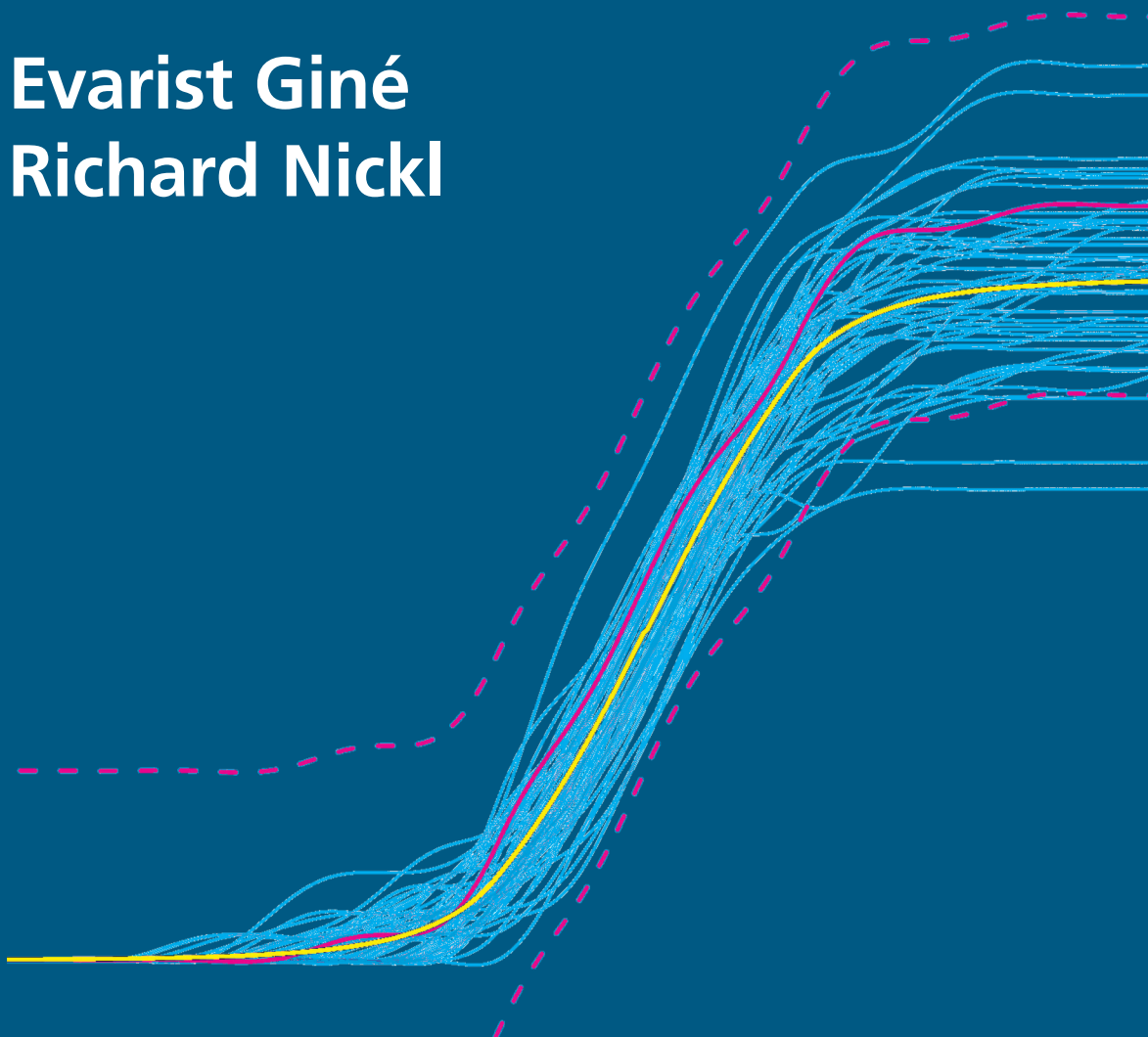
Cambridge Series in Statistical and Probabilistic Mathematics

Mathematical Foundations of Infinite-Dimensional Statistical Models

Evarist Giné
Richard Nickl



Mathematical Foundations of Infinite-Dimensional Statistical Models

In nonparametric and high-dimensional statistical models, the classical Gauss-Fisher-Le Cam theory of the optimality of maximum likelihood estimators and Bayesian posterior inference does not apply, and new foundations and ideas have been developed in the past several decades. This book gives a coherent account of the statistical theory in infinite-dimensional parameter spaces. The mathematical foundations include self-contained 'mini-courses' on the theory of Gaussian and empirical processes, on approximation and wavelet theory, and on the basic theory of function spaces. The theory of statistical inference in such models - hypothesis testing, estimation and confidence sets – is then presented within the minimax paradigm of decision theory. This includes the basic theory of convolution kernel and projection estimation, but also Bayesian nonparametrics and nonparametric maximum likelihood estimation. In a final chapter the theory of adaptive inference in nonparametric models is developed, including Lepski's method, wavelet thresholding, and adaptive inference for self-similar functions.

Winner of the 2017 PROSE Award for Mathematics

EVARIST GINÉ (1944–2015) was Head of the Department of Mathematics at the University of Connecticut. Giné was a distinguished mathematician who worked on mathematical statistics and probability in infinite dimensions. He was the author of two books and more than 100 articles.

RICHARD NICKL is Professor of Mathematical Statistics in the Statistical Laboratory within the Department of Pure Mathematics and Mathematical Statistics at the University of Cambridge.

CAMBRIDGE SERIES IN STATISTICAL AND PROBABILISTIC MATHEMATICS

This series of high-quality upper-division textbooks and expository monographs covers all aspects of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization and mathematical programming. The books contain clear presentations of new developments in the field and of the state of the art in classical methods. While emphasising rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

A complete list of books in the series can be found at www.cambridge.org/statistics.
Recent titles include the following:

14. *Statistical Analysis of Stochastic Processes in Time*, by J. K. Lindsey
15. *Measure Theory and Filtering*, by Lakhdar Aggoun and Robert Elliott
16. *Essentials of Statistical Inference*, by G. A. Young and R. L. Smith
17. *Elements of Distribution Theory*, by Thomas A. Severini
18. *Statistical Mechanics of Disordered Systems*, by Anton Bovier
19. *The Coordinate-Free Approach to Linear Models*, by Michael J. Wichura
20. *Random Graph Dynamics*, by Rick Durrett
21. *Networks*, by Peter Whittle
22. *Saddlepoint Approximations with Applications*, by Ronald W. Butler
23. *Applied Asymptotics*, by A. R. Brazzale, A. C. Davison and N. Reid
24. *Random Networks for Communication*, by Massimo Franceschetti and Ronald Meester
25. *Design of Comparative Experiments*, by R. A. Bailey
26. *Symmetry Studies*, by Marlos A. G. Viana
27. *Model Selection and Model Averaging*, by Gerda Claeskens and Nils Lid Hjort
28. *Bayesian Nonparametrics*, edited by Nils Lid Hjort et al.
29. *From Finite Sample to Asymptotic Methods in Statistics*, by Pranab K. Sen, Julio M. Singer and Antonio C. Pedrosa de Lima
30. *Brownian Motion*, by Peter Mörters and Yuval Peres
31. *Probability* (Fourth Edition), by Rick Durrett
32. *Stochastic Processes*, by Richard F. Bass
33. *Regression for Categorical Data*, by Gerhard Tutz
34. *Exercises in Probability* (Second Edition), by Loïc Chaumont and Marc Yor
35. *Statistical Principles for the Design of Experiments*, by R. Mead, S. G. Gilmour and A. Mead
36. *Quantum Stochastics*, by Mou-Hsiung Chang
37. *Nonparametric Estimation under Shape Constraints*, by Piet Groeneboom and Geurt Jongbloed
38. *Large Sample Covariance Matrices*, by Jianfeng Yao, Zhidong Bai and Shurong Zheng

Mathematical Foundations of Infinite-Dimensional Statistical Models

Evarist Giné

Richard Nickl
University of Cambridge

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108994132
DOI: 10.1017/9781009022811

First published 2016
Revised edition 2021

Printed in the United Kingdom by TJ Books Limited, Padstow Cornwall

A catalogue record for this publication is available from the British Library.

ISBN 978-1-108-99413-2 Paperback

A la meva esposa Rosalind

Dem Andenken meiner Mutter Reingard, 1940–2010

Contents

Preface

The classical theory of statistics was developed for parametric models with *finite-dimensional parameter spaces*, building on fundamental ideas of C. F. Gauss, P. S. Laplace, R. A. Fisher and L. Le Cam, among others. It has been successful in providing modern science with a paradigm for making statistical inferences, in particular, in the 'frequentist large sample size' scenario. A comprehensive account of the mathematical foundations of this classical theory is given in the monograph by A. van der Vaart, *Asymptotic Statistics* (Cambridge University Press, 1998).

The last three decades have seen the development of statistical models that are infinite (or 'high') dimensional. The principal target of statistical inference in these models is a function or an infinite vector f that itself is not modelled further parametrically. Hence, these models are often called, in some abuse of terminology, *nonparametric models*, although f itself clearly also is a parameter. In view of modern computational techniques, such models are tractable and in fact attractive in statistical practice. Moreover, a mathematical theory of such nonparametric models has emerged, originally driven by the Russian school in the early 1980s and since then followed by a phase of very high international activity.

This book is an attempt to describe some elements of the mathematical theory of statistical inference in such *nonparametric*, or infinite-dimensional, models. We will first establish the main probabilistic foundations: the theory of Gaussian and empirical processes, with an emphasis on the 'nonasymptotic concentration of measure' perspective on these areas, including the pathbreaking work by M. Talagrand and M. Ledoux on concentration inequalities for product measures. Moreover, since a thorough understanding of infinite-dimensional models requires a solid background in functional analysis and approximation theory, some of the most relevant results from these areas, particularly the theory of wavelets and of Besov spaces, will be developed from first principles in this book.

After these foundations have been laid, we turn to the statistical core of the book. Comparing nonparametric models in a very informal way with classical parametric models, one may think of them as models in which the number of parameters that one estimates from the observations is *growing proportionally to sample size n* and has to be carefully selected by the statistician, ideally in a data-driven way. In practice, nonparametric modelling is often driven by the honesty of admitting that the traditional assumption that n is large compared to the number of unknown parameters is too strong. From a mathematical point of view, the frequentist theory that validates statistical inferences in such models undergoes a radical shift: leaving the world of finite-dimensional statistical models behind implies that the likelihood function no longer provides 'automatically optimal' statistical methods ('maximum likelihood estimators') and that extreme care has to be exercised when

constructing inference procedures. In particular, the Gauss–Fisher–Le Cam efficiency theory based on the Fisher information typically yields nothing informative about what optimal procedures are in nonparametric statistics, and a new theoretical framework is required. We will show how the minimax paradigm can serve as a benchmark by which a theory of optimality in nonparametric models can be developed. From this paradigm arises the 'adaptation' problem, whose solution has been perhaps one of the major achievements of the theory of nonparametric statistics and which will be presented here for nonparametric function estimation problems. Finally, likelihood-based procedures can be relevant in nonparametric models as well, particularly after some regularisation step that can be incorporated by adopting a 'Bayesian' approach or by imposing qualitative a priori shape constraints. How such approaches can be analysed mathematically also will be shown here.

Our presentation of the main statistical materials focusses on function estimation problems, such as density estimation or signal in white-noise models. Many other nonparametric models have similar features but are formally different. Our aim is to present a unified statistical theory for a canonical family of infinite-dimensional models, and this comes at the expense of the breadth of topics that could be covered. However, the mathematical mechanisms described here also can serve as guiding principles for many nonparametric problems not covered in this book.

Throughout this book, we assume familiarity with material from real and functional analysis, measure and probability theory on the level of a US graduate course on the subject. We refer to the monographs by G. Folland, *Real Analysis* (Wiley, 1999), and R. Dudley, *Real Analysis and Probability* (Cambridge University Press, 2002), for relevant background. Apart from this, the monograph is self-contained, with a few exceptions and 'starred sections' indicated in the text.

This book would not have been possible without the many colleagues and friends from whom we learnt, either in person or through their writings. Among them, we would like to thank P. Bickel, L. Birgé, S. Boucheron, L. Brown, T. Cai, I. Castillo, V. Chernozhukov, P. Dawid, L. Devroye, D. Donoho, R. Dudley, L. Dümbgen, U. Einmahl, X. Fernique, S. Ghosal, A. Goldenshluger, Y. Golubev, M. Hoffmann, I. Ibragimov, Y. Ingster, A. Iouditski, I. Johnstone, G. Kerkyacharian, R. Khasminskii, V. Koltchinskii, R. Latala, M. Ledoux, O. Lepski, M. Low, G. Lugosi, W. Madych, E. Mammen, D. Mason, P. Massart, M. Nussbaum, D. Picard, B. Pötscher, M. Reiß, P. Rigollet, Y. Ritov, R. Samworth, V. Spokoiny, M. Talagrand, A. Tsybakov, S. van de Geer, A. van der Vaart, H. van Zanten, J. Wellner, H. Zhou and J. Zinn.

We are grateful to A. Carpentier, I. Castillo, U. Einmahl, D. Gauthier, D. Heydecker, K. Ray, J. Söhl and B. Szabò for proofreading parts of the manuscript and providing helpful corrections.

Moreover, we are indebted to Diana Gillooly of Cambridge University Press for her support, patience and understanding in the process of this book project since 2011.

R.N. would also like to thank his friends N. Berestycki, C. Damböck, R. Dawid and M. Neuber for uniquely stimulating friendships that have played a large role in the intellectual development that led to this book (and beyond).

Outline and Reading Guide

In principle, all the chapters of this book can be read independently. In particular, the chapters on Gaussian and empirical processes, as well as the one on function spaces and approximation theory, are mostly self-contained. A reader interested primarily in the 'statistical chapters' (5 through 8) may choose to read those first and then turn to the mathematical foundations laid out in Chapters 2 through 4 later, when required. A short outline of the contents of each chapter is given in the following paragraphs:

Chapter 1 introduces the kinds of statistical models studied in this book. In particular, we will discuss why many common 'regular' regression models with normally distributed error terms can be mathematically accommodated within one Gaussian function estimation problem known as the *Gaussian white noise model*.

Chapters 2 and 3 lay the probabilistic foundations of much of the statistical theory that follows: one chapter on Gaussian processes and one on empirical processes. The Gaussian theory is mostly classical, presented with a focus on statistically relevant materials, such as the isoperimetric inequality for Gaussian measures and its consequences on concentration, as well as a study of suprema of Gaussian processes. The theory for empirical measures reflects the striking recent developments around the concentration-of-measure phenomenon. Effectively, here, the classical role of the central limit theorem in statistics is replaced by nonasymptotic concentration properties of product measures, as revealed in fundamental work by Talagrand, Ledoux, Massart and others. This is complemented by a treatment of abstract empirical process theory, including metric entropy methods, Vapnik-Červonenkis classes and uniform central limit theorems.

Chapter 4 develops from first principles some key aspects of approximation theory and its functional analytic foundations. In particular, we give an account of wavelet theory and of Besov spaces, with a focus on results that are relevant in subsequent chapters.

Chapter 5 introduces basic linear estimation techniques that are commonly used in nonparametric statistics, based on convolution kernels and finite-dimensional projection operators. Tools from Chapters 3 and 4 are used to derive a variety of probabilistic results about these estimators that will be useful in what follows.

Chapter 6 introduces a theoretical paradigm – the *minimax paradigm* – that can be used to objectively measure the performance of statistical methods in nonparametric models. The basic information-theoretic ideas behind it are developed, and it is shown how statistical inference procedures – estimators, tests and confidence sets – can be analysed and compared from a minimax point of view. For a variety of common nonparametric models, concrete constructions of minimax optimal procedures are given using the results from previous chapters.

Chapter 7 shows how the likelihood function can still serve as a successful guiding principle in certain nonparametric problems if a priori information is used carefully. This can be done by imposing certain qualitative constraints on the statistical model or by formally adopting a Bayesian approach which then can be analysed from a frequentist point of view. The key role of the Hellinger distance in this theory (as pointed out in work by Le Cam, Birgé, van de Geer, van der Vaart and others) is described in some detail.

Chapter 8 presents the solution to the nonparametric adaptation problem that arises from the minimax paradigm and gives a theory of statistical inference for 'fully automatic' statistical procedures that perform well over maximal collections of nonparametric statistical models. Surprising differences are shown to arise when considering the existence of adaptive estimation procedures in contrast to the existence of associated adaptive confidence sets. A resolution of this discrepancy can be obtained by considering certain nonparametric models of 'self-similar' functions, which are discussed in some detail and for which a unified theory of optimal statistical inference can be developed.

Each chapter is organised in several sections, and historical notes complementing each section can be found at the end of each chapter – these are by no means exhaustive and only indicate our understanding of the literature.

At the end of each section, exercises are provided: these, likewise, complement the main results of the text and often indicate interesting applications or extensions of the materials presented.

Postscript

It is a terrible tragedy that Evarist Giné passed away shortly after we completed the manuscript. His passion for mathematics was exceeded only by his love for his wife, Rosalind; his daughters, Núria and Roser; and his grandchildren, Liam and Mireia. He mentioned to me in September 2014, when I last met him in Cambridge (MA), that perhaps he wanted to dedicate this book to all of them, but in an e-mail to me in January 2015, he mentioned explicitly that he wanted it to be for Rosalind. I have honoured his decision; however, I know that with this last work he wanted to thank all of them for having been his wonderful family – who continue his infectious passion into new generations.

I am myself deeply grateful to my father, Harald, for all his support and inspiration throughout my life in all domains. I dedicate this book to the memory of my mother, Reingard, in loving gratitude for all her courage and everything she has done for me. And of course, insofar as this book relates to the future, it is for Ana and our son, Julian, with love and affection.

Postscript (2020)

In this paperback edition a large number of (mostly minor) corrections have been incorporated. I would like to thank the various readers and students, specifically Kweku Abraham, who pointed them out to me.

1

Nonparametric Statistical Models

In this chapter we introduce and motivate the statistical models that will be considered in this book. Some of the materials depend on basic facts developed in subsequent chapters – mostly the basic Gaussian process and Hilbert space theory. This will be hinted at when necessary.

Very generally speaking, a *statistical model* for a random observation Y is a family

$$\{P_f : f \in \mathcal{F}\}$$

of probability distributions P_f, each of which is a candidate for having generated the observation Y. The parameter f belongs to the *parameter space* $\mathcal{F}$. The problem of *statistical inference* on f, broadly speaking, can be divided into three intimately connected problems of using the observation Y to

(a) *Estimate* the parameter f by an estimator $T(Y)$,
(b) *Test hypotheses* on f based on test functions $\Psi(Y)$ and/or
(c) *Construct confidence sets* $C(Y)$ that contain f with high probability.

To interpret inferential results of these kinds, we will typically need to specify a distance, or loss function on $\mathcal{F}$, and for a given model, different loss functions may or may not lead to very different conclusions.

The statistical models we will introduce in this chapter are, on the one hand, conceptually closely related to each other in that the parameter space $\mathcal{F}$ is infinite or high dimensional and the loss functions relevant to the analysis of the performance of statistical procedures are similar. On the other hand, these models are naturally divided by the different probabilistic frameworks in which they occur – which will be either a *Gaussian noise model* or an *independent sampling model*. These frameworks are asymptotically related in a fundamental way (see the discussion after Theorem 1.2.1). However, the most effective probabilistic techniques available are based on a direct, nonasymptotic analysis of the Gaussian or product probability measures that arise in the relevant sampling context and hence require a separate treatment.

Thus, while many of the statistical intuitions are common to both the sampling and the Gaussian noise models and in fact inform each other, the probabilistic foundations of these models will be laid out independently.

1.1 Statistical Sampling Models

Let X be a random experiment with associated sample space $\mathcal{X}$. We take the mathematical point of view of probability theory and model X as a random variable, that is, as a measurable mapping defined on some underlying probability space that takes values in the measurable space $(\mathcal{X},\mathcal{A})$, where $\mathcal{A}$ is a σ-field of subsets of $\mathcal{X}$. The law of X is described by the probability measure P on $\mathcal{A}$. We may typically think of $\mathcal{X}$ equal to $\mathbb{R}^d$ or a measurable subset thereof, equipped with its Borel σ-field $\mathcal{A}$.

The perhaps most basic problem of statistics is the following: consider repeated outcomes of the experiment X, that is, a random sample of independent and identically distributed (i.i.d.) copies $X_1,\ldots,X_n$ from X. The joint distribution of the X_i equals the product probability measure $P^n = \otimes_{i=1}^n P$ on $(\mathcal{X}^n,\mathcal{A}^n)$. The goal is to recover P from the n observations. 'Recovering P' can mean many things. Classical statistics has been concerned mostly with models where P is explicitly parameterised by a finite-dimensional parameter, such as the mean and variance of the normal distribution, or the 'parameters' of the usual families of statistical distributions (gamma, beta, exponential, Poisson, etc.). Recovering P then simply means to use the observations to make inferences on the unknown parameter, and the fact that this parameter is finite dimensional is crucial for this traditional paradigm of statistical inference, in particular, for the famous likelihood principle of R. A. Fisher. In this book, we will follow the often more realistic assumption that no such parametric assumptions are made on P. For most sample spaces $\mathcal{X}$ of interest, this will naturally lead to models that are infinite dimensional, and we will investigate how the theory of statistical inference needs to be developed in this situation.

1.1.1 Nonparametric Models for Probability Measures

In its most elementary form, without imposing any parameterisations on P, we can simply consider the problem of making inferences on the unknown probability measure P based on the sample. Natural loss functions arise from the usual metrics on the space of probability measures on $\mathcal{X}$, such as the total variation metric

$$\|P-Q\|_{TV} = \sup_{A\in\mathcal{A}} |P(A)-Q(A)|$$

or weaker metrics that generate the topology of weak convergence of probability measures on $\mathcal{X}$. For instance, if $\mathcal{X}$ itself is endowed with a metric d, we could take the bounded Lipschitz metric

$$\beta_{(\mathcal{X},d)}(P,Q) = \sup_{f\in BL(1)} \left| \int_{\mathcal{X}} f\,(dP-dQ) \right|$$

for weak convergence of probability measures, where

$$BL(M) = \left\{ f:\mathcal{X}\to\mathbb{R},\ \sup_{x\in\mathcal{X}} |f(x)| + \sup_{x\neq y} \frac{|f(x)-f(y)|}{d(x,y)} \le M \right\}, \quad 0<M<\infty.$$

If $\mathcal{X}$ has some geometric structure, we can consider more intuitive loss functions. For example, if $\mathcal{X}=\mathbb{R}$, we can consider the cumulative distribution function

$$F(x) = P(X\le x), \quad x\in\mathbb{R},$$

or, if X takes values in $\mathbb{R}^d$, its multivariate analogue. A natural distance function on distribution functions is simply the supremum-norm metric ('Kolmogorov distance')

$$\|F_P - F_Q\|_\infty = \sup_{x\in\mathbb{R}} |F_P(x) - F_Q(x)|.$$

Since the indicators $\{1_{(-\infty,x]} : x \in \mathbb{R}\}$ generate the Borel σ-field of $\mathbb{R}$, we see that, on $\mathbb{R}$, the statistical parameter P is characterised entirely by the functional parameter F, and vice versa. The parameter space is thus the infinite-dimensional space of all cumulative distribution functions on $\mathbb{R}$.

Often we will know that P has some more structure, such as that P possesses a probability-density function $f : \mathbb{R} \to [0,\infty)$, which itself may have further properties that will be seen to influence the complexity of the statistical problem at hand. For probability-density functions, a natural loss function is the L^1-distance

$$\|f_P - f_Q\|_1 = \int_{\mathbb{R}} |f_P(x) - f_Q(x)|dx$$

and in some situations also other L^p-type and related loss functions. Although in some sense a subset of the other, the class of probability densities is more complex than the class of probability-distribution functions, as it is not described by monotonicity constraints and does not consist of functions bounded in absolute value by 1. In a heuristic way, we can anticipate that estimating a probability density is harder than estimating the distribution function, just as the preceding total variation metric is stronger than any metric for weak convergence of probability measures (on nontrivial sample spaces $\mathcal{X}$). In all these situations, we will see that the theory of statistical inference on the parameter f significantly departs from the usual finite-dimensional setting.

Instead of P, a particular functional $\Phi(P)$ may be the parameter of statistical interest, such as the moments of P or the quantile function F^{-1} of the distribution function F – examples for this situation are abundant. The nonparametric theory is naturally compatible with such functional estimation problems because it provides the direct plug-in estimate $\Phi(T)$ based on an estimator T for P. Proving closeness of T to P in some strong loss function then gives access to 'many' continuous functionals Φ for which $\Phi(T)$ will be close to $\Phi(P)$, as we shall see later in this book.

1.1.2 Indirect Observations

A common problem in statistical sampling models is that some systematic measurement errors are present. A classical problem of this kind is the statistical regression problem, which will be introduced in the next section. Another problem, which is more closely related to the sampling model from earlier, is where one considers observations in $\mathbb{R}^d$ of the form

$$Y_i = X_i + \varepsilon_i, \quad i = 1,\dots,n, \tag{1.1}$$

where the X_i are i.i.d. with common law P_X, and the ε_i are random 'error' variables that are independent of the X_i and have law P_ε. The law P_ε is assumed to be known to the observer – the nature of this assumption is best understood by considering examples: the attempt is to model situations in which a scientist, for reasons of cost, complexity or lack of precision of the involved measurement device, is forced to observe Y_i instead of the

realisations X_i of interest. The observer may, however, have very concrete knowledge of the source of the error, which could, for example, consist of light emissions of the Milky Way interfering with cosmic rays from deeper space, an erratic optical device through which images are observed (e.g., a space telescope which cannot be repaired except at very high cost) or transmissions of signals through a very busy communication channel. Such situations of implicit measurements are encountered frequently in the applied sciences and are often called *inverse problems*, as one wishes to 'undo' the errors inflicted on the signal in which one is interested. The model (1.1) gives a simple way to model the main aspects of such statistical inverse problems. It is also known as the *deconvolution model* because the law of the Y_i equals

$$P_Y = P_X * P_\varepsilon,$$

the convolution of the two probability measures P_X, P_ε, and one wishes to 'deconvolve' P_ε.

As earlier, we will be interested in inference on the underlying distribution P_X of the signal X when the statistical model for P_X is infinite dimensional. The loss functions in this problem are thus typically the same as in the preceding subsection.

1.2 Gaussian Models

The randomness in the preceding sampling model was encoded in a general product measure P^n describing the joint law of the observations. Another paradigm of statistical modelling deals with situations in which the randomness in the model is described by a Gaussian (normal) distribution. This paradigm naturally encompasses a variety of nonparametric models, where the infinite-dimensional character of the problem does not necessarily derive from the probabilistic angle but from a functional relationship that one wishes to model.

1.2.1 Basic Ideas of Regression

Perhaps the most natural occurrence of a statistical model in the sciences is the one in which observations, modelled here as numerical values or vectors, say, (Y_i, x_i), arise according to a functional relationship

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.2}$$

where n is the number of observations (sample size), f is some function of the x_i and the ε_i are random noise. By 'random noise', we may mean here either a probabilistic model for certain measurement errors that we believe to be intrinsic to our method of making observations, or some innate stochastic nature of the way the Y_i are generated from the $f(x_i)$. In either case, we will model the ε_i as random variables in the sense of axiomatic probability theory – the question of the genuine physical origin of this random noise will not concern us here. It is sometimes natural to assume also that the x_i are realisations of random variables X_i – we can either take this into account explicitly in our analysis or make statements conditional on the observed values $X_i = x_i$.

The function f often will be unknown to the observer of observations (Y_i, x_i), and the goal is to recover f from the (Y_i, x_i). This may be of interest for various reasons, for instance, for predicting new values Y_{n+1} from $f(x_{n+1})$ or to gain quantitative and qualitative understanding of the functional relationship $Y_i = f(x_i)$ under consideration.

In the preceding context, a statistical model in the broad sense is an a priori specification of both a parameter space for the functions f that possibly could have generated (1.2) and a family of probability measures that describes the possible distributions of the random variables ε_i. By 'a priori', we mean here that this is done independently of (e.g., before) the observational process, reflecting the situation of an experimentalist.

