2} \right)$$
as in Lemma 2.40.
- $\mathcal{H}_\mu$ is the intersection of all linear subspaces of $\mathcal{V}$ that have full μ-measure.

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

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

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

Furthermore, equivalence holds if and only if

(a) $\operatorname{ran} C_\mu^{1/2} = \operatorname{ran} C_\nu^{1/2}$*;*

(b) $m_\mu - m_\nu \in \operatorname{ran} C_\mu^{1/2} = \operatorname{ran} C_\nu^{1/2}$*; and*

(c) $T := (C_\mu^{-1/2} C_\nu^{1/2})(C_\mu^{-1/2} C_\nu^{1/2})^* - I$ *is Hilbert–Schmidt in* $\operatorname{ran} C_\mu^{1/2}$.

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

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

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

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

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

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

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

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

for all $k_i \in \mathcal{H}_i$, $k_j \in \mathcal{H}_j$, so that C is decomposed[2] *in block form as*

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

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

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

2.8 Interpretations of Probability

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

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

representation of one's degree of belief in the truth of a proposition E. The frequentist's point of view is *objective*; the Bayesian's is *subjective*; both use the same mathematical machinery of probability measures to describe the properties of the function μ.

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

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

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

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

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

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

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

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

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

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

$$\widehat{\theta}^{\mathrm{MAP}} := \operatorname*{arg\,max}_{\theta \in \mathbb{R}} \pi(\theta|d)$$

or the *maximum likelihood estimator*

$$\widehat{\theta}^{\mathrm{MLE}} := \operatorname*{arg\,max}_{\theta \in \mathbb{R}} L(d|\theta).$$

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

It is also worth noting that there is a significant community that, in addition to being frequentist or Bayesian, asserts that selecting a single probability measure is too precise a description of uncertainty. These 'imprecise probabilists' count such distinguished figures as George Boole and John Maynard Keynes among their ranks, and would prefer to say that $\frac{1}{2} - 2^{-100} \leq \mathbb{P}[\text{heads}] \leq \frac{1}{2} + 2^{-100}$ than commit themselves to the assertion that $\mathbb{P}[\text{heads}] = \frac{1}{2}$; imprecise probabilists would argue that the former assertion can be verified, to a prescribed level of confidence, in finite time, whereas the latter cannot. Techniques like the use of *lower and upper probabilities* (or *interval probabilities*) are popular in this community, including sophisticated generalizations like Dempster–Shafer theory; one can also consider *feasible sets of probability measures*, which is the approach taken in Chapter 14.

2.9 Bibliography

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

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

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

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

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

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

2.10 Exercises

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

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

(a) Show that C is conjugate-symmetric and positive semi-definite. For what collection of vectors in $\mathbb{C}^n$ is C the Gram matrix?

(b) Show that if the support of X is all of $\mathbb{C}^n$, then C is positive definite. Hint: suppose that C has non-trivial kernel, construct an open half-space H of $\mathbb{C}^n$ such that $X \notin H$ almost surely.

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

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

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

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

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

i.e., for every integrable function f,

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

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

Exercise 2.5. For $i = 1$, 2, let $X_i \sim \mathcal{N}(m_i, C_i)$ independent Gaussian random variables taking values in Hilbert spaces $\mathcal{H}_i$, and let $T_i \colon \mathcal{H}_i \to \mathcal{K}$ be a bounded linear map taking values in another Hilbert space $\mathcal{K}$, with adjoint $T_i^* \colon \mathcal{K} \to \mathcal{H}_i$. Show that $T_1 X_1 + T_2 X_2$ is a Gaussian random variable in $\mathcal{K}$ with

$$T_1 X_1 + T_2 X_2 \sim \mathcal{N}\big(T_1 m_1 + T_2 m_2, T_1 C_1 T_1^* + T_2 C_2 T_2^*\big).$$

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

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

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

is self-adjoint and positive-definite.

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

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

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

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

and

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

Hence derive the *Woodbury formula*

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

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

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

and C is self-adjoint and positive definite.

(a) Show that C_{11} and C_{22} are self-adjoint and positive-definite.

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

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

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

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