Probability tools, tricks, and miracles

Version: 6 June 2026

Dedicated in gratitude to John Hartigan—my friend, colleague, and mentor.

©David Pollard (http://www.stat.yale.edu/∼pollard/)

Contents

1 What and why **6**
1.1 How I chose material for this book 6
1.2 Naming and notation 8
1.3 The Chapters 10

2 It's just Calculus and convexity **15**
2.1 Introduction 15
2.2 Polynomial approximations 15
2.3 Convexity of the log MGF 18
2.4 Conjugate convex functions 20
2.5 Approximation of factorials 22
2.6 Distances between probability measures 28
2.7 Bounding sums by integrals 29
2.8 Problems 29
2.9 Notes 34

3 The moment generating function method **37**
3.1 Tail bounds from the moment generating function 37
*3.2 Behavior of the tail bound near the origin 39
3.3 Normal 41
*3.4 Global behavior of the normal hazard rate 46
3.5 Poisson 52
3.6 Gamma and chi-squared 54
3.7 Binomial 57
3.8 Sampling and the hypergeometric 60
3.9 Problems 63
3.10 Notes 68

4 Poisson-Binomial **74**
4.1 The extreme Binomial 74
4.2 The PGF of the Poisson-Binomial 76
4.3 Polynomials with real roots 76
4.4 Sampling and the Poisson-Binomial 79
4.5 Mode(s) of the Binomial 82
4.6 Mode(s) of the Poisson-Binomial 83
4.7 Binomial tails versus Poisson-Binomial tails 87
4.8 The median of the Poisson-Binomial 92
4.9 Problems 92
4.10 Notes 94

5 Orlicz spaces **98**
5.1 What is an ORLICZ space? 98
5.2 Facts about ORLICZ functions 100
5.3 ORLICZ functions that grow exponentially fast 104
5.4 Norms of maxima of finitely many variables 106
5.5 Conjugacy and the YOUNG inequality 108
*5.6 Comparison of $\mathcal{L}^p$ and $\mathcal{L}^\Psi$ 112
5.7 Problems 115
5.8 Notes 118

6 Multivariate normal **122**
6.1 Introduction 122
6.1.1 Subgaussian concentration 123
6.1.2 Comparison inequalities 125
6.2 The path method 126
6.3 Concentration of LIPSCHITZ functionals 128
6.3.1 The PISIER-MAUREY method 129
6.3.2 The smart path method 130
6.3.3 The stochastic calculus method 132
*6.4 The gaussian isoperimetric inequality 136
6.5 Tail bound comparisons 137
*6.6 An application of the GORDON inequality 142
6.7 A generalized FERNIQUE inequality 145
6.8 Problems 147
6.9 Notes 151

7 Subgaussian distributions **156**
7.1 Definition of subgaussian 156
7.2 Characterizations of subgaussianity 158
7.3 The HOEFFDING inequality for independent summands 160
7.4 From independence to martingales 162
7.5 A cautionary example 168
7.6 Problems 169
7.7 Notes 171

8 Between subgaussian and subexponential **173**
8.1 A BENNETT inequality for independent summands 173
8.2 Quadratics in independent subgaussians 175
8.3 A BENNETT inequality for martingales 178
*8.4 Concentration of random polynomials 181
8.5 Problems 189
8.6 Notes 194

9 Measurability woes **197**
9.1 The difficulty 197
9.2 Convergence in distribution, traditional style 199
9.3 Stochastic processes as random functions 201
9.4 Essential supremum and infimum 204
9.5 Upper/lower integrals and measurable cover functions 205
9.6 DOOB-separability: reduction to countable collections 207
9.7 Convergence in distribution without BOREL measurability 209
9.8 An oscillation condition for convergence in distribution 213
9.9 Problems 217
9.10 Notes 219

10 Chaining **222**
10.1 Maximal inequalities 222
10.2 What is chaining? 223
10.3 Chaining frameworks 226
10.4 Covering and packing numbers 229
10.5 A reminder about ORLICZ norms 231
10.6 Chaining with packing numbers 233
10.7 Oscillation and continuity of sample paths 237
*10.8 An example of a classical chaining argument 239
10.9 Problems 240
10.10 Notes 242

11 Majorizing measures and partitions **247**
11.1 Overview 247
11.2 Partitions, nested and not nested 249
11.3 Small facts about majorizing measures 251
*11.4 Majorizing measures as smoothers 253
11.5 Partitions with weights 256
*11.6 From majorizing measures to weighted frameworks 258
11.7 Talagrand and admissible partitions 262
*11.8 From majorizing measures to admissible partitions 265
11.9 An illuminating example 269
11.10 A comparison of chaining methods 272
11.11 Problems 273
11.12 Notes 276

12 Majorizing measures for gaussian processes **279**
12.1 Introduction 279
12.2 From functional to nested partitions with weights 281
*12.3 Uniformly continuous gaussian sample paths 285
*12.4 From functional to admissible partitions 286
12.5 Beyond gaussian 292
12.6 Problems 292

12.7 Notes 293

13 Symmetrization **295**
13.1 Symmetry? 295
13.2 Symmetrization for general stochastic processes 303
13.3 Empirical processes? 305
13.4 Maximal inequalities for empirical processes: norm-like 307
13.5 Maximal inequalities for empirical processes: tail-probabilty 310
13.6 Oscillation control for an empirical process 311
13.7 Problems 317
13.8 Notes 319

14 VC Sets **323**
14.1 An introductory example 323
14.2 VC-dimension, shatter-dimension 326
*14.3 Proof of the Shatter Theorem 332
*14.4 Subsets of a discrete hypercube as graphs 335
*14.5 Haussler's improvement of the packing bound 338
14.6 Problems 342
14.7 Notes 344

15 Fat shattering/surrounding **347**
15.1 Signed-sum processes 347
15.2 Surround dimension 348
15.3 Fattening up the concept of surround dimension 352
15.4 Lattice cubes surrounded by well separated vectors 353
15.5 Packing for subsets of $[0,1]^n$ 357
15.6 Weak ℓ^q spaces 358
15.7 Packing for subsets of an ℓ^q ball 360
*15.8 Chaining bounds via surround dimension 362
15.9 Problems 364
15.10 Notes 369

Index **374**

Chapter 1
What and why

1.1 How I chose material for this book

Once upon a time I decided to update the empirical process material in my books Pollard (1984) and Pollard (1990), with the help of some of the probability theory contained in UGMTP=Pollard (2001). I was aiming to create a brief document that would be useful to anyone who wanted to understand empirical process theory, including the improvements made in the last thirty or so years.

> **Remark.** Actually much of the material in UGMTP was originally intended to be part of an overly ambitious project called "Asymptopia", which covered much of the material that ended up in Pttm, the book you are now reading.

As the years went by I slowly realized that I had no hope of covering all the new material. I also realized that many of the ideas that I had first encountered in an empirical process context were actually useful more broadly. That made me rethink my goal: I would try for a book that could be helpful to anyone wanting to understand some of the most important ideas without needing to study all the intimidatingly large literature—essentially the position I found myself in nearly fifty years ago.

It seemed to me that such a book should include a lot of motivating material and heuristics. It should also contain some explanation for how complicated mathematical ideas can develop from much simpler precursors. Even though some experts might regard those precursors as mere special cases of powerful modern theories, I think that some knowledge of the history of an idea can help us to understand current theory. Such understanding becomes particularly useful when modern theory needs to be slightly modified to handle problems that don't quite satisfy the modern assumptions. Impatient readers can always just skip over those Sections of the book (or the numerous ***Remarks*** and the ***Notes*** at the end of each Chapter) that discuss material sometimes regarded as obsolete.

The first part of the book (Chapters 2 through 5) deals with some basic inequalities and well-known Calculus and convexity tricks. (I say well-known even though it sometimes took many years before I learned about them.) Here I am thinking of a trick as a clever idea for solving a particular problem, perhaps by repurposing an old idea from some apparently unrelated area. If tricks turn out to be more generally useful they sometimes become standard tools that anyone interested in an area should know.

Think of Calculus, for example, which started in the 17th century as an area of active research by the leading mathematical minds of the day then eventually became a subject taught even to secondary school students.

Some particularly clever tricks were initially so surprising and unexpected that they were regarded as minor miracles. Two examples come to mind: Charles Stein's method for normal approximation (touched on in §3.4); and Fernique's conjecture (Chapter 11) that existence of a majorizing measure was both sufficient and necessary for a centered gaussian process to have bounded sample paths, an idea vigorously developed and extended by Michel Talagrand. Maybe those miracles will one day be regarded as standard probability tools.

I strongly believe in the value of seeing how an idea can develop from an initial brilliant insight (a 'miracle') to a 'trick' (a method known to, and often developed by, a few experts) to a 'tool' (a method that is well understood and can be applied in settings, even those unsuspected by its inventors). Indeed, one of the great pleasures in my research life has been the progression from first wondering how anyone could have such a clever idea to the final realization that everything depended on only a few stunning insights. I decided it was worthwhile to share these pleasures with my intended audience. That means you will encounter many digressions exposing a lot of my stumblings and misconceptions along my path to a better understanding.

Chapters 6 through 8 contain results about a few general families of distributions: multivariate normal (a source of some amazing results in the past half century); the subgaussian (whose tails decrease like $\exp(-cx^2)$, for some $c > 0$, as $x \to \infty$); and the subexponential (whose tails start out like subgaussians but become more like $\exp(-cx)$ further out). A surprising amount of modern asymptotic theory relies heavily on facts about these families.

Chapter 9 describes some classical ways to overcome the sorts of measurabilty difficulties that greatly complicated the early empirical process literature.

Chapters 10 through 15 focus on results more obviously related to the empirical process theory of the late twentieth century.

While writing the book my views on what sort of theory is most valuable changed a lot. I went from an enthusiastic developer of limit theory to an equally enthusiastic student of the inequalities that usually underlie the limit theorems. To me it now seems strange to assume that the limits described by those theorems automatically provide reasonable approximations for finite samples, even when the theorems give no information about the point at which we get suitably close to the limit.

My decision to devote most attention to underlying ideas meant that I had to omit many beautiful theorems and skip over many applications, giving instead only a few pointers to relevant literature in the copious

Chapter Notes. I hope those Notes will help the interested reader to locate some of the missing ideas. I also hope that they will make obvious how much is borrowed from other authors. In particular, I have tried to dig into many of the early sources in order to understand how some fancy modern ideas evolved from simple beginnings.

1.2 Naming and notation

The most obvious surprise for some readers will be my use of the de Finetti notation, whereby sets are identified with their indicator functions and integrals are treated as linear functionals, à la Bourbaki. In particular, the expectation of a random variable becomes $\mathbb{P}X$ instead of the traditional $\mathbb{E}X$. See UGMTP Chapters 1 and 2 for my attempt to explain why this approach is far superior to using a notation inherited from a time before probabilities became identified as measures. I like to compare the older notation to an exercise in long multiplication carried out using Roman numerals.

Concepts named after humans

Many significant results are routinely attributed to the wrong author. For example, according to Dudley (2003):

(i) (page 381) "Jensen's inequality" was first proved for $\mathbb{R}^1$, if X is a simple function and f has a first derivative everywhere, by Hölder (1889)

(ii) (page 182) Cauchy (1821, p. 373) proved inequality 5.1.4 for finite sums, in other words where μ is counting measure on a finite set. Bunyakovsky (1859) proved it for Riemann integrals. H. A. Schwarz (1885, p. 351) rediscovered it much later. The inequality was commonly known as the "Schwarz inequality" or, more recently, as the "Cauchy-Schwarz" inequality.