A systematic use and study of such models was undertaken in the early nineteenth century by Carl Friedrich Gauss, who was mostly interested in predicting astronomical observations. When the model is translated into the preceding formalisation, Gauss effectively assumed that the x_i are vectors $(x_{i1},\dots,x_{ip})^T$ and thought of f as a linear function in that vector, more precisely,

$$f(x_i) = x_{i1}\theta_i + \dots x_{ip}\theta_p, \quad i = 1,\dots,n,$$

for some real-valued parameters $\theta_j, j = 1,\dots,p$. The parameter space for f is thus the Euclidean space $\mathbb{R}^p$ expressed through all such linear mappings. In Gauss's time, the assumption of linearity was almost a computational necessity.

Moreover, Gauss modelled the random noise ε_i as independent and identically distributed samples from a normal distribution $N(0,\sigma^2)$ with some variance σ^2. His motivation behind this assumption was twofold. First, it is reasonable to assume that $E(\varepsilon_i) = 0$ for every i. If this expectation were nonzero, then there would be some deterministic, or 'systematic', measurement error $e_i = E(\varepsilon_i)$ of the measurement device, and this could always be accommodated in the functional model by adding a constant $x_{10} = \dots = x_{n0} = 1$ to the preceding linear relationship. The second assumption that ε_i has a normal distribution is deeper. If we think of each measurement error ε_i as the sum of many 'very small', or infinitesimal, independent measurement errors $\varepsilon_{ik}, k = 1,2,\dots$, then, by the central limit theorem, $\varepsilon_i = \sum_k \varepsilon_{ik}$ should be approximately normally distributed, regardless of the actual distribution of the ε_{ik}. By the same reasoning, it is typically natural to assume that the ε_i are also independent among themselves. This leads to what is now called the *standard Gaussian linear model*

$$Y_i = f(x_i) + \varepsilon_i \equiv \sum_{j=1}^{p} x_{ij}\theta_j + \varepsilon_i, \qquad \varepsilon_i \sim^{i.i.d.} N(0,\sigma^2), \quad i = 1,\dots,n, \tag{1.3}$$

which bears this name both because Gauss studied it and, since the $N(0,\sigma^2)$ distribution is often called the *Gaussian distribution*, because Gauss first made systematic use of it. The unknown parameter (θ,σ^2) varies in the $(p+1)$-dimensional parameter space

$$\Theta \times \Sigma = \mathbb{R}^p \times (0,\infty).$$

This model constitutes perhaps *the* classical example of a *finite-dimensional model*, which has been studied extensively and for which a fairly complete theory is available. For instance, when p is smaller than n, the least-squares estimator of Gauss finds the value $\hat{\theta} \in \mathbb{R}^p$ which solves the optimisation problem

$$\min_{\theta\in\mathbb{R}^p} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} x_{ij}\theta_j \right)^2$$

and hence minimises the Euclidean distance of the vector $Y = (Y_1,\dots,Y_n)^T$ to the p-dimensional subspace spanned by the p vectors $(x_{1j},\dots,x_{nj})^T, j = 1,\dots,p$.

1.2.2 Some Nonparametric Gaussian Models

We now give a variety of models that generalise Gauss's ideas to infinite-dimensional situations. In particular, we will introduce the Gaussian white noise model, which serves as a generic surrogate for a large class of nonparametric models, including even non-Gaussian ones, through the theory of equivalence of experiments (discussed in the next section).

Nonparametric Gaussian Regression

Gauss's model and its theory basically consist of two crucial assumptions: one is that the ε_i are normally distributed, and the other is that the function f is linear. The former assumption was argued to be in some sense natural, at least in a measurement-error model (see also the remarks after Theorem 1.2.1 for further justification). The latter assumption is in principle quite arbitrary, particularly in times when computational power does not constrain us as much any longer as it did in Gauss's time. A nonparametric approach therefore attempts to assume as little structure of f as possible. For instance, by the *nonparametric regression model with fixed, equally spaced design on* $[0,1]$, we shall understand here the model

$$Y_i = f(x_i) + \varepsilon_i, \qquad x_i = \frac{i}{n}, \quad \varepsilon_i \sim^{i.i.d.} N(0,\sigma^2), \quad i = 1,\dots,n. \tag{1.4}$$

where f is any function defined on $[0,1]$. We are thus sampling the unknown function f at an equally spaced grid of $[0,1]$ that, as $n \to \infty$, grows dense in the interval $[0,1]$ as $n \to \infty$.

The model immediately generalises to bounded intervals $[a,b]$, to 'approximately' equally spaced designs $\{x_i : i = 1,\dots,n\} \subset [a,b]$ and to multivariate situations, where the x_i are equally spaced points in some hypercube. We note that the assumption that the x_i are equally spaced is important for the theory that will follow – this is natural as we cannot hope to make inference on f in regions that contain no or too few observations x_i.

Other generalisations include the *random design regression model*, in which the x_i are viewed as i.i.d. copies of a random variable X. One can then either proceed to argue conditionally on the realisations $X_i = x_i$, or one takes this randomness explicitly into account by making probability statements under the law of X and ε simultaneously. For reasonable design distributions, this will lead to results that are comparable to the fixed-design model – one way of seeing this is through the equivalence theory for statistical experiments (see after Theorem 1.2.1).

A priori it may not be reasonable to assume that f has any specific properties other than that it is a continuous or perhaps a differentiable function of its argument. Even if we assumed that f has infinitely many continuous derivatives the set of all such f would be infinite dimensional and could never be fully captured by a p-dimensional parameter space. We thus have to expect that the theory of statistical inference in this nonparametric model will be different from the one in Gauss's classical linear model.

The Gaussian White Noise Model

For the mathematical development in this book we shall work with a mathematical idealisation of the regression model (1.4) in continuous time, known as the *Gaussian white noise model*, and with its infinite sequence space analogue. While perhaps at first appearing more complicated than the discrete model, once constructed, it allows for a clean

and intuitive mathematical exposition that mirrors all the main ideas and challenges of the discrete case with no severe loss of generality.

Consider the following stochastic differential equation:

$$dY(t) \equiv dY_f^{(n)}(t) = f(t)dt + \frac{\sigma}{\sqrt{n}} dW(t), \qquad t \in [0,1], \quad n \in \mathbb{N}, \tag{1.5}$$

where $f \in L^2 \equiv L^2([0,1])$ is a square integrable function on $[0,1]$, $\sigma > 0$ is a dispersion parameter and dW is a *standard Gaussian white noise process*. When we observe a realisation of (1.5), we shall say that we observe the function or signal f in Gaussian white noise, at the noise level, or a signal-to-noise ratio $\sigma/\sqrt{n}$. We typically think of n large, serving as a proxy for sample size, and of $\sigma > 0$ a fixed known value. If σ is unknown, one can usually replace it by a consistent estimate in the models we shall encounter in this book.

The exact meaning of dW needs further explanation. Heuristically, we may think of dW as a weak derivative of a standard Brownian motion $\{W(t) : t \in [0,1]\}$, whose existence requires a suitable notion of stochastic derivative that we do not want to develop here explicitly. Instead, we take a 'stochastic process' approach to define this stochastic differential equation, which for statistical purposes is perfectly satisfactory. Let us thus agree that 'observing the trajectory (1.5)' will simply mean that we observe a realisation of the Gaussian process defined by the application

$$g \mapsto \int_0^1 g(t) dY^{(n)}(t) \equiv \mathbb{Y}_f^{(n)}(g) \sim N\left(\langle f, g\rangle, \frac{\|g\|_2^2}{n}\right), \tag{1.6}$$

where g is any element of the Hilbert space $L^2([0,1])$ with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|_2$. Even more explicitly, we observe all the $N(\langle f, g\rangle, \|g\|_2^2/n)$ variables, as g runs through $L^2([0,1])$. The randomness in the equation (1.5) comes entirely from the additive term dW, so after translating by $\langle f, g\rangle$ and scaling by $1/\sqrt{n}$, this means that dW is defined through the Gaussian process obtained from the action

$$g \mapsto \int_0^1 g(t) dW(t) \equiv \mathbb{W}(g) \sim N(0, \|g\|_2^2), \quad g \in L^2([0,1]). \tag{1.7}$$

Note that this process has a diagonal covariance in the sense that for any *finite* set of orthonormal vectors $\{e_k\} \subset L^2$ we have that the family $\{\mathbb{W}(e_k)\}$ is a multivariate standard normal variable, and as a consequence of the Kolmogorov consistency theorem (Proposition 2.1.10), $\mathbb{W}$ and $\mathbb{Y}^{(n)}$ indeed define Gaussian processes on L^2.

The fact that the model (1.5) can be interpreted as a Gaussian process indexed by L^2 means that the natural sample space $\mathcal{Y}$ in which dY from (1.5) takes values is the 'path' space $\mathbb{R}^{L^2([0,1])}$. This space may be awkward to work with in practice. In Section 6.1.1 we shall show that we can find more tractable choices for $\mathcal{Y}$ where dY concentrates with probability 1.

Gaussian Sequence Space Model

Again, to observe the stochastic process $\{\mathbb{Y}_f^{(n)}(g) : g \in L^2\}$ just means that we observe $\mathbb{Y}_f^{(n)}(g)$ for all $g \in L^2$ simultaneously. In particular, we may pick any orthonormal basis $\{e_k : k \in \mathbb{Z}\}$ of L^2, giving rise to an observation in the *Gaussian sequence space model*

$$Y_k \equiv Y_{f,k}^{(n)} = \langle f, e_k\rangle + \frac{\sigma}{\sqrt{n}} g_k, \qquad k \in \mathbb{Z}, \quad n \in \mathbb{N}, \tag{1.8}$$

where the g_k are i.i.d. of law $\mathbb{W}(e_k) \sim N(0, \|e_k\|_2^2) = N(0,1)$. Here we observe all the basis coefficients of the unknown function f with additive Gaussian noise of variance σ^2/n. Note that since the $\{e_k : k \in \mathbb{Z}\}$ realise a sequence space isometry between L^2 and the sequence space ℓ_2 of all square-summable infinite sequences through the mapping $f \mapsto \langle f, e_k \rangle$, the law of $\{Y_{f,k}^{(n)} : k \in \mathbb{Z}\}$ completely characterises the finite-dimensional distributions, and thus the law, of the process $\mathbb{Y}_f^{(n)}$. Hence, models (1.5) and (1.8) are observationally equivalent to each other, and we can prefer to work in either one of them (see also Theorem 1.2.1).

We note that the random sequence $Y = (Y_k : k \in \mathbb{Z})$ itself does not take values in ℓ_2, but we can view it as a random variable in the 'path' space $\mathbb{R}^{\ell_2}$. A more tractable, separable sample space on which $(Y_k : k \in \mathbb{Z})$ can be realised is discussed in Section 6.1.1.

A special case of the Gaussian sequence model is obtained when the space is restricted to n coefficients

$$Y_k = \theta_k + \frac{\sigma}{\sqrt{n}} g_k, \quad k = 1, \ldots, n, \tag{1.9}$$

where the θ_k are equal to the $\langle f, e_k \rangle$. This is known as the *normal means model*. While itself a finite-dimensional model, it cannot be compared to the standard Gaussian linear model from the preceding section as its dimension increases as fast as n. In fact, for most parameter spaces that we will encounter in this book, the difference between model (1.9) and model (1.8) is negligible, as follows, for instance, from inspection of the proof of Theorem 1.2.1.

Multivariate Gaussian Models

To define a Gaussian white noise model for functions of several variables on $[0,1]^d$ through the preceding construction is straightforward. We simply take, for $f \in L^2([0,1]^d)$,

$$dY(t) = f(t)dt + \frac{\sigma}{\sqrt{n}} dW(t), \qquad t \in [0,1]^d, \quad n \in \mathbb{N}, \quad \sigma > 0, \tag{1.10}$$

where dW is defined through the action

$$g \mapsto \int_{[0,1]^d} g(t) dW(t) \equiv \mathbb{W}(g) \sim N(0, \|g\|_2^2) \tag{1.11}$$

on elements g of $L^2([0,1]^d)$, which corresponds to multivariate stochastic integrals with respect to independent Brownian motions $W_1(t_1), \ldots, W_d(t_d)$. Likewise, we can reduce to a sequence space model by taking an orthonormal basis $\{e_k : k \in \mathbb{Z}^d\}$ of $L^2([0,1]^d)$.

1.2.3 Equivalence of Statistical Experiments

It is time to build a bridge between the preceding abstract models and the statistically more intuitive nonparametric fixed-design regression model (1.4). Some experience with the preceding models reveals that a statistical inference procedure in any of these models constructively suggests a procedure in the others with comparable statistical properties. Using a suitable notion of distance between statistical experiments, this intuition can be turned into a theorem, as we show in this subsection. We present results for Gaussian regression models; the general approach, however, can be developed much further to show that even highly non-Gaussian models can be, in a certain sense, asymptotically equivalent to the standard Gaussian white noise model (1.5). This gives a general justification for a

rigorous study of the Gaussian white noise model in itself. Some of the proofs in this subsection require material from subsequent chapters, but the main ideas can be grasped without difficulty.

The Le Cam Distance of Statistical Experiments

We employ a general notion of distance between statistical experiments $\mathcal{E}^{(i)}, i = 1,2$, due to Le Cam. Each experiment $\mathcal{E}^{(i)}$ consists of a sample space $\mathcal{Y}_i$ and a probability measure $P_f^{(i)}$ defined on it, indexed by a common parameter $f \in \mathcal{F}$. Let $\mathcal{T}$ be a measurable space of 'decision rules', and let

$$L : \mathcal{F} \times \mathcal{T} \to [0, \infty)$$

be a 'loss function' measuring the performance of a decision procedure $T^{(i)}(Y^{(i)}) \in \mathcal{T}$ based on observations $Y^{(i)}$ in experiment i. For instance, $T^{(i)}(Y^{(i)})$ could be an estimator for f so that $\mathcal{T} = \mathcal{F}$ and $L(f,T) = d(f,T)$, where d is some metric on $\mathcal{F}$, but other scenarios are possible. The risk under $P_f^{(i)}$ for this loss is the $P_f^{(i)}$-expectation of $L(f, T^{(i)}(Y^{(i)}))$, denoted by $R^{(i)}(f, T^{(i)}, L)$. Define also

$$|L| = \sup\{L(f,T) : f \in \mathcal{F}, T \in \mathcal{T}).$$

The *Le Cam distance* between two experiments is defined as

$$\Delta_{\mathcal{F}}(\mathcal{E}^{(1)}, \mathcal{E}^{(2)}) \equiv \max\Bigg[\sup_{T^{(2)}} \inf_{T^{(1)}} \sup_{f, L:|L|=1} \left| R^{(1)}(f, T^{(1)}, L) - R^{(2)}(f, T^{(2)}, L) \right|, \tag{1.12}$$

$$\sup_{T^{(1)}} \inf_{T^{(2)}} \sup_{f, L:|L|=1} \left| R^{(1)}(f, T^{(1)}, L) - R^{(2)}(f, T^{(2)}, L) \right| \Bigg].$$

If this quantity equals zero, this means that any decision procedure $T^{(1)}$ in experiment $\mathcal{E}^{(1)}$ can be translated into a decision procedure $T^{(2)}$ in experiment $\mathcal{E}^{(2)}$, and vice versa, and that the statistical performance of these procedures in terms of the associated risk $R^{(i)}$ will be the same for any bounded loss function L. If the distance is not zero but small, then, likewise, the performance of the corresponding procedures in both experiments will differ by at most their Le Cam distance.

Some useful observations on the Le Cam distance are the following: if both experiments have a common sample space $\mathcal{Y}^{(1)} = \mathcal{Y}^{(2)} = \mathcal{Y}$ equal to a complete separable metric space, and if the probability measures $P_f^{(1)}, P_f^{(2)}$ have a common dominating measure μ on $\mathcal{Y}$, then

$$\Delta_{\mathcal{F}}(\mathcal{E}^{(1)}, \mathcal{E}^{(2)}) \leq \sup_{f \in \mathcal{F}} \int_{\mathcal{Y}} \left| \frac{dP_f^{(1)}}{d\mu} - \frac{dP_f^{(2)}}{d\mu} \right| d\mu \equiv \|P^{(1)} - P^{(2)}\|_{1,\mu,\mathcal{F}}. \tag{1.13}$$

This follows from the fact that in this case we can always use the decision rule $T^{(2)}(Y)$ in experiment $\mathcal{E}^{(1)}$ and vice versa and from

$$|R^{(1)}(f,T,L) - R^{(2)}(f,T,L)| \leq \int_{\mathcal{Y}} |L(f,T(Y))| |dP_f^{(1)}(Y) - dP_f^{(2)}(Y)| \leq |L| \|P^{(1)} - P^{(2)}\|_{1,\mu,\mathcal{F}}.$$

The situation in which the two experiments are not defined on the sample space needs some more thought. Suppose, in the simplest case, that we can find a bi-measurable isomorphism B of $\mathcal{Y}^{(1)}$ with $\mathcal{Y}^{(2)}$, independent of f, such that

$$P_f^{(2)} = P_f^{(1)} \circ B^{-1}, \qquad P_f^{(1)} = P_f^{(2)} \circ B \quad \forall f \in \mathcal{F}.$$

Then, given observations $Y^{(2)}$ in $\mathcal{Y}^{(2)}$, we can use the decision rule $T^{(2)}(Y^{(2)}) \equiv T^{(1)}(B^{-1}(Y^{(2)}))$ in $\mathcal{E}^{(2)}$, and vice versa, and the risks $R^{(i)}$ in both experiments coincide by the image measure theorem. We can conclude in this case that

$$\Delta_{\mathcal{F}}(\mathcal{E}^{(1)}, \mathcal{E}^{(2)}) = \Delta_{\mathcal{F}}(\mathcal{E}^{(1)}, B^{-1}(\mathcal{E}^{(2)})) = 0. \tag{1.14}$$

In the absence of such a bijection, the theory of sufficient statistics can come to our aid to bound the Le Cam distance. Let again $\mathcal{Y}^{(i)}, i = 1,2$, be two sample spaces that we assume to be complete separable metric spaces. Let $\mathcal{E}^{(1)}$ be the experiment giving rise to observations $Y^{(1)}$ of law $P_f^{(1)}$ on $\mathcal{Y}^{(1)}$, and suppose that there exists a mapping $S : \mathcal{Y}^{(1)} \to \mathcal{Y}^{(2)}$ independent of f such that

$$Y^{(2)} = S(Y^{(1)}), \qquad Y^{(2)} \sim P_f^{(2)} \quad \text{on } \mathcal{Y}^{(2)}.$$

Assume, moreover, that $S(Y^{(1)})$ is a sufficient statistic for $Y^{(1)}$; that is, the conditional distribution of $Y^{(1)}$ given that we have observed $S(Y^{(1)})$ is independent of $f \in \mathcal{F}$. Then

$$\Delta_{\mathcal{F}}(\mathcal{E}^{(1)}, \mathcal{E}^{(2)}) = 0. \tag{1.15}$$

The proof of this result, which is an application of the *sufficiency principle* from statistics, is left as Exercise 1.1.

Asymptotic Equivalence for Nonparametric Gaussian Regression Models

We can now give the main result of this subsection. We shall show that the experiments

$$Y_i = f(x_i) + \varepsilon_i, \qquad x_i = \frac{i}{n}, \quad \varepsilon_i \sim^{i.i.d.} N(0, \sigma^2), \quad i = 1, \dots, n, \tag{1.16}$$

and

$$dY(t) = f(t)dt + \frac{\sigma}{\sqrt{n}} dW(t), \qquad t \in [0,1], \quad n \in \mathbb{N}, \tag{1.17}$$

are asymptotically ($n \to \infty$) equivalent in the sense of Le Cam distance. In the course of the proofs, we shall show that any of these models is also asymptotically equivalent to the sequence space model (1.8). Further models that can be shown to be equivalent to (1.17) are discussed after the proof of the following theorem.

We define classes

$$\mathcal{F}(\alpha, M) = \left\{ f : [0,1] \to \mathbb{R}, \ \sup_{x \in [0,1]} |f(x)| + \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x - y|^{\alpha}} \leq M \right\},$$
$$0 < \alpha \leq 1, \quad 0 < M < \infty,$$

of α-Hölderian functions. Moreover, for $(x_i)_{i=1}^n$ the design points of the fixed-design regression model (1.16) and for f any bounded function defined on $[0,1]$, let $\pi_n(f)$ be the unique function that interpolates f at the x_i and that is piecewise constant on each interval $(x_{i_1}, x_i] \subset [0,1]$.

Theorem 1.2.1 *Let $(\mathcal{E}_n^{(i)} : n \in \mathbb{N}), i = 1,2,3$, equal the sequence of statistical experiments given by $i = 1$ the fixed-design nonparametric regression model (1.16); $i = 2$, the standard Gaussian white noise model (1.17); and $i = 3$, the Gaussian sequence space model (1.8),*

respectively. Then, for $\mathcal{F}$ any family of bounded functions on $[0,1]$, for $\pi_n(f)$ as earlier and for any $n \in \mathbb{N}$,

$$\Delta_{\mathcal{F}}(\mathcal{E}_n^{(2)}, \mathcal{E}_n^{(3)}) = 0, \quad \Delta_{\mathcal{F}}(\mathcal{E}_n^{(1)}, \mathcal{E}_n^{(2)}) \le \sqrt{\frac{n\sigma^2}{2}} \sup_{f \in \mathcal{F}} \|f - \pi_n(f)\|_2. \tag{1.18}$$

In particular, if $\mathcal{F} = \mathcal{F}(\alpha, M)$ for any $\alpha > 1/2, M > 0$, then all these experiments are asymptotically equivalent in the sense that their Le Cam distance satisfies, as $n \to \infty$,

$$\Delta_{\mathcal{F}}(\mathcal{E}_n^{(i)}, \mathcal{E}_n^{(j)}) \to 0, \quad i,j \in \{1,2,3\}. \tag{1.19}$$

Proof In the proof we shall say that two experiments are equivalent if their Le Cam distance is exactly equal to zero. The first claim in (1.18) immediately follows from (1.14) and the isometry between $L^2([0,1])$ and ℓ_2 used in the definition of the sequence space model (1.8).

Define $\mathcal{V}_n$ to equal the n-dimensional space of functions $f : [0,1] \to \mathbb{R}$ that are piecewise constant on the intervals

$$I_{in} = (x_{i-1}, x_i] = \left(\frac{i-1}{n}, \frac{i}{n} \right], \quad i = 1, \dots, n.$$

The indicator functions $\phi_{in} = 1_{I_{in}}$ of these intervals have disjoint support, and they form an orthonormal basis of $\mathcal{V}_n$ for the inner product

$$\langle f, g \rangle_n = \sum_{j=1}^{n} f(x_j) g(x_j),$$

noting that $\sum_{j=1}^{n} \phi_{in}^2(x_j) = 1$ for every i. Given bounded $f : [0,1] \to \mathbb{R}$, let $\pi_n(f)$ be the $\langle \cdot, \cdot \rangle_n$-projection of f onto $\mathcal{V}_n$. Since

$$\langle f, \phi_{in} \rangle_n = \sum_{j=1}^{n} f(x_j) \phi_{in}(x_j) = f(x_i) \ \forall i,$$

we see

$$\pi_n(f)(t) = \sum_{i=1}^{n} f(x_i) \phi_{in}(t), \quad t \in [0,1],$$

so this projection interpolates f at the design points x_i, that is, $\pi_n(f)(x_j) = f(x_j)$ for all j. Note that the functions $\{\sqrt{n}\phi_{in} : i = 1, \dots, n\}$ also form a basis of $\mathcal{V}_n$ in the standard $L^2([0,1])$ inner product $\langle \cdot, \cdot \rangle$. This simultaneous orthogonality property will be useful in what follows.

Observing $Y_i = f(x_i) + \varepsilon_i$ in $\mathbb{R}^n$ from model (1.16) with bounded f is, by (1.14), equivalent to observations in the n-dimensional functional space $\mathcal{V}_n$ given by

$$\sum_{i=1}^{n} Y_i \phi_{in}(t) = \sum_{i=1}^{n} f(x_i) \phi_{in}(t) + \sum_{i=1}^{n} \varepsilon_i \phi_{in}(t), \quad t \in [0,1]. \tag{1.20}$$

We immediately recognise that $\sum_{i=1}^{n} f(x_i) \phi_{in}$ is the interpolation $\pi_n(f)$ of f at the x_i. Moreover, the error process is a scaled white noise process restricted to the space $\mathcal{V}_n$: indeed, its $L^2([0,1])$ action on $h \in \mathcal{V}_n$ is given by

$$\int_0^1 \sum_{i=1}^{n} \varepsilon_i \phi_{in}(t) h(t) dt = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \langle h, \sqrt{n}\phi_{in} \rangle \sim N\left(0, \frac{\sigma^2}{n} \sum_{i=1}^{n} \langle h, \sqrt{n}\phi_{in} \rangle^2\right) = N\left(0, \frac{\sigma^2}{n} \|h\|_2^2\right)$$

using Parseval's identity and that the $\sqrt{n}\phi_{in}$ form an $L^2([0,1])$ orthonormal basis of $\mathcal{V}_n$. If Π_n is the $L^2([0,1])$ projector onto $\mathcal{V}_n$ spanned by the $\{\sqrt{n}\phi_{in}\}$, then one shows, by the same arguments, that this process can be realised as a version of the Gaussian process defined on L^2 by the action $h \mapsto \mathbb{W}(\Pi_n(h))$, where $\mathbb{W}$ is as in (1.7). In other words, it equals the L^2-projection of the standard white noise process dW onto the finite-dimensional space $\mathcal{V}_n$, justifying the notation

$$\frac{\sigma}{\sqrt{n}} dW_n(t) \equiv \sum_{i=1}^{n} \varepsilon_i \phi_{in}(t) dt.$$

To summarise, (1.16) is equivalent to model (1.20), which itself can be rewritten as

$$d\tilde{Y}(t) \equiv \pi_n(f)(t) + \frac{\sigma}{\sqrt{n}} dW_n(t), \quad t \in [0,1]. \tag{1.21}$$

Next, consider the model

$$d\bar{Y}(t) = \pi_n(f)(t) + \frac{\sigma}{\sqrt{n}} dW(t), \quad t \in [0,1], \tag{1.22}$$

which is the standard white noise model (1.17) but with f replaced by its interpolation $\pi_n(f)$ at the design points x_i. Since $\pi_n(f) \in \mathcal{V}_n$, we have $\Pi_n(\pi_n(f)) = \pi_n(f)$, and since $dW_n = \Pi_n(dW) \in \mathcal{V}_n$, the statistics

$$d\tilde{Y} = \Pi_n(d\bar{Y}) = \left\{ \int_0^1 h(t) d\tilde{Y}(t) : h \in \mathcal{V}_n \right\}$$

are sufficient for $d\bar{Y}$, so by (1.15) the models (1.21) and (1.22) are equivalent. [To use (1.15) rigorously, we interpret $d\tilde{Y}, d\bar{Y}$ as tight random variables in a large enough, separable Banach space (see Section 6.1.1).]

To prove the second claim in (1.18), we relate (1.22) to (1.17), that is, to

$$dY(t) = f(t) + \frac{\sigma}{\sqrt{n}} dW(t), \quad t \in [0,1].$$

Both experiments have the same sample space, which in view of Section 6.1.1 we can take to be, for instance, the space of continuous functions on $[0,1]$, and the standard white noise $\mathbb{W}$ gives a common dominating measure P_0^Y on that space for the corresponding probability measures $P_f^Y, P_{\pi_n(f)}^Y$. In view of (1.13) and using Proposition 6.1.7a) combined with (6.16), we see that the Le Cam distance is bounded by

$$\sup_{f \in \mathcal{F}} \|P_f^Y - P_{\pi_n(f)}^Y\|_{1,\mu,\mathcal{F}}^2 \le \frac{n}{\sigma^2} \sup_{f \in \mathcal{F}} \|f - \pi_n(f)\|_2^2, \tag{1.23}$$

which gives (1.18). Finally, for (1.19), uniformly in $f \in \mathcal{F}(\alpha, M)$,

$$\|f - \pi_n(f)\|_2^2 = \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} (f(x) - f(x_i))^2 dx \le M^2 \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} |x - x_i|^{2\alpha} dx$$

$$\le M^2 n^{-2\alpha} \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} dx = O(n^{-2\alpha}),$$

so for $\alpha > 1/2$, the quantity in (1.23) converges to zero, completing the proof. ■

In the preceding theorem the Hölder classes $\mathcal{F}(\alpha, M)$ could be replaced by balls in the larger Besov-Sobolev spaces $B^{\alpha}_{2\infty}$ (defined in Chapter 4) whenever $\alpha > 1/2$. The condition on α, however, cannot be relaxed, as we discuss in the notes.