(iii) (page 182) Hölder (1889) made clear that he was indebted to a paper of L. J. Rogers (1888)

And from Dudley (2014, p. 445):

(iv) W. H. Young (1912) proved his inequality (. . .) for smooth functions Φ. Birnbaum and Orlicz (1931) apparently began the theory of "Orlicz spaces," and Luxemburg (1955) defined the norms $\|\cdot\|_\Phi$; see Luxemburg and Zaanen (1956). Krasnosel'skii and Rutitskii (1961) wrote a book on the topic.

The list goes on. Why not replace gaussian by demoivrian? (cf. Stigler, 1986, pp.70–77 for the history) Why not refer to Kolmogorov's extension

theorem as the Daniell-Kolmogov extension theorem (Daniell, 1919, Kolmogorov, 1933)? Why not the Cauchy-Bunyakovsky-Schwarz inequality? Why not the "Blackwell-Stein-Sherman-Cartier-Meyer-Strassen" theorem? (Compare with historical remarks by Strassen, 1965, p423.) Eventually I resisted the urge to meddle with names and settled for JENSEN, CAUCHY-SCHWARZ, HÖLDER and gaussian.

Remark. If gaussian and euclidean instead of Gaussian and Euclidean shock you, you will be horrified by the International System of Units, which works with ampere, hertz, joule, kelvin, newton, and tesla.

Product space constructions

Often we deal with a large collection of random objects, all defined on a single problem space $(\Omega, \mathcal{F}, \mathbb{P})$. Usually the choice of Ω is far less important than the joint distributions of the objects, although sometimes it helps to specify it as a particular product space. Sometimes we are also interested in the conditional distributions of the objects, which can lead to subtle measurability difficulties if the objects take values in exotic sets like $\mathbb{R}^T$ with T uncountably infinite. However, the conditioning sometimes reduces to a FUBINI exercise. For these reasons I will often (especially in Chapter 13) replace a generic $(\Omega, \mathcal{F}, \mathbb{P})$ by particular product spaces.

If J is a finite set of size n then it can be identified with another finite set, $[\![n]\!] := \{k \in \mathbb{Z} : 1 \le k \le n\}$. More explicitly, there exists a bijection ψ between J and $[\![n]\!]$, in which case $J = \{\psi(1), \ldots, \psi(n)\}$ or $\{j_1, \ldots, j_n\}$ if $\psi(i) = j_i$. It seldom matters which bijection we use. However, there are some situations where the bijections just get in the way. For example, Kolmogorov's extension theorem for $\mathbb{R}^T$ involves a big collection of finite dimensional distribution $\{P_J : J \text{ a finite, non-empty subset of } T\}$. With bijections, P_J becomes $P_{j_1,\ldots,j_n}$ if $|J| = n$. But we could also write it as $P_{\sigma(j_1),\ldots,\sigma(j_n)}$ with σ a permutation map. That leads to a requirement that the relationship between $P_{j_1,\ldots,j_n}$ and $P_{\sigma(j_1),\ldots,\sigma(j_n)}$ be spelled out, becoming a totally superfluous assumption of theorem. To see the effect, compare the Kolmogorov (1933, §II.3) (or Billingsley, 1968, §36) statement of the theorem with Dudley (2003, Thm 12.1.2). For similar reasons I regard $\mathbb{R}^n$ as shorthand for $\mathbb{R}^{[\![n]\!]}$.

For an explanation of other strange symbols, consult the table of "Symbols and acronyms" at the start of the Index.

Mea culpa

Finally, let me explain my slightly erratic use of grammatical tense. When writing material discussing heuristics or proofs I usually slipped into teaching mode: "the next Theorem will show that ..." and "Section 3 explained the need for ...". However, I treated Problems and summaries of Chapters as historical objects: "Problem 5 shows why ...". References to UGMTP

(my 20001 Probability book) usually appear in present tense. I am sure you will find many examples where I violated these conventions.

1.3 The Chapters

Here is a more detailed description of the next 14 Chapters.

Chapter 2: It's just calculus and convexity

Introduces some basic mathematical tools. Amongst my favorites are the painless method for proving facts such as: the function $x \mapsto (e^x - 1 - x)/x^2$ is stricting increasing and convex over the whole real line. This approach replaces some gruesome brute force arguments that appeared in the literature. Equally clean is the "exponential tilting" argument for proving that the logarithm of a moment generating is a convex function, a result that turns out to be lurking behind many seemingly unrelated results in the literature.

The conjugate of a convex function is defined in §2.4. This concept turned out to be more important than I had initially expected.

The gamma function and $n!$ are discussed in §2.5.

Chapter 3: The moment generating function method

For every random variable X with MGF $M_X(\lambda) := \mathbb{P}e^{\lambda X}$ and every positive λ, the pointwise inequality $\{X \geq x\} \leq e^{\lambda(X-x)}$ integrates to give the tail bound

$$\mathbb{P}\{X \geq x\} \leq \inf_{\lambda>0} e^{-\lambda x} M_X(\lambda) = \exp\left(\inf_{\lambda>0} \log M_X(\lambda) - \lambda x\right),$$

a bound that fits neatly with the general idea of a convex conjugate of the convex function $\log M_X(\lambda)$.

Tail probability bounds are worked out by this method for several well known distributions: normal, POISSON, GAMMA, and BINOMIAL. The resulting inequalities represent the prototypes for a host of tail bounds from the literature derived by first establishing upper bounds for a MGF.

More surprisingly, convexity arguments show that the tail probabilities for POISSON-BINOMIAL and the HYPERGEOMETRIC distributions are smaller than the MGF bound for BINOMIAL tail probabilities.

Chapter 4: Poisson-Binomial

This Chapter was inspired by a handful of suggestive facts: the median of $\text{BIN}(n, p)$ lies in the range $np \pm 1$; tail bounds for independent sampling from $P_1, \ldots, P_n$ are often smaller than corresponding bounds for sampling from $P := (P_1 + \cdots + P_n)/n$ and, in particular, $\mathbb{P}\{Y \geq k\} \leq \mathbb{P}\{W \geq k\}$ for $k \in \mathbb{N}_0$ with $\mu + 1 \leq k \leq n$ if $Y \sim \text{PBIN}(p_1, \ldots, p_n)$ with $\sum_i p_i = \mu$ and

$W \sim \text{BIN}(n, \mu/n)$; the HYPERGEOMETRIC distribution is actually a special case of the POISSON-BINOMIAL. The method of proof uses an extension of a trick invented by Chebyshev.

Chapter 5: Orlicz spaces

The ORLICZ norms generalize the familar $\mathcal{L}^p$ norms. They provide a convenient way to control increments of stochastic processes, with the gaussian case the leading example. They offer a particularly simple way to control maxima of finite sets of random variables, which will be helpful for the chaining arguments in Chapters 10 through 12. Conjugacy arguments give an inequality, due to W. H. Young, that generalizes the HÖLDER inequality.

Chapter 6: Multivariate normal

Many beautiful results arose as generalizations of facts about the MVN. Using the ***path method***, this Chapter develops some theory, including the amazing fact that if $X \sim N(\mu, V)$ then the distribution of $f(X)$, for a LIPSCHITZ function f, concentrates near $\mathbb{P}f(X)$. The same method leads to several most useful inequalities (due to Slepian, Gordon, and Fernique) that reduce comparisons of tail bounds for $N(0, V_1)$ and $N(0, V_2)$ to simpler comparisons between V_1 and V_2.

Chapter 7: Subgaussian distributions

As shown in Chapter 3, if $X \sim N(0, 1)$ then $\mathbb{P}\{X \geq r\}$ decreases faster than $\exp(-r^2/2)$ as $r \to \infty$. This rate of decrease is shared by a large family of other distributions. The main results are two such exponential tail bounds due to Hoeffding, for sums of independent bounded random variables and for martingales. The so-called ***bounded difference inequalities***, which have proved most useful for treating combinatorial problems, are derived from the martingale HOEFFDING inequality.

Chapter 8: Between subgaussian and subexponential

The BENNETT inequalities (both for independent summands and for martingales) sharpen subgaussian inequalities by making better use of variances, or upper bounds on variances. These tail bounds are slightly sharper than the SUBGAMMA tail bounds from §3.6 and the generalization of the HANSON-WRIGHT equality in §8.2, which also makes use of a very clever decoupling trick that really surprised me. The final Section of the Chapter tries (with only partial success) to apply the martingale BENNETT inequality to aid in the proof of the KIM-VU concentration inequality for random polynomials.

Chapter 9: Measurability woes

Stochastic processes with uncountably infinite index sets create measure-theoretic difficulties. Several strategies—such as DOOB-separabilty and the introduction of upper and lower integrals—have been developed to work around those difficulties. These techniques are illustrated by means of necessary and sufficient conditions for "convergence in distribution" for sequences of stochastic processes that don't have "distributions" in the traditional sense.

Chapter 10: Chaining

The method known as "chaining" seems to have a reputation as an abstruse and complicated idea. This Chapter attempts to persuade you that it really is quite simple once some of the tricks are explained. Covering- and packing-numbers make their first appearance. The highpoint of the Chapter is the proof (in less than a page) of a quite robust sufficient condition for a stochastic process to have uniformly continuous sample paths.

Chapter 11: Chaining via partitions

The Chapter starts by describing majorizing measures (MM's), whose existence Xavier Fernique conjectured to be necessary for a gaussian process to have bounded sample paths. It then explains some of the history of MM's from an initial role as a smoothers of sample paths, through its role in attaching weights to the edges of the tree in a chaining framework, to its equivalence with Michel Talagrand's concept of admissible partitions. The Chapter ends with a variation on a classical example that nicely illustrates the subtle differences between what can be achieved by covering/packing methods and what can be achieved by MM/admissible partition methods.

Chapter 12: Majorizing measures for gaussian processes

Talagrand proved Fernique's conjecture about gaussian processes and the existence of majorizing measures. His initial proofs were very deep and difficult. Then he came up with a simpler approach that showed bounded gaussian sample paths implied existence of weighted chaining frameworks that were equivalent to existence of majorizing measures. This Chapter describes some of my efforts to understand just a tiny fraction of Talagrand's ideas.

Chapter 13: Symmetrization

The word "symmetrization" usually refers to a method for making some object more symmetrical. This Chapter tries first to argue the symme-

try is not the real reason for the idea's usefulness. Then it picks out some highlights in the development of symmetrization as a general tool. (The earliest example cited involves R. A. Fisher's work on experimental design in the 1930's.) Then it describes how symmetrization became an important tool in empirical process theory before evolving into several methods that produce maximal inequalities for sums of independent stochastic processes.