The theory of asymptotic equivalence can be taken much further, to include results like the one preceding for random design regression experiments in possibly multivariate settings and with possibly non-Gaussian noise ε. The theory also extends to non-Gaussian settings that are not of regression type: one can show that nonparametric models for probability or spectral densities, or ergodic diffusions, are asymptotically equivalent to a suitable Gaussian white noise model. We discuss relevant references in the notes.

Asymptotic equivalence theory, which is a subject in its own, justifies that the Gaussian white noise model is, in the sense of the Le Cam distance, a canonical limit experiment in which one can develop some main theoretical ideas of nonparametric statistics. For Gaussian regression problems, the closeness of the experiments involved is in fact of a nonasymptotic nature, as shown by Theorem 1.2.1, and in this book we thus shall concentrate on the white noise model as the natural continuous surrogate for the standard fixed-design regression model. For other, non-Gaussian models, such as density estimation, asymptotic equivalence theory is, however, often overly simplistic in its account of the probabilistic structure of the problem at hand, and for the purposes of this book, we hence prefer to stay within the product-measure setting of Section 1.1, such that a nonasymptotic analysis is possible.

Exercises

1.1 Prove (1.15). [*Hint*: Use the fact that the proof of the standard sufficiency reduction principle extends to complete separable metric spaces (see Le Cam 1986).]

1.3 Notes

The modern understanding of statistical inference as consisting of the three related branches of estimation, testing and confidence statements probably goes back, in its most fundamental form, to the work of Fisher (1922; 1925a, b), who considered mostly parametric (finite-dimensional) statistical models. The need to investigate nonparametric statistical models was realised not much later, roughly at the same time at which the axiomatic approach to probability theory was put forward by Kolmogorov (1933). Classic papers on fully nonparametric sampling models for the cumulative distribution function are, for instance, Glivenko (1933), Cantelli (1933), Kolmogorov (1933a), and Smirnov (1939). More recent developments will be reviewed in later chapters of this book.

The linear regression model with normally distributed errors was initiated by Gauss (1809), who used it successfully in the context of observational astronomy. Gauss most likely was the first to use the least-squares algorithm, although Legendre and even some others can claim priority as well. The history is reviewed, for example, in Plackett (1972) and Stigler (1981).

Nonparametric regression models were apparently not studied systematically before the 1960s; see Nadaraya (1964) and Watson (1964). The Gaussian white noise model and its sequence space analogue were systematically developed in the 1970s and later by the Russian school – we refer to the seminal monograph by Ibragimov and Khasminskii (1981). The asymptotic equivalence theory for statistical experiments was developed by Le Cam; we refer to his fundamental book Le Cam (1986) and also to Le Cam and Yang (1990). Landmark contributions in nonparametric asymptotic equivalence theory are the papers Brown and Low (1996) and Nussbaum (1996), who

treated univariate regression models with fixed design and density estimation, respectively. The necessity of the assumption $\alpha \geq 1/2$ is the subject of the paper by Brown and Zhang (1998). Asymptotic equivalence for random design regression is somewhat more involved: the univariate case is considered in Brown et al. (2002), and the general, multivariate random design regression case is considered in Reiß (2008). Further important results include asymptotic equivalence for nonparametric regression with non-Gaussian error distributions in Grama and Nussbaum (2002), asymptotic equivalence for spectral density estimation in Golubev, Nussbaum and Zhou (2010), and asymptotic equivalence for ergodic diffusions in Dalalyan and Reiß (2006).

2

Gaussian Processes

This chapter develops some classical theory and fundamental tools for Gaussian random processes. We start with the basic definitions of Gaussian processes indexed by abstract parameter spaces and, by way of introduction to the subject, derive some elementary yet powerful properties. We present the isoperimetric and log-Sobolev inequalities for Gaussian measures in $\mathbb{R}^n$ and apply them to establish concentration properties for the supremum of a Gaussian process about its median and mean, which are some of the deepest and most useful results on Gaussian processes. Then we introduce Dudley's metric entropy bounds for moments of suprema of (sub-) Gaussian processes as well as for their a.s. modulus of continuity. The chapter also contains a thorough discussion of convexity and comparison properties of Gaussian measures and of reproducing kernel Hilbert spaces and ends with an exposition of the limit theory for suprema of stationary Gaussian processes.

2.1 Definitions, Separability, 0-1 Law, Concentration

We start with some preliminaries about stochastic processes, mainly to fix notation and terminology. Then these concepts are specialised to Gaussian processes, and some first properties of Gaussian processes are developed. The fundamental observation is that a Gaussian process X indexed by a set T induces an intrinsic distance d_X on T ($d_X(s,t)$ is the L^2-distance between $X(s)$ and $X(t)$), and all the probabilistic information about X is contained in the metric or pseudo-metric space (T,d). This is tested on some of the first properties, such as the 0-1 law and the existence of separable versions of X. One of the main properties of Gaussian processes, namely, their concentration about the mean, is introduced; this subject will be treated in the next section, but a first result on it, which is not sharp but that has been chosen for its simplicity, is given in this section.

2.1.1 Stochastic Processes: Preliminaries and Definitions

Let $(\Omega, \Sigma, \Pr)$ be a probability space, and let T be a set. A stochastic process X indexed by T and defined on the probability space $(\Omega, \Sigma, \Pr)$ is a function $X : T \times \Omega \mapsto \mathbb{R}$, $(t,\omega) \mapsto X(t,\omega)$ such that, for each $t \in T$, $X(t,\cdot)$ is a random variable. Then, for any finite set $F \subset T$, the maps $\Omega \mapsto \mathbb{R}^F$ given by $\omega \mapsto \{X(t,\omega) : t \in F\}$ are also measurable, and their probability laws $\mu_F = \Pr \circ \{X(t,\cdot) : t \in F\}^{-1}$ are the *finite-dimensional distributions* (or finite-dimensional marginal distributions or finite-dimensional marginals) of X. If $F \subset G \subset T$ and G is finite and π_{GF} is the natural projection from $\mathbb{R}^G$ onto $\mathbb{R}^F$, then, obviously, the *consistency*

conditions $\mu_F = \mu_G \circ \pi_{GF}^{-1}$ are satisfied ($\pi_{GF}(\{X(t) : t \in G\}) = \{X(t) : t \in F\}$). Conversely, the Kolmogorov consistency theorem shows that any collection of Borel probability measures μ_F on $\mathbb{R}^F$, indexed by the finite subsets $F \subset T$ and satisfying the consistency conditions, is the collection of finite-dimensional distributions of a stochastic process X indexed by T. In other words, a consistent family of probability measures μ_F, $F \subset T$, F finite, defines a unique probability measure μ on the cylindrical σ-algebra $\mathcal{C}$ of $\mathbb{R}^T$ such that $\mu_F = \mu \circ \pi_{TF}^{-1}$. (The cylindrical σ-algebra $\mathcal{C}$ is the σ-algebra generated by the cylindrical sets with finite-dimensional base, $\pi_{TF}^{-1}(A)$, $A \in \mathcal{B}(\mathbb{R}^F)$, $F \subset T$, F finite.) Then the map $X : T \times \mathbb{R}^T \mapsto \mathbb{R}$, $(t,x) \mapsto x(t)$, is a process defined on the probability space $(\mathbb{R}^T, \mathcal{C}, \mu)$. If μ is the probability measure on $(\mathbb{R}^T, \mathcal{C})$ defined by the finite-dimensional distributions of a process X, then we say that μ is the *probability law of* X (which can be thought of as a 'random variable' taking values on the measurable space $(\mathbb{R}^T, \mathcal{C})$). See almost any probability textbook, for example, Dudley (2002).

Definition 2.1.1 Two processes X and Y of index set T are said to be a *version* of each other if both have the same finite-dimensional distributions $\mathcal{L}(X(t_1), \ldots, X(t_n)) = \mathcal{L}(Y(t_1), \ldots, Y(t_n))$ for all $n \in \mathbb{N}$ and $t_i \in T$ or, what is the same, if both have the same probability law on $(\mathbb{R}^T, \mathcal{C})$. They are said to be a *strict version* or a *modification* of each other if $\Pr\{X(t) = Y(t)\} = 1$ for all t.

It is convenient to recall the definition of pseudo-distance and pseudo-metric space. A pseudo-distance d on T is a nonnegative symmetric function of two variables $s, t \in T$ that satisfies the triangle inequality but for which $d(s,t) = 0$ does not necessarily imply $s = t$. A pseudo-metric space (T,d) is a set T equipped with a pseudo-distance d. Clearly, a pseudo-metric space becomes a metric space by taking the quotient with respect to the equivalence relation $s \simeq t$ iff $d(s,t) = 0$. For instance, the space $\mathcal{L}^p$ of functions is a pseudo-metric space for the L^p (pseudo-)norm, and the space of equivalence classes, L^p, is a metric space for the same norm. One only seldom needs to distinguish between the two.

If the index set T of a process X is a metric or pseudo-metric space (T,d), we say that X is *continuous in probability* if $X(t_n) \to X(t)$ in probability whenever $d(t_n, t) \to 0$. In this case, if T_0 is a d-dense subset of T, the law of the process on $(\mathbb{R}^T, \mathcal{C})$ is determined by the finite-dimensional distributions $\mathcal{L}(X(t_1), \ldots, X(t_n))$ for all $n \in \mathbb{N}$ and $t_i \in T_0$.

Here are two more definitions of interest.

Definition 2.1.2 A process $X(t)$, $t \in T$, (T,d) a metric or pseudo-metric space, is separable if there exists $T_0 \subset T$, T_0 countable, and $\Omega_0 \subset \Omega$ with $\Pr(\Omega_0) = 1$ such that for all $\omega \in \Omega_0$, $t \in T$ and $\varepsilon > 0$,

$$X(t,\omega) \in \overline{\{X(s,\omega) : s \in T_0 \cap B_d(t,\varepsilon)\}},$$

where $B_d(t,\varepsilon)$ is the open d-ball about t of radius ε. X is *measurable* if the map $(\Omega \times T, \Sigma \otimes \mathcal{T}) \to \mathbb{R}$ given by $(\omega, t) \longrightarrow X(\omega, t)$ is jointly measurable, where $\mathcal{T}$ is the σ-algebra generated by the d-balls of T.

By definition, if $X(t)$, $t \in T$, is separable, then there are points from T_0 in any neighborhood of t, $t \in T$; hence (T,d) is separable; that is, (T,d) possesses a countable dense subset. Note that if X is separable, then $\sup_{t \in T} X(t) = \sup_{s \in T_0} X(s)$ a.s., and the latter, being a countable supremum, is measurable; that is, suprema over uncountable sets are measurable. The same holds for $|X(t)|$.

Often we require the *sample paths* $t \mapsto X(t,\omega)$ to have certain properties for almost every ω, notably, to be bounded or bounded and uniformly continuous ω a.s.

Definition 2.1.3 A process $X(t)$, $t \in T$, is sample bounded if it has a version $\tilde{X}$ whose sample paths $t \mapsto \tilde{X}(t,\omega)$ are almost all uniformly bounded, that is, $\sup_{t\in T} |\tilde{X}(t)| < \infty$ a.s. If (T,d) is a metric or pseudo-metric space, then X is sample continuous (more properly, sample bounded and uniformly continuous) if it has a version $\tilde{X}(t)$ whose sample paths are almost all bounded and uniformly d-continuous.

Note that if X is sample continuous, then the finite-dimensional distributions of X are the marginals of a probability measure μ defined on the cylindrical σ-algebra $\mathcal{C} \cap C_u(T,d)$ of $C_u(T,d)$, the space of bounded uniformly continuous functions on (T,d), $\mathcal{L}(X(t_1),\ldots,X(t_k)) = \mu \circ (\delta_{t_1},\ldots,\delta_{t_k})^{-1}$, $t_i \in T$, $k < \infty$ (here and in what follows, δ_t is unit mass at t). The vector space $C_u(T,d)$, equipped with the supremum norm $\|f\|_\infty = \sup_{t\in T} |f(t)|$, is a Banach space, that is, a complete normed space for which the vector space operations are continuous. The Banach space $C_u(T,d)$ is separable if (and only if) (T,d) is totally bounded, and in this case, $C_u(T,d)$ is isometric to $C(\bar{T},d)$, where $(\bar{T},d)$ is the completion of (T,d), which is compact. Then, assuming (T,d) totally bounded, we have $\|f\|_\infty = \sup_{t\in T_0} |f(t)|$, where T_0 is any countable dense subset of T; in particular, the closed balls of $C_u(T,d)$ are measurable for the cylindrical σ-algebra: $\{f : \|f - f_0\|_\infty \le r\} = \cap_{t\in T_0}\{f : |f(t) - f_0(t)| \le r\}$. This implies that the open sets are also measurable because, by separability of $C_u(T,d)$, every open set in this space is the union of a countable number of closed balls. This proves that the Borel and the cylindrical σ-algebras of $C_u(T,d)$ coincide if (T,d) is totally bounded. Hence, in this case, the finite-dimensional distributions of X are the marginal measures of a Borel probability measure μ on $C_u(T,d)$. Since $C_u(T,d)$ is separable and complete (for the supremum norm), the probability law μ of X is tight in view of the following basic result that we shall use frequently in this book (see Exercise 2.1.6 for its proof). Recall that a probability measure μ is tight if for all $\varepsilon > 0$ there is K compact such that $\mu(K^c) < \varepsilon$.

Proposition 2.1.4 (**Oxtoby-Ulam**) *If μ is a Borel probability measure on a complete separable metric space, then μ is tight.*

In general, given a Banach space B, a *B-valued random variable X* is a Borel measurable map from a probability space into B. Thus, the preceding considerations prove the following proposition. It is convenient to introduce an important Banach space: given a set T, $\ell_\infty(T) \subset \mathbb{R}^T$ will denote the set of bounded functions $x : T \mapsto \mathbb{R}$. Note that this is a Banach space if we equip it with the supremum norm $\|x\|_T = \sup_{t\in T} |x(t)|$ and that the inclusion of $C_u(T)$ into $\ell_\infty(T)$ is isometric. Observe that $\ell_\infty(T)$ is separable for the supremum norm if and only if T is finite.

Proposition 2.1.5 *If (T,d) is a totally bounded metric or pseudo-metric space and $X(t)$, $t \in T$, is a sample continuous process, then X has a version which is a $C_u(T,d)$-valued random variable, and its probability law is a tight Borel measure with support contained in $C_u(T,d)$ and hence a tight Borel probability measure on $\ell_\infty(T)$.*

Example 2.1.6 (**Banach space–valued random variables as sample continuous processes.**) Let B be a separable Banach space, let B^* be its dual space and let B_1^* denote the

closed-unit ball of B_1^* about the origin. Then there exists a countable set $D \subset B_1^*$ such that $\|x\| = \sup_{f\in D} f(x)$ for all $x \in B$: if $\{x_i\} \subset B$ is a countable dense subset of B and $f_i \in B_1^*$ are such that $f_i(x_i) = \|x_i\|$ (note that f_i exists by the Hahn-Banach theorem), then $D = \{f_i\}$ is such a set. The inclusion $B \mapsto C_u(D, \|\cdot\|)$, where $\|\cdot\|$ is the norm on B_1^*, is an isometric imbedding, and every B-valued random variable X defines a process $f \mapsto f(X)$, $f \in D$, with all its sample paths bounded and uniformly continuous. Hence, any results proved for sample bounded and uniformly continuous processes indexed by totally bounded metric spaces do apply to Banach space–valued random variables for B separable.

If $X(t)$, $t \in T$, is a sample bounded process, then its probability law is defined on the cylindrical σ-algebra of $\ell_\infty(T)$, $\Sigma = \mathcal{C} \cap \ell_\infty(T)$. Since $\ell_\infty(T)$ is a metric space for the supremum norm, it also has another natural σ-algebra, the Borel σ-algebra. We conclude with the interesting fact that if the law of the bounded process X extends to a tight Borel measure on $\ell_\infty(T)$, then X is sample continuous with respect to a metric d for which (T,d) is totally bounded.

Proposition 2.1.7 *Let $X(t)$, $t \in T$, be a sample bounded stochastic process. Then the finite-dimensional probability laws of X are those of a tight Borel probability measure on $\ell_\infty(T)$ if and only if there exists on T a pseudo-distance d for which (T,d) is totally bounded and such that X has a version with almost all its sample paths uniformly continuous for d.*

Proof Let us assume that the probability law of X is a tight Borel measure μ on $\ell_\infty(T)$; let K_n, $n \in \mathbb{N}$, be an increasing sequence of compact sets in $\ell_\infty(T)$ such that $\mu\big(\cup_{n=1}^\infty K_n\big) = 1$; and set $K = \cup_{n=1}^\infty K_n$. Define a pseudo-metric d as

$$d(s,t) = \sum_{n=1}^{\infty} 2^{-n}\big(1 \wedge d_n(s,t)\big),$$

where

$$d_n(s,t) = \sup\big\{|f(t) - f(s)| : f \in K_n\big\}.$$

To prove that (T,d) is totally bounded, given $\varepsilon > 0$, let m be such that $\sum_{n=m+1}^\infty 2^{-n} < \varepsilon/4$. Since the set $\cup_{n=1}^m K_n$ is compact, it is totally bounded, and therefore, it contains a finite subset $\{f_1, \ldots, f_r\}$ which is $\varepsilon/4$ dense in $\cup_{n=1}^m K_n$ for the supremum norm; that is, for each $f \in \cup_{n=1}^m K_n$, there is $i \le r$ such that $\|f - f_i\|_\infty \le \varepsilon/4$. Since $\cup_{n=1}^m K_n$ is a bounded subset of $\ell_\infty(T)$ (as it is compact), it follows that the subset $A = \{(f_1(t), \ldots, f_r(t)) : t \in T\}$ of $\mathbb{R}^r$ is bounded, hence precompact, hence totally bounded, and therefore there exists a finite set $T_\varepsilon = \{t_i : 1 \le i \le N\}$ such that for each $t \in T$ there is $i = i(t) \le N$ such that $\max_{1\le s\le r} |f_s(t) - f_s(t_i)| \le \varepsilon/4$. It follows that T_ε is ε dense in T for the pseudo-metric d: for $n \le m$, $t \in T$ and $t_i = t_{i(t)}$, we have

$$d_n(t,t_i) = \sup_{f\in K_n} |f(t) - f(t_i)| \le \max_{s\le r} |f_s(t) - f_s(t_i)| + \varepsilon/2 \le \frac{3\varepsilon}{4}$$

and therefore

$$d(t,t_i) \le \frac{\varepsilon}{4} + \sum_{n=1}^{m} 2^{-n} d_n(t,t_i) \le \varepsilon,$$

proving that (T,d) is totally bounded.

Next, since $\mu(K) = 1$, the identity map of $(\ell_\infty(T), \mathcal{B}, \mu)$ is a version of X with almost all its trajectories in K. Thus, to prove that X has a version with almost all its sample paths bounded and uniformly d-continuous, it suffices to show that the functions from K have these properties. If $f \in K_n$, then $|f(s) - f(t)| \le d_n(s,t) \le 2^n d(s,t)$ for all $s,t \in T$ with $d(s,t) < 2^{-n}$, proving that f is uniformly continuous, and f is bounded because K_n is bounded.

Conversely, let $X(t)$, $t \in T$, be a process with a version whose sample paths are almost all in $C_u(T,d)$ for a distance or pseudo-distance d on T for which (T,d) is totally bounded, and let us continue denoting X such a version (recall the notation $C_u(T,d)$ as the space of bounded uniformly continuous functions on (T,d)). Then X is a random variable taking values in $C_u(T,d)$, and its marginal laws correspond to a Borel probability measure on $C_u(T,d)$ (see the argument following Definition 2.1.3). But since (T,d) is precompact, $C_u(T,d)$ is separable, and the law of X is in fact a tight Borel measure by the Oxtoby-Ulam theorem (Proposition 2.1.4). But a tight Borel probability measure on $C_u(T,d)$ is a tight Borel measure on $\ell_\infty(T)$ because the inclusion of $C_u(T,d)$ into ℓ_∞ is continuous. ■

2.1.2 Gaussian Processes: Introduction and First Properties

We now look at Gaussian processes. Recall that a finite-dimensional random vector or a multivariate random variable $Z = (Z_1, \dots, Z_n)$, $n \in \mathbb{N}$, is an n-dimensional Gaussian vector, or a multivariate normal random vector, or its coordinates are jointly normal, if the random variables $\langle a, Z\rangle = \sum_{i=1}^n a_i Z_i$, $a = (a_1, \dots, a_n) \in \mathbb{R}^n$, are normal variables, that is, variables with laws $N(m(a), \sigma^2(a))$, $\sigma(a) \ge 0$, $m \in \mathbb{R}$. If $m = m(a) = 0$ for all $a \in \mathbb{R}^n$, we say that the Gaussian vector is *centred*.

Definition 2.1.8 A stochastic process $X(t)$, $t \in T$, is a Gaussian process if for all $n \in \mathbb{N}$, $a_i \in \mathbb{R}$ and $t_i \in T$, the random variable $\sum_{i=1}^n a_i X(t_i)$ is normal or, equivalently, if all the finite-dimensional marginals of X are multivariate normal. X is a centred Gaussian process if all these random variables are normal with mean zero.

Definition 2.1.9 A covariance Φ on T is a map $\Phi : T \times T \to \mathbb{R}$ such that for all $n \in \mathbb{N}$ and $t_1, \dots, t_n \in T$, the matrix $(\Phi(t_i, t_j))_{i,j=1}^n$ is symmetric and nonnegative definite (i.e., $\Phi(t_i, t_j) = \Phi(t_j, t_i)$ and $\sum_{i,j} a_i a_j \Phi(t_i, t_j) \ge 0$ for all a_i).

The following is a consequence of the Kolmogorov consistency theorem.

Proposition 2.1.10 *Given a covariance Φ on T and a function f on T, there is a Gaussian process $X(t)$ such that $E(X(t)) = f(t)$ and $E[(X(t) - f(t))(X(s) - f(s))] = \Phi(s,t)$ for all $s,t \in T$. Φ is called the covariance of the process and f its expectation, and we say that X is a centred Gaussian process if and only if $f \equiv 0$.*

Proof If $F \subset T$ is finite, take $\mu_F = N((f(t) : t \in F), \Phi|_{F \times F})$. It is easy to see that the set $\{\mu_F : F \subset T,\ F \text{ finite}\}$ is a consistent system of marginals. Hence, by the Kolmogorov consistency theorem, there is a probability on $(\mathbb{R}^T, \mathcal{C})$, hence a process, with $\{\mu_F\}$ as its set of finite-dimensional marginals. ■

Example 2.1.11 A basic example of a Gaussian process is the *isonormal* or *white noise process* on a separable Hilbert space H, where $\{X(h) : h \in H\}$ has a covariance diagonal

for the inner product $\langle \cdot, \cdot \rangle$ of H: $EX(h) = 0$ and $EX(h)X(g) = \langle h, g \rangle_H$ for all $g, h \in H$. The existence of this process does not even require the Kolmogorov consistency theorem but only the existence of an infinite sequence of random variables (i.e., the existence of an infinite product probability space): if $\{g_i\}$ is a sequence of independent $N(0,1)$ random variables and $\{\psi_i\}$ is an orthonormal basis of H, the process defined by linear and continuous extension of $\tilde{X}(\psi_i) = g_i$ (i.e., by $\tilde{X}(\sum a_i \psi_i) = \sum a_i g_i$ whenever $\sum a_i^2 < \infty$) is clearly a version of X. Note for further use that if $V \subset L^2(\Omega, \Sigma, \Pr)$ is the closed linear span of the sequence $\{g_i\}$, then the map $\tilde{X} : H \mapsto V$ is an isometry.

From now on, all our Gaussian processes will be *centred*, even if sometimes we omit mentioning it. If X is a centred Gaussian process on T, the L^2-pseudo-distance between $X(t)$ and $X(s)$ defines a pseudo-distance d_X on T

$$d_X^2(s,t) := E\,(X(t) - X(s))^2 = \Phi(t,t) + \Phi(s,s) - 2\Phi(s,t)$$

that we call the *intrinsic distance* of the process. With this pseudo-metric, T is isometric to the subspace $\{X(t) : t \in T\}$ of $L^2(\Omega, \Sigma, \Pr)$. Clearly, a centred Gaussian process X is continuous in probability for the pseudo-distance d_X; in particular, its probability law in $(\mathbb{R}^T, \mathcal{C})$ is determined by the finite-dimensional marginals based on subsets of any d_X-dense subset T_0 of T.

It is important to note that the probability law of a centred Gaussian process X is completely determined by its intrinsic distance d_X (or by the covariance Φ). Thus, all the probabilistic information about a centred Gaussian process is contained in the metric (or pseudo-metric) space (T, d_X). This is a very distinctive feature of Gaussian processes.

Here is a first, albeit trivial, example of the exact translation of a property of the metric space (T, d_X) into a probabilistic property of X, actually, necessarily of a version of X.

Proposition 2.1.12 *For a Gaussian process X indexed by T, the following are equivalent:*

1. *The pseudo-metric space (T, d_X) is separable, and*
2. *X, as a process on (T, d_X), has a separable, measurable (strict) version.*

Proof If point 2 holds, let $\bar{X}$ be a separable and measurable version of X (in particular, $d_{\bar{X}} = d_X$), and let T_0 be a countable set as in the definition of separability. Then, as mentioned earlier, the very definition of separability implies that $T_0 \cap B_{d_X}(t, \varepsilon) \neq \emptyset$ for all $t \in T$ and $\varepsilon > 0$. Thus, T_0 is dense in (T, d_X), and therefore, (T, d_X) is separable.

Assume now that (T, d_X) is separable, and let T_0 be a countable d_X-dense subset of T. Also assume, as we may by taking equivalence classes, that $d_X(s,t) \neq 0$ for all $s, t \in T_0$, $s \neq t$. If $T_0 = \{s_i : i \in \mathbb{N}\}$, define, for each n, the following partition of T:

$$C_n(s_m) = B\left(s_m, 2^{-n}\right) \setminus \bigcup_{k<m} B\left(s_k, 2^{-n}\right), \quad m \in \mathbb{N}.$$

For each $t \in T$, let $s_n(t)$ be the only $s \in T_0$ such that $t \in C_n(s)$, and define $X_n(t) = X(s_n(t))$. Now $X_n(t, \omega)$ is jointly measurable because $X_n^{-1}(A) = \bigcup_{i \in \mathbb{N}} [C_n(s_i) \times \{\omega : X(s_i, \omega) \in A\}]$. Since, for any $t \in T$, $\Pr\{|X_n(t) - X(t)| > 1/n\} \leq n^2\ E\ (X(s_n(t)) - X(t))^2 \leq n^2/2^{2n}$, it follows by Borel-Cantelli that $X_n(t) \to X(t)$ a.s.

Define $\bar{X}(t, \omega) = \limsup_n X_n(t, \omega)$, which, for each t, is ∞ at most on a set of measure 0. Then the process $\bar{X}(t, \omega)$ is measurable because it is a lim sup of measurable functions. Also,

for each t, $\bar{X}(t) = X(t)$ on a set of measure 1; that is, $\bar{X}$ is a strict version of X. Next we show that $\bar{X}$ is separable. Given $r \in \mathbb{N}$, there exists n_r large enough that $d_X(s_r, s_l) > 1/2^{n_r}$ for all $l < r$; hence, for $n \geq n_r$, $X_n(s_r) = X(s_r)$. This shows that $\bar{X}(s) = X(s)$ for all $s \in T_0$. Then, for all $\omega \in \Omega$,

$$\bar{X}(t,\omega) = \limsup X_n(t,\omega) = \limsup X(s_n(t),\omega) = \limsup \bar{X}(s_n(t),\omega),$$

proving that $\bar{X}$ is separable. ■

Just as with normal random variables, Gaussian processes also satisfy the Gaussian stability property, namely, that if two Gaussian processes with index set T are independent, then their sum is a Gaussian process with covariance the sum of covariances (and mean the sum of means); in particular, if X and Y are independent and equally distributed Gaussian processes (meaning that they have the same finite-dimensional marginal distributions or, what is the same, the same law on the cylindrical σ-algebra $\mathcal{C}$ of $\mathbb{R}^T$), then the process $\alpha X + \beta Y$ has the same law as $(\alpha^2 + \beta^2)^{1/2} X$. This property has many consequences, and here is a nice instance of its use.

Theorem 2.1.13 (0-1 law) *Let $F \subset \mathbb{R}^T$ be a $\mathcal{C}$-measurable linear subspace, and let X be a (centred) Gaussian process indexed by T. Then*

$$\Pr\{X \in F\} = 0 \text{ or } 1.$$

Proof Let X_1 and X_2 be independent copies of X. Define sets

$$A_n = \{X_1 + nX_2 \in F\} \quad \text{and} \quad B_n = \{X_2 \notin F\} \cap A_n, \quad n \in \mathbb{N}.$$

Since $X_1 + nX_2$ is a version of $\sqrt{1+n^2}X$ and F is a vector space, we have

$$\begin{aligned}
\Pr\{B_n\} &= \Pr\{A_n\} - \Pr[A_n \cap \{X_2 \in F\}] \\
&= \Pr\{X \in F\} - \Pr\{X_1 + n\, X_2 \in F,\, X_2 \in F\} \\
&= \Pr\{X \in F\} - \Pr\{X_1 \in F, X_2 \in F\} \\
&= \Pr\{X \in F\} - [\Pr\{X \in F\}]^2.
\end{aligned}$$

Clearly, $B_n \cap B_m = \emptyset$ if $n \neq m$; hence, since by the preceding equalities $\Pr\{B_n\}$ does not depend on n, it follows that $\Pr\{B_n\} = 0$ for all n. But then, again by the same inequalities, $\Pr\{X \in F\}$ can only be 0 or 1. ■

Corollary 2.1.14 *Let X be a centred Gaussian process on T and $\|\cdot\|$ be a $\mathcal{C}$-measurable pseudo-norm on $\mathbb{R}^T$. Then*

$$P\{\|X\| < \infty\} = 0 \text{ or } 1.$$

Proof The set $\{x \in \mathbb{R}^T : \|x\| < \infty\} = \cup_n \{x \in \mathbb{R}^T : \|x\| < n\}$ is a measurable vector space, and the 0-1 law yields the result. ■

Example 2.1.15 If X is Gaussian, separable and centred, then there exists $T_0 \subset T$, T_0 countable, such that $\sup_{t\in T} |X(t)| = \sup_{t\in T_0} |X(t)|$ a.s, but $\|x\|_{T_0} := \sup_{t\in T_0} |x(t)|$ is a measurable pseudo-norm, and hence it is finite with probability 0 or 1.

Example 2.1.16 *The B-valued Gaussian variables where B is a separable Banach space constitute a very general and important class of Gaussian processes, and we define them now. Given a separable Banach space B, a B-valued random variable X is centred Gaussian if $f(X)$ is a mean zero normal variable for every $f \in B^*$, the topological dual of B. By linearity, this is equivalent to the statement that $f_1(X), \ldots, f_n(X)$ are jointly centred normal for every $n \in \mathbb{N}$ and $f_i \in B^*$. In particular, if X is a B-valued centred Gaussian random variable, then the map $X : B^* \mapsto \mathcal{L}^2(\Omega, \Sigma, \Pr)$, defined by $X(f) = f(X)$, is a centred Gaussian process. If $B = E$ has dimension d, X is centred Gaussian iff the coordinates of X in a basis of E are jointly normal with mean zero (hence, the same is true for the coordinates of X in any basis).*

Now we turn to a very useful property of Gaussian processes X, namely, that *the supremum norm of a Gaussian process concentrates about its mean, as well as about its median, with very high probability*, in fact as if it were a real normal variable with variance the largest variance of the individual variables $X(t)$. This result is a consequence of an even deeper result, the isoperimetric inequality for Gaussian measures, although it has simpler direct proofs, particularly if one is allowed some latitude and does not aim at the best result. Here is one such proof that uses the stability property in an elegant and simple way.

We should recall that a function $f : V \mapsto \mathbb{R}$, where V is a metric space, is Lipschitz with Lipschitz constant $c = \|f\|_{\mathrm{Lip}}$ if $c := \sup_{x \neq y} |f(x) - f(y)|/d(x,y) < \infty$. Rademacher's theorem asserts that if $f : \mathbb{R}^n \mapsto \mathbb{R}$ is Lipschitz, then it is a.e. differentiable and the essential supremum of the norm of its derivative is bounded by its Lipschitz constant $\|f\|_{\mathrm{Lip}}$. We remark that although we will use this result in the theorem that follows, it is not needed for its application to a concentration of maxima of jointly normal variables because one can compute by hand the derivative of the Lipschitz function $x \mapsto \max_{i \leq d} |x_i|$, $x \in \mathbb{R}^d$.

Theorem 2.1.17 *Let $(B, \|\cdot\|_B)$ be a finite-dimensional Banach space, and let X be an B-valued centred Gaussian random variable. Let $f : B \mapsto \mathbb{R}$ be a Lipschitz function. Let $\Psi : \mathbb{R} \mapsto \mathbb{R}$ be a nonnegative, convex, measurable function. Then the following inequality holds:*

$$E[\Psi(f(X) - Ef(X))] \leq E\left[\Psi\left(\frac{\pi}{2}\langle f'(X), Y\rangle\right)\right], \tag{2.1}$$

where Y is an independent copy of X (X and Y have the same probability law and are independent), and $\langle \cdot, \cdot \rangle$ denotes the duality action of B^* on B.

Proof Since the range of X is a full subspace, we may assume without loss of generality that B equals the range of X (i.e., the support of the law of X is B). This has the effect that the law of X and Lebesgue measure on B are mutually absolutely continuous (as the density of X is strictly positive on its supporting subspace). For $\theta \in [0, 2\pi)$, define $X(\theta) = X\sin\theta + Y\cos\theta$. Then $X'(\theta) = X\cos\theta - Y\sin\theta$, and notice that $X(\theta)$ and $X'(\theta)$ are (normal and) independent: it suffices to check covariances, and if $f, g \in B^*$, we have

$$E[f(X(\theta))g(X'(\theta))] = E(f(X)g(X))\sin\theta\cos\theta - E(f(Y)g(Y))\sin\theta\cos\theta = 0.$$

In other words, the joint probability laws of X and Y and of $X(\theta)$ and $X'(\theta)$ coincide.

Since for any increasing sequence θ_i

$$\sum |f(X(\theta_i)) - f(X(\theta_{i-1}))| \le \|f\|_{\mathrm{Lip}} \sum \|X(\theta_i) - X(\theta_{i-1})\|$$
$$\le \|f\|_{\mathrm{Lip}}(\|X\| + \|Y\|) \sum |\theta_i - \theta_{i-1}|,$$

it follows that the function $\theta \mapsto f(X(\theta))$ is absolutely continuous, and therefore, we have

$$f(X) - f(Y) = f(X(\pi/2)) - f(X(0)) = \int_0^{\pi/2} \frac{d}{d\theta} f(X(\theta))\, d\theta.$$

Using convexity of Ψ, Fubini's theorem and the preceding, we obtain

$$E\Psi(f(X) - Ef(X)) = E\Psi(f(X) - Ef(Y)) \le E\Psi(f(X) - f(Y))$$
$$= E\Psi\left(\int_0^{\pi/2} \frac{d}{d\theta} f(X(\theta)) d\theta\right) \le \frac{2}{\pi} E \int_0^{\pi/2} \Psi\left(\frac{\pi}{2} \frac{d}{d\theta} f(X(\theta))\right) d\theta$$
$$= \frac{2}{\pi} \int_0^{\pi/2} E\Psi\left(\frac{\pi}{2} \frac{d}{d\theta} f(X(\theta))\right) d\theta.$$

Now f is m a.e. differentiable with a bounded derivative by Rademacher's theorem, where m is Lebesgue measure on B, and since $\mathcal{L}(X(\theta))$ is absolutely continuous with respect to Lebesgue measure for every $\theta \in [0, \pi/2)$ ($X(\theta)$ has the same support as X), f' exists a.s. relative to the law of $X(\theta)$. Since $X'(\theta)$ exists for each θ, it follows from the chain rule that given θ, $df(X(\theta))/d\theta = \langle f'(X(\theta)), X'(\theta)\rangle$ a.s. Then, since $\mathcal{L}(X, Y) = \mathcal{L}(X(\theta), X'(\theta))$, we have

$$E\Psi\left(\frac{\pi}{2} \frac{d}{d\theta} f(X(\theta))\right) = E\Psi\left(\frac{\pi}{2} \langle f'(X), Y\rangle\right),$$

which, combined with the preceding string of inequalities, proves the theorem. ■

Remark 2.1.18 It turns out, as we will see in the next section, that Lipschitz functions are the natural tool for extracting concentration results from isoperimetric inequalities, on the one hand, and on the other, as we will see now, the supremum norm of a vector in $\mathbb{R}^n$ is a Lipschitz function, so concentration inequalities for Lipschitz functions include as particular cases concentration inequalities for the supremum norm and for other norms as well.

Example 2.1.19 (**Concentration for the maximum of a finite number of jointly normal variables**) To estimate the distribution of $\max_{i \le n} |g_i|$ for a finite sequence $g_1, \ldots, g_n$ of jointly normal variables using the preceding theorem, we take $B = \ell_\infty^n$, which is $\mathbb{R}^n$ with the norm $f(x) = \max_{i \le n} |x_i|$, where $x = (x_1, \ldots, x_n)$, which we take as our function f, and we take $X = (g_1, \ldots, g_n)$. f is obviously Lipschitz, so the previous theorem will apply to it. We also have that for each $1 \le i \le n$, $f(x) = x_i$ on the set $\{x : x_i > |x_j|, 1 \le j \le n, j \ne i\}$ and $f(x) = -x_i$ on $\{x : -x_i > |x_j|, 1 \le j \le n, j \ne i\}$. It follows that m a.s. the gradient of f has all but one coordinate equal to zero, and this coordinate is 1 or -1. If $g_i \ne \pm g_j$ for $i \ne j$, which we can assume without loss of generality (by deleting repeated coordinates without changing the maximum), then this also holds a.s. for the law of X. Let $\sigma_i^2 = Eg_i^2$ and $\sigma^2 = \max_{i \le n} \sigma_i^2$. For almost every $X = x$ fixed, $\langle f'(x), Y\rangle$ is $\pm g_i$ for some i, that is, in law, the same as $\sigma_i g$, g standard normal. Therefore, if we assume that the function Ψ is as in the preceding

theorem and that, moreover, it is even and nondecreasing on $[0,\infty)$, then, letting E_Y denote integration with respect to the variable Y only, the preceding observation implies that, X a.s.,

$$E_Y \Psi\left(\frac{\pi}{2}\langle f'(x), Y\rangle\right) \le E\Psi\left(\frac{\pi}{2}\sigma g\right).$$

We conclude that for any $n \in \mathbb{N}$, if $g_1,\dots,g_n$ are jointly normal random variables and if $\sigma^2 = \max_{i\le n} Eg_i^2$, then for any nonnegative, even, convex function Ψ nondecreasing on $[0,\infty)$,

$$E\Psi\left(\max_{i\le n}|g_i| - E\max_{i\le n}|g_i|\right) \le E\Psi\left(\frac{\pi}{2}\sigma g\right), \tag{2.2}$$

where g denotes a standard normal random variable.

Now $Ee^{t|g|} \le E(e^{tg}+e^{-tg}) = 2e^{t^2/2}$. Thus, if $\Psi_\lambda(x) = e^{\lambda|x|}$, we have

$$E\Psi_\lambda\left(\frac{\pi}{2}\sigma g\right) \le 2e^{\lambda^2\pi^2\sigma^2/8}$$

and, by (2.2) and Chebyshev's inequality,

$$\Pr\left\{\left|\max_{i\le n}|g_i| - E\max_{i\le n}|g_i|\right| > u\right\} \le 2e^{-\lambda u+\lambda^2\pi^2\sigma^2/8}, \quad u \ge 0.$$

With $\lambda u/2 = \lambda^2\pi^2\sigma^2/8$, that is, $\lambda = 4u/(\pi^2\sigma^2)$, this inequality gives the following approximate concentration inequality about its mean for the maximum of any finite number of normal random variables:

$$\Pr\left\{\left|\max_{i\le n}|g_i| - E\max_{i\le n}|g_i|\right| > u\right\} \le 2e^{-\frac{4}{\pi^2}\frac{u^2}{2\sigma^2}}, \quad u \ge 0. \tag{2.3}$$

The last inequality and the one in the next theorem are suboptimal: the factor $4/\pi^2$ in the exponent is superfluous, as we will see in two of the sections that follow. We can translate (2.2) and (2.3) into a concentration inequality for the supremum norm of a separable Gaussian process (and draw as well some consequences).

Theorem 2.1.20 *Let $\{X(t),\ t\in T\}$ be a separable centred Gaussian process such that*

$$\Pr\{\sup_{t\in T}|X(t)| < \infty\} > 0.$$

Let Ψ be an even, convex, measurable function, nondecreasing on $[0,\infty)$. Let g be $N(0,1)$. Then,

a. $\sigma = \sigma(X) := \sup_{t\in T}\left(EX^2(t)\right)^{1/2} < \infty$ and $E\sup_{t\in T}|X(t)| < \infty$ and
b. The following inequalities hold:

$$E\Psi\left(\sup_{t\in T}|X(t)| - E\sup_{t\in T}|X(t)|\right) \le E\Psi\left(\frac{\pi}{2}\sigma g\right)$$

and

$$\Pr\left\{\left|\sup_{t\in T}|X(t)| - E\sup_{t\in T}|X(t)|\right| > u\right\} \le 2e^{-(Ku^2/2\sigma^2)},$$

where $K = \frac{4}{\pi^2}$.

(As mentioned earlier, the optimal constant K in this theorem will be shown to be 1.)

Proof By assumption and the 0-1 law (Theorem 2.1.13; see the example following Corollary 2.1.14), $\sup_{t\in T}|X(t)|<\infty$ a.s. Let $0<z_{1/2}<1$ be such that $\Pr\{|g|>z_{1/2}\}=1/2$, and let $M<\infty$ be such that $\Pr\left\{\sup_{t\in T}|X(t)|>M\right\}<1/2$. Then, for each t,

$$1/2>\Pr\{|X(t)|>M\}=\Pr\{|g|>M/(EX(t)^2)^{1/2}\},$$

which implies that $\sigma=\sup_{t\in T}(EX^2(t))^{1/2}\le M/z_{1/2}<\infty$.

Let $T_0=\{t_i\}_{i=1}^{\infty}$ be a countable set such that $\sup_{t\in T}|X(t)|=\sup_{t\in T_0}|X(t)|$. For every $n\in\mathbb{N}$, we have, by inequality (2.3),

$$\Pr\left\{\left|\max_{i\le n}|X(t_i)|-E\max_{i\le n}|X(t_i)|\right|>\sigma u\right\}\le 2e^{-2u^2/\pi^2}.$$

Since $\sup_{t\in T}|X(t)|<\infty$ a.s., this variable has a finite median m, and also for all n,

$$\Pr\left\{\max_{i\le n}|X(t_i)|\le m\right\}\ge\frac{1}{2}.$$

If u_0 is such that $2e^{-2u_0^2/\pi^2}<1/2$, these two inequalities imply that for all $n\in\mathbb{N}$, the intersection of the two sets $\left\{x:\left|E\max_{i\le n}|X(t_i)|-x\right|\le\sigma u_0\right\}$ and $\{x:x\le m\}$ is not empty and hence that $E\max_{i\le n}|X(t_i)|\le m+\sigma u_0<\infty$. a) is proved.

We have $\sup_{t\in T}|X(t)|=\lim_{n\to\infty}\max_{i\le n}|X(t_i)|$ a.s. and, by monotone convergence, also in $L^1(\Pr)$. Hence, the first inequality in (b) follows by inequality (2.2), continuity of Ψ and Fatou's lemma. The second inequality follows from the first by Chebyshev's inequality in the same way as (2.3) follows from (2.2). ■

Exercises

In Exercises 2.1 to 2.4 we write $\|X\|$ for $\sup_{t\in T}|X(t)|$, and X denotes a separable, centred Gaussian process such that $\Pr\left\{\sup_{t\in T}|X(t)|<\infty\right\}>0$. Also, for any random variable ξ, $\|\xi\|_p$ will denote its L^p-norm.

2.1.1 Prove that there exists $\alpha>0$ such that $Ee^{\alpha\|X\|^2}<\infty$.

2.1.2 Use results from this section to show that for all $p\ge 1$,

$$(E\|X\|^p)^{1/p}\le K\sqrt{p}E\|X\|$$

for a universal constant $K<\infty$. *Hint*: Integrating the exponential inequality in Theorem 2.1.20 with respect to $pt^{p-1}dt$ yields $\|\|X\|-E\|X\|\|_p\le c\sigma\|g\|_p$, where g is standard normal and c a universal constant. Check that $\|g\|_p$ is of the order of $\sqrt{p}$.

2.1.3 Prove that the median m of $\|X\|$, satisfies $KE\|X\|\le m\le 2E\|X\|$ for a universal constant $K>0$. *Hint*: The second inequality is obvious, and the first is contained in the proof of Theorem 2.1.20.

2.1.4 Prove that if X_n are separable, centred Gaussian processes such that $\Pr\{\|X_n(t)\|<\infty\}>0$, then $\|X_n\|\to 0$ in pr. iff $\|X_n\|\to 0$ in L^p for some $p\ge 1$ iff $\|X_n\|\to 0$ in L^p for all $p\ge 1$. *Hint*: L^p convergences for different p are equivalent by Exercise 2.1.2, and the equivalence extends to convergence in probability by Exercise 2.1.2 and the *Paley-Zygmund argument* as follows: for any $0<\tau<1$,

$$E\|X\|\le\tau E\|X\|+E(\|X\|I_{\|X\|>\tau E\|X\|})\le\tau E\|X\|+\left(E\|X\|^2\right)^{1/2}\left(\Pr\{\|X\|>\tau E\|X\|\}\right)^{1/2},$$

so

$$\Pr\{\|X\|>\tau E\|X\|\}\ge\left[\frac{(1-\tau)E\|X\|}{\left(E\|X\|^2\right)^{1/2}}\right]^2\ge K(1-\tau)^2$$

for a universal constant K. Thus, if $\|X_n\| \to 0$ in probability, then $E\|X_n\| \to 0$.

2.1.5 Let B be a separable Banach space, and let X be a B-valued Gaussian (centred) random variable. Show that the previous theorems apply to $\|X\|$ where now $\|\cdot\|$ is the Banach space norm. *Hint*: Use Example 2.1.6.

2.1.6 Prove Proposition 2.1.4. *Hint*: Recall that a subset of a complete separable metric space S is compact if and only if it is closed and totally bounded. Given $\varepsilon > 0$, by separability, for each n there exists a finite collection $\{F_{n,k}\}_{k=1}^{k_n}$ of closed sets of diameter not exceeding n^{-1} and such that $\mu\left(\cup_{k=1}^{k_n} F_{n,k}\right)^c < \varepsilon/2^n$. The set $K = \cap_{n=1}^{\infty} \cup_{k=1}^{k_n} F_{n,k}$ is compact and satisfies $\mu(K^c) < \varepsilon$.

2.2 Isoperimetric Inequalities with Applications to Concentration

The Gaussian isoperimetric inequality, in its simplest form, identifies the half-spaces as the sets of $\mathbb{R}^n$ with the smallest Gaussian perimeter among those with a fixed Gaussian measure, where the Gaussian measure in question is the standard one, that is, the probability law of n independent standard normal random variables, and where the Gaussian perimeter of a set is taken as the limit of the measure of the difference of an ε-enlargement of the set and the set itself divided by ε. The proof of this theorem was obtained originally by translating the isoperimetric inequality on the sphere to the Gaussian setting by means of Poincaré's lemma, which states that the limiting distribution of the orthogonal projection onto a Euclidean space of fixed dimension n of the uniform distribution on the sphere of $\mathbb{R}^{m+1}$ with radius $\sqrt{m}$ is the standard Gaussian measure of $\mathbb{R}^n$. The isoperimetric inequality on the sphere is a deep result that goes back to P. Lévy and E. Schmidt, ca. 1950 (although the equivalent isoperimetric problem on the plane goes back to the Greeks–recall, for instance, 'Dido's problem'). The Gaussian isoperimetric inequality does imply best possible concentration inequalities for Lipschitz functions on $\mathbb{R}^n$ and for functions on $\mathbb{R}^{\mathbb{N}}$ that are Lipschitz 'in the direction of ℓ_2', although concentration inequalities have easier proofs, as seen in the preceding section and as will be seen again in further sections. The Gaussian isoperimetric inequality in general Banach spaces requires the notion of reproducing kernel Hilbert space and will be developed in a further section as well. This section contains proofs as short as we could find of the isoperimetric inequalities on the sphere and for the standard Gaussian measure on $\mathbb{R}^n$, $n \leq \infty$, with applications to obtain the best possible concentration inequality with respect to the standard Gaussian measure for Lipschitz functions f about their medians and for the supremum norm of a separable Gaussian process X when $\sup_{t\in T} |X(t)| < \infty$ a.s.

2.2.1 The Isoperimetric Inequality on the Sphere

Let $S^n = \left\{x \in \mathbb{R}^{n+1} : \|x\|^2 = \sum_{i=1}^{n+1} x_i^2 = 1\right\}$, where $x = (x_1, \ldots, x_{n+1})$; let p be an arbitrary point in S^n that we take to be the north pole, $p = (0, \ldots, 0, 1)$; and let μ be the uniform probability distribution on S^n (equal to the normalized volume element – surface area for S^2 – equal also to the normalized Haar measure of the rotation group). Let d be the geodesic distance on S^n, defined, for any two points, as the length of the shortest segment of the great circle joining them.

A closed *cap* centred at a point $x \in S^n$ is a geodesic closed ball around x, that is, a set of the form $C(x, \rho) := \{y : d(x, y) \leq \rho\}$. Here ρ is the radius of the cap, and clearly, the μ-measure

of a cap is a continuous function of its radius, varying between 0 and 1. Often we will not specify the centre or the radius of $C = C(x, \rho)$, particularly if the centre is the north pole.

The isoperimetric inequality on the sphere states that the caps are the sets of shortest perimeter among all the measurable sets of a given surface area. What we will need is an equivalent formulation, in terms of neighbourhoods of sets, defined as follows: the closed ε neighbourhood of a set A is defined as $A_\varepsilon = \{x : d(x,A) \le \varepsilon\}$, with the distance between a point and a set being defined, as usual, by $d(x,A) = \inf\{d(x,y) : y \in A\}$. The question is: among all measurable subsets of the sphere with surface area equal to the surface area of A, find sets B for which the surface areas of their neighbourhoods B_ε, $0 < \varepsilon < 1$, are smallest. The following theorem shows that an answer are the caps (they are in fact *the* answer, but uniqueness will not be considered: we are only interested in the value of $\inf \mu(A_\varepsilon)$, $\varepsilon > 0$).

Theorem 2.2.1 *Let $A \neq \emptyset$ be a measurable subset of S^n, and let C be a cap such that $\mu(C) = \mu(A)$. Then, for all $\varepsilon > 0$,*

$$\mu(C_\varepsilon) \le \mu(A_\varepsilon). \tag{2.4}$$

The proof is relatively long, and some prior digression may help. The idea is to construct transformations $A \mapsto A^*$ on measurable subsets of the sphere that preserve area, that is, $\mu(A) = \mu(A^*)$, and decrease perimeter, a condition implied by $\mu((A^*)_\varepsilon) \le \mu(A_\varepsilon) = \mu((A_\varepsilon)^*)$, $\varepsilon > 0$, because the perimeter of A is the limit as $\varepsilon \to 0$ of $\mu(A_\varepsilon \setminus A)/\varepsilon$. Then iterating transformations that satisfy these two properties should eventually produce the solution, in our case a cap. Or, more directly, one may obtain a cap using a more synthetic compactness argument instead of iteration. In the sense that A^* concentrates the same area as A on a smaller perimeter, A^* is closer to the solution of the problem than A is. A^* is called a *symmetrisation* of A.

Proof If $\mu(A) = 0$, then C consists of a single point, and (2.4) holds. Next, we observe that by regularity of the measure μ, it suffices to prove the theorem for A compact. By regularity, there exist A^m compact, $A^m \subset A$, A^m increasing and such that $\mu(A^m) \nearrow \mu(A)$. Let C^m be caps with the same centre as C and with $\mu(C^m) = \mu(A^m)$. Since the measure of a cap is a continuous one-to-one function of its geodesic radius, we also have $\mu(C^m_\varepsilon) \nearrow \mu(C_\varepsilon)$, and if the theorem holds for compact sets, then

$$\mu(A_\varepsilon) \ge \lim \mu(A^m_\varepsilon) \ge \lim \mu(C^m_\varepsilon) = \mu(C_\varepsilon),$$

and the theorem holds in general. Thus, we will assume that A is compact and that $\mu(A) \neq 0$. We divide the proof into several parts.

Part 1: Construction and main properties of the symmetrisation operation. Given an n-dimensional subspace $H \subset \mathbb{R}^{n+1}$ that does not contain the point p, let $\sigma = \sigma_H$ be the reflection about H; that is, if $x = u + v$ with $u \in H$ and v orthogonal to H, then $\sigma(x) = u - v$. Clearly, σ is an isometry (so it preserves μ-measure), and it is involutive; that is, $\sigma^2 = \sigma$. It also satisfies a property that, together with the preceding two, is crucial for the symmetrisation operation to work, namely, that if x and y are on the same half-space with respect to H, then

$$d(x,y) \le d(x,\sigma(y)). \tag{2.5}$$

To see this, observe that the geodesic distance is an increasing function of the Euclidean distance, so it suffices to prove (2.5) for the Euclidean distance. Changing orthogonal coordinates if necessary, we may and do assume that $H = \{x : x_{n+1} = 0\}$, so if x and y are in the same hemisphere, then $\text{sign}(x_{n+1}) = \text{sign}(y_{n+1})$, which implies that the $(n+1)$th coordinate of $x - y$ is dominated in absolute value by the $(n+1)$th coordinate of $x - \sigma(y)$, whereas the first n coordinates of these two vectors coincide. Hence, $\sum_{i=1}^{n+1}(x_i - y_i)^2 \le \sum_{i=1}^{n+1}(x_i - \sigma(y)_i)^2$.

H divides S^n into two open hemispheres, and we denote by S_+ the open hemisphere that contains p, S_- the other hemisphere, and $S_0 = S^n \cap H$. The symmetrisation of A with respect to $\sigma = \sigma_H$, $s_H(A) = A^*$ is defined as

$$s_H(A) = A^* := [A \cap (S_+ \cup S_0)] \cup \{a \in A \cap S_- : \sigma(a) \in A\} \cup \{\sigma(a) : a \in A \cap S_-, \sigma(a) \notin A\}. \tag{2.6}$$

Note that A^* is obtained from A by reflecting towards the northern hemisphere every $a \in A \cap S_-$ for which $\sigma(A)$ is not already in A. It is easy to see (Exercise 2.2.1) that if A is compact, then so is A^* and that if C is a cap with centre at p or at any other point in the northern hemisphere, then $C^* = C$. Next, observe that the three sets in the definition are disjoint and that, σ being an isometry, the measure of the third set equals $\mu\{a \in A \cap S_- : \sigma(a) \notin A\}$, which implies that

$$\mu(A^*) = \mu(A), \quad A \in \mathcal{B}(S^{n+1}). \tag{2.7}$$

This is one of the two properties of the symmetrisation operation that we need.

We now show that the ε-neighbourhoods of A^* are less massive than those of A (thus making A^* 'closer' to being a cap than A is), actually, we prove more, namely, that for all $A \in \mathcal{B}(S^n)$ and $\varepsilon > 0$, then

$$(A^*)_\varepsilon \subseteq (A_\varepsilon)^*, \quad \text{hence} \quad \mu((A^*)_\varepsilon) \le \mu((A_\varepsilon)^*) = \mu(A_\varepsilon). \tag{2.8}$$

To see this, let $x \in (A^*)_\varepsilon$ and let $y \in A^*$ be such that $d(x,y) \le \varepsilon$ (such a $y \in A^*$ exists by compactness). Then, using (2.5) and that σ is an involutive isometry, we obtain, when x and y lay on different half-spaces,

$$d(\sigma(x), y) = d(x, \sigma(y)) \le d(\sigma(x), \sigma(y)) = d(x,y) \le \varepsilon.$$

Thus, since $y \in A^*$ implies that either $y \in A$ or $\sigma(y) \in A$, in either case we have that both $x \in A_\varepsilon$ and $\sigma(x) \in A_\varepsilon$; hence, $x \in (A_\varepsilon)^*$. If x and y are in S_-, then y and $\sigma(y)$ are both in A, and therefore, by the last identity earlier, $x \in A_\varepsilon$ and $\sigma(x) \in A_\varepsilon$; hence, $x \in (A_\varepsilon)^*$ in this case as well. If x and y are in S_+, then either y or $\sigma(y)$ is in A; hence, either x or $\sigma(x)$ is in A_ε, which together with $x \in S_+$ implies that $x \in (A_\varepsilon)^*$. The cases where x and/or y are in S_0 are similar, even easier, and they are omitted. The inclusion in (2.8) is proved, and the inequality there follows from the inclusion and from (2.7).