Chapter 14: VC sets

This Chapter introduces the concept of a vc-class of sets, named in honor of Vapnik & Červonenkis. It starts with the concrete example of closed half-spaces, in $\mathbb{R}^2$ where elementary methods suffice. Dudley's bound on packing numbers for vc-classes is proved. The idea is then generalized to a concept of shatter-dimension $\text{SDIM}(\mathcal{X})$ for subsets $\mathcal{X}$ of $\{0,1\}^n$. The Shatter Theorem, which provides a polynomial bound (of degree equal to $\text{SDIM}(\mathcal{X})$) on the cardinality of $\mathcal{X}$, is proved by means of a downshifting argument. The set $\mathcal{X}$ is turned into a graph by defining edges as pairs from $\mathcal{X}$ whose HAMMING distance equals 1. Using that graph some beautiful refinements of the packing bound, due to Haussler, are proved—again using downshifting argument.

Chapter 15: Fat shattering

The Chapter focuses on signed-sum processes indexed by bounded subsets of $\mathbb{R}^n$: for each $x = (x_1, \ldots, x_n)$ in $\mathcal{X}$ there is a random variable $\pm x_1 + \cdots \pm x_n$, with the $\pm$'s chosen independently according to the uniform distribution on $\{-1, +1\}$. Combinatorial properties for $\mathcal{X}$, analogous to SDIM idea from Chapter 14, are shown to provide useful bounds on covering/packing numbers. Most of the Chapter focuses on a beatiful idea (due to Roman Vershynin and friends) that reduces the study of packing numbers to an exercise in lattice-point counting, which is assisted by combinatorial arguments. The probabilistic method plays a leading role.

***

Originally I had wanted to include three more Chapters: on Talagrand's inductive method for establishing concentration inequalities; the entropy method for establishing concentration inequalities; and the bracketing method for establishing maximal inequalities. Unfortunately they are still in an embarrasingly disorganized state.

References

Billingsley, P. (1968). *Probability Measures* (2nd ed.). New York: Wiley.

Daniell, P. J. (1919). Integrals in an infinite number of dimensions. *Annals of Mathematics (series 2) 20*, 281–288.

Dudley, R. M. (2003). *Real Analysis and Probability* (2nd ed.), Volume 74 of *Cambridge studies in advanced mathematics*. Cambridge University Press.

Dudley, R. M. (2014). *Uniform Central Limit Theorems* (2nd ed.), Volume 142 of *Cambridge studies in advanced mathematics*. Cambridge University Press. (First edition, 1999).

Kolmogorov, A. N. (1933a). *Foundations of the Theory of Probability*. Springer/Chelsea. (The Chelsea 1950 translation of 'Grundbegriffe der Wahrscheinlichkeitsrechnung', published in 1933 by Springer.).

Kolmogorov, A. N. (1933b). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer-Verlag. Second English Edition, *Foundations of Probability* 1950, published by Chelsea, New York.

Luxemburg, W. A. J. (1955). *Banach function spaces*. Ph. D. thesis, TU Delft. http://resolver.tudelft.nl/uuid:252868f8-d63f-42e4-934c-20956b86783f.

Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*, Volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Hayward, CA: Institute of Mathematical Statistics.

Pollard, D. (2001). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, Massachusetts: Harvard University Press.

Strassen, V. (1965). The existence of probability measures with given marginals. *Annals of Mathematical Statistics 36*, 423–439.

Chapter 2
It's just Calculus and convexity

SECTION 2.1 *explains why you can safely just skim through this Chapter on a first pass.*

SECTION 2.2 *Describes a simple trick using* TAYLOR *expansion to bound functions of the form* $2(f(x) - f(0) - x\dot{f}(0))/x^2$. *Two important special cases are described.*

SECTION 2.3 *explains why the logarithm of a moment generating function (MGF) is always convex.*

SECTION 2.4 *describes briefly a procedure known known in the literature by several different names, explaining how it relates to a minimization method (Chapter 3) for deriving probability tail bounds. The magical* ψ_{benn} *makes its first appearance.*

SECTION 2.5 *waxes historical about the* STIRLING *formula and binomial coefficients. It describes a few tricks that pop up often in the literature on high-dimensional statistics.*

SECTION 2.6 *is an advertisement for the convex function* ψ_{benn}. *It has been a hobby of mine to find hidden appeals to the helpful properties of that function buried inside complex arguments.*

SECTION 2.7 *explains why many important quantities are written as integrals.*

2.1 Introduction

You could safely skim quickly through this Chapter, just to get an idea of how much fancy modern theory depends on only a little bit of calculus with magical inequalities coming via convexity arguments, sometimes in disguise. Later in the book I'll tell you when something from here is needed, at which point you might want to read some parts of this Chapter more carefully. I prefer having the elementary—but very useful—arguments collected together in one place rather than having them buried inside complicated proofs scattered all over the place.

2.2 Polynomial approximations

Where would Mathematics be without approximation of functions by polynomials?

For my purposes, the integral form of the TAYLOR theorem is the most useful. Suppose g is a real-valued function defined at least on an interval J of the real line that contains both 0 and a point x. If g is twice continuously differentiable then

$$g(x) - g(0) = \int_0^1 \frac{\partial g(xt)}{\partial t}\, dt = x \int_0^1 \dot{g}(xt)\, dt \qquad \text{for } x \in J$$

and

$$\begin{aligned} R_1(x,g) &= g(x) - g(0) - x\dot{g}(0) \\ &= x \int_0^1 \dot{g}(xt) - \dot{g}(0)\, dt \\ &= x^2 \int_0^1 \int_0^1 \{0 \le s \le t \le 1\} \ddot{g}(xs)\, ds\, dt \\ &= \tfrac{1}{2} x^2 \Delta(x,g) \qquad \text{where } \Delta(x,g) := \int_0^1 2(1-s)\ddot{g}(xs)\, ds \end{aligned} \tag{<1>}$$

And so on, if higher-order derivatives exist (see Problem [1]).

Remark. The assumptions on g could be relaxed slightly by using some fancy facts about absolute continuity.

The behaviors of $R_1(x,g)$ and $\Delta(x,g)$ are controlled by $\ddot{g}$. For example, if $\ddot{g}$ is non-negative then so are $\Delta(\cdot)$ and $R_1(\cdot)$, and if $\ddot{g}$ is continuous then so are $\Delta(\cdot)$ and $R_1(\cdot)$. More interestingly, if $\ddot{g}$ is a nondecreasing function then so is $\Delta(\cdot)$, and if $\ddot{g}$ is a convex function then so is $\Delta(\cdot)$.

We can interpret the factor $2(1-s)$ in the definition of Δ as the density (with respect to LEBESGUE measure) of a probability measure on the interval $(0,1)$, with expected value

$$\int_0^1 s2(1-s)\, ds = 1/3.$$

If $\ddot{g}$ is convex then the JENSEN inequality gives

$$\Delta(x,g) \ge \ddot{g}\left(x \int_0^1 2s(1-s) \right) = \ddot{g}(x/3). \tag{<2>}$$

This inequality accounts for a few factors of $1/3$ that pop up in well known inequalities.

My main interest in these calculus tricks originally came from studying two special examples, which will be important throughout this book. (Hence the funny fonts.) The two functions are

$$\mathbb{f}(x) = e^x - 1 - x \qquad \text{for } x \in \mathbb{R} \tag{<3>}$$

$$\mathbb{h}(x) = (1+x)\log(1+x) - x \qquad \text{for } x \ge -1. \tag{<4>}$$

The convention $0 \log 0 = 0$ gives $\mathbb{h}(-1) = 1$ and, for reasons that will be explained in the next Section, it sometimes helps to extend the domain of $\mathbb{h}$ to the whole real line by defining $\mathbb{h}(x) = +\infty$ when $x < -1$,

The first and second derivatives,

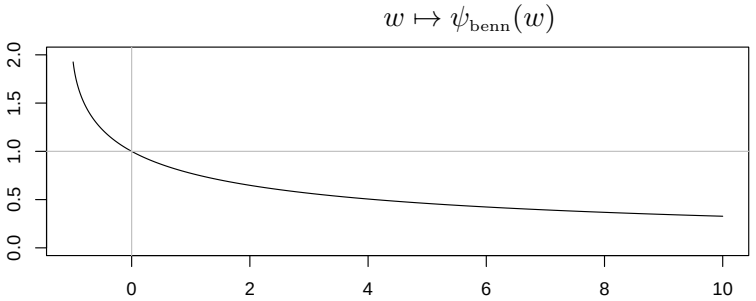
$$\dot{\mathbb{f}}(x) = e^x - 1 \quad \text{AND} \quad \ddot{\mathbb{f}}(x) = e^x$$
$$\dot{\mathbb{h}}(x) = \log(1+x) \quad \text{AND} \quad \ddot{\mathbb{h}}(x) = (1+x)^{-1},$$

show that both functions are non-negative and convex, with each achieving its minimum value of 0 at $x = 0$. From <1> we have representations

<5> $$\mathbb{f}(x) = \tfrac{1}{2}x^2 \mathbb{\Delta}(x) \qquad \text{where } \mathbb{\Delta}(x) := \int_0^1 2(1-s)e^{xs} ds$$

<6> $$\mathbb{h}(x) = \tfrac{1}{2}x^2 \psi_{\text{benn}}(x) \qquad \text{where } \psi_{\text{benn}}(x) := \int_0^1 \frac{2(1-s)}{(1+xs)} ds \text{ for } x \geq -1.$$

The function $\mathbb{\Delta}$ is continuous, convex, and strictly increasing. The function ψ_{benn} is continuous, convex, and strictly decreasing on $[-1, \infty)$.



Moreover, by the JENSEN inequality,

<7> $$\psi_{\text{benn}}(x) \geq \ddot{\mathbb{h}}(x/3) = (1 + x/3)^{-1}.$$

Compare with the fact that $\psi_{\text{benn}}(x)$ actually decreases like $2x^{-1}\log(x)$ as x tends to infinity.

As you will see in Chapter 3, the function ψ_{benn} lies hidden inside a handful of very useful tail bounds. Inequality <7> will also have some surprising consequences: in Section 2.6, in the context of the PINSKER inequality; and in Chapter 8, to explain the relationship between the BENNETT and BERNSTEIN exponential inequalities.

Remarks. The subscript 'benn' stands for Bennett, who established a now famous exponential tail bound (Bennett, 1962, equation 8b), although he did not write it using the ψ_{benn} function. This bound, and its martingale extension, are described in Chapter 8.

The monotonicity and convexity of $\mathbb{\Delta}$ are easy to check for $x \geq 0$ from the series expansion

$$\mathbb{\Delta}(x) = 2 \sum\nolimits_{k \geq 2} x^{k-2}/k! \,.$$

However, even the monotonicity is less obvious when x is negative. Brute force calculation of the derivative,

$$d\triangle(x)/dx = \frac{(x-2)e^x + x + 2}{x^3/2},$$

didn't help me much when I first encountered $\triangle$ in the Chow and Teicher (1978, Section 11.1) book. A later edition of that book did use the shorter proof, almost as above. See also the proof by Freedman (1975, Lemma 3.1).

2.3 Convexity of the log MGF

For a random variable X, the function $M_X(\lambda) = \mathbb{P}\exp(\lambda X)$ for $\lambda \in \mathbb{R}$ is called the moment generating function (MGF) for X. It is always finite at the origin because $M_X(0) = \mathbb{P}1 = 1$.The expectation is well defined and nonnegative for all real λ, although it might take the value $+\infty$. For example, if X has a CAUCHY distribution then $M_X(\lambda) = +\infty$ for all nonzero λ; and if X has a standard exponential distribution then

$$M_X(\lambda) = \int_0^\infty e^{\lambda x - x}\, dx = \begin{cases} (1-\lambda)^{-1} & \text{for } \lambda < 1 \\ \infty & \text{otherwise} \end{cases}.$$

The function M_X inherits convexity from the exponential function. More precisely, the set

$$\text{EPI}(M_X) = \{(\lambda, t) \in \mathbb{R}^2 : t \geq M_X(\lambda)\}$$

is convex and closed as a subset of $\mathbb{R}^2$. The function M_X is a closed, proper ($M_X(\lambda) > -\infty$), convex function, in the sense described by Rockafellar (1970, Section 7). Moreover, the restriction ofthe function M_X to the convex set $\mathcal{D}_X := \{\lambda \in \mathbb{R} : M_X(\lambda) < \infty\}$ is continuous and convex in the usual sense. If $\text{int}(\mathcal{D}_X)$, the interior of $\mathcal{D}_X$, is non-empty then M_X is infinitely differentiable with

<8> $$M_X^{(k)}(\lambda) = \mathbb{P}\left(X^k e^{\lambda X}\right) \qquad \text{for } k = 1, 2, \ldots \text{ if } \lambda \in \text{int}(\mathcal{D}_X).$$

See Problem [3] for the (easy) proofs.

Remark. The differentiability can fail at the boundary of $\mathcal{D}_X$. For example, if X has the distribution with density $x^{-2}\{x \geq 1\}$ with respect to Lebesgue measure then $M_X(\lambda) < \infty$ iff $\lambda \leq 0$. The left-hand derivative at $\lambda = 0$ is infinite: the ratio $(M_X(0) - M_X(-h))/h$ tends to $+\infty$ as h decreases to zero.

The function $L_X := \log M_X$ inherits from M_X the lower semi-continuity and infinite differentiability on the interior of $\mathcal{D}_X$. More surprisingly, L_X is also a proper, closed convex function. The convexity follows from the

HÖLDER inequality: for nonnegative random variables Y_1 and Y_2: for positive constants α_i with $\alpha_1 + \alpha_2 = 1$,

$$\mathbb{P}\left(Y^{\alpha_1} Y_2^{\alpha_2}\right) \le (\mathbb{P}Y)^{\alpha_1} \left(\mathbb{P}Y_2\right)^{\alpha_2}.$$

Taking $Y_i = \exp(\lambda_i X)$ we get

$$M_X(\alpha_1\lambda_1 + \alpha_2\lambda_2) \le M_X(\lambda_1)^{\alpha_1} M_X(\lambda_2)^{\alpha_2}.$$

By taking logs we get

$$L_X(\alpha_1\lambda_1 + \alpha_2\lambda_2) \le \alpha_1 L_X(\lambda_1) + \alpha_2 L_X(\lambda_2).$$

Remark. If you worry about things like $\log(6\infty) = \log(6) + \log(\infty)$ you could instead note that the restriction of L_X to $\mathcal{D}_X$ is convex in the usual sense, with $L_X(\lambda) = \infty$ outside $\mathcal{D}_X$, which is equivalent to convexity of EPI(L_X) as a subset of $\mathbb{R}^2$.

Formula <8> leads to interesting expressions for the first two derivatives of L_X on $\operatorname{int}(\mathcal{D}_X)$:

$$\dot{L}(\lambda) = \frac{\dot{M}_X(\lambda)}{M_X(\lambda)} = \mathbb{P}X \frac{e^{\lambda X}}{M_X(\lambda)}$$
$$\ddot{L}(\lambda) = \frac{\ddot{M}_X(\lambda)}{M_X(\lambda)} - \left(\frac{\dot{M}_X(\lambda)}{M_X(\lambda)}\right)^2 = \mathbb{P}X^2 \frac{e^{\lambda X}}{M_X(\lambda)} - \left(\mathbb{P}X \frac{e^{\lambda X}}{M_X(\lambda)}\right)^2.$$

If we define a probability measure $\mathbb{P}_\lambda$ by its density, $d\mathbb{P}_\lambda/d\mathbb{P} = e^{\lambda X}/M_X(\lambda)$, then

<9> $$\dot{L}(\lambda) = \mathbb{P}_\lambda X \qquad \text{AND} \qquad \ddot{L}(\lambda) = \mathbb{P}_\lambda \left(X - \mathbb{P}_\lambda X\right)^2 = \operatorname{var}_\lambda(X) \ge 0,$$

with strict inequality except in the degenerate case where P concentrates at a single point. Thus L is strictly convex except in the degenerate case. In particular, if $0 \in \operatorname{int}(\mathcal{D}_X)$ then $\dot{L}(0) = \mathbb{P}X$, so that Taylor expansion gives

<10> $$L(\lambda) = \lambda \mathbb{P}X + \tfrac{1}{2}\lambda^2 \operatorname{var}_{\lambda^*}(X) \qquad \text{with } \lambda^* \text{ lying between 0 and } \lambda.$$

Remark. The method for constructing $\mathbb{P}_\lambda$ is often called ***exponential tilting***, a technique introduced by Esscher (1932). It plays a large role in the theory of large deviations, started by Cramér (1938) (who cited Esscher).

I was quite surprised when I first learned how often <9> lurks behind useful facts. Problem [7] presents an example, which I learned from one of my colleagues only last year.

Remark. Bounds on the $\text{var}_{\lambda^*}(X)$, for some hard to calculate λ^* that depends on λ, lead to bounds on L and tail bounds for $X - \mathbb{P}X$. The proof of the HOEFFDING inequality in Chapter 7 uses the representation <10> to control the MGF of a bounded random variable.

2.4 Conjugate convex functions

Many probability bounds involve a minimization problem of the form

$$\text{for each } w \in \mathbb{R} \text{ minimize over } \lambda \in J \text{ the function } f(\lambda) - \lambda w,$$

where J is a nonempty convex subset of the real line, and $f : J \to \mathbb{R}$. Of course a convex J is some sort of interval, such as $\mathbb{R}$ itself or $(-\infty, 3]$ or $[-7, 3)$, and so on. The role of J can be ignored if we extend f to a function taking values in $\mathbb{R} \cup \{\infty\}$ by defining $f(\lambda) = +\infty$ for $\lambda \notin J$. While we are at it, we should allow for the possibility that there might be no minimizing value by defining

<11> $$g(w) = \inf_{\lambda \in \mathbb{R}} \Big(f(\lambda) - \lambda w \Big) \qquad \text{for } w \in \mathbb{R}.$$

Equivalently,

$$-g(w) = \sup_{\lambda \in \mathbb{R}} \Big(\lambda w - f(\lambda) \Big) \qquad \text{for } w \in \mathbb{R}.$$

The function $f^* = -g$ is called the ***convex conjugate*** of f. It is also called the FENCHEL-LEGENDRE transformation. The name 'conjugate' was applied by Fenchel (1949). See also Kiselman (2019), which includes a translation of a letter from Fenchel that explains the connection with earlier work of Legendre and others.

Remark. Note that for any function $\ell : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$,

$$\ell(w) + f(\lambda) \geq w\lambda \qquad \text{for all } w, \lambda \in \mathbb{R}$$

if and only if $\ell(w) \geq \mathbb{f}^*(w)$ for all w.

The calculation of the infimum in <11> is easier when f itself is convex. It is then usually easy to crank out g by brute force calculus by finding where the derivative is positive or negative or zero. In Chapter 3 you will encounter several such cases where $f(\lambda) = \log \mathbb{P}e^{\lambda X}$ for a random variable X.

It is also convenient that g, as an infimum of linear functions of w, always turns out to be concave and f^* is always convex. The convexity can also be seen from consideration of the ***epigraph***:

$$\begin{aligned} \text{epi}(f^*) &= \{(t, w) \in \mathbb{R}^2 : t \geq f^*(w)\} \\ &= \cap_{\lambda \in \mathbb{R}} \{(t, w) \in \mathbb{R}^2 : t - \lambda w \geq -f(\lambda)\}, \end{aligned}$$

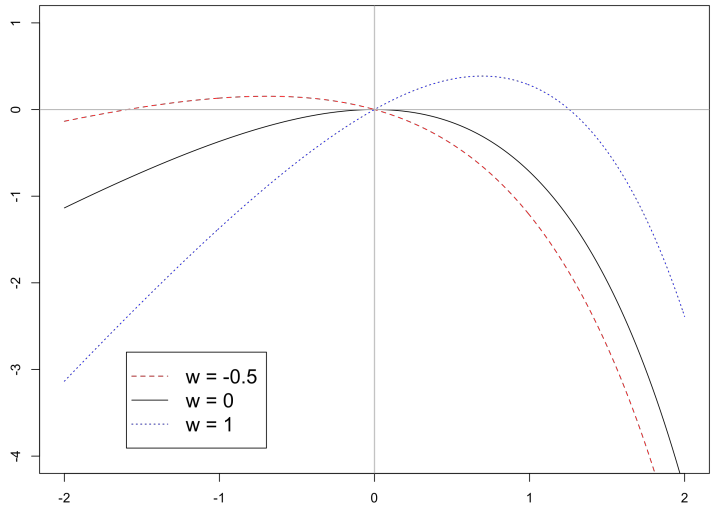
an intersection of closed half-spaces. The epigraph is closed as a (nonempty) subset of $\mathbb{R}^2$, not just convex. That makes f^* a closed, proper convex function in the sense used by Rockafellar (1970, Section 7).

Remark. For a proper, convex function with $J = \{\lambda : f(\lambda) < \infty\}$ the closedness property just involves the values of f at the endpoints of J, if any. This subtlety need not concern us.

The general theory is interesting because $f^{**} = (f^*)^*$ happens to be the largest closed, convex function for which $f(w) \geq f^{**}(w)$ for all w (see Rockafellar (1970, pages 103-4). In particular, if f itself is closed and convex then $f^{**} = f$. Such a property plays a central role role in the study of convex duality (Rockafellar, 1974).

Remark. Unfortunately these wonderful duality facts about f^{**} don't seem to help much for the material in this book. I should perhaps look more carefully at the pages of Boucheron et al. (2000) pointed at by index entries of the form "duality ...".

<12> **Example.** Consider the convex function $\mathbb{f}(\lambda) = e^\lambda - 1 - \lambda$, for $\lambda \in \mathbb{R}$, from Section 2.2. It is convex on the real line with derivative 0 at the origin. For each $w \in \mathbb{R}$ the concave function $G_w(\lambda) := \lambda w - \mathbb{f}(\lambda) = \lambda(w+1) - e^\lambda + 1$ has λ-derivative $G'_w(\lambda) = (1+w) - e^\lambda$.



If $w > -1$ the maximum of G_w is achieved when $G'_w(\lambda) = 0$, that is, at $\lambda = \log(1+w)$. For $w = -1$ we have $G_{-1}(\lambda) = 1 - e^\lambda$, which approaches its supremum of 1 as $\lambda \to \infty$. For $w < -1$, we have $G_w(\lambda) \uparrow +\infty$ as $\lambda \downarrow -\infty$. In summary,

<13>
$$\mathbb{f}^*(w) = \begin{cases} (1+w)\log(1+w) - w & \text{if } w > -1\text{; achieved at } \lambda = \log(1+w) \\ 1 & \text{if } w = -1\text{; approached as } \lambda \to \infty \\ +\infty & \text{if } w < -1\text{; approached as } \lambda \to \infty \end{cases}$$

That is, $\mathbb{f}^* = \mathbb{h}$, as a map from $\mathbb{R}$ to $(-\infty, +\infty]$.