Part 2: Preparation for the compactness argument. Let $(\mathcal{K}, h)$ denote the set of nonempty compact subsets of S^n equipped with the Hausdorff distance, defined as $h(A,B) = \inf\{\varepsilon : A \subseteq B_\varepsilon,\ B \subseteq A_\varepsilon\}$, $A, B \in \mathcal{K}$. $(\mathcal{K}, h)$ is a compact metric space (Exercise 2.2.2). Given a compact nonempty set $A \subseteq S^n$, let $\mathcal{A}$ be the minimal closed subset of $\mathcal{K}$ that contains A and is preserved by s_H for all n-dimensional subspaces H of $\mathbb{R}^{n+1}$ that do not contain the north pole p (meaning that if $A \in \mathcal{K}$, then $s_H(A) \in \mathcal{K}$ for all H with $p \notin H$). $\mathcal{A}$ exists and is

nonempty because $\mathcal{K}$ is a closed $\{s_H\}$-invariant collection of sets that contains A. Also note that since $(\mathcal{K},h)$ is compact and $\mathcal{A}$ closed, $\mathcal{A}$ is compact. We have

Claim: If $B \in \mathcal{A}$, then (a) $\mu(B) = \mu(A)$, and (b) for all $\varepsilon > 0$, $\mu(B_\varepsilon) \le \mu(A_\varepsilon)$.
Proof of the claim. It suffices to show that the collection of closed sets $\mathcal{F}$ satisfying a) and b) is preserved by s_H for all H not containing p and is a closed subset of $\mathcal{K}$ because then $\mathcal{A} \subseteq \mathcal{F}$ follows by minimality of $\mathcal{A}$. That $s_H(\mathcal{F} \Rightarrow \subseteq \mathcal{F}$ follows from (2.7) and (2.8). Let now $B^n \in \mathcal{F}$ and $h(B^n,B) \to 0$. Let $\varepsilon > 0$ be fixed. Given $\delta > 0$, there exists n_δ such that $B \subseteq B^n_\delta$ for all $n \ge n_\delta$; hence, $B_\varepsilon \subseteq B^n_{\delta+\varepsilon}$ and $\mu(B_\varepsilon) \le \mu(B^n_{\delta+\varepsilon}) \le \mu(A_{\delta+\varepsilon})$. Letting $\delta \searrow 0$ shows that B satisfies condition (b). Letting $\varepsilon \searrow 0$ in condition (b) for B shows that $\mu(B) \le \mu(A)$. Using that for all n large enough we also have $B^n \subseteq B_\delta$, we get that $\mu(A) = \mu(B^n) \le \mu(B_\delta)$ and, letting $\delta \searrow 0$, that $\mu(A) \le \mu(B)$, proving condition (a). The claim is proved.

Part 3: Completion of the proof of Theorem 2.2.1. Clearly, because of the claim about $\mathcal{A}$, it suffices to show that if C is the cap centred at p such that $\mu(A) = \mu(C)$, then $C \in \mathcal{A}$.

Define $f(B) = \mu(B \cap C)$, $B \in \mathcal{A}$. We show first that f is upper semicontinuous on $\mathcal{A}$. If $h(B^n,B) \to 0$, then, given $\delta > 0$, for all n large enough, $B^n \subseteq B_\delta$, which, as is easy to see, implies that $B^n \cap C \subseteq (B \cap C_\delta)_\delta$. Hence, $\limsup_n \mu(B^n \cap C) \le \mu((B \cap C_\delta)_\delta)$, but because B and C are closed, if $\delta_n \searrow 0$, then $\cap_n (B \cap C_{\delta_n})_{\delta_n} = B \cap C$, thus obtaining $\limsup_n \mu(B^n \cap C) \le \mu(B \cap C)$.

Since f is upper semicontinuous on $\mathcal{A}$ and $\mathcal{A}$ is compact, f attains its maximum at some $B \in \mathcal{A}$. The theorem will be proved if we show that $C \subseteq B$. Assume that $C \not\subset B$. Then, since $\mu(C) = \mu(A) = \mu(B)$ and both C and B are closed, we have that both $B \setminus C$ and $C \setminus B$ have positive μ-measure. Thus, the Lebesgue density theorem, which holds on S^n (see Exercise 2.2.3 for definitions and a sketch of the proof), implies that there exist points of density $x \in B \setminus C$ and $y \in C \setminus B$. Let H be the subspace of dimension n orthogonal to the vector $x - y$, and let us keep the shorthand notation σ for the reflection with respect to H, D^* for $s_H(D)$, S_+, S_- for the two hemispheres determined by H, and S_0 for $S^n \cap H$. Then $\sigma(y) = x$. Since $y \in C$ and $x \notin C$, we have both, that p is not in H (the reflection of a point in C with respect to a hyperplane through p is necessarily in C) and that y is closer to p than x is; that is, $d(y,p) \le d(x,p) = d(\sigma_H(y),p)$. Then it follows from this last obsesrvation and (2.5) that $y \in S_+$ and $x \in S_-$.

Let $x \in (B \cap C)^*$. Then, if $x \in B \cap C \cap (S_+ \cup S_0)$ or if $x \in B \cap C \cap S_-$ and $\sigma(x) \in B \cap C$, we obviously have $x \in B^* \cap C$. Now, if $z \in C \cap S_-$, then $\sigma(z) \in C$ (as $\sigma(z)$ is closer to p than z is); hence, if $x = \sigma(z)$ with $z \in B \cap C \cap S_-$ and $\sigma(z) \notin B \cap C$, then $\sigma(z)$ is not in B and therefore $x \in B^* \cap C$. We conclude that $(B \cap C)^* \subseteq B^* \cap C$ and, in particular, that

$$\mu(B \cap C) = \mu((B \cap C)^*) \le \mu(B^* \cap C). \tag{2.9}$$

By definition of density point, for $\delta > 0$ small enough, $C(x,\delta) \subset S_-$, $\sigma(C(x,\delta)) = C(y,\delta) \subset S_+$, $\mu((B \setminus C) \cap C(x,\delta)) \ge 2\mu(C(x,\delta))/3$, and $\mu((C \setminus B) \cap C(y,\delta)) \ge 2\mu(C(y,\delta))/3$. Then the set

$$D = ((B \setminus C) \cap C(x,\delta)) \cap \sigma((C \setminus B) \cap C(y,\delta))$$

satisfies

$$\mu(D) \ge \mu(C(x,\delta))/3 > 0, \quad D \subset (B \setminus C) \cap S_- \quad \text{and} \quad \sigma(D) \subset C \setminus B. \tag{2.10}$$

The inclusions in (2.10) imply that $\sigma(D) \subset B^* \cap C$ and $\sigma(D) \cap (B \cap C)^* = \emptyset$ (as $z \in (B \cap C)^*$ implies either $z \in B \cap C$ or $\sigma(z) \in B \cap C$). This together with (2.9) and $\mu(D) > 0$ proves

$$\mu(B^* \cap C) \geq \mu((B \cap C)^* \cup \sigma(D)) = \mu((B \cap C)^*) + \mu(D) > \mu((B \cap C)^*),$$

which, because $B^* \in \mathcal{A}$, contradicts the fact that f attains it maximum at B. ■

2.2.2 The Gaussian Isoperimetric Inequality for the Standard Gaussian Measure on $\mathbb{R}^{\mathbb{N}}$

In this subsection we translate the isoperimetric inequality on the sphere to an isoperimetric inequality for the probability law γ_n of n independent $N(0,1)$ random variables by means of Poincaré's lemma, which states that this measure can be obtained as the limit of the projection of the uniform distribution on $\sqrt{m}S^{n+m}$ onto $\mathbb{R}^n$ when $m \to \infty$. We also let $n \to \infty$.

In what follows, g_i, $i \in \mathbb{N}$, is a sequence of independent $N(0,1)$ random variables, and as mentioned earlier, $\gamma_n = \mathcal{L}(g_1, \ldots, g_n)$. We call γ_n the *standard Gaussian measure on* $\mathbb{R}^n$. We also set $\gamma = \mathcal{L}(\{g_i\}_{i=1}^{\infty})$, the law of the process $i \mapsto g_i$, $i \in \mathbb{N}$, a probability measure on the cylindrical σ-algebra $\mathcal{C}$ of $\mathbb{R}^{\mathbb{N}}$, which we also refer to as the *standard Gaussian measure on* $\mathbb{R}^{\mathbb{N}}$.

Here is the Gaussian isoperimetric problem: for a measurable subset A of $\mathbb{R}^n$, and $\varepsilon > 0$, define its Euclidean neighbourhoods A_ε as $A_\varepsilon := \{x \in \mathbb{R}^n : d(x,A) \leq \varepsilon\} = A + \varepsilon O_n$, where d denotes Euclidean distance and O_n is the closed d-unit ball centred at $0 \in \mathbb{R}^n$. The problem is this: given a Borel set A, find among the Borel sets $B \subset \mathbb{R}^n$ with the same γ_n-measure as A those for which the γ_n-measure of the neighbourhood B_ε is smallest, for all $0 < \varepsilon < 1/2$. The solution will be shown to be the affine half-space ($\{x : \langle x, u \rangle \leq \lambda\}$, u any unit vector, $\lambda \in \mathbb{R}$) of the same measure as A. Note that $\gamma_n\{x : \langle x, u \rangle \leq \lambda\} = \gamma_1\{x \leq \lambda\}$.

Prior to stating and proving the main results, we describe the relationship between the uniform distribution on the sphere of increasing radius and dimension and the standard Gaussian measure on $\mathbb{R}^n$.

Lemma 2.2.2 (**Poincaré's lemma**) *Let μ_{n+m} be the uniform distribution on $\sqrt{m}S^{n+m}$, the sphere of $\mathbb{R}^{n+m+1}$ of radius $\sqrt{m}$ and centred at the origin. Let π_m be the orthogonal projection $\mathbb{R}^{n+m+1} \mapsto \mathbb{R}^n = \{x \in \mathbb{R}^{n+m+1} : x_i = 0, n < i \leq n+m+1\}$, and let $\tilde{\pi}_m$ be the restriction of π_m to $\sqrt{m}S^{n+m}$. Let $\nu_m = \mu_{n+m} \circ \tilde{\pi}_m^{-1}$ be the projection onto $\mathbb{R}^n$ of μ_{n+m}. Then ν_m has a density f_m such that if ϕ_n is the density of γ_n, $\lim_{m\to\infty} f_m(x) = \phi_n(x)$ for all $x \in \mathbb{R}^n$. Therefore,*

$$\gamma_n(A) = \lim_{m\to\infty} \mu_{n+m}(\tilde{\pi}_m^{-1}(A)) \tag{2.11}$$

for all Borel sets A of $\mathbb{R}^n$.

Proof Set $G_n := (g_1, \ldots, g_n)$ and $G_{n+m+1} := (g_1, \ldots, g_{n+m+1})$. The rotational invariance of the standard Gaussian law on Euclidean space implies that μ_{n+m} is the law of the vector $\sqrt{m}G_{n+m+1}/|G_{n+m+1}|^{1/2}$. Hence, ν_m is the law of $\sqrt{m}G_n/|G_{n+m+1}|^{1/2}$. This allows for computations with normal densities that we only sketch. For any measurable set A of $\mathbb{R}^n$,

$$\nu_m(A) = \frac{1}{(2\pi)^{(n+m+1)/2}} \int_{\mathbb{R}^{m+1}} \int_{\tilde{A}(y)} e^{-(|z|^2+|y|^2)/2} dz\, dy$$

where $z \in \mathbb{R}^n$ and $y \in \mathbb{R}^{m+1}$, and $\tilde{A} = \left\{ z \in \mathbb{R}^n : \sqrt{m/(|z|^2+|y|^2)}\, z \in A \right\}$. Make the change of variables $z \mapsto x$, $x = \sqrt{m/(|z|^2+|y|^2)}\, z$ or $z = |y|x/\sqrt{m-|x|^2}$, $|x| \le \sqrt{m}$. Its Jacobian is $\partial(z)/\partial(x) = m|y|^n/(m-|x|^2)^{1+n/2}$, thus obtaining

$$\begin{aligned} \nu_m(A) &= \frac{1}{(2\pi)^{(n+m+1)/2}} \int_A I(|x|^2 < m) \frac{m}{(m-|x|^2)^{n/2+1}} \int_{\mathbb{R}^{m+1}} |y|^n \exp\left(-\frac{1}{2} \frac{m|y|^2}{m-|x|^2} \right) dy\, dx \\ &= \frac{E(|G_{m+1}|^n)}{m^{n/2}} \frac{1}{(2\pi)^{n/2}} \int_A \left(1 - \frac{|x|^2}{m} \right)^{(m-1)/2} I(|x|^2 < m)\, dx. \end{aligned}$$

Hence, the density of ν_m is $f_m(x) = C_{n,m}(2\pi)^{-n/2}(1-|x|^2/m)^{(m-1)/2} I(|x|^2 < m)$, $x \in \mathbb{R}^n$. Clearly, $(2\pi)^{-n/2}(1-|x|^2/m)^{(m-1)/2} I(|x|^2 < m) \to (2\pi)^{-n/2} e^{-|x|^2/2}$ for all x as $m \to \infty$. Moreover, since for $0 \le a < m$ and $m \ge 2$ we have $1 - a/m \le e^{-a/2(m-1)}$, it follows that $(1-|x|^2/m)^{(m-1)/2} I(|x|^2 < m)$ is dominated by the integrable function $e^{-|x|^2/4}$. Thus, by the dominated convergence theorem, $f_m(x)/C_{n,m} \to (2\pi)^{-n/2} e^{-|x|^2/2}$ in L^1, which implies that $C_{n,m}^{-1} \to 1$, proving the lemma. (Alternatively, just show that $C_{n,m} = E(|G_{m+1}|^n)/m^{n/2} \to 1$ as $m \to \infty$ by taking limits on well-known expressions for the moments of chi-square random variables.) Now the limit (2.11) for any Borel set follows by dominated convergence. ■

Theorem 2.2.3 *For $n < \infty$, let γ_n be the standard Gaussian measure of $\mathbb{R}^n$, let A be a measurable subset of $\mathbb{R}^n$, and let H be a half-space $H = \{x \in \mathbb{R}^n : \langle x, u \rangle \le a\}$, u a unit vector, such that $\gamma_n(H) = \gamma_n(A)$ and hence with $a := \Phi^{-1}(\gamma_n(A))$, where Φ denotes the standard normal distribution function. Then, for all $\varepsilon > 0$,*

$$\gamma_n(H + \varepsilon O_n) \le \gamma_n(A + \varepsilon O_n), \tag{2.12}$$

which, by the definition of a, is equivalent to

$$\gamma_n(A + \varepsilon O_n) \ge \Phi(\Phi^{-1}(\gamma_n(A)) + \varepsilon). \tag{2.13}$$

Proof First, we check the behaviour of distances under $\tilde{\pi}_m$. If d_{n+m} denotes the geodesic distance of $\sqrt{m}S^{n+m}$, it is clear that the projection $\tilde{\pi}_m$ is a contraction from the sphere onto $\mathbb{R}^n$; that is, $|\tilde{\pi}_m(x) - \tilde{\pi}_m(y)| \le d_{n+m}(x,y)$ for any $x, y \in \sqrt{m}S^{n+m}$. Moreover, if in the half-space $H_b := \{x \in \mathbb{R}^n : \langle x, u \rangle \le b\}$, we have $-\sqrt{m} < b < \sqrt{m}$; then its pre-image $\tilde{\pi}^{-1}(H_b)$ is a nonempty cap, and for $0 < \varepsilon < \sqrt{m} - b$, we have $(\tilde{\pi}^{-1}(H_b))_\varepsilon = \tilde{\pi}^{-1}(H_b + \tau(b,\varepsilon)O_n) = \tilde{\pi}^{-1}(H_{b+\tau(b,\varepsilon)})$, where

$$b + \tau = \sqrt{m} \cos\left(\cos^{-1} \frac{b}{\sqrt{m}} \pm \frac{\varepsilon}{\sqrt{m}} \right),$$

which, taking limits in the addition formula for the cosine, immediately gives $\lim_{m\to\infty} \tau(b,\varepsilon) = \varepsilon$.

Let now $b < a = \Phi^{-1}(\gamma_n(A))$ so that $H_b = \{x : \langle x, u \rangle \le b\} \subset H$. Then, by Poincaré's lemma,

$$\lim_m \mu_{n+m}(\tilde{\pi}_m^{-1}(A)) = \gamma_n(A) > \gamma_n(H_b) = \lim_m \mu_{n+m}(\tilde{\pi}_m^{-1}(H_b)),$$

so for all m large enough, we have both $b \in (-\sqrt{m}, \sqrt{m})$, such that $\tilde{\pi}_m^{-1}(H_b)$ is a nonempty cap in the sphere, and $\mu_{n+m}(\tilde{\pi}_m^{-1}(A)) \ge \mu_{n+m}(\tilde{\pi}_m^{-1}(H_b))$. Then the isoperimetric inequality for μ_{n+m} (Theorem 2.2.1) yields that for each $\varepsilon > 0$, $b + \varepsilon < \sqrt{m}$, for all m large enough,

$$\mu_{n+m}\left((\tilde{\pi}_m^{-1}(A))_\varepsilon \right) \ge \mu_{n+m}\left((\tilde{\pi}_m^{-1}(H_b))_\varepsilon \right) = \mu_{n+m}\left(\tilde{\pi}_m^{-1}(H_{b+\tau(b,\varepsilon)}) \right),$$

so by Poincaré's lemma again,

$$\gamma_n(A+\varepsilon O_n) \geq \limsup_m \mu_{n+m}\left((\tilde{\pi}_m^{-1}(A))_\varepsilon\right) \geq \limsup_m \mu_{n+m}\left((\tilde{\pi}_m^{-1}(H_{b+\tau(b,\varepsilon)})\right) = \gamma_n(H_{b+\varepsilon}).$$

Since this holds for all $b < a$, it also holds with b replaced by a. ■

Theorem 2.2.3 extends to infinite dimensions, as will be shown in Theorem 2.6.12. An extension to the standard Gaussian measure on $\mathbb{R}^{\mathbb{N}}$, that is, for the law γ of a sequence of independent standard normal random variables, can be obtained directly. Before stating the theorem, it is convenient to make some topological and measure-theoretic considerations. The distance $\rho(x,y) = \sum_{k=1}^{\infty} \min(|x_k - y_k|, 1)/2^k$ metrises the product topology of $\mathbb{R}^{\mathbb{N}}$, and $(\mathbb{R}^{\mathbb{N}}, \rho)$ is a separable and complete metric space, as is easy to see. That is, $\mathbb{R}^{\mathbb{N}}$ is a Polish space (a topological space that admits a metric for which it is separable and complete). Then the cylindrical σ-algebra $\mathcal{C}$ coincides with the Borel σ-algebra of $\mathbb{R}^{\mathbb{N}}$, and any finite cylindrical (hence Borel) measure is tight (Radon). The product space $\mathbb{R}^{\mathbb{N}} \times \ell_2$ is also Polish, and for each $t \in \mathbb{R}$, the map $f_t : \mathbb{R}^{\mathbb{N}} \times \ell_2 \mapsto \mathbb{R}^{\mathbb{N}}$, $f_t(x,y) = x + ty$ is continuous. Then the image of f_t is universally measurable, that is, measurable for any Radon measure, in particular, in our case, measurable for any finite measure on the cylindrical σ-algebra $\mathcal{C}$ of $\mathbb{R}^{\mathbb{N}}$. See, for example, theorem 13.2.6 in section 13.2 in Dudley (2002).

Theorem 2.2.4 *Let A be a Borel set of $\mathbb{R}^{\mathbb{N}}$ (i.e., $A \in \mathcal{C}$), and let γ be the probability law of $(g_i : i \in \mathbb{N})$, g_i independent standard normal. Let O denote the unit ball about zero of $\ell_2 \subset \mathbb{R}^{\mathbb{N}}$, $O = \{x \in \mathbb{R}^{\mathbb{N}} : \sum_i x_i^2 \leq 1\}$. Then, for all $\varepsilon > 0$,*

$$\gamma(A + \varepsilon O) \geq \Phi(\Phi^{-1}(\gamma(A)) + \varepsilon). \tag{2.14}$$

The proof is indicated in Exercises 2.2.5 through 2.2.7.

2.2.3 Application to Gaussian Concentration

We would like to translate the isoperimetric inequality in Theorem 2.2.4 into a concentration inequality for functions of $\{g_i\}_{i=1}^n$ about their medians, that is, into a bound for $\gamma\{|f(x) - M| > \varepsilon\}$ for all $\varepsilon > 0$. The following definition describes the functions for which such a translation is almost obvious.

Definition 2.2.5 A function $f : \mathbb{R}^{\mathbb{N}} \mapsto \mathbb{R}$ is Lipschitz in the direction of ℓ_2, or ℓ_2-Lipschitz for short, if it is measurable and if

$$\|f\|_{\text{Lip2}} := \sup\left\{\frac{|f(x) - f(y)|}{|x-y|} : x, y \in \mathbb{R}^{\mathbb{N}}, x \neq y, x - y \in \ell_2\right\} < \infty,$$

where $|x-y|$ is the ℓ_2 norm of $x-y$.

For a measurable function f on $\mathbb{R}^{\mathbb{N}}$, we denote by M_f the median of f with respect to the Gaussian measure γ, defined as $M_f = \inf\{t : \gamma\{x : f(x) \leq t\} > 1/2\}$. Then $\gamma(f \leq M_f) \geq 1/2$ and $\gamma(f \geq M_f) \geq 1/2$, and M is the largest number satisfying these two inequalities.

Theorem 2.2.6 *If f is an ℓ_2-Lipschitz function on $\mathbb{R}^{\mathbb{N}}$, and if M_f is its median with respect to γ, then*

$$\gamma\{x : f(x) \geq M_f + \varepsilon\} \leq (1 - \Phi(\varepsilon/\|f\|_{\mathrm{Lip2}})),$$
$$\gamma\{x : f(x) \leq M_f - \varepsilon\} \leq (1 - \Phi(\varepsilon/\|f\|_{\mathrm{Lip2}})), \tag{2.15}$$

in particular

$$\gamma\{x : |f(x) - M_f| \geq \varepsilon\} \leq 2(1 - \Phi(\varepsilon/\|f\|_{\mathrm{Lip2}})) \leq e^{-\varepsilon^2/2\|f\|_{\mathrm{Lip2}}}, \tag{2.16}$$

for all $\varepsilon > 0$.

Proof Let $A^+ = \{x \in \mathbb{R}^{\mathbb{N}} : f(x) \geq M_f\}$ and $A^- = \{x \in \mathbb{R}^{\mathbb{N}} : f(x) \leq M_f\}$. Then $\gamma(A^+) \geq 1/2$, $\gamma(A^-) \geq 1/2$. Moreover, if $x \in A^+ + \varepsilon O$, then there exists $h \in O$ such that $x - \varepsilon h \in A^+$; hence, $f(x - \varepsilon h) \geq M_f$ and $f(x) + \varepsilon\|f\|_{\mathrm{Lip2}} \geq f(x - \varepsilon h) \geq M_f$; that is, $A^+ + \varepsilon O \subset \{x : f(x) \geq M_f - \varepsilon\|f\|_{\mathrm{Lip2}}\}$. Then the Gaussian isoperimetric inequality (2.14) for $A = A^+$ gives (recall $\Phi^{-1}(1/2) = 0$)

$$\gamma\{f < M_f - \varepsilon\|f\|_{\mathrm{Lip2}}\} \leq 1 - \gamma(A^+ + \varepsilon O) \leq 1 - \Phi(\varepsilon),$$

which is the second inequality in (2.15). Likewise, $A^- + \varepsilon O \subset \{x : f(x) \leq M_f + \varepsilon\|f\|_{\mathrm{Lip2}}\}$, and the isoperimetric inequality applied to A^+ gives the first inequality in (2.15). Finally, (2.16) follows by combination of the previous two inequalities and a known bound for the tail probabilities of a normal variable (Exercise 2.2.8). ■

Let now $X(t)$, $t \in T$, be a separable centred Gaussian process such that $\Pr\{\sup_{t\in T}|X(t)| < \infty\} > 0$. Then $\sup_{t\in T}|X(t)| = \sup_{t\in T_0}|X(t)| < \infty$ a.s., where $T_0 = \{t_k\}_{k=1}^{\infty}$ is a countable subset of T (see Example 2.1.15). Ortho-normalizing (in $L^2(\Pr)$), the jointly normal sequence $\{X(t_k)\}$ yields $X(t_k) = \sum_{i=1}^{k} a_{ki}g_i$, where g_i are independent standard normal variables, and $\sum_{i=1}^{k} a_{ki}^2 = EX^2(t_k)$. Then the probability law of the process $X(t_k)$, $k \in \mathbb{N}$, coincides with the law of the random variable defined on the probability space $(\mathbb{R}^{\mathbb{N}}, \mathcal{C}, \gamma)$, $\tilde{X} : \mathbb{R}^{\mathbb{N}} \mapsto \mathbb{R}$, $\tilde{X}(t_k, x) = \sum_{i=1}^{k} a_{ki}x_i$. This is so because the coordinates of $\mathbb{R}^{\mathbb{N}}$, considered as random variables on the probability space $(\mathbb{R}^{\mathbb{N}}, \mathcal{C}, \gamma)$, are i.i.d. $N(0,1)$. Now define a function $f : \mathbb{R}^{\mathbb{N}} \mapsto \mathbb{R}$ by

$$f(x) = \sup_k \left|\sum_{i=1}^{k} a_{ki}x_i\right|.$$

The probability law of f under γ is the same as the law of $\sup_{t\in T_0}|X(t)|$, which, in turn, is the same as the law of $\sup_{t\in T}|X(t)|$. Moreover, if $h \in O$, the unit ball of ℓ_2, by Cauchy-Schwarz,

$$|f(x+h) - f(x)|^2 = \sup_k \left|\sum_{i=1}^{k} a_{ki}h_i\right|^2 \leq \sup_k \left[\sum_{i=1}^{k} a_{ki}^2 \sum_{i=1}^{k} h_i^2\right] \leq \sup_k \sum_{i=1}^{k} a_{ki}^2 = \sup_k EX^2(t_k).$$

Therefore,

$$\|f\|_{\mathrm{Lip2}} \leq \sigma^2(X), \quad \text{where } \sigma^2 = \sigma^2(X) := \sup_{t\in T} EX^2(t).$$

Recall from an argument at the beginning of the proof of Theorem 2.1.20 that for the processes X we are considering here, $\sigma^2 < \infty$ and the median $M < \infty$. Then Theorem 2.2.6 applies to the function f and gives the following concentration inequality:

Theorem 2.2.7 (The Borell-Sudakov-Tsirelson concentration inequality for Gaussian processes) *Let $X(t)$, $t \in T$, be a centred separable Gaussian process such that $\Pr\{\sup_{t\in T} |X(t)| < \infty\} > 0$, and let M be the median of $\sup_{t\in T} |X(t)|$ and σ^2 the supremum of the variances $EX^2(t)$. Then, for all $u > 0$,*

$$\Pr\left\{\sup_{t\in T} |X(t)| > M + u\right\} \le 1 - \Phi(u/\sigma), \quad \Pr\left\{\sup_{t\in T} |X(t)| < M - u\right\} \le 1 - \Phi(u/\sigma), \tag{2.17}$$

and hence,

$$\Pr\left\{\left|\sup_{t\in T} |X(t)| - M\right| > u\right\} \le 2(1 - \Phi(u/\sigma)) \le e^{-u^2/2\sigma^2}. \tag{2.18}$$

Inequality (2.18) is also true with the median M of $\sup_{t\in T} |X(t)|$ replaced by the expectation $E\left(\sup_{t\in T} |X(t)|\right)$, as we will see in Section 2.5 as a consequence of the Gaussian logarithmic Sobolev inequality (other proofs are possible; see Section 2.1 for a simple proof of a weaker version). But such a result, in its sharpest form, does not seem to be obtainable from (2.18). However, notice that if we integrate in (2.18) and let g be a $N(0,1)$ random variable, we obtain

$$\left|E \sup_{t\in T} |X(t)| - M\right| \le E\left|\sup_{t\in T} |X(t)| - M\right| \le \sigma E|g| = \sqrt{2/\pi}\ \sigma, \tag{2.19}$$

an inequality which is interesting in its own right and which gives, by combining with the same (2.18),

$$\Pr\left\{\left|\sup_{t\in T} |X(t)| - E \sup_{t\in T} |X(t)|\right| > u + \sqrt{2/\pi}\sigma\right\} \le e^{-u^2/2\sigma^2}, \tag{2.20}$$

which is of the right order for large values of u.