The general duality theory asserts that

$$\mathbb{f}(\lambda) = \mathbb{h}^*(\lambda) = \sup_{w\in\mathbb{R}} w\lambda - \mathbb{h}(w) = \sup_{w\ge -1} w\lambda - \mathbb{h}(w).$$

Let me check.

For $\lambda \in \mathbb{R}$ the supremum defining $\mathbb{h}^*(\lambda)$ need only run over $w \ge -1$ because $\lambda w - \mathbb{h}(w) = -\infty$ for $w < -1$.

The derivative of the concave function $w \mapsto \lambda w - \mathbb{h}(w)$ is $\lambda - \log(1+w)$, which tends to $+\infty$ as $w \to -1$ and $-\infty$ as $w \to \infty$. For the maximizing w we have $\lambda = \log(1+w)$, which gives

$$\mathbb{h}^*(\lambda) = \lambda(e^\lambda - 1) - \lambda e^\lambda + (e^\lambda - 1) = e^\lambda - 1 - \lambda.$$

□ That is, $\mathbb{h}^* = \mathbb{f}$, as asserted by the general theory.

2.5 Approximation of factorials

Abraham de Moivre (1756, page 244), before deriving his famous approximation to the Binomial distribution, made a comment about a constant B (which we now know is equal to $\sqrt{2\pi}$), which he had approximated by adding up some terms of a slowly converging infinite series. He went on:

> ... seeing at the same time that what I had done answered my purpose tolerably well, I desisted from proceeding farther till my worthy and learned Friend Mr. *James Stirling*, who had applied himself after me to that inquiry, found that the Quantity B did denote the Square-root of the Circumference of a Circle whose Radius is Unity ...

Yes, it was the Stirling of the formula that is now often cited in the exquisitely sharp form: for $n \in \mathbb{N}$,

<14> $$n! = Cn^{n+1/2}e^{-n+r_n} \qquad \text{where } b_n := \frac{1}{12n+1} < r_n < \frac{1}{12n} =: a_n,$$

where C is a universal constant. In fact, a simple CLT argument (see below) shows that C equals $\sqrt{2\pi}$.

Remark. In my opinion, the main value of the inequality $b_n < r_n$ comes from the reassurance it gives that the upper bound is astoundingly sharp: the difference $a_n - b_n$ between the upper and lower bounds for $\log n!$ is of order $O(n^{-2})$. If instead we used the bounds $0 < r_n < a_n$ the corresponding difference would be of order $O(n^{-1})$, still very respectable.

Proof (of <14>). The formula can be rewritten as

<15> $$b_n < r_n = d_n - \log C < a_n \qquad \text{or} \qquad d_n - a_n < \log C < d_n - b_n$$
$$\text{where } d_n := \log\left(n!e^n/n^{n+1/2}\right) = \log(n!) - (n+\tfrac{1}{2})\log n + n.$$

This assertion will follow from the inequalities

$$<16> \qquad \frac{\theta}{(n+\theta)(n+13\theta)} = b_n - b_{n+1} < d_n - d_{n+1} < a_n - a_{n+1} = \frac{\theta}{n(n+1)},$$

where $\theta := 1/12$. Indeed, these inequalities show that $\{d_n - a_n : n \in \mathbb{N}\}$ is an increasing sequence and $\{d_n - b_n : n \in \mathbb{N}\}$ is a decreasing sequence with $d_n - a_n < d_n - b_n$, which ensures existence of a finite constant c for which $d_n - a_n \uparrow c$ and $d_n - b_n \downarrow c$. Inequality <15> will hold with $C = e^c$.

The proof of <16> is elementary and quite elegant. First write everything in terms of $T_n = t_n^{-1} = 2n+1$. By direct calculation,

$$\begin{aligned}
d_n - d_{n+1} &= (n+1/2)\log\left(\frac{n+1}{n}\right) - 1 = \tfrac{1}{2}T_n \log\left(\frac{T_n+1}{T_n-1}\right) - 1 \\
&= \frac{1}{2t_n}\log\left(\frac{1+t_n}{1-t_n}\right) - 1 \\
&= \frac{t_n^2}{3} + \frac{t_n^4}{5} + \frac{t_n^6}{6} + \dots \qquad \text{by TAYLOR expansion} \\
&< \frac{t_n^2}{3} + \frac{t_n^4}{3} + \frac{t_n^6}{3} + \dots = \frac{t_n^2}{3}\left(1-t_n^2\right)^{-1} = \frac{1}{3}\left(T_n^2-1\right)^{-1} \\
&= \frac{\theta}{n(n+1)} = a_n - a_{n+1}.
\end{aligned}$$

Similarly,

$$d_n - d_{n+1} > \frac{t_n^2}{3} = \frac{1}{3T_n^2} = \frac{\theta}{n^2+n+1/4} > b_n - b_{n+1},$$

the final inequality coming from

$$\begin{aligned}
(n^2+14n\theta+13\theta^2) - (n^2+12n\theta+36\theta^2) &= 2n\theta - 23\theta^2 \\
&\ge \theta(2-23/12) = \theta^2.
\end{aligned}$$

It remains only to show that C equals $\sqrt{2\pi}$. Here we can make use of a charming connection between C and $p_n(i) := \mathbb{P}\{X_n = i\} = e^{-n}n^i/i!$ for $X_n \sim \text{POISSON}(n)$. The maximum of $p_n(i)$ is achieved at $i = n$, with

$$\sqrt{n}p_n(n) = e^{-n}n^{n+1/2}/n! = e^{-r_n}/C \to 1/C \qquad \text{as } n \to \infty.$$

Thus $C = \sqrt{2\pi}$ iff $\sqrt{n}p_n(n) \to 1/\sqrt{2\pi}$. By the central limit theorem we have $Z_n := n^{-1/2}(X_n - n) \rightsquigarrow \gamma := N(0,1)$, which suggests that

$$p_n(n) = \mathbb{P}\{n < X_n \le n+1\} = \mathbb{P}\{0 < Z_n \le n^{-1/2}\} \approx \gamma(0, n^{-1/2}] \approx \frac{1}{\sqrt{2\pi n}}.$$

The usual CLT does not quite justify all this handwaving but the basic idea works if we replace the $n^{-1/2}$ by a small positive number δ. With

$k := k_n(\delta) := \lfloor \delta\sqrt{n} \rfloor$ we then have

$$\mathbb{P}\{0 < Z_n \le \delta\} = \sum_{i=1}^{k} p_n(n+i) \to \gamma(0,\delta] \qquad \text{as } n \to \infty, \text{ by CLT,}$$
$$\text{and } \gamma(0,\delta]/\delta \to (2\pi)^{-1/2} \qquad \text{as } \delta \to 0.$$

For $1 \le i \le k$ we have

$$\frac{p_n(n+i)}{p_n(n)} = \frac{n^i n!}{(n+i)!} > \prod_{i=1}^{k} (1+i/n)^{-1} \ge (1+k/n)^{-k} \qquad \text{for } k = \lfloor \delta\sqrt{n} \rfloor.$$

Thus

$$1 \ge \sum_{i=1}^{k} \frac{p_n(n+i)}{k p_n(n)} \ge \exp\left(-k \log(1+k/n)\right) \to e^{-\delta^2} \qquad \text{as } n \to \infty,$$

which rearranges to

$$(\delta + o(1))\sqrt{n} p_n(n) \ge \gamma(0,\delta] + o(1) \ge (\delta + o(1))\sqrt{n} p_n(n) \left(e^{-\delta^2} + o(1)\right),$$

with the $o(1)$ converging to zero as $n \to \infty$. In the limit we have

$$e^{\delta^2}\gamma(0,\delta]/\delta \ge \limsup_n \sqrt{n} p_n(n) \ge \liminf_n \sqrt{n} p_n(n) \ge \gamma(0,\delta]/\delta$$

□ for each $\delta > 0$. It follows that $\sqrt{n} p_n(n) \to 1/\sqrt{2\pi}$, as hoped for.

The factorial function can be extended to a continuous domain by means of the ***gamma function*** defined by

<17> $$\Gamma(x) := \int_0^\infty r^{x-1} e^{-r}\, dr \qquad \text{for } x > 0.$$

It has the following properties.

(i) $\log \Gamma$ is convex (by either of the methods from Section 2.3).

(ii) $\Gamma(x+1) = x\Gamma(x)$ for $x > 0$ (integration-by-parts).

(iii) $\Gamma(n+1) = n!$ for $n \in \mathbb{N}$ (from (ii) and $\Gamma(1) = 1$).

(iv) $\inf_{x>0} \Gamma(x) \approx 0.89$, achieved near 1.46 (according to **R**).

(v) $\Gamma(x) = \sqrt{2\pi} \exp\left((x-1/2)\log x - x + R(x)\right)$ with $0 < R(x) < (12x)^{-1}$ for all $x > 0$. See Problem [6].

Approximation (v) will be most helpful in Chapter 15.

<18> **Example.** A lot of asymptotic statistical theory involves the binomial coefficients:

$$\binom{n}{i} = \frac{n!}{i!(n-i)!} \qquad \text{for } i = 0, 1, \ldots, n.$$

To avoid an overabundance of parentheses, for $0 < i < n$ define $\eta = i/n = 1 - \gamma$ and $\epsilon_{i,n} = r_i + r_{n-i} - r_n$. Then

$$\begin{aligned}\binom{n}{i} &= \frac{n!}{i!(n-i)!} = \frac{n^{n+1/2}\exp(\epsilon_{i,n})}{\sqrt{2\pi}(n\eta)^{i+1/2}(n\gamma)^{n-i+1/2}} \qquad \text{by <14>} \\ &= \frac{1}{\sqrt{2\pi n\eta\gamma}} \exp\left(-n\Big[\eta\log\frac{1}{\eta} + \gamma\log\frac{1}{\gamma}\Big] + \epsilon_{i,n}\right).\end{aligned}$$

The $\epsilon_{i,n}$ is usually ignored if both i and $n - i$ are large. The term within square brackets in the exponent is called the ***entropy*** of the $\text{BER}(\eta)$ distribution, a quantity much loved by information theorists (Cover and Thomas, 2012, Example 2.1.1). When multiplied by n it gives the entropy of the $\text{BIN}(n, \eta)$ distribution.