Theorem 2.2.7, or even (2.20), expresses the remarkable fact that the supremum of a Gaussian process $X(t)$, centred at its mean or at its median, has tail probabilities not worse than those of a normal variable with the largest of the variances $EX^2(t)$, $t \in T$. In particular, if we knew the size of $E\sup_{t\in T} |X(t)|$, we would have a very exact knowledge of the distribution of $\sup_{t\in T} |X(t)|$. This will be the object of the next two sections.

We complete this section with simple applications of Theorem 2.2.7 to integrability and moments of the supremum of a Gaussian processes.

Corollary 2.2.8 *Let $X(t)$, $t \in T$, be a Gaussian process as in Theorem 2.2.7. Let M and σ also be as in this theorem, and write $\|X\| := \sup_{t\in T} |X(t)|$ to ease notation. Then there exists $K < \infty$ such that with the same hypothesis and notation as in the preceding corollary, for all $p \ge 1$,*

$$(E\|X\|^p)^{1/p} \le 2E\|X\| + (E|g|^p)^{1/p}\sigma \le K\sqrt{p}E\|X\|$$

for some absolute constant K.

Proof Just integrate inequality (2.18) with respect to $pt^{p-1}dt$ and then use that $M \le 2E\|X\|$ (by Chebyshev) and that $\sigma \le \sqrt{\pi/2}\sup_{t\in T} E|X(t)|$. See Exercise 2.1.2. ■

Corollary 2.2.9 *Let $X(t)$, $t \in T$, be a Gaussian process as in Theorem 2.2.7, and let $\|X\|$, M and σ be as in Corollary 2.2.8. Then*

$$\lim_{u\to\infty} \frac{1}{u^2} \log \Pr\{\|X\| > u\} = -\frac{1}{2\sigma^2}$$

and

$$Ee^{\lambda\|X\|^2} < \infty \text{ if and only if } \lambda < \frac{1}{2\sigma^2}.$$

Proof The first limit follows from the facts that the first inequality in (2.17) can be rewritten as

$$\frac{1}{(u-M)^2} \log \Pr\{\|X\| > u\} \le -\frac{1}{\sigma^2}$$

and that $\Pr\{\|X\| > u\} \ge \Pr\{|X(t)| > u\}$ for all $t \in T$ (as, for a $N(0,1)$ variable g, we do have $u^{-2} \log \Pr\{|g| > u/a\} \to -1/2a^2$, e.g., by l'Hôpital's rule). For the second statement, just apply the first limit to $Ee^{\lambda\|X\|} = 1 + \int_0^\infty \int_0^{\lambda\|X\|^2} e^v dv \, d\mathcal{L}(\|X\|)(u) = 1 + \int_0^\infty e^v \Pr\{\|X\| > \sqrt{v/\lambda}\} dv$. ■

Exercises

2.2.1 Prove that if A is closed, so is $s_H(A)$ for any subspace H of dimension n. *Hint*: Conveniently enlarge some of the components in the definition of $s_H(A)$ to make them compact and still keep the same union.

2.2.2 Prove that $(\mathcal{K}, h)$, the space of nonempty compact subsets of S^n with the Hausdorff distance, is a compact metric space. *Hint*: Show that the map $\mathcal{K} \mapsto C(S^n)$, $A \mapsto d(\cdot, A)$, is an isometry between $(\mathcal{K}, h)$ and its image in $(C(S^n), \|\cdot\|_\infty)$ and that this image is compact in $C(S^n)$ (note that $x \mapsto d(x, A)$ is bounded and Lipschitz or see Beers (1993)).

2.2.3 Show that the Lebesgue density theorem holds in S^n for the uniform measure; that is, show that if $\mu(E) > 0$, then μ-almost all points of E satisfy $\lim_{\rho\to 0} [\mu(E \cap C(x,\rho))]/[\mu(C(x,\rho))] = 1$. *Hint*: First adapt the usual proof of the Vitali covering theorem to the sphere, using that $L_n < \infty$ such that any cap of radius 2ρ can be covered by L_n caps of radius ρ. Then use the Vitali covering theorem to show that if for each $0 < \alpha < 1$, A_α is the set of those points in E for which $\liminf_{\rho\to 0} [\mu(E \cap C(x,\rho))]/[\mu(C(x,\rho))] < \alpha < 1$, then $\mu(A_\alpha) = 0$ as follows: if G is an open set containing A_α with $\mu(G) < \mu(A_\alpha)/\alpha$, let $\mathcal{V}$ be the set of caps $C(x,\rho)$ that satisfy $[\mu(E \cap C(x,\rho))]/[\mu(C(x,\rho))] < \alpha$ and are contained in G; get a Vitali subcover and show that its total measure, which is at most $\mu(G)$, is larger than or equal than $\mu(A_\alpha)/\alpha$, a contradiction. Or refer to Mattila (1995).

2.2.4 Prove that for $n \ge 2$, if $\mu(A) \ge 1/2$, then $\mu(A_\varepsilon) \ge 1 - (\pi/8)^{1/2} e^{-(n-1)\varepsilon^2/2}$, where μ is the uniform probability measure on S^n.

2.2.5 Let $\pi_n : \mathbb{R}^{\mathbb{N}} \mapsto \mathbb{R}^n$ be the projection $\pi_n(x) = \pi_n(x_k : k \in \mathbb{N}) = (x_1, \ldots, x_n)$. Then show that (a) $\gamma_n = \gamma \circ \pi_n^{-1}$, (b) if $K \subset \mathbb{R}^{\mathbb{N}}$ is compact, then $K = \cap_{n=1}^\infty \pi_n^{-1}(\pi_n(K))$, and (c) $K + tO$, where O is the closed unit ball of ℓ_2, is compact if K is.

2.2.6 Use Theorem 2.2.3 and Exercise 2.2.5 to prove Theorem 2.2.4 in the particular case where A is a compact set.

2.2.7 Since $\mathbb{R}^{\mathbb{N}}$ is Polish, it follows that γ is tight (Proposition 2.1.4). Use this and Exercise 2.2.2 to prove Theorem 2.2.4 for any $A \in \mathcal{C}$.

In the remaining exercises, the process X is as in Theorem 2.2.7.

2.2.8 Let Φ be the $N(0,1)$ distribution function. Then, for all $u \geq 0$, show that $2(1-\Phi(u)) \leq e^{-u^2/2}$. *Hint*: Use the well-known bound $\int_u^\infty e^{-t^2/2}dt \leq u^{-1}e^{-u^2/2}$ for $u \geq \sqrt{2/\pi}$ and differentiation for $0 \leq u \leq \sqrt{2/\pi}$.

2.2.9 Prove the analogue of Theorem 2.2.7 for $\sup_{t\in T} X(t)$ and its median.

2.2.10 Show that $\Pr\{\sup_{t\in T}|X(t)| > u\} \leq 2\Pr\{\sup_{t\in T} X(t) > u\}$.

2.2.11 Show that: (a) The random variable $\sup_{t\in T}|X(t)|$ has a unique median, meaning that M is the only number for which both $\Pr\{\sup_{t\in T}|X(t)| \geq M\}$ and $\Pr\{\sup_{t\in T}|X(t)| \leq M\}$ are larger than or equal to $1/2$. In particular, the distribution function of $\sup_{t\in T}|X(t)|$ is continuous at M. *Hint*: The second equation in (2.15) implies that no number below the largest median of f for the measure γ can be a median; now apply this to the appropriate f. (b) Use the same reasoning to conclude that if $M_a = \inf\{u : \Pr\{\sup_{t\in T}|X(t)| \leq u\} > a\}$, $0 < a < 1$, then the distribution function of $\sup_{t\in T}|X(t)|$ is continuous at M_a.

2.2.12 Let B be a Banach space whose norm $\|\cdot\|$ satisfies the following: there exists a countable subset D of the unit ball of its (topological) dual space B^* such that $\|x\| = \sup_{f\in D}|f(x)|$ for all $x \in B$. For instance, this is true for separable Banach spaces as well as for ℓ_∞. Define a Gaussian random variable X with values in B as a map from some probability space $(\Omega, \Sigma, \Pr)$ into B such that $f(X)$ is a centred normal random variable for every $f \in B^*$. Prove that if $\|X\|$ is finite almost surely, if M is a median of $\|X\|$ and $\sigma^2 = \sup_{f\in D} Ef^2(X)$, then

$$\Pr\{|\|X\| - M| > u\} \leq 2(1-\Phi(u/\sigma)) \leq e^{-u^2/2\sigma^2}. \tag{2.21}$$

2.2.13 Let B be a Banach space as in Exercise 2.2.12, and let X be a centred Gaussian B-valued random variable. Use Exercise 2.2.12 to show that the distribution function $F_{\|X\|}$ of $\|X\|$ is continuous at M_a for all $0 < a < 1$, where M_a is as defined in Exercise 2.2.11 with the supremum of the process replaced by $\|X\|$.

2.3 The Metric Entropy Bound for Suprema of Sub-Gaussian Processes

In this section we define sub-Gaussian processes and obtain the celebrated Dudley's entropy bound for their supremum norm. We are careful about the constants, as they are of some consequence in statistical estimations, at the expense of making the 'chaining argument' (proof of Theorem 2.3.6) slightly more complicated than it could be. Combined with concentration inequalities, these bounds yield good estimates of the distribution of the supremum of a Gaussian process. They also constitute sufficient conditions for sample boundedness and sample continuity of Gaussian and sub-Gaussian processes and provide moduli of continuity for their sample paths which are effectively sharp in light of Sudakov's inequality derived in the next section.

A square integrable random variable ξ is said to be *sub-Gaussian* with parameter $\sigma > 0$ if for all $\lambda \in \mathbb{R}$,

$$Ee^{\lambda\xi} \leq e^{\lambda^2\sigma^2/2}.$$

Developing the two exponentials, dividing by $\lambda > 0$ and by $\lambda < 0$ and letting $\lambda \to 0$ in each case yield $E\xi = 0$; that is, sub-Gaussian random variables are automatically centred. Then, if in the two developments once the expectation term is cancelled, we divide by λ^2 and let $\lambda \to 0$, we obtain $E\xi^2 \leq \sigma^2$.

Aside from normal variables, perhaps the main examples of sub-Gaussian variables are the linear combinations of independent *Rademacher* (or symmetric *Bernoulli*) random

variables $\xi = \sum_{i=1}^n a_i\varepsilon_i$, where ε_i are independent identically distributed and $\Pr\{\varepsilon_i = 1\} = \Pr\{\varepsilon_i = -1\} = 1/2$. To see that these variables are sub-Gaussian, just note that by Taylor expansion, if ε is a Rademacher variable,

$$Ee^{\lambda\varepsilon} = (e^{\lambda} + e^{-\lambda})/2 \le e^{\lambda^2/2}, \quad \lambda \in \mathbb{R},$$

so that, by independence,

$$Ee^{\lambda\sum a_i\varepsilon_i} \le e^{\lambda^2\sum a_i^2/2}.$$

Both for Gaussian and for linear combinations of independent Rademacher variables, $\sigma^2 = E\xi^2$.

The distributions of sub-Gaussian variables have *sub-Gaussian tails*: Chebyshev's inequality in exponential form, namely,

$$\Pr\{\xi \ge t\} = \Pr\left\{e^{\lambda\xi} \ge e^{\lambda t}\right\} \le e^{\lambda^2\sigma^2/2 - \lambda t}, \quad t > 0,\ \lambda > 0,$$

with $\lambda = t/\sigma^2$ and applied as well to $-\xi$, gives that if ξ is sub-Gaussian for σ^2, then

$$\Pr\{\xi \ge t\} \le e^{-t^2/2\sigma^2} \quad \text{and} \quad \Pr\{\xi \le -t\} \le e^{-t^2/2\sigma^2}, \quad \text{hence,}$$
$$\Pr\{|\xi| \ge t\} \le 2e^{-t^2/2\sigma^2}, \quad t > 0. \tag{2.22}$$

The last inequality in (2.22) in the case of linear combinations of independent Rademacher variables is called *Hoeffding's inequality*. Of course, we can be more precise about the tail probabilities of normal variables: simple calculus gives that for all $t > 0$,

$$\frac{t}{t^2+1}e^{-t^2/2} \le \int_t^\infty e^{-u^2/2}du \le \min\left(t^{-1}, \sqrt{\pi/2}\right)e^{-t^2/2}, \tag{2.23}$$

(see Exercise 2.2.8).

Back to the inequalities (2.22), we notice that if they hold for ξ, then ξ/c enjoys square exponential integrability for some $0 < c < \infty$: if $c^2 > 2\sigma^2$, then

$$Ee^{\xi^2/c^2} - 1 = \int_0^\infty 2te^{t^2}\Pr\{|\xi| > ct\}\,dt \le \frac{2}{c^2/2\sigma^2 - 1} < \infty. \tag{2.24}$$

The collection of random variables ξ on $(\Omega, \Sigma, \Pr)$ that satisfy this integrability property constitutes a vector space, denoted by $L^{\psi_2}(\Omega, \Sigma, \Pr)$, and the functional

$$\|\xi\|_{\psi_2} = \inf\{c > 0 : E\psi_2(|\xi|/c) \le 1\},$$

where $\psi_2(x) := e^{x^2} - 1$ (a convex function which is zero at zero) is a pseudo-norm on it for which L^{ψ_2}, with identification of a.s. equal functions, is a Banach space (Exercise 2.3.5). With this definition, inequality (2.24) shows that

$$\Pr\{|\xi| \ge t\} \le 2e^{-t^2/2\sigma^2} \quad \text{for all } t > 0 \quad \text{implies} \quad \|\xi\|_{\psi_2} \le \sqrt{6}\sigma. \tag{2.25}$$

To complete the set of relationships developed so far, suppose that $\xi \in L^{\psi_2}$ and $E\xi = 0$, and let us show that ξ is sub-Gaussian. We have

$$Ee^{\lambda\xi} - 1 \le E\sum_{k=2}^\infty |\lambda^k\xi^k|/k \le \frac{\lambda^2}{2}E\left(\xi^2 e^{|\lambda\xi|}\right).$$

Now we estimate the exponent $|\lambda\xi|$ on the region $|\xi| > 2\lambda\|\xi\|_{\psi_2}^2$ and on its complement to obtain, after multiplying and dividing by $\|\xi\|_{\psi_2}^2$ and using that $a < e^{a/2}$ for all $a > 0$,

$$\frac{\lambda^2}{2}E\left(\xi^2 e^{|\lambda\xi|}\right) \le \frac{\lambda^2\|\xi\|_{\psi_2}^2}{2}e^{2\lambda^2\|\xi\|_{\psi_2}^2}E\left(\frac{\xi^2}{\|\xi\|_{\psi_2}^2}e^{\xi^2/2\|\xi\|_{\psi_2}^2}\right)$$
$$\le \lambda^2\|\xi\|_{\psi_2}^2 e^{2\lambda^2\|\xi\|_{\psi_2}^2}Ee^{\xi^2/\|\xi\|_{\psi_2}^2}/2 \le \lambda^2\|\xi\|_{\psi_2}^2 e^{2\lambda^2\|\xi\|_{\psi_2}^2}.$$

Using $1 + a \le e^a$, the last two bounds give

$$Ee^{\lambda\xi} \le e^{3\lambda^2\|\xi\|_{\psi_2}^2}, \tag{2.26}$$

showing that ξ is sub-Gaussian with $\sigma \le \sqrt{6}\|\xi\|_{\psi_2}$. If ξ is symmetric, just developing the exponential gives the better inequality $Ee^{\lambda\xi} \le e^{\lambda^2\|\xi\|_{\psi_2}^2/2}$.

We collect these facts:

Lemma 2.3.1 *If ξ is sub-Gaussian for a constant $\sigma > 0$, then it satisfies the sub-Gaussian tail inequalities (2.22), and therefore, $\xi \in L^{\psi_2}$, with $\|\xi\|_{\psi_2} \le \sqrt{6}\sigma$. Conversely, if ξ is in L^{ψ_2} and is centred, then it is sub-Gaussian for the constant $\sigma \le \sqrt{6}\|\xi\|_{\psi_2}$, and in particular, it also satisfies the inequalities (2.22) for $\sigma = \sqrt{6}\|\xi\|_{\psi_2}$.*

In other words, ignoring constants, for ξ centred, the conditions (a) $\xi \in L^{\psi_2}$ and (b) ξ satisfies the sub-Gaussian tail inequalities (2.22) for some σ_1 and (c) ξ is sub-Gaussian for some σ_2 are all equivalent.

Lemma 2.3.1 extends to random variables whose tail probabilities are bounded by a constant times the sub-Gaussian probabilities in (2.22) as follows.

Lemma 2.3.2 *Assume that*

$$\Pr\{|\xi| \ge t\} \le 2Ce^{-t^2/2\sigma^2}, \quad t > 0, \tag{2.27}$$

for some $C \ge 1$ and $\sigma > 0$, a condition implied by the Laplace transform condition

$$Ee^{\lambda\xi} \le Ce^{\lambda^2\sigma^2/2}, \quad \lambda \in \mathbb{R}. \tag{2.28}$$

Then ξ also satisfies

$$\|\xi\|_{\psi_2} \le \sqrt{2(2C+1)}\sigma. \tag{2.29}$$

Moreover, if in addition $E\xi = 0$, then also

$$Ee^{\lambda\xi} \le e^{3\lambda^2(2(2C+1))\sigma^2}, \quad \lambda \in \mathbb{R}, \tag{2.30}$$

that is, ξ is sub-Gaussian with constant $\tilde{\sigma}^2 = 12(2C+1)\sigma^2$.

Proof The proof of inequality (2.22) shows that (2.28) implies (2.27). The preceding proof showing that (2.22) implies (2.25), with only formal changes, proves that (2.27) implies (2.29). Finally, inequality (2.30) follows from (2.29) and (2.26). ■

This lemma is useful in that showing that a variable ξ is sub-Gaussian reduces to proving the tail probability bounds (2.27) for some $C > 1$, which may be easier than proving them for $C = 1$.

Lemma 2.3.1 (or, more precisely, the inequalities that make it possible) has many important consequences on the size of maxima of sub-Gaussian stochastic processes. The simplest examples of such processes are finite collections of sub-Gaussian variables. The following lemma contains a maximal inequality for variables in $\xi_i \in L^{\psi_2}$ not necessarily centred, and it applies by Lemma 2.3.1 to finite collections of sub-Gaussian variables.

Lemma 2.3.3 *Let $\xi_i \in L^{\psi_2}$, $i = 1, \ldots, N$, $2 \le N < \infty$. Then*

$$\left\| \max_{i \le N} |\xi_i| \right\|_{\psi_2} \le 4\sqrt{\log N} \max_{i \le N} \|\xi_i\|_{\psi_2}, \tag{2.31}$$

and, in particular, there exist $K_p < \infty$, $1 \le p < \infty$, such that

$$\left\| \max_{i \le N} |\xi_i| \right\|_{L^p} \le K_p \sqrt{\log N} \max_{i \le N} \|\xi_i\|_{\psi_2}. \tag{2.32}$$

Proof To prove (a), we may assume that $\max \|\xi_i\|_{\psi_2} = 1$. Then the definition of the ψ_2 norm together with the exponential Chebyshev's inequality gives

$$\begin{aligned} E \max_{i \le N} e^{\xi_i^2/(16 \log N)} &= \int_0^\infty \Pr\left\{ \max_{i \le N} e^{\xi_i^2/(16 \log N)} \ge t \right\} dt \\ &\le e^{1/8} + \sum_{i=1}^N \int_{e^{1/8}}^\infty \Pr\left\{ e^{\xi_i^2/(16 \log N)} \ge t \right\} dt \\ &\le e^{1/8} + 2N \int_{e^{1/8}}^\infty e^{-8(\log N)(\log t)} dt = e^{1/8} + 2N \int_{e^{1/8}}^\infty t^{-8 \log N} dt \\ &= e^{1/8} \left(1 + \frac{2}{8(\log N) - 1} \right) < 2, \end{aligned}$$

proving (2.31). For part (b), use that $\|\zeta\|_{L^{2k}} \le (k)^{1/2k} \|\zeta\|_{L^{\psi_2}}$ for any random variable $\zeta \in L^{\psi_2}$ (as observed earlier) and part (a) to obtain inequality (2.32). ■

It is convenient to have sensible values of K_p at hand, particularly for $p = 1$. The method to obtain the following bound is quite simple and general: let Φ be a nonnegative, strictly increasing, convex function on a finite or infinite interval I, and let ξ_i, $1 \le i \le N$, be random variables taking values in I and such that $E\Phi(\xi_i) < \infty$. We then have, by Jensen's inequality and the properties of Φ,

$$\begin{aligned} \Phi\left(E \max_{i \le N} \xi_i \right) &\le E\Phi\left(\max_{i \le N} \xi_i \right) = E \max_{i \le N} \Phi(\xi_i) \\ &\le \sum_{i=1}^N E\Phi(\xi_i) \le N \max_{i \le N} E\Phi(\xi_i), \end{aligned} \tag{2.33}$$

and, inverting Φ,

$$E \max_{i \le N} \xi_i \le \Phi^{-1}\left(N \max_{i \le N} E\Phi(\xi_i) \right). \tag{2.34}$$

Lemma 2.3.4 *For any $N \geq 1$, if ξ_i, $i \leq N$, are sub-Gaussian random variables admitting constants σ_i, then*

$$E \max_{i \leq N} \xi_i \leq \sqrt{2 \log N} \max_{i \leq N} \sigma_i, \quad E \max_{i \leq N} |\xi_i| \leq \sqrt{2 \log 2N} \max_{i \leq N} \sigma_i. \tag{2.35}$$

Proof We take $\Phi(x) = e^{\lambda x}$ in (2.34). Since ξ_i is sub-Gaussian, we have $E\Phi(\xi_i) \leq e^{\lambda^2 \sigma_i^2/2}$, and (2.34) gives

$$E \max_{i \leq N} \xi_i \leq \frac{\log N}{\lambda} + \frac{1}{2} \lambda \max_{i \leq N} \sigma_i^2.$$

The first inequality in the lemma follows by minimizing in λ in this inequality (*i.e.*, by taking $\lambda = (2 \log N)^{1/2} / \max_{i \leq N} \sigma_i$). The second inequality follows by applying the first to the collection of $2N$ random variables $\eta_i = \xi_i$, $\eta_{n+i} = -\xi_i$, $1 \leq i \leq N$. ∎

We now consider more general sub-Gaussian processes.

Definition 2.3.5 A centred stochastic process $X(t)$, $t \in T$, is sub-Gaussian with respect to a distance or pseudo-distance d on T if its increments satisfy the sub-Gaussian inequality, that is, if

$$E e^{\lambda (X(t) - X(s))} \leq e^{\lambda^2 d^2(s,t)/2} \ \lambda \in \mathbb{R}, \quad s, t \in T. \tag{2.36}$$

If instead of condition (2.36) the centred process X satisfies

$$E e^{\lambda (X(t) - X(s))} \leq C e^{\lambda^2 d^2(s,t)/2} \quad \text{or} \quad \Pr\{|X(t) - X(s)| \geq u\} \leq 2C e^{-u^2/2d^2(s,t)},$$

for all $\lambda \in \mathbb{R}$, $u > 0$ and $s, t \in T$ and some $C > 1$, then, by Lemma 2.3.2, X is sub-Gaussian for the distance $\tilde{d}(s,t) := \sqrt{12(2C+1)} d$. Then all the results that follow for sub-Gaussian processes apply as well to processes X satisfying this condition, and the effects on the results themselves of the dilation of the distance d can be easily quantified.

Gaussian processes, that is, processes $X(t)$ such that for every finite set of indices $t_1, \ldots, t_k$, $k < \infty$, the vectors $(X(t_i) : 1 \leq i \leq k)$ are multivariate normal and are sub-Gaussian with respect to the L^2-distance $d_X(s,t) = \|X(t) - X(s)\|_{L^2}$. *Randomized empirical processes* constitute another important class of examples. Let $(S, \mathcal{S}, P)$ be a probability space, and let $X_i : S^{\mathbb{N}} \mapsto S$, $i \in \mathbb{N}$, be the coordinate functions (which are i.i.d. with law P). Given a collection $\mathcal{F}$ of measurable functions on $(S, \mathcal{S})$, the empirical measures indexed by $\mathcal{F}$ and based on $\{X_i\}$ are defined as

$$\left\{ P_n(f) := \frac{1}{n} \sum_{i=1}^{n} f(X_i) : f \in \mathcal{F} \right\}, \quad n \in \mathbb{N},$$

and a related process that has turned out to be an excellent tool in the study of empirical measures is the randomized empirical process, defined for each $n \in \mathbb{N}$ as

$$\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(X_i) : f \in \mathcal{F} \right\},$$

where $\{\varepsilon_i\}$ is a sequence of independent Rademacher variables, independent of the variables X_i. Since linear combinations of independent Rademacher variables are sub-Gaussian with respect to their variance, it follows that randomized empirical processes are *sub-Gaussian with respect to $d(f,g) = \|f - g\|_{L^2(P_n)}$ conditionally on the variables X_i.*

Here are two useful observations about sub-Gaussian processes: if X is a Gaussian process with respect to d, then the definition immediately implies that

$$E(X(t)-X(s))^2 \le d^2(s,t)$$

(as observed earlier, just after the definition of sub-Gaussian variables). Moreover, since for any s,t, $(X(t)-X(s))/d(s,t)$ is a sub-Gaussian variable with variance not exceeding 1, Lemma 2.3.4 implies that if F is a finite subset of $T \times T$ of cardinality N, then

$$E \max_{(s,t)\in F} |X(t)-X(s)| \le \sqrt{2\log 2N} \max_{(s,t)\in F} d(s,t). \tag{2.37}$$

Inequalities analogous to those in Lemma 2.3.3 for these maxima hold as well.

Given a sub-Gaussian process $X(t)$, $t \in T$, it is of great interest to determine the (stochastic) size of $\sup_{t\in T} |X(t)|$ or of $\sup_{s,t\in T, d_X(s,t)\le\delta} |X(t)-X(s)|$ or whether X has a version with bounded sample paths or with uniformly d_X-continuous sample paths (or perhaps continuous in another metric). For Gaussian processes, these questions should and have been answered exclusively in terms of the properties of the metric space (T,d), and for sub-Gaussian processes, properties of this metric space do provide good control of these quantities and good sufficient conditions for sample boundedness and continuity. Although there are much more refined analyses (see the notes at the end of the section), we will develop only the very neat and useful chaining method based on Dudley's metric entropy. The reason for not presenting this subject in more generality is that it is not needed in this book.

The following theorem indicates a way to control $\sup_{t\in T} |X(t)|$ based on a combination of the bound in Lemma 2.3.4 with the size of the (pseudo-) metric space (T,d), measured in terms of the size of the most economical coverings. Given a metric or pseudo-metric space (T,d), for any $\varepsilon > 0$, its *covering number* $N(T,d,\varepsilon)$ is defined as the smallest number of closed d-balls of radius ε needed to cover T, formally, if $B(t,\varepsilon) := \{s \in T : d(s,t) \le \varepsilon\}$,

$$N(T,d,\varepsilon) := \min\left\{n : \text{ there exist } t_1,\dots,t_n \in T \text{ such that } T \subseteq \cup_{i=1}^n B(t_i,\varepsilon)\right\},$$

where we take the minimum of the empty set to be infinite. The *packing numbers*

$$D(T,d,\varepsilon) := \max\left\{n : \text{ there exist } t_1,\dots,t_n \in T \text{ such that } \min_{1\le i,j\le n} d(t_i,t_j) > \varepsilon\right\}$$

are sometimes useful and are equivalent to the covering numbers: it is easy to see (and we will use it without explicit mention) that, for all $\varepsilon > 0$,

$$N(T,d,\varepsilon) \le D(T,d,\varepsilon) \le N(T,d,\varepsilon/2).$$

The logarithm of the covering number of (T,d) is known as its *metric entropy*.