For many purposes a somewhat cruder upper bound for $\binom{n}{i}$ suffices. For example, if $1 \le i \le d \le n$,

$$\binom{n}{i} = \frac{n(n-1)\ldots(n-i+1)}{i!} < \Big(\frac{n}{d}\Big)^i \frac{d^i}{i!} < \Big(\frac{n}{d}\Big)^d \frac{d^i}{i!},$$

implying

$$\sum\nolimits_{0\le i\le d}\binom{n}{i} \le 1 + \sum\nolimits_{1\le i\le d}\Big(\frac{n}{d}\Big)^d \frac{d^i}{i!} < \Big(\frac{en}{d}\Big)^d.$$

This inequality plays a central role in the ultra-famous theory initiated by Vapnik & Červonenkis. See §14.2. □

<19> **Example.** If $X \sim \text{BIN}(n, p)$ and $q = 1 - p$ then, again writing $\eta = 1 - \gamma$ for i/n, we have

$$\begin{aligned}\mathbb{P}\{X = i\} &= \binom{n}{i} p^i q^{n-i} = \binom{n}{i}\exp\left(n\eta\log p + n\gamma\log q\right) \\ &= \frac{1}{\sqrt{2\pi n\eta\gamma}}\exp\left(-n\Big[\eta\log\frac{p}{\eta} + \gamma\log\frac{q}{\gamma}\Big] + \epsilon_{i,n}\right).\end{aligned}$$

The quantity

$$D_{KL}(\eta \,||\, p) := \eta\log\left(\eta/p\right) + \gamma\log\left(\gamma/q\right), \qquad \text{for } q = 1-p \text{ and } \gamma = 1-\eta,$$

is called the KULLBACK-LEIBLER distance (or KL-divergence) between the $\text{BER}(\eta)$ and $\text{BER}(p)$ distributions. Multiplication by n gives the KL-divergence between the $\text{BIN}(n, \eta)$ and the $\text{BIN}(n, p)$ distributions.

The KL-divergence is also closely related to the function ψ_{benn}, defined (in <6>) by $\frac{1}{2}t^2\psi_{\text{benn}}(t) = \hbar(t) = (1+t)\log(1+t) - t$. If $i = np + x$ then $\eta = p(1 + x/(np))$ and $\gamma = q(1 - x/(nq))$ then $D(\text{BIN}(n,\eta) \,||\, \text{BIN}(n,p))$ equals

$$\begin{aligned} & np\left(1+\frac{x}{np}\right)\log\left(1+\frac{x}{np}\right) + nq\left(1-\frac{x}{nq}\right)\log\left(1-\frac{x}{nq}\right) \\ &= np \times \hbar\left(\frac{x}{np}\right) + nq \times \hbar\left(\frac{-x}{nq}\right) + np\left(\frac{x}{np}\right) + nq\left(\frac{-x}{nq}\right) \\ &= \frac{x^2 g(x,n,p)}{2npq} \qquad \text{where } g(x,n,p) := q\psi_{\text{benn}}\left(\frac{x}{np}\right) + p\psi_{\text{benn}}\left(\frac{-x}{nq}\right). \end{aligned}$$

In summary, for $0 < i = np + x < n$ and $X \sim \text{BIN}(n,p)$ we have

<20> $$\mathbb{P}\{X = i\} = \frac{1}{\sqrt{2\pi npq}} \exp\left(-\mathfrak{Q}(x)\right) \times \Lambda(x,n,p)$$

where

$$\begin{aligned} \mathfrak{Q}(x) &:= np \times \hbar\left(\frac{x}{np}\right) + nq \times \hbar\left(\frac{-x}{nq}\right) = \frac{x^2 g(x,n,p)}{2npq} \\ g(x,n,p) &= q\psi_{\text{benn}}\left(\frac{x}{np}\right) + p\psi_{\text{benn}}\left(\frac{-x}{nq}\right) \\ \log\Lambda(x,n,p) &:= r_{np+x} + r_{nq-x} - r_n - \tfrac{1}{2}\log\left[\left(1 + x/(np)\right)\left(1 - x/(nq)\right)\right]. \end{aligned}$$

The constraint $0 < i < n$ is only needed to ensure that there are no divisions by 0 in the definition of Λ.

If $\min(np, nq)$ is large and $x/\min(np, nq)$ is small then $\Lambda(x,n,p) \approx 1$ and $g(x,n,p) \approx 1$, so that

$$\mathbb{P}\{X = i\} \approx \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{x^2}{2npq}\right).$$

This result should remind you of the normal approximation to the $\text{BIN}(n,p)$, but not just as an approximation for tail probabilities. A more careful handling of the errors of approximation would lead to a so-called ***local limit theorem*** for the Binomial distribution (Petrov, 1975, Chapter 7).

If you don't like great gobs of algebra you could safely skip the rest of this Example. I am just about to make a small point about a refinement of the normal approximation to the Binomial, with another advertisement for the methods from Section 2.2.

The appearance of the convex function ψ_{benn} (from <6>) in the definition of $g(x,n,p)$ tell us something surprising about departures from normality. By convexity,

<21> $$g(x,n,p) \geq \psi_{\text{benn}}\left(q\frac{x}{np} + p\frac{-x}{nq}\right) = \psi_{\text{benn}}\left(\frac{(q-p)x}{npq}\right).$$

If $p > 1/2$ then $g(x, n, p) > 1$ for $x > 0$. We get something smaller than the normal exponent $-x^2/(2npq)$ in the right tail beyond the mean np. This effect can be attributed to the skewness, $\mathbb{P}(X - np)^3/\text{var}^{3/2}(X) = npq(q-p)/(npq)^{3/2}$, which is negative if $p > 1/2$. To a first approximation, the $\text{Bin}(n, p)$ is symmetric around its mean np. To a finer approximation, if $p > 1/2$ then the probabilities are slightly smaller in the left tail than in the right tail.

By the principle of no free lunches, if skewness gives us a nicer right tail then it will probably give us an uglier left tail. That is in fact the case but it cannot be deduced directly from inequality <21>. Instead we need to identify some higher order terms in the series expansion of $\mathcal{Q}(x)$. From

$$\dot{\mathbb{h}}(x) = \log(1+x) \qquad \ddot{\mathbb{h}}(x) = 1/(1+x)$$
$$\mathbb{h}^{(3)}(x) = -1/(1+x)^2 \qquad \mathbb{h}^{(4)}(x) = 2/(1+x)^3$$

and Problem [1] we get

$$\mathbb{h}(x) = \frac{x^2}{2} - \frac{x^3}{6} + \frac{x^4}{3}\int_0^1 \frac{(1-s)^3}{(1+sx)^3}\,ds \qquad \text{for } x > -1$$

and

$$\begin{aligned}
\mathcal{Q}(x) &= nq\mathbb{h}(x/np) + np\mathbb{h}(-x/nq) \\
&= np\left(\frac{x^2}{2(np)^2} - \frac{x^3}{6(np)^3} + \frac{x^4}{3(np)^4}\int_0^1 \frac{(1-s)^3}{(1+sx/np)^3}ds\right) \\
&\quad + nq\left(\frac{x^2}{2(nq)^2} + \frac{x^3}{6(nq)^3} + \frac{x^4}{3(nq)^4}\int_0^1 \frac{(1-s)^3}{(1-sx/nq)^3}ds\right) \\
&= \frac{x^2}{2npq} + \frac{x^3(q-p)}{6n^2p^2q^2} \\
&\qquad + \frac{x^4}{3}\int_0^1 (1-s)^3\left[(np+sx)^{-3} + (nq-sx)^{-3}\right]\,ds.
\end{aligned}$$

The pattern is clearer if we consider terms $i = np + x$ with $x = y\sqrt{npq}$, that is, if we scale by the standard error $\sigma = \sqrt{npq}$ for X:

$$\begin{aligned}
\mathcal{Q}(y\sigma) &= \frac{y^2}{2} + \frac{y^3(q-p)}{6\sqrt{npq}} \\
&\qquad + \frac{y^4p^2q^2}{3n}\int_0^1 (1-s)^3\left[(p+sy\sigma/n)^{-3} + (q-sy\sigma/n)^{-3}\right]\,ds.
\end{aligned}$$

The y^2 term contributes the normal approximation; the y^3 term contributes a skewness perturbation of order $n^{-1/2}$; and the y^4 term scoops up the leftovers, which are of order n^{-1}. If you like this kind of thing then you will love Petrov (1975, Chapter 6). □

2.6 Distances between probability measures

Here is another example of ψ_{benn} popping up in unexpected places.

Modern statistical theory makes extensive use of various measures of 'distance' between probability measures P and Q defined on the same sigma-field. To keep the following discussion simple I'll assume that both probabilities are defined on the same finite set $\mathcal{X}$ with $Q\{x\} > 0$ for each $x \in \mathcal{X}$. Define

$$\rho(x) = \frac{P\{x\}}{Q\{x\}} \qquad \text{for } x \in \mathcal{X}.$$

For each real-valued function g on $\mathcal{X}$ think of its Q expectation as the linear functional

$$Qg = Qg(x) := \sum\nolimits_{x\in\mathcal{X}} Q\{x\}g(x).$$

Remark. The quantity $\rho(x)$ is sometimes called the ***likelihood ratio***. If you desire greater generality, feel free to extend the arguments to general sigma-fields with P absolutely continuous with respect to Q with density (RADON-NIKODYM derivative) $\rho = dP/dQ$. You will also need to worry about integrability. I would even forgive you for adopting the hideous notation $\mathbb{E}_Q g$ for Qg, but the even more hideous $\mathbb{E}_{x\sim Q}g(x)$ is beyond the pale, in my humble opinion.

For each convex function f on $\mathbb{R}^+$ for which $f(1) = 0$, the f-divergence is defined as $D_f(P \mid\mid Q) := Qf(\rho(x))$. Three particularly popular special cases correspond to the functions $f(x) = |x-1|$ (the L^1 distance), and $f(x) = x\log x$ (the KULLBACK-LEIBLER divergence) and $f(x) = (x-1)^2$ (the χ^2-divergence):

$$\begin{aligned}
\|P - Q\|_1 &:= Q|\rho(x) - 1| = \sum\nolimits_x |P\{x\} - Q\{x\}| \\
D_{KL}(P \mid\mid Q) &:= Q\rho(x)\log(\rho(x)) = \sum\nolimits_x P\{x\}\log\left(P\{x\}/Q\{x\}\right) \\
D_{\chi^2}(P \mid\mid Q) &:= Q\left(\rho(x)-1\right)^2 = \sum\nolimits_x \frac{(P\{x\} - Q\{x\})^2}{Q\{x\}}
\end{aligned}$$

Of these, only the L^1 distance is a metric.

The relationship between the three divergences becomes clearer when they are written as integrals involving $\delta(x) = \rho(x) - 1$:

$$\begin{aligned}
\|P - Q\|_1 &:= Q|\delta| \\
D_{KL}(P \mid\mid Q) &:= Q\mathbb{h}(\delta) = \tfrac{1}{2}Q\left(\delta^2\psi_{\text{benn}}(\delta)\right) \\
D_{\chi^2}(P \mid\mid Q) &:= Q\delta^2
\end{aligned}$$

As before, $\mathbb{h}(x) = (1+x)\log(1+x) - x$. The extra $-x$ does not change the integral because $Q\delta = Q\rho - 1 = 1 - 1$.

The divergences are related by the inequalities

$$D_{\chi^2}(P \,||\, Q) \geq D_{KL}(P \,||\, Q) \geq \tfrac{1}{2} \|P - Q\|_1^2 .$$

I don't know if anyone has claimed parentage for the first inequality, which follows from the fact that $\psi_{\text{benn}}(\delta) \leq 2$. The second inequality, which is often called the **pinsker** inequality, follows from CAUCHY-SCHWARZ:

$$\begin{aligned}
\tfrac{1}{2}(Q|\delta|)^2 &= \tfrac{1}{2} Q \left(\frac{|\delta|}{\sqrt{1+\delta/3}} \sqrt{1+\delta/3} \right)^2 \\
&\leq \tfrac{1}{2} Q \left(\frac{\delta^2}{1+\delta/3} \right) Q(1+\delta/3) \qquad \text{by CAUCHY-SCHWARZ} \\
&\leq \tfrac{1}{2} Q \left(\delta^2 \psi_{\text{benn}}(\delta) \right) \qquad \text{by <7> and } Q\delta = 0.
\end{aligned}$$

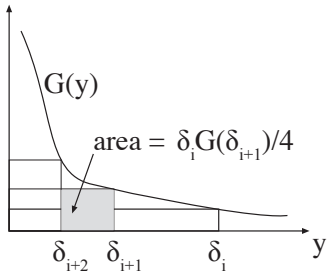
I learned of this neat argument from Kemperman (1969).