Theorem 2.3.6 *Let (T,d) be a pseudo-metric space, and let $X(t)$, $t \in T$, be a stochastic process sub-Gaussian with respect to the pseudo-distance d, that is, one whose increments satisfy condition (2.36). Then, for all finite subsets $S \subseteq T$ and points $t_0 \in T$, the following inequalities hold:*

$$E \max_{t\in S} |X(t)| \le E|X(t_0)| + 4\sqrt{2} \int_0^{D/2} \sqrt{\log 2N(T,d,\varepsilon)}\, d\varepsilon, \tag{2.38}$$

where D is the diameter of (T,d), and

$$E\max_{\substack{s,t\in S\\ d(s,t)\le\delta}}|X(t)-X(s)| \le (16\sqrt{2}+2)\int_0^{\delta}\sqrt{\log 2N(T,d,\varepsilon)}\,d\varepsilon, \tag{2.39}$$

for all $\delta>0$, where the integrals are taken to be 0 if $D=0$.

Proof If the d-diameter D of T is zero, or if $\int_0^{D/2}\sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon=\infty$, there is nothing to prove. Thus, we assume that $D>0$ and that the entropy integral is finite, in which case (T,d) is totally bounded and, in particular, $D<\infty$. By taking $(X(t)-X(t_0))/((1+\delta)D)$ instead of $X(t)$ and $d/((1+\delta)D)$ for any small δ instead of d, we may assume that $X(t_0)=0$ and $1/2<D<1$. Given $S\subset T$ finite, since $d(s,t)=0$ implies $X(t)=X(s)$ a.s., we can identify points of S at d-distance zero from each other; that is, we can assume that d is a proper distance on S. We also can assume that S has cardinality at least 2. Since S is finite, there is $k_1\in\mathbb{N}$ such that for each $t\in T$, the ball $B(t,2^{-k_1})$ contains at most one point from S. Set $T_{k_1}=S$, which has cardinality at most $N(T,d,2^{-k_1})$, set $T_0=\{t_0\}$ and for $1\le k<k_1$, let T_k be a set of centres of $N(T,d,2^{-k})$ d-balls of radius 2^{-k} covering T. For each $s\in S$, we construct a chain $(\pi_{k_1}(s),\pi_{k_1-1}(s),\dots,\pi_0(s))$ with links $\pi_k(s)\in T_k$, $0\le k\le k_1$, as follows: $\pi_{k_1}(s)=s$ and, given $\pi_k(s)$, $k_1\ge k>0$, $\pi_{k-1}(s)$ is taken to be a point in T_{k-1}, for which the ball $B(\pi_{k-1}(s),2^{-(k-1)})$ contains $\pi_k(s)$, this being done in such a way that $\pi_{k-1}(s)$ depends only on $\pi_k(s)$ in the sense that if $\pi_k(s)=\pi_k(t)$, then $\pi_{k-1}(s)=\pi_{k-1}(t)$. Note that $\pi_0(s)=t_0$ for all s. In particular, for each $0\le k\le k_1$, the number of 'subchains' $(\pi_k(s),\pi_{k-1}(s),\dots,\pi_0(s))$, $s\in S$, is exactly $\mathrm{Card}\{\pi_k(s):s\in S\}\le N(T,d,2^{-k})$. In particular, for $k=1,\dots,k_1$,

$$\mathrm{Card}\{(X(\pi_k(s))-X(\pi_{k-1}(s))):s\in S\}=\mathrm{Card}\{\pi_k(s):s\in S\}\le N(T,d,2^{-k}).$$

Moreover, since $\pi_k(s)\in B(\pi_{k-1}(s),2^{-(k-1)})$,

$$\left[E(X(\pi_k(s))-X(\pi_{k-1}(s)))^2\right]^{1/2}\le d(\pi_k(s),\pi_{k-1}(s))\le 2^{-k+1},\quad k=1,\dots,k_1.$$

Hence, by inequality (2.37),

$$E\max_{s\in S}|X(\pi_k(s))-X(\pi_{k-1}(s))|\le 2^{-k+1}\sqrt{2\log 2N(T,d,2^{-k})},\quad k=1,\dots,k_1.$$

(Note that $N(T,d,2^{-k})\ge 2$ for $k\ge 1$ because $D>1/2$, so this inequality holds even if $\mathrm{Card}(\pi_k(s):s\in S\}=1$.) Therefore, noting that $X(\pi_0(s))=X(t_0)=0$ and $X(\pi_{k_1}(s))=X(s)$, we have

$$\begin{aligned}E\max_{s\in S}|X(s)| &\le \sum_{k=1}^{k_1}E\max_{s\in S}|X(\pi_k(s))-X(\pi_{k-1}(s))|\\ &\le \sum_{k=1}^{\infty}2^{-k+1}\sqrt{2\log 2N(T,d,2^{-k})}\\ &\le 4\int_0^{1/2}\sqrt{2\log 2N(T,d,\varepsilon)}\,d\varepsilon.\end{aligned}$$

Replacing $X(t)$ by $(X(t)-X(t_0))/D$ and d by $d/(1+\delta)D$ and letting $\delta\to 0$, we obtain inequality (2.38).

Given $\delta < \operatorname{diam}(T)$, let $k(\delta) = \min\{k \in \mathbb{N} : 2^{-k} \le \delta\}$. Define

$$U = \left\{(x,y) \in T_{k(\delta)} \times T_{k(\delta)} : \exists\, u,v \in S, d(u,v) \le \delta, \pi_{k(\delta)}(u) = x, \pi_{k(\delta)}(v) = y\right\},$$

and given $(x,y) \in U$, fix $u_{x,y},\ v_{x,y} \in S$, such that $\pi_{k(\delta)}(u_{x,y}) = x$, $\pi_{k(\delta)}(v_{x,y}) = y$, $d(u_{x,y}, v_{x,y}) \le \delta$. For $s,t \in S$ such that $d(s,t) \le \delta$, obviously, $(x,y) := (\pi_{k(\delta)}(s), \pi_{k(\delta)}(t)) \in U$, and we can write

$$\begin{aligned}|X(t) - X(s)| &\le |X(t) - X(\pi_{k(\delta)}(t))| + |X(\pi_{k(\delta)}(t)) - X(v_{x,y})| + |X(v_{x,y}) - X(u_{x,y})| + |X(u_{x,y})\\ &\quad - X(\pi_{k(\delta)}(s))| + |X(\pi_{k(\delta)}(s)) - X(s)|\\ &\le \sup_{(x,y)\in U} |X(u_{x,y}) - X(v_{x,y})| + 4 \max_{r\in S} |X(r) - X(\pi_{k(\delta)}(r))|.\end{aligned}$$

Since $\operatorname{Card}(U) \le (N(T,d,2^{-k(\delta)}))^2$ and, for $(x,y) \in U$, $d(u_{x,y}, v_{x,y}) \le \delta$, inequality (2.37) gives

$$E \sup_{(x,y)\in U} |X(u_{x,y}) - X(v_{x,y})| \le \delta\sqrt{2\log 2N^2(T,d,2^{-k(\delta)})}.$$

Next, the proof of (2.38) gives

$$E \max_{r\in S} |X(r) - X(\pi_{k(\delta)}(r))| \le \sum_{k>k(\delta)} 2^{-k+1}\sqrt{2\log 2N(T,d,2^{-k})}.$$

We then conclude from the last three inequalities that

$$\begin{aligned}E \max_{\substack{s,t\in S\\ d(s,t)\le\delta}} |X(t) - X(s)| &\le 2\delta\sqrt{\log\sqrt{2}N(T,d,2^{-k(\delta)})} + 4\sum_{k>k(\delta)} 2^{-k+1}\sqrt{2\log 2N(T,d,2^{-k})}\\ &\le (16\sqrt{2}+2)\int_0^\delta \sqrt{\log 2N(T,d,\varepsilon)}\, d\varepsilon. \quad \blacksquare\end{aligned}$$

Theorem 2.3.6 implies the existence of versions of $X(t)$ whose sample paths are bounded and uniformly continuous for d, actually, that this holds for all the separable versions of X, and they do exist (recall Proposition 2.1.12 complemented by Exercise 2.3.6, and note that the entropy condition obviously implies that (T,d) is a separable pseudo-metric space). For the next theorem, recall the definition of sample bounded and sample continuous processes (Definition 2.1.3).

Theorem 2.3.7 *Let (T,d) be a metric or pseudo-metric space, and let $X(t)$, $t \in T$, be a sub-Gaussian process relative to d. Assume that*

$$\int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\, d\varepsilon < \infty. \tag{2.40}$$

Then

(a) $X(t)$, $t \in T$, is sample d-continuous (in particular, X admits a separable version), and
(b) Any separable version of $X(t)$, $t \in T$, that we keep denoting by $X(t)$ has almost all its sample paths bounded and uniformly d-continuous, and satisfies the inequalities

$$E \sup_{t\in T} |X(t)| \le E|X(t_0)| + 4\sqrt{2}\int_0^{D/2} \sqrt{\log 2N(T,d,\varepsilon)}\, d\varepsilon, \tag{2.41}$$

where $t_0 \in T$, D is the diameter of (T,d) and

$$E \sup_{\substack{s,t\in T\\ d(s,t)\le\delta}} |X(t)-X(s)| \le (16\sqrt{2}+2)\int_0^\delta \sqrt{\log 2N(T,d,\varepsilon)}\, d\varepsilon, \tag{2.42}$$

for all $\delta > 0$.

Proof The entropy condition implies that (T,d) is totally bounded, in particular, separable. Then, if T_0 is a countable dense set and $T_n \nearrow T_0$, T_n finite, the monotone convergence theorem together with inequality (2.38) implies that both this inequality holds for $\sup_{t\in T_0}|X(t)|$ and this random variable is almost surely finite. Likewise, monotone convergence also proves inequality (2.39) for T_0 and, in particular, that for any sequence $\delta_n \searrow 0$,

$$E \sup_{\substack{s,t\in T_0\\ d(s,t)\le\delta_n}} |X(t)-X(s)| \searrow 0.$$

This implies not only that these random variables are finite a.s. but also that $\sup_{s,t\in T_0, d(s,t)\le\delta_n}|X(t)-X(s)| \searrow 0$ a.s. Hence, there exists a set $\Omega_0 \subseteq \Omega$ with $\Pr(\Omega_0)=1$ such that the restriction $X|_{T_0}$ of X to T_0 has bounded and d-uniformly continuous sample paths $t \mapsto X(t,\omega)$, $t \in T_0$, for all $\omega \in \Omega_0$. If we extend each of these paths to T by continuity, we obtain a separable version $\tilde{X}$ of the process X with almost all its sample paths bounded and d-uniformly continuous and such that the inequalities (2.38) and (2.39) hold for $\sup_{t\in T}|\tilde{X}(t)|$ and $\sup_{s,t\in T, d(s,t)\le\delta}|\tilde{X}(t)-\tilde{X}(s)|$, respectively (as these suprema equal the corresponding suprema over T_0 for all $\omega \in \Omega_0$). This proves part (a) and the inequalities in part (b) for the version just constructed. Now, if $\bar{X}$ is any separable version of X and T_0 is the countable set from Definition 2.1, we can apply to them the same reasoning as earlier and conclude part (b). ■

The chaining argument also can be adapted to obtain a metric entropy bound on the *modulus of continuity* of a sample continuous Gaussian or sub-Gaussian process.

Theorem 2.3.8 (Dudley's theorem) *If $X(t)$, $t \in T$, is a Gaussian process for a pseudo-metric d such that (T,d) has positive d-diameter and satisfies the metric entropy condition (2.40), then, for any separable version of X (still denoted by X), we have, with the convention $0/0=0$,*

$$E\left[\sup_{s,t\in T}\frac{|X(t)-X(s)|}{\int_0^{d(s,t)}\sqrt{\log N(T,d,\varepsilon)}\, d\varepsilon}\right] < \infty. \tag{2.43}$$

Proof The main part of the proof consists of showing that

$$\sup_{s,t\in T}\frac{|X(t)-X(s)|}{\int_0^{d(s,t)}\sqrt{\log N(T,d,\varepsilon)}\, d\varepsilon} < \infty \ a.s. \tag{2.44}$$

Once this is proved, (2.43) will follow from general properties of Gaussian processes. The proof of (2.44) (which in fact applies also to sub-Gaussian processes) consists of a

delicate chaining argument. Set $H(\varepsilon) = \log N(T,d,\varepsilon)$. Instead of discretising at $\varepsilon = 2^{-k}$ as in the proof of Theorem 2.3.6, we define $\varepsilon_1 = 1$ and, inductively, $\delta_k \searrow 0$ and $\varepsilon_k \searrow 0$ as

$$\delta_k = 2\inf\{\varepsilon : H(\varepsilon) \le 2H(\varepsilon_k)\}, \quad \varepsilon_{k+1} = \min(\varepsilon_k/3, \delta_k), \quad k \in \mathbb{N}.$$

Then, since $\varepsilon_{k+1} \le \varepsilon_k/3$, we have $\varepsilon_k \le 3(\varepsilon_k - \varepsilon_{k+1})/2$; also, if $\varepsilon_{k+1} = \delta_k$, then $H(\varepsilon_{k+1}) \le H(2\delta_k/3) \le 2H(\varepsilon_k)$, so $\int_{\varepsilon_{k+1}}^{\varepsilon_k} H^{1/2}(x)dx \le 2\varepsilon_k H^{1/2}(\varepsilon_k)$, whereas if $\varepsilon_{k+1} = \varepsilon_k/3$, then $\int_{\varepsilon_{k+1}}^{\varepsilon_k} H^{1/2}(x)dx \le 2\varepsilon_{k+1}H^{1/2}(\varepsilon_{k+1})$. This gives, for each n,

$$\frac{2}{3}\sum_{k=n}^{\infty} \varepsilon_k H^{1/2}(\varepsilon_k) \le \sum_{k=n}^{\infty} (\varepsilon_k - \varepsilon_{k+1}) H^{1/2}(\varepsilon_k) \le \int_0^{\varepsilon_n} H^{1/2}(x)dx \le 4\sum_{k=n}^{\infty} \varepsilon_k H^{1/2}(\varepsilon_k), \tag{2.45}$$

and the sums converge because, by (2.40), so does the integral. We also have, for each k,

$$H(\varepsilon_{k+2}) \ge H(\varepsilon_{k+1}/3) \ge H(\delta_k/3) \ge 2H(\varepsilon_k). \tag{2.46}$$

Finally, $\{\delta_k\}$ relates to $\{\varepsilon_k\}$ as follows: by definition, if $\tau < \delta_k/2$, then $H(\tau) > 2H(\varepsilon_k) \ge H(\varepsilon_k)$ so that $\delta_k \le 2\varepsilon_k$, which gives

$$\varepsilon_{k+1} \le \delta_k \le 6\varepsilon_{k+1}. \tag{2.47}$$

For each k, let T_k be a set of cardinality $N(\delta_k) = N(T,d,\delta_k)$ and δ_k-dense in T for d, and let $G_k = \{(s,t) : s \in T_{k-1}, t \in T_k\}$. Then $\mathrm{Card}(T_k) = e^{H(\delta_k)} \le e^{2H(\varepsilon_k)}$ by definition of δ_k, and $\mathrm{Card}(G_k) \le e^{4H(\varepsilon_k)}$. Then the sub-Gaussian tail bound (2.22) combined with the bound on the cardinality of G_k gives

$$\sum_k \Pr\left\{\max_{s\in T_{k-1}, t\in T_k} \frac{|X(t)-X(s)|}{d(s,t)} \ge 3H^{1/2}(\varepsilon_k)\right\} \le 2\sum_k e^{4H(\varepsilon_k) - 9H(\varepsilon_k)/2} \le 2\sum_k e^{-H^{1/2}(\varepsilon_k)/2},$$

which is finite because by (2.46) this last series is dominated by the sum of two convergent geometric series. Hence, by the Borel-Cantelli lemma, there exists $n_0(\omega) < \infty$ a.s. such that

$$\frac{|X(t,\omega) - X(s,\omega)|}{d(s,t)} \le 3H^{1/2}(\varepsilon_n), \quad \text{for all } (s,t) \in G_n \text{ and } n \ge n_0(\omega). \tag{2.48}$$

Next, given $n \in \mathbb{N}$, and $t \in T$, let $\pi_n(t) \in T_n$ be such that $d(t,\pi_n(t)) < \delta_n$. The metric entropy being finite, any separable version of X has almost all its sample paths continuous by Theorem 2.3.7; hence, there is a set of measure one Ω_1 such that if $\omega \in \Omega_1$, both $n_0(\omega) < \infty$ and $X(\pi_k(t),\omega)$ converges to $X(t,\omega)$ for all $t \in T$ (actually, there is no need to invoke this theorem because it is easy to see that $\{X(\pi_k(t),\omega))\}$ is a Cauchy sequence for all ω such that $n_0(\omega) < \infty$ by (2.48) and finiteness of the entropy integral. Hence, we can take a version of X such that, for these ω, $X(t,\omega) = \lim X(\pi_n(t),\omega))$.

Then, if $n \geq n_0(\omega)$ and $\varepsilon_{n-1} < d(s,t) \leq \varepsilon_n$, $s,t \in T$, the preceding two observations and the fact that $d(\pi_k(s),\pi_k(t)) \leq d(s,t) + 2\delta_k$, give

$$\begin{aligned}
|X(t,\omega) - X(s,\omega)| &\leq |X(\pi_n(t),\omega) - X(\pi_n(s),\omega)| + \sum_{k=n}^{\infty} |X(\pi_k(t),\omega) - X(\pi_k(t),\omega)| \\
&\quad + \sum_{k=n}^{\infty} |X(\pi_k(s),\omega) - X(\pi_k(s),\omega)| \\
&\leq 3(d(s,t) + 2\delta_n) H^{1/2}(\varepsilon_n) + 12 \sum_{k=n}^{\infty} \delta_k H^{1/2}(\varepsilon_{k+1}) \\
&\leq 39 d(s,t) H^{1/2}(d(s,t)) + 108 \int_0^{d(s,t)} H^{1/2}(x)dx \\
&\leq 147 \int_0^{d(s,t)} H^{1/2}(x)dx,
\end{aligned}$$

where, besides (2.48) and the convergence of $X(\pi_k(t))$, we have used (2.47) and (2.45). Thus, the modulus $\int_0^{d(s,t)} H^{1/2}(x)dx$ for $X(t,\omega)$ is valid for $d(s,t) \leq \varepsilon_{n_0(\omega)}$, and hence, by total boundedness of T, it is valid for all $d(s,t)$ and for all $\omega \in \Omega_1$. This proves (2.44)

Next, we show how (2.44) implies (2.43). Set

$$U = \{u = (u_1,u_2) : u_1,u_2 \in T, d(u_1,u_2) \neq 0\},$$

and define on U the pseudo-metric $D(u,v) = d(u_1,v_1) + d(u_2,v_2)$. Then (U,D) is a separable metric or pseudo-metric space because (T,d) is separable by Proposition 2.1.12. Consider the Gaussian process

$$Y(u) = \frac{X(u_2) - X(u_1)}{J(d(u_1,u_2))}, \quad u \in U,$$

where $J(x) = \int_0^x \sqrt{\log N(T,d,\varepsilon)}d\varepsilon$, and note that $J(x) > 0$ for all $x > 0$ (as the diameter of T is not zero). This is a Gaussian process on U with bounded sample paths (by (2.44)). It also has continuous paths for D because

$$\begin{aligned}
|Y(u) - Y(u^0)| &\leq \frac{|X(u_2) - X(u_1) - (X(u_2^0) - X(u_1^0))|}{J((d(u_1^0,u_2^0))} \\
&\quad + \left(\sup_{s,t\in T} |X(t) - X(s)|\right) \left| \frac{1}{J((d(u_1,u_2))} - \frac{1}{J((d(u_1^0,u_2^0))} \right|
\end{aligned}$$

tends to zero as $u \to u^0$ in the D-distance because of (a) the sample continuity of X, (b) the first part of the theorem, (c) the continuity of $J(x)$ and (d) $J(d(u_1^0,u_2^0) > 0$. In particular, Y is a separable Gaussian process on (U,D) with bounded sample paths; hence,

$$E \sup_{u\in U} |Y(u)| < \infty$$

by part (a) of Theorem 2.1.20, proving (2.43). ■

In fact, Theorem 2.1.20 yields more than just first-moment integrability in (2.43) once (2.44) is proved, namely, square exponential integrability.

Exercises

2.3.1 The main ingredient in the basic estimates of Theorem 2.3.6 is clearly the first maximal inequality in (2.37) (hence, Lemma 2.3.4). Replace this inequality with the maximal inequality (2.31) for the ψ_2-norm from Lemma 2.3.3 in the proof of Theorem 2.3.6 to obtain that if $X(t)$, $t \in T$, is a sub-Gaussian process for a pseudo-distance d for which (T,d) satisfies the entropy condition (2.40), then the following inequalities hold for any separable version of X:

$$\left\| \sup_{t\in T} |X(t)| \right\|_{\psi_2} \le \|X(t_0)\|_{\psi_2} + 16\sqrt{6} \int_0^D \sqrt{\log N(T,d,\varepsilon)}\, d\varepsilon,$$

where $t_0 \in T$ is arbitrary and D is the d-diameter of T, and

$$\left\| \sup_{\substack{s,t\in T \\ d(s,t)\le\delta}} |X(t)-X(s)| \right\|_{\psi_2} \le 128\sqrt{3} \int_0^\delta \sqrt{\log N(T,d,\varepsilon)}\, d\varepsilon,$$

for any $\delta > 0$ In particular, these inequalities also hold for the L^p-norms of these random variables, $p < \infty$, possibly with different constants.

2.3.2 Brownian motion on $[0,1]$ is defined as a centred Gaussian process $X(t)$ with continuous sample paths and such that $X(0) = 0$ a.s., $E(X(s)-X(t))^2 = |t-s|$, $s,t \in [0,1]$. Prove the existence of Brownian motion, and show that $\sup_{s,t\in[0,1]} |X(t)-X(s)|/\sqrt{|t-s||\log|t-s||} < \infty$ almost surely.

2.3.3 For real random variables X_i, give an upper bound for $E\sup_{t\in\mathbb{R}} \left|1/\sqrt{n}\sum_{i=1}^n \varepsilon_i I(X_i \le t)\right|$, $n \in \mathbb{N}$; in particular, prove that $E\sup_{t\in\mathbb{R}} \left|1/n\sum_{i=1}^n \varepsilon_i I(X_i \le t)\right| \to 0$ (Glivenko-Cantelli theorem). *Hint*: Conditionally on $\{X_i\}$, take $d^2(s,t) = 1/n\sum_{i=1}^n (I(X_i \le t) - I(X_i \le s)|)^2$, and notice that if $X_{(i)}$, $i = 1,\dots,n$, are the order statistics, $d(s,t) = 0$ if (and only if) both s and t belong to one of the sets $(-\infty, X_{(1)}]$, $(X_{(n)},\infty)$ or $(X_{(i)}, X_{(i+1)}]$, $i = 1,\dots,n-1$. Note also that $d(s,t) \le 1$ for all s,t. Deduce that $N(\mathbb{R},d,\varepsilon) \le n+1$ for all $\varepsilon > 0$ and that $D \le 1$. The bound follows from this estimate and the entropy integral bound.

2.3.4 (Alternate proof of inequality (2.39) with a slightly larger constant.) Define $V = \{(s,t) \in T\times T : d(s,t) \le \delta\}$ and on V the process $Y(u) = X(t_u) - X(s_u)$, where $u = (s_u,t_u) \in V$. Take on V the pseudo-distance $\rho(u,v) := \|Y(u)-Y(v)\|_{\psi_2}$. One has that $Y(v)$ is sub-Gaussian for $\sqrt{6}\rho$ on V, that $2\max_{u\in V}\|Y(u)\|_{\psi_2} \le 2\sqrt{6}\delta$ and that $\rho(u,v) \le \sqrt{6}(d(t_u,t_v)+d(s_u,s_v))$, all by Lemma 2.3.1. Thus, one can apply inequality (2.38) to Y for ρ, using that the first of the preceding two inequalities gives a bound for the ρ-diameter of V and that the second implies $N(V,\rho,4\sqrt{6}\varepsilon) \le N^2(T,d,\varepsilon)$.

2.3.5 Use the fact that the function $e^{x^2} - 1$ is convex and zero at zero to show that $\|\cdot\|_{\psi_2}$ is a (pseudo-)norm on the space L^{ψ_2} of all the random variables $\xi : \Omega \mapsto \mathbb{R}$ such that $Ee^{\lambda\xi^2} < \infty$ for some $\lambda > 0$ (with identification of a.s. equal functions). Show that the resulting normed space is complete.

2.3.6 Show that Proposition 2.1.12 holds true for sub-Gaussian processes.

2.3.7 Show that a separable stochastic process $X(t)$, $t \in T$, is sample continuous on (T,d) iff there exists a Borel probability measure on $C_u(T,d)$, the Banach space of bounded and uniformly continuous functions on (T,d), whose finite-dimensional marginals $\mu\circ(\delta_{t_1},\dots,\delta_{t_n})^{-1}$ are the marginals $\mathcal{L}(X(t_1),\dots,X(t_n))$, for all $t_i \in T$, $i \le n$, $n \in \mathbb{N}$.

2.3.8 Prove the following inequality, which is a qualitative improvement on (2.31) as it does not assume a finite number of variables: there exists a universal constant $K < \infty$ such that

$$\left\| \sup_k \frac{|\xi_k|}{\psi_2^{-1}(k)} \right\|_{\psi_2} \le K \sup_k \|\xi_k\|_{\psi_2}$$

with $\|\xi_k\|_{\psi_2}$ replaced by $\|\xi_k\|_{L^2}$ if the variables ξ_i are normal. *Hint*: Assume $\|\xi_k\|_{\psi_2} \le 1$. Then using a union bound,

$$\Pr\left\{\exp\left[\sup_{k\ge 9}\left(\frac{|\xi_k|}{\sqrt{6\log k}}\right)^2\right] > t\right\} \le \sum_{k=9}^{\infty} \Pr\left\{e^{|\xi_k|^2} > e^{6(\log k)(\log t)}\right\},$$

and then apply inequality (2.22) together with the fact that for $t \ge 3/2$ and $k \ge 9$, $\log(kt) \le 3(\log k)(\log t)$. Use the resulting bound to show that

$$E\exp\left[\sup_{k\ge 9}\left(|\xi_k|/(\sqrt{6\log k})\right)^2\right] < 2.$$

2.3.9 Let X_i, $i \le n$, be separable centred Gaussian processes such that $E\|X_i\|_\infty < \infty$ (where $\|\cdot\|_\infty$ denotes the supremum norm), and let σ_i^2 and M_i be, respectively, their sup of second moments and median. Prove that

$$E\max_{i\le n}\|X_i\|_\infty \le \max_{i\le n} E\|X_i\|_\infty + (8\sqrt{\log n} + \sqrt{2/\pi})\max_{i\le n}\sigma_i.$$

Hint: By Theorem 2.2.7 and Lemma 2.3.2, the variables $|\|X_i\| - M_i|$ have ψ_2-norm bounded by $2\sigma_i$, and the result then follows from Lemma 2.3.3 and inequality (2.19).

2.3.10 Show that there exists $K < \infty$ such that if $Y(t)$, $t \in T$, is a centred Gaussian process such that $d_Y^2(s,t) = E(Y(t) - Y(s))^2 \le d^2(s,t)$ and (T,d) is totally bounded, then

$$E\sup_{d(s,t)<\delta}|Y(t)-Y(s)| \le K\left[\sup_{t\in T} E\sup_{s\in T: d(s,t)<\delta}|Y(t)-Y(s)| + \delta(\log N(T,d,\delta))^{1/2}\right].$$

Hint: Let U be the set of centres of $N(T,d,\delta)$ d-balls of radius δ covering T. Apply the result in Exercise 2.3.9 to the processes $Y_u = Y - Y(u)$, $u \in U$, and inequality (2.35) to $\max_{u,v\in U: d(u,v)<3\delta}|Y(u)-Y(v)|$.