2.7 Bounding sums by integrals

Many approximation schemes (particularly those that will be discussed in Chapter 10) bound useful quantities by sums of the form $\sum_{i=0}^{k} \delta_i G(\delta_{i+1})$, with $G(\cdot)$ a decreasing function on $\mathbb{R}^+$. If the δ_i's decrease geometrically, say $\delta_i = \delta_0/2^i$, then the sum can be bounded above by an integral,

$$\sum_{i=0}^{k} \delta_i G(\delta_{i+1}) \leq 4 \int_{\delta_{k+2}}^{\delta_1} G(r)\, dr \leq 4 \int_0^{\delta_1} G(r)\, dr.$$

If G is integrable the replacement of the lower terminal by 0 costs little.



In the early empirical process literature, some authors chose the δ_i's to make the $G(\delta_i)$'s increase geometrically, for similar reasons. (Use horizontal slices.)

In delicate situations, sometimes the $\{\delta_i\}$ sequence needs to chosen in a more cunning way. See Section Dudley (1973, p 74), for example.

2.8 Problems

[1] Suppose g is a real-valued function, defined at least on an interval J of the

real line that contains both 0 and a point x, that is $(k+1)$-times continuously differentiable. For x in the interior of J show that

$$\begin{aligned} R_k(x,g) &:= g(x) - \left[g(0) + x\dot{g}(0) + \cdots + x^k g^{(k)}(0)/(k)!\right] \\ &= x^{k+1} \int \cdots \int \{0 \le t_{k+1} \le t_k \le \cdots \le t_1 \le 1\} g^{(k)}(t_{k+1}x)\, dt_{k+1} \ldots dt_1 \\ &= x^{k+1} \int_0^1 \frac{(1-s)^k}{k!} g^{(k+1)}(sx)\, ds. \end{aligned}$$

[2] By splitting into real and imaginary parts, show that

$$e^{ix} - \left(1 + ix + (ix)^2/2! + \cdots + (ix)^k/k!\right) = x^{k+1} \int_0^1 \frac{(1-s)^k}{k!} i^{k+1} e^{isx}\, ds.$$

for all real x. Deduce that

$$\left|e^{ix} - \left(1 + ix + (ix)^2/2! + \cdots + (ix)^k/k!\right)\right| \le \frac{|x|^{k+1}}{(k+1)!},$$

an inequality that is useful for arguments involving FOURIER transforms.

[3] Suppose $M_X(\lambda) = \mathbb{P}e^{\lambda X}$ for $\lambda \in \mathbb{R}$ and $\mathcal{D}_X = \{\lambda \in \mathbb{R} : M_X(\lambda) < \infty\}$.

(i) Show that $M_X(\lambda)$ is lower semi-continuous at each point of $\mathbb{R}$. Equivalently, EPI(M_X) is a closed subset of $\mathbb{R}^2$. Hint: If $\lambda_n \to \lambda \in \mathbb{R}$, the FATOU lemma tells you something about $\liminf_n \mathbb{P}e^{\lambda_n X}$. Argue similarly if $(t_n, \lambda_n) \in$ EPI(M_X) and $(t_n, \lambda_n) \to (t, \lambda) \in \mathbb{R}^2$. (As Rockafellar (1970, page 52) argued, lower semi-continuity of M_X is equivalent to EPI(M_X) being closed as a subset of $\mathbb{R}^2$.)

(ii) Show that the restriction of M_X to $\mathcal{D}_X$ is continuous as a function on $\mathcal{D}_X$. Hint: Suppose $[\lambda, \gamma] \subset \mathcal{D}_X$ and $\lambda_n = \alpha_n\lambda + (1-\alpha_n)\gamma$. Convexity of M_X gives

$$M_X(\lambda_n) \le \alpha_n M_X(\lambda) + (1-\alpha_n) M_X(\gamma). \qquad \text{for } 0 \le t \le 1.$$

Take the lim sup as α_n tends to 0 to deduce that $\limsup M_X(\lambda_n) \le M_X(\lambda)$, which together with lower semi-continuity proves continuity from the right at λ. Argue similarly if $\alpha_n \to 1$ for continuity from the left at γ.

(iii) Suppose λ is an interior point of $\mathcal{D}_X$. Choose $\delta > 0$ so that $[\lambda - 2\delta, \lambda + 2\delta] \subset \mathcal{D}_X$. For $|h| < \delta$, establish the domination bound

$$\begin{aligned} |e^{(\lambda+h)X} - e^{\lambda X}|/|h| &\le \left|\int_0^1 X e^{\lambda X} e^{shX} ds\right| \\ &\le e^{\lambda X}\left(e^{\delta X} + e^{-\delta X}\right)^2/\delta, \end{aligned}$$

which is integrable. Deduce via a dominated convergence argument that $M'(\lambda) = \mathbb{P}(Xe^{\lambda X})$.

(iv) Argue similarly to show that M_X is infinitely differentiable on the interior of $\mathcal{D}_X$, with $M_X^{(k)}(\lambda) = \mathbb{P}\left(X^k e^{\lambda X}\right)$.

[4] For $k \in \mathbb{N}$ prove that $e^k \geq k^k/k! \geq 1$. Deduce that $(k/e)^k \leq k! \leq k^k$ and $k/e \leq (k!)^{1/k} \leq k$. Compare with the sharper bounds given by the STIRLING formula: $(k!)^{1/k} = \rho_k k/e$, where $\rho_k = \left(\sqrt{2\pi k}e^{r_k}\right)^{1/k}$ is a decreasing function of k that converges to 1 as $k \to \infty$.

k	1	2	3	4	5	6	7	8	9	10
ρ_k	2.718	1.922	1.646	1.504	1.416	1.356	1.313	1.279	1.253	1.231

[5] (The BOHR-MOLLERUP theorem, apparently first stated in a 1922 textbook by H. Bohr and J. Mollerup. The following derivation is based on the exposition by Artin, 1964.) Suppose f is a strictly positive function on $(0, \infty)$ for which:

(a) $f(x+1) = xf(x)$ for each $x > 0$.

(b) $g := \log f$ is convex.

(c) $f(1) = 1$.

Prove that $f(x) = \Gamma(x)$ for all $x > 0$ by arguing as follows.

(i) Note that $f(x+n) = f(x)\rho_n(x)$ and $\Gamma(x+n) = \Gamma(x)\rho_n(x)$ for $x > 0$ and $n \in \mathbb{N}$, where $\rho_n(x) := x(x+1)\ldots(x+n-1)$. In particular, $f(1+n) = n! = \Gamma(1+n)$ for each $n \in \mathbb{N}_0$.

(ii) Convexity of g tells us that the slope $H(a,b) := (g(b) - g(a))/(b-a)$, for $0 < a < b$, is an increasing function of each argument. In particular, for an integer $n \geq 2$ and $0 < x \leq 1$ we have

$$g(n) - g(n-1) = H(n-1,n) \leq H(n,n+x) = \Big(g(n+x) - g(n)\Big)/x \leq H(n,n+1) = g(n+1) - g(n).$$

Thus

$$(n-1)^x = \left(\frac{f(n)}{f(n-1)}\right)^x \leq \frac{f(n+x)}{f(n)} = \frac{f(x)\rho_n(x)}{(n-1)!} \leq n^x,$$

implying

$$(n-1)^x(n-1)!/\rho_n(x) \leq f(x) \leq n^x(n-1)!/\rho_n(x) \qquad \text{for } n \geq 2.$$

The first inequality also holds with n replaced by $n+1$. Thus

$$\frac{n^x n!}{\rho_{n+1}(x)} \leq f(x) \leq \frac{n^x(n-1)!}{\rho_n(x)} = \frac{n^x n!}{\rho_{n+1}(x)}\frac{x+n}{n} \qquad \text{for } n \geq 2.$$

(iii) The same inequalities also hold with $f(x)$ replaced by $\Gamma(x)$, because the analogs of (a), (b), and (c) also hold for Γ. Taking the ratios we then get

$$n/(x+n) \le f(x)/\Gamma(x) \le (x+n)/n,$$

which in the limit as $n \to \infty$ gives $f(x) = \Gamma(x)$ for $0 < x \le 1$. The factorizations in (i) extend the equality to all $x > 0$.

[6] The STIRLING formula,

$$n! = \sqrt{2\pi}\exp\left((n+1/2)\log n - n + r_n\right) \qquad \text{with } (12n=1)^{-1} < r_n < (12n)^{-1},$$

has a continuous analog: for $x > 0$,

$$\Gamma(x) = \sqrt{2\pi}\exp\left((x-1/2)\log x - x + R(x)\right) \qquad \text{with } 0 < R(x) < (12x)^{-1}.$$

Verify the following assertions to establish this approximation.

(i) Define $h(x) := (x-1/2)\log x - x$ and $f(x) := e^{h(x)+R(x)}$ for an $R(x)$ to be determined. We have $f(x+1) = xf(x)$ iff

$$\begin{aligned} R(x) - R(x+1) &= h(x+1) - h(x) - \log x \\ &= g(x) := (x+1/2)\log\left(1+x^{-1}\right) - 1 \\ &= (2y)^{-1}\log\left(\frac{1+y}{1-y}\right) - 1 \qquad \text{where } y := 1/(2x+1) \\ &= y^2/3 + y^4/5 + \dots \qquad \text{by TAYLOR for } \log(1\pm y) \\ &\le \frac{1}{3}(y^2 + y^4 + \dots) = \frac{y^2}{3(1-y^2)} \\ &= (12x(x+1))^{-1} = (12x)^{-1} - (12x+12)^{-1}. \end{aligned}$$

(ii) Define

$$R(x) := \sum\nolimits_{i\in\mathbb{N}_0} g(x+i) = \frac{1}{3(2x+1)^2} + \frac{1}{5(2x+1)^4} + \dots,$$

a convergent sum of positive convex functions. Thus $0 < R(x) \le (12x)^{-1}$ and R is convex.

(iii) The function h is convex because $\ddot{h}(x) = x^1 + (2x^2)^{-1} > 0$. Thus $\log f(x) := h(x) + R(x)$ is convex. By the BOHR-MOLLERUP theorem (an idea due to Artin, 1964), $Cf(x) = \Gamma(x)$ for all $x > 0$, for some constant C. It remains only to show that $C = \sqrt{2\pi}$. To that end, note that STIRLING gives

$$Cf(n+1) = \Gamma(n+1) = n! = \sqrt{2\pi}\exp\left((n-1/2)\log n - n + \log n + r_n\right)$$

which rearranges to

$$\log(C/\sqrt{2\pi}) = h(n) + \log n + r_n - h(n+1) - R(n+1) = r_n - g(n) - R(n+1).$$

That is, the constant $\log(C/\sqrt{2\pi})$ equals $r_n - R(n) = o(1)$ as $n \to \infty$. It follows that $C = \sqrt{2\pi}$.