2.4 Anderson's Lemma, Comparison and Sudakov's Lower Bound

In this section we deal with the general question of how comparison of the distributions of the supremum of two Gaussian processes follows from comparison of their covariances or of their induced metric structures. Perhaps the most important results of this kind are Anderson's inequality regarding the probability, relative to a centred Gaussian measure on $\mathbb{R}^n$, of a convex symmetric set and its translates, and Slepian's lemma that allows comparing the distributions of the suprema of $X(t)$ and $Y(t)$ if the covariance of one of the processes dominates the other. Anderson's lemma is related to the fact that centred Gaussian measures on $\mathbb{R}^n$ are log-concave.

These results have several important consequences, and we will examine two particularly interesting ones, the Khatri-Sidak inequality and Sudakov's inequality, that compare, for a jointly normal variable $(g_1,\dots,g_n)$, the distribution of $\max_{1\le i\le n}|g_i|$ with the maximum of related independent normal random variables. Sudakov's inequality shows that Dudley's entropy bound is effectively sharp and, in this sense, complements it.

2.4.1 Anderson's Lemma

A set C in a vector space is convex and symmetric if $\sum_{i=1}^n \lambda_i x_i \in C$ whenever $x_i \in C$ and $\lambda_i \in \mathbb{R}$ satisfy $\sum_{i=1}^n |\lambda_i| = 1$, $n < \infty$. Example: Balls centred at the origin in Banach spaces,

$\{x : \|x\| \le c\}$. Anderson's lemma states that for a centred Gaussian measure μ on $\mathbb{R}^n$, if C is a measurable, convex, symmetric set, then

$$\mu(C+x) \le \mu(C),$$

for all $x \in \mathbb{R}^n$. Suppose now that $X = Y + Z$, where Y and Z are two independent centred Gaussian random vectors in $\mathbb{R}^n$, which holds if and only if the difference of covariances $C_X - C_Y$ is nonnegative definite. Then

$$\Pr\{X \in C\} = \int \Pr\{Y \in C - z\} d\mathcal{L}(Z)(z) \le \Pr\{Y \in C\}.$$

This inequality is stronger than $E\|Y+Z\|^p \ge E\|Y\|^p$ for all $p \ge 1$, which follows from it and also from Jensen's inequality. Both Anderson's inequality and its corollary on comparison of Gaussian probabilities are quite useful. The modern proof of Anderson's lemma uses the Brunn-Minkowski inequality, or inequalities similar to it, expressing the log-concavity of the function $A \mapsto m(A)$, where m is Lebesgue measure and, as a consequence (of a slightly stronger inequality) of $A \mapsto \mu(A)$, μ-Gaussian and centred.

We start with the Brunn-Minkowski inequality for Lebesgue measure in $\mathbb{R}$. Given two sets A and B in a vector space, their Minkowski addition is $A + B = \{x + y : x \in A, y \in B\}$, and λA is defined as $\lambda A = \{\lambda x : x \in A\}$. In this subsection, m will stand for Lebesgue measure on $\mathbb{R}^n$ for any n.

Lemma 2.4.1 *Let A and B be Borel measurable sets in $\mathbb{R}$. Then*

$$m(A+B) \ge m(A) + m(B).$$

Proof Note that $A + B$ is Lebesgue measurable as it is the image by a continuous function of the Borel set $A \times B$, hence, analytic. Regularity of m by compact sets reduces the problem to A and B compact. Since m is invariant by translations, neither side of the inequality changes if we translate the sets A and/or B; hence, by taking $A + \{-\sup A\}$ and $B + \{-\inf B\}$ instead of A and B, we can assume $A \subset \{x \le 0\}$, $B \subseteq \{x \ge 0\}$ and $A \cap B = \{0\}$. But then $m(A+B) \ge m(A \cup B) = m(A) + m(B)$. ■

Theorem 2.4.2 (**Précopa-Leindler theorem**) *Let f, g, φ be Lebesgue measurable functions on $\mathbb{R}^n$ taking values in $[0, \infty]$ and satisfying, for some $0 < \lambda < 1$ and all $u, v \in \mathbb{R}^n$,*

$$\varphi(\lambda u + (1-\lambda)v) \ge f^{\lambda}(u) g^{1-\lambda}(v). \tag{2.49}$$

Then

$$\int \varphi \, dm \ge \left(\int f \, dm \right)^{\lambda} \left(\int g \, dm \right)^{1-\lambda}. \tag{2.50}$$

Proof The proof is by induction on the dimension n. Assume that $n = 1$. We can divide both sides of inequality (2.49) by $\|f\|_\infty^\lambda \|g\|_\infty^{1-\lambda}$; that is, we can assume without loss of generality that $\|f\|_\infty = \|g\|_\infty = 1$. Then, for $0 \le t < 1$, the sets $\{x : f(x) \ge t\}$ and $\{x : g(x) \ge t\}$ are not empty, and we have

$$\lambda\{f \ge t\} + (1-\lambda)\{g \ge t\} \subseteq \{\varphi \ge t\},$$

since, by (2.49), if $f(u) \ge t$ and $g(v) \ge t$, then $\varphi(\lambda u + (1-\lambda)v) \ge t$. But then, by Lemma 2.4.1,

$$m\{\varphi \ge t\} \ge \lambda m\{f \ge t\} + (1-\lambda) m\{g \ge t\}.$$

Integrating with respect to t and using the concavity of the logarithm, we obtain

$$\int \varphi \, dm \geq \lambda \int f \, dm + (1-\lambda) \int g \, dm \geq \left(\int f \, dm \right)^{\lambda} \left(\int g \, dm \right)^{1-\lambda},$$

proving the theorem for $n = 1$. Assume now that the result holds for $n-1$, and let φ, f, g, λ be as in the statement of the theorem. Fix a coordinate, say, $x_n = x$, and consider $\varphi_x : \mathbb{R}^{n-1} \mapsto [0, \infty]$, defined by $\varphi_x(t) = \varphi(t,x)$, and likewise define f_x and g_x. Then, for x_1, x_2 such that $x = \lambda x_1 + (1-\lambda)x_2$ and for any $u, v \in \mathbb{R}^{n-1}$,

$$\varphi_x(\lambda u + (1-\lambda)v) = \varphi(\lambda(u,x_1) + (1-\lambda)(v,x_2)) \geq f^{\lambda}(u,x_1) g^{1-\lambda}(v,x_2) = f_{x_1}^{\lambda}(u) g_{x_2}^{1-\lambda}(v).$$

Hence, induction gives

$$\int_{\mathbb{R}^{n-1}} \varphi_x \, dm \geq \left(\int_{\mathbb{R}^{n-1}} f_{x_1} \, dm \right)^{\lambda} \left(\int_{\mathbb{R}^{n-1}} g_{x_2} \, dm \right)^{1-\lambda},$$

and (2.50) now follows by application of the very same result in dimension one. ■

We sketch in Exercise 2.4.1 how to obtain the Brunn-Minkowski inequality from Theorem 2.4.2. Of course, we are primarily interested in using this theorem to prove that centred Gaussian measures are logarithmically concave.

Theorem 2.4.3 (Log-concavity of Gaussian measures in $\mathbb{R}^n$) *Let μ be a centred Gaussian measure on $\mathbb{R}^n$. Then, for any Borel sets A, B in $\mathbb{R}^n$ and $0 \leq \lambda \leq 1$, we have*

$$\mu(\lambda A + (1-\lambda)B) \geq (\mu(A))^{\lambda} (\mu(B))^{1-\lambda}. \tag{2.51}$$

Proof Let μ be a centred Gaussian measure on $\mathbb{R}^n$. Then μ is supported by a subspace $V \subset \mathbb{R}^n$, and the density of the restriction of μ to V with respect to Lebesgue measure on V is $\phi(x) = ce^{-|\Gamma x|^2/2}$, where $\Gamma : V \mapsto V$ is the positive square root of the inverse of the restriction to V of the covariance of μ and is a strictly positive definite operator. It is easy to see, for example, by diagonalising Γ, that the function $x \mapsto \log \phi(x) = -|\Gamma x|^2$, $x \in V$, is concave and therefore that

$$\phi(\lambda u + (1-\lambda)v) \geq \phi^{\lambda}(u) \phi^{1-\lambda}(v), \quad u, v \in V. \tag{2.52}$$

Now, if A and B are Borel sets of $\mathbb{R}^n$, we define, on V,

$$\varphi = \phi I_{\lambda(A \cap V) + (1-\lambda)(B \cap V)}, \; f = \phi I_{A \cap V}, \; g = \phi I_{B \cap V}.$$

Note that the set $\lambda(A \cap V) + (1-\lambda)(B \cap V)$ is the image by a continuous function of a Borel set on $V \times V$; hence, it is measurable for the completion of any Borel measure on V (e.g., Dudley (2002), section 13.2)). Inequality (2.52) shows that these functions satisfy the hypothesis (2.49) with $\mathbb{R}^n$ replaced by V. Hence, Theorem 2.4.2 applies to give

$$\int_{\lambda(A \cap V) + (1-\lambda)(B \cap V)} \phi \, dm \geq \left(\int_{A \cap V} \phi \, dm \right)^{\lambda} \left(\int_{B \cap V} \phi \, dm \right)^{1-\lambda},$$

where m is Lebesgue measure on V. This inequality implies the theorem because

$$\begin{aligned} \mu(\lambda A + (1-\lambda)B) &= \mu[(\lambda A + (1-\lambda)B) \cap V] \\ &\geq \mu(\lambda(A \cap V) + (1-\lambda)(B \cap V)) = \int_{\lambda(A \cap V) + (1-\lambda)(B \cap V)} \phi \, dm, \end{aligned}$$

and $\mu(A) = \int_{A \cap V} \phi \, dm$ and likewise for $\mu(B)$. ■

An immediate consequence of this theorem is Anderson's inequality for *any* centred Gaussian measure on $\mathbb{R}^n$.

Theorem 2.4.4 (Anderson's lemma) *Let $X = (g_1, \ldots, g_n)$ be a centred jointly normal vector in $\mathbb{R}^n$, and let C be a measurable convex symmetric set of $\mathbb{R}^n$. Then, for all $x \in \mathbb{R}^n$,*

$$\Pr\{X + x \in C\} \leq \Pr\{X \in C\}. \tag{2.53}$$

Proof Let $\mu = \mathcal{L}(X)$. Let $A = C + x$, $B = C - x$ and $\lambda = 1/2$ in (2.51), and note that by symmetry of μ and symmetry of C, $\mu(A) = \mu(B)$, so we obtain $\mu(C) \geq \mu(C + x)$, which is (2.53). ■

The assumption of measurability for C in the statement of the preceding theorem is superfluous because the boundary of a convex set C has μ-measure zero (whereas obviously its closure and its interior are measurable), but in applications, C is usually open or closed and hence measurable.

Theorem 2.4.4 extends to infinite dimensions, both for B-valued random variables, B separable (next theorem) and processes (Exercise 2.4.3).

Theorem 2.4.5 *Let B be a separable Banach space, let X be a B-valued centred Gaussian random variable and let C be a closed, convex, symmetric subset of B. Then, for all $x \in B$,*

$$\Pr\{X + x \in C\} \leq \Pr\{X \in C\}.$$

In particular, $\Pr(\|X\| \leq \varepsilon) > 0$, *for all* $\varepsilon > 0$.

Proof By the Hahn-Banach separation theorem in locally convex topological spaces, there exists a set $D_C \subset B^*$ such that $C = \cap_{f \in D_C}\{|f| \leq 1\}$. Then $C^c = \cup_{f \in D_C}\{|f| > 1\}$. Since C^c is separable, its topology has a countable base, and therefore, this covering admits a countable subcovering; that is, there exists a countable subset $T_C \subset D_C$ such that $C^c = \cup_{f \in T_C}\{|f| > 1\}$ or $C = \cap_{f \in T_C}\{|f| \leq 1\}$. Then, if $T_n \nearrow T_C$, T_n finite, we have

$$\Pr\{X \in C\} = \Pr\{\sup_{f \in T_C} |f(X)| \leq 1\} = \lim_{n \to \infty} \Pr\left\{\max_{f \in T_n} |f(X)| \leq 1\right\}$$

$$\geq \lim_{n \to \infty} \Pr\left\{\max_{f \in T_n} |f(X + x)| \leq 1\right\} = \Pr\{X + x \in C\},$$

where the inequality follows from Theorem 2.4.4 applied to the Gaussian vector $(f(X) : f \in T_n)$ and the convex set $\{x \in \mathbb{R}^{\mathrm{Card}(T_n)} : |x_i| \leq 1, i = 1, \ldots, \mathrm{Card}(T_n)\}$. For the last claim, apply the first part to closed balls $C_i = \{x : \|x - x_i\| \leq \varepsilon\}$ for x_i a countable dense subset of B. ■

Anderson's lemma applies to the comparison of the probabilities that $X = Y + Z$ and Y fall in convex symmetric sets C, where Y and Z are independent centred Gaussian $\mathbb{R}^n$-valued random vectors (Exercise 2.4.2), and gives

$$\Pr\{X \in C\} \leq \Pr\{Y \in C\}.$$

Here is another application of Anderson's lemma, in the version of Exercise 2.4.5, to comparison of Gaussian processes, concretely, to proving the simplest yet useful instance of the famous Gaussian correlation conjecture, known as the *Khatri-Sidak inequality*. The

Gaussian correlation conjecture itself states that for symmetric convex sets A, B, if X and Y are arbitrary centred Gaussian vectors, $\Pr\{X \in A, Y \in B\} \geq \Pr\{X \in A\}\Pr\{Y \in B\}$; that is, the independent case gives the smallest probability of the intersection of two symmetric convex sets.

Corollary 2.4.6 (Khatri-Sidak inequality) *Let $n \geq 2$, and let $g_1, \ldots, g_n$ be jointly normal centred random variables. Then, for all $x \geq 0$,*

$$\Pr\{\max_{1\leq i\leq n} |g_i| \leq x\} \geq \Pr\{|g_1| \leq x\}\Pr\{\max_{2\leq i\leq n} |g_i| \leq x\},$$

and hence, iterating,

$$\Pr\{\max_{1\leq i\leq n} |g_i| \leq x\} \geq \prod_{i=1}^{n} \Pr\{|g_i| \leq x\}.$$

Proof Note that $\Pr\{\max_{2\leq i\leq n} |g_i| \leq x\} = \lim_{t\to\infty} \Pr\{\max_{2\leq i\leq n} |g_i| \leq x, |g_1| \leq t\}$. Hence, it suffices to show that for any convex symmetric subset A of $\mathbb{R}^{n-1}$, the function

$$f(t)/g(t) := \Pr\{|g_1| \leq t, (g_2, \ldots, g_n) \in A\}/\Pr\{|g_1| \leq t\}$$

is monotone decreasing. Let ϕ_1 denote the density of g_1, and set $X = (g_2, \ldots, g_n)$. Since

$$\begin{aligned}\Pr\{X \in A | |g_1| \leq t\} &= \int_{-t}^{t} \Pr\{X \in A | g_1 = u\} d\mathcal{L}(g_1 | |g_1| \leq t)(u) \\ &= \int_{-t}^{t} \Pr\{X \in A | g_1 = u\}\phi_1(u)du / \Pr\{|g_1| \leq t\},\end{aligned}$$

we have (using symmetry of the different laws) that

$$f(t) = \int_{-t}^{t} \Pr\{X \in A | g_1 = u\}\phi_1(u)du, \; f'(t) = 2\Pr\{X \in A | g_1 = t\}\phi_1(t)$$

and that, by Exercises 2.4.5 and 2.4.6,

$$\Pr\{X \in A | |g_1| \leq t\} \geq \Pr\{X \in A | g_1 = t\}.$$

These two observations give

$$\begin{aligned}g^2(t)(f/g)'(t) &= 2\phi_1(t)\Pr\{X \in A | g_1 = t\}\Pr\{|g_1| \leq t\} - 2\Pr\{|g_1| \leq t, (g_2, \ldots, g_n) \in A\}\phi_1(t) \\ &= 2\phi_1(t)\Pr\{|g_1| \leq t\}\left[\Pr\{X \in A | g_1 = t\} - \Pr\{X \in A | |g_1| \leq t\}\right] \leq 0.\end{aligned}$$

Thus, the function f/g is monotone decreasing, proving the corollary. ■

2.4.2 Slepian's Lemma and Sudakov's Minorisation

Before proving the basic comparison result, it is convenient to consider a useful identity regarding derivatives of the multidimensional normal density. Let $f(C,x) = ((2\pi)^n \det C)^{-1/2} e^{-xC^{-1}x^T/2}$ be the $N(0,C)$ density in $\mathbb{R}^n$, where $C = (C_{ij})$ is an $n \times n$ symmetric strictly positive definite matrix $x = (x_1, \ldots, x_n)$ and x^T is the transpose of x. Consider f as a function of the real variables C_{ij}, $1 \leq i \leq j \leq n$, and x_i, $1 \leq i \leq n$. Then

$$\frac{\partial f(C,x)}{\partial C_{ij}} = \frac{\partial^2 f(C,x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(C,x)}{\partial x_j \partial x_i}, \quad 1 \leq i < j \leq n. \tag{2.54}$$

To see this, just note that by the inversion formula for characteristic functions,

$$f(C,x) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-ixu^T} e^{-uCu^T/2} du$$

and that differentiation under the integral sign is justified by dominated convergence, so the three partial derivatives in (2.54) are all equal to $-x_i x_j f(C,x)$.

We can now prove the following comparison result:

Theorem 2.4.7 *Let $X = (X_1,\dots,X_n)$ and $Y = (Y_1,\dots,Y_n)$ be centred normal vectors in $\mathbb{R}^n$ such that $EX_i^2 = EY_j^2 = 1$, $1 \le i,j \le n$. Set, for each $1 \le i < j \le n$, $C_{ij}^1 = E(X_iX_j)$, $C_{ij}^0 = E(Y_iY_j)$ and $\rho_{ij} = \max\{|C_{ij}^0|, |C_{ij}^1|\}$. Then, for any $\lambda_i \in \mathbb{R}$,*

$$\Pr\bigcap_{i=1}^n \{X_i \le \lambda_i\} - \Pr\bigcap_{i=1}^n \{Y_i \le \lambda_i\} \le \frac{1}{2\pi} \sum_{1\le i<j\le n} (C_{ij}^1 - C_{ij}^0)^+ \frac{1}{(1-\rho_{ij}^2)^{1/2}} \exp\left(-\frac{(\lambda_i^2+\lambda_j^2)/2}{1+\rho_{ij}}\right). \tag{2.55}$$

Moreover, if $\mu_i \le \lambda_i$ and $v = \min\{|\lambda_i|, |\mu_i| : i = 1,\dots n\}$, then

$$\left|\Pr\bigcap_{i=1}^n \{\mu_i \le X_i \le \lambda_i\} - \Pr\bigcap_{i=1}^n \{\mu_i \le Y_i \le \lambda_i\}\right| \le \frac{2}{\pi} \sum_{1\le i<j\le n} |C_{ij}^1 - C_{ij}^0| \frac{1}{(1-\rho_{ij}^2)^{1/2}} \times \exp\left(-\frac{v^2}{1+\rho_{ij}}\right). \tag{2.56}$$

Proof We may assume that the covariances of X and Y are invertible (so that both X and Y have densities): just take, if necessary, $X_\varepsilon = (1-\varepsilon^2)^{1/2}X + \varepsilon G$, $Y_\varepsilon = (1-\varepsilon^2)^{1/2}Y + \varepsilon G$ instead, where G is a standard normal random vector on $\mathbb{R}^n$ independent of X and Y. Then the result for X_ε and Y_ε implies the result for X and Y by letting $\varepsilon \to 0$. Moreover, since both the hypotheses and the conclusions of the theorem involve the probability laws of X and Y but not their joint law, we may also assume that X and Y are independent.

Under these two assumptions, define $X(t) = t^{1/2}X + (1-t)^{1/2}Y$. Then $X(0) = Y$, $X(1) = X$ and $C^t := \mathrm{Cov}(X(t)) = tC^1 + (1-t)C^0$. This curve in $\mathbb{R}^{n(n-1)/2}$ has a neighbourhood consisting only of (symmetric) strictly positive definite matrices. Let f_t denote the density of $X(t)$, and define

$$F(t) = \int_{-\infty}^{\lambda_1} \cdots \int_{-\infty}^{\lambda_n} f_t(x)dx, \tag{2.57}$$

which can be easily seen to be in $C([0,1])$. Then the left-hand side of (2.55) is precisely

$$F(1) - F(0) = \int_0^1 F'(t)\, dt.$$

We can still differentiate under the integral sign in (2.57), and since by (2.54)

$$\frac{df_t}{dt} = \sum_{1\le i<j\le n} \frac{\partial f_t}{\partial C_{ij}} \frac{dC_{ij}}{dt} = \sum_{1\le i<j\le n} (C_{ij}^1 - C_{ij}^0) \frac{\partial^2 f_t}{\partial x_i \partial x_j},$$

we obtain

$$F'(t) = \sum_{1\le i<j\le n} (C_{ij}^1 - C_{ij}^0) \int_{-\infty}^{\lambda_1} \cdots \int_{-\infty}^{\lambda_n} \frac{\partial^2 f_t}{\partial x_i \partial x_j} dx.$$

Integrating $\partial f_t/(\partial x_i \partial x_j)$ with respect to x_i and x_j, we obtain $f_t(x')$, where $x'_k = x_k$ if $k \neq i,j$, $x'_i = \lambda_i$, $x'_j = \lambda_j$. Moreover, we can bound the integrals with respect to the other coordinates, $\int_{-\infty}^{\lambda_k}$, by integrals over $\mathbb{R}$ and obtain

$$\int_{-\infty}^{\lambda_1} \cdots \int_{-\infty}^{\lambda_n} \frac{\partial^2 f_t}{\partial x_i \partial x_j} dx \leq \int_{\mathbb{R}^{n-2}} f_t(x_1, \ldots, x_{i-1}, \lambda_i, x_{i+1}, \ldots, x_{j-1}, \lambda_j, x_{j+1}, \ldots, x_n) dx.$$

This last integral is just the evaluation at the point (λ_i, λ_j) of the joint density of $X_i(t)$ and $X_j(t)$, that is, the density of the centred normal probability law in $\mathbb{R}^2$ with covariance

$$\begin{pmatrix} 1 & C_{ij}^t \\ C_{ij}^t & 1 \end{pmatrix},$$

$$\frac{1}{2\pi(1-(C_{ij}^t)^2)^{1/2}} \exp\left(-\frac{\lambda_i^2 - 2C_{ij}^t \lambda_i \lambda_j + \lambda_j^2)}{2(1-(C_{ij}^t)^2)}\right).$$

Replacing C_{ij}^t with its absolute value and noting that the minimum of the function of u, $(a^2 - 2uab + b^2)/(1-u)$ on $[0,\infty)$, is attained at $u = 0$, to obtain $(\lambda_i^2 - 2C_{ij}^t \lambda_i \lambda_j + \lambda_j^2)/(2(1-(C_{ij}^t)^2)) \geq (\lambda_i^2 + \lambda_j^2)/2(1+|C_{ij}^t|)$, and then using that $\rho_{ij} \geq |C_{ij}^t|$, we see that the quantity in the last display is dominated by

$$\frac{1}{2\pi(1-\rho_{ij}^2)^{1/2}} \exp\left(-\frac{(\lambda_i^2 + \lambda_j^2)/2}{1+\rho_{ij}}\right).$$

This shows that

$$F'(t) \leq \frac{1}{2\pi} \sum_{1 \leq i < j \leq n} (C_{ij}^1 - C_{ij}^0)^+ \frac{1}{(1-\rho_{ij}^2)^{1/2}} \exp\left(-\frac{(\lambda_i^2 + \lambda_j^2)/2}{1+\rho_{ij}}\right)$$

and that this is a bound for its integral over $[0,1]$ as well, that is, for $F(1) - F(0)$, proving (2.55).

To prove (2.56), we define

$$\tilde{F}(t) = \int_{\mu_1}^{\lambda_1} \cdots \int_{\mu_n}^{\lambda_n} f_t(x) dx$$

and proceed as before to obtain, as a result of the double integration $\int_{\mu_i}^{\lambda_i} \int_{\mu_j}^{\lambda_j} (\partial^2 f_t)/(\partial x_i \partial x_j)$, the sum of four functions of $n-2$ variables, two of them obtained from f_t by, respectively, setting $(x_i, x_j) = (\lambda_i, \lambda_j)$ and $(x_i, x_j) = (\mu_i, \mu_j)$ and the other two from $-f_t$ by, respectively, setting $(x_i, x_j) = (\lambda_i, \mu_j)$ and $(x_i, x_j) = (\mu_i, \lambda_j)$. Then, on integrating over $\mathbb{R}^{n-2}$ as earlier (instead of between μ_k and λ_k for each $k \neq i,j$), we obtain

$$|\tilde{F}'(t)| \leq \frac{4}{2\pi} \sum_{1 \leq i < j \leq n} |C_{ij}^1 - C_{ij}^0| \frac{1}{(1-\rho_{ij}^2)^{1/2}} \exp\left(-\frac{v^2}{1+\rho_{ij}}\right),$$

which yields inequality (2.56) by integrating between 0 and 1. ■

In this section we need a little less, in fact, only the following consequence of Theorem 2.4.7:

Theorem 2.4.8 (Slepian's lemma) *Let $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_n)$ be centred jointly normal vectors in $\mathbb{R}^n$ such that*

$$E(X_i X_j) \le E(Y_i Y_j) \quad \text{and} \quad EX_i^2 = EY_i^2 \quad \text{for } 1 \le i,j \le n. \tag{2.58}$$

Then, for all $\lambda_i \in \mathbb{R}$, $i \le n$,

$$\Pr\left(\bigcup_{i=1}^{n} \{Y_i > \lambda_i\}\right) \le \Pr\left(\bigcup_{i=1}^{n} \{X_i > \lambda_i\}\right), \tag{2.59}$$

and therefore,

$$E \max_{i \le n} Y_i \le E \max_{i \le n} X_i. \tag{2.60}$$

Proof Under assumptions (2.58), the right-hand side of (2.55) is less than or equal to zero, so (2.59) follows from Theorem 2.4.7. Inequality (2.60) follows from (2.58) by integration by parts ($E|\xi| = \int_0^\infty \Pr\{|\xi| > \lambda\} d\lambda$). ■

Remark 2.4.9 Sometimes one wishes to compare expected values of the maximum of the absolute values, and to this end, the following may be useful: for X_i symmetric, for any $i_0 \in \{1, \dots, n\}$,

$$E \max_{i \le n} X_i \le E \max_{i \le n} |X_i| \le E|X_{i_0}| + E \max_{i,j} |X_i - X_j| \le E|X_{i_0}| + 2E \max_{i \le n} X_i,$$

where the last inequality follows because

$$E \max_{i,j} |X_i - X_j| = E \max_{i,j} (X_i - X_j) = E \max_i X_i + E \max_j (-X_j) = 2E \max_i X_i.$$

It is also worth mentioning that for any real random variable with mean zero, $E \max_i (X_i + Z) = EZ + E \max_i X_i = E \max_i X_i$.

The following corollary of Slepian's lemma is sometimes easier to apply than Theorem 2.4.8 because it does not require $EX_i^2 = EY_i^2$, $i \le n$.

Corollary 2.4.10 *Let $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_n)$ be two centred, jointly normal vectors in $\mathbb{R}^n$, and assume that*

$$E(Y_i - Y_j)^2 \le E(X_i - X_j)^2, \quad i,j \in \{1, \dots, n\}.$$

Then

$$E \max_{i \le n} Y_i \le 2E \max_{i \le n} X_i.$$

Proof Replacing X_i by $X_i - X_1$ and Y_i by $Y_i - Y_1$, we may assume that $X_1 = Y_1 = 0$ (see the preceding remark), which in particular implies that $EY_i^2 \le EX_i^2$. Set $\sigma_X^2 = \max_{i \le n} EX_i^2$, and let $\bar{X}$ and $\bar{Y}$ be Gaussian vectors whose coordinates are defined by

$$\bar{X}_i = X_i + (\sigma_X^2 + EY_i^2 - EX_i^2)^{1/2} g, \quad \bar{Y}_i = Y_i + \sigma_X g, \quad i = 1, \dots, n,$$

where g is standard normal and independent of X and Y. Then

$$E\bar{X}_i^2 = E\bar{Y}_i^2 = EY_i^2 + \sigma_X^2$$

and

$$E(\bar{Y}_i - \bar{Y}_j)^2 = E(Y_i - Y_j)^2 \le E(X_i - X_j)^2 \le E(\bar{X}_i - \bar{X}_j)^2.$$