[7] (See Wu and Zhou (2021, §1.1) for an example of how the results from this Problem can be used.) Suppose $X \sim P$ and $Z \sim \gamma_1 := N(0,1)$ are independent. Define $U = \alpha X + \beta Z$ for positive constants α and β. Define $h(u) = \mathbb{P}(X \mid U = u)$. Follow these steps to show that

(a) h is an increasing function

(b) If P is symmetric around 0 then h is an odd function.

(i) Explain why the result is trivial if either of α or β is zero. Deduce that, without loss of generality, we may suppose $\alpha = \beta = 1$.

(ii) For suitably integrable f, show that

$$\mathbb{P}f(X, X+Z) = P^x \int_{\mathbb{R}} f(x,u) m(x,u) \phi(u)\, du$$

where $m(x,u) = \phi(x-u)/\phi(u) = \exp(xu - x^2/2)$. Deduce that: the joint distribution $\mathbb{M}$ for (X,U) has density m with respect to $P \otimes \gamma_1$; the marginal distribution (with respect to γ_1) for U is $q(u) = P^x m(x,u)$; and the conditional density (with respect to P) for X given $U = u$ is $m(x \mid u) = m(x,u)/q(u)$.

(iii) Assuming the usual conditions for differentiating inside P integrals, deduce that $u \mapsto \log q(u)$ is convex, so that its derivative, $q'(u)/q(u)$, is an increasing function.

(iv) Use the fact that $\partial m(x,u) = x m(x,u)$ to show that

$$h(u) = P^x x m(x,u)/q(u) = P^x \frac{\partial m(x,u)}{\partial u}/q(u) = q'(u)/q(u),$$

an increasing function.

(v) Note that $m(-x,-u) = m(x,u)$ for all x and u. For symmetric P deduce that

$$q(-u) = P^x m(x,-u) = P^x m(-x,-u) = q(u)$$

and

$$q(-u)h(-u) = q(u) P^x x m(x,-u) = q(u) P^x (-x) m(-x,-u) = -q(u)h(u),$$

which implies that h is an odd function.

2.9 Notes

See Stigler (1986, pages 70-76) for an illuminating discussion of the interaction between de Moivre and Stirling. I learned the STIRLING formula, as in <14>, from Feller (1968, page 52). He cited Robbins (1955), who commented that the only novelty in his proof was in the derivation of the inequality $b_n < r_n$; he acknowledged that the standard parts of his argument were taken from Darmois (1928, pages 315-317). Robbins also commented that the "editor has pointed out that the inequalities ... permit the following brief proof", which was a "modification of that attributed to Cesàro by A. Fisher, *Mathematical theory of probabilities*, New York, 1936, pp. 93-95." (The 1955 Mathematical Monthly listed F. A. Ficken as the editor of the section containing Robbins's note.) The Cesàro book was published in 1884. Even though I cannot read Italian, I feel the proof on page 270 is very similar to the one presented by Feller. As they say, there is nothing new under the sun.

Feller (1968, footnote to page 53) cited Feller (1967) for the identification of the limit of $\{e^{d_n}\}$ as $\sqrt{2\pi}$. In another footnote Feller (1968, Section VII.2) recovered the $\sqrt{2\pi}$ from a local limit theorem for the $\text{BIN}(n, 1/2)$. The argument in Section 2.5 for determining C is essentially that of de Moivre (1756), applied to the POISSON instead of the BINOMIAL. See also Pitman (1993, Section 2.3) for a beautiful exposition of de Moivre's method. At the end of Section 3.6 of the same textbook, Pitman noted that de Moivre's method also works with the HYPERGEOMETRIC distribution. It can be also be applied to other distributions P on the integers for which the ratios $P\{i+1\}/P\{i\}$ have a tractable form. Similar ideas also play a role in the the study of log-concave discrete distributions and even in the approximations developed by Chen and Stein (Barbour, Holst, and Janson, 1992, Section 9.2).

Initially I included Section 2.4 mostly because I feared I had missed some subtle point made by Boucheron, Lugosi, and Massart (2013, Sections 2.2), who seemed to be suggesting that the FENCHEL-LEGENDRE transform plays a role far more important than just as a name for a minimization procedure. To me it didn't seem very helpful just to know that a particular minimization of a convex function was an example of a technique named after famous mathematicians. Eventually my opinion changed when it became clearer to me that FENCHEL-LEGENDRE conjugates were lurking in the background for many useful ideas. See, for example, the discussion of the YOUNG inequality in §5.5. The general theory also extends easily to functions on $\mathbb{R}^d$, where it plays an important role in the study of large deviations for sums of independent random vectors—see Dembo and Zeitouni (1998, Chapter 2).

There is currently an archive of Fenchel's papers at

`http://web.math.ku.dk/arkivet/fenchel/wfenpapr.htm`

There is also a most informative note ("Werner Fenchel, a pioneer in convexity theory and a migrant scientist" by Christer Oscar Kiselman) about Fenchel and duality at

`http://www.cb.uu.se/~kiselman/bibliography.html`

The note is listed on the web site as having appeared in *Normat. Nordisk matematisk tidskrift* 61, No. 2-4, 133-152, but I have only seen the online version.

References

Artin, E. (1964). *The Gamma Function*. Holt, Rinehart and Winston. (Translated from the 1931 German original.).

Barbour, A. D., L. Holst, and S. Janson (1992). *Poisson Approximation*. Oxford University Press.

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association 57*, 33–45.

Boucheron, S., G. Lugosi, and P. Massart (2000). A sharp concentration inequality with applications. *Random Structures and Algorithms 16*, 277–292.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Chow, Y. S. and H. Teicher (1978). *Probability Theory: Independence, Interchangeability, Martingales*. New York: Springer.

Cover, T. M. and J. A. Thomas (2012). *Elements of Information Theory*. John Wiley & Sons.

Cramér, H. (1938). Sur un nouveau théorème-limite de la théorie des probabilités. In *Colloque consacré à la théorie des probabilités, Actualités scientifiques et industrielles 736*, pp. 2–23. Hermann & Cie, Paris. Available at arXiv:1802.05988v3 in the form of a TeXed version of the original paper in French paired with an English translation by Hugo Touchette. Reprinted in: H. Cramér, Collected Works, A. Martin-Löf (Ed.), Vol. II, Springer, Berlin, 1994, p. 895-913.

Darmois, G. (1928). *Statistique Mathématique*. Doin & cie.

de Moivre, A. (1756). *The Doctrine of Chances: or, A Method of Calculating the Probabilities of Events in Play*. New York: Chelsea. Third edition (fuller, clearer, and more correct than the former), reprinted in 1967. First edition 1718.

Dembo, A. and O. Zeitouni (1998). *Large Deviations Techniques and Applications* (2 ed.). Springer-Verlag.

Dudley, R. M. (1973). Sample functions of the Gaussian process. *Annals of Probability 1*, 66–103.

Esscher, F. (1932). On the probability function in the collective theory of risk. *Scandinavian Actuarial Journal 1932*(3), 175–195.

Feller, W. (1967). A direct proof of Stirling's formula. *The American Mathematical Monthly 74*(10), 1223–1225.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (third ed.), Volume 1. New York: Wiley.

Fenchel, W. (1949). On conjugate convex functions. *Canadian Journal of Mathematics 1*, 73–77.

Freedman, D. A. (1975). On tail probabilities for martingales. *Annals of Probability 3*(1), 100–118.

Kemperman, J. H. B. (1969). On the optimum rate of transmitting information. In *Probability and Information Theory*. Springer-Verlag. Lecture Notes in Mathematics, 89, pages 126–169.

Kiselman, C. O. (2019). Werner Fenchel, a pioneer in convexity theory and a migrant scientist. *Normat: Nordisk matematisk tidskrift 61*(2–4), 133–152.

Petrov, V. V. (1975). *Sums of Independent Random Variables*. Springer-Verlag. Enlish translation from 1972 Russian edition.

Pitman, J. (1993). *Probability*. Springer.

Robbins, H. (1955). A remark on Stirling's formula. *The American Mathematical Monthly 62*(1), 26–29.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton, New Jersey: Princeton Univ. Press.

Rockafellar, R. T. (1974). *Conjugate Duality and Optimization*, Volume 16 of *Regional Conference Series in Applied Mathematics*. Siam.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, Massachusetts: Harvard University Press.

Wu, Y. and H. H. Zhou (2021). Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations. *Mathematical Statistics and Learning 4*(3), 143–220.

Chapter 3
The moment generating function method

SECTION 3.1 *introduces the* MGF *(moment generating function) method for bounding tail probabilities.*

*SECTION 3.2 *describes the behavior of log(*MGF*) near the origin.*

SECTION 3.3 *illustrates the* MGF *method for the simplest case, the normal distribution. The normal is the prototype for the subgaussian family of distributions, which will be discussed in Chapter 7.*

*SECTION 3.4 *derives some global properties for the standard normal hazard function $\phi(x)/\bar{\Phi}(x)$. The value of the global approach is illustrated by a brief discussion of Charles Stein's miraculous method for developing normal approximations.*

SECTION 3.5 *derives tail bounds for the* POISSON *distribution. The omnipresent convex function $\psi_{\text{benn}}()$ puts in an appearance. The Poisson is the prototype for the* BENNETT *inequalities.*

SECTION 3.6 *establishes tail bounds for the gamma distribution, the prototype for distributions of type* SUBGAMMA*. The Examples for* SUBGAMMA *include a* BERNSTEIN *inequality.*

SECTION 3.7 *establishes very good bounds for the tails of the* BINOMIAL *distribution. These bounds also work for* POISSON-BINOMIAL *distributions and other sums of independent, $[0,1]$-valued random variables.*

SECTION 3.8 *shows that the tail bounds derived by the* MGF *method for the hypergeometric distribution are smaller than the bounds for the corresponding* BINOMIAL*.*

3.1 Tail bounds from the moment generating function

Much modern statistical theory relies on a handful of probabilistic inequalities, often in the form of bounds on tail probabilites or concentration inequalities. This Chapter introduces one of the main methods for establishing such bounds. By way of illustration, the method is applied to derive bounds for several well studied cases, which provide the prototypes for a handful of very useful tail bounds.

The method uses the MGF, $M_X(\lambda) := \mathbb{P}e^{\lambda X} = e^{L_X(\lambda)}$, to get upper bounds for $\mathbb{P}\{X \geq x\}$. Remember from §2.3 that L_X is infinitely differentiable and convex on the set $\{\lambda \in \mathbb{R} : M_X(\lambda) < \infty\}$.

From the fact that the exp() function is everywhere nonnegative and $\exp(\lambda(X-x)) \geq 1$ when $X \geq x$ and $\lambda \geq 0$ we have

<1> $$\mathbb{P}\{X \geq x\} \leq \inf_{\lambda \geq 0} \mathbb{P} e^{\lambda(X-x)} = \inf_{\lambda \geq 0} e^{-\lambda x} M_X(\lambda) = \exp\left(\inf_{\lambda \geq 0}\Big(L_X(\lambda) - \lambda x\Big)\right).$$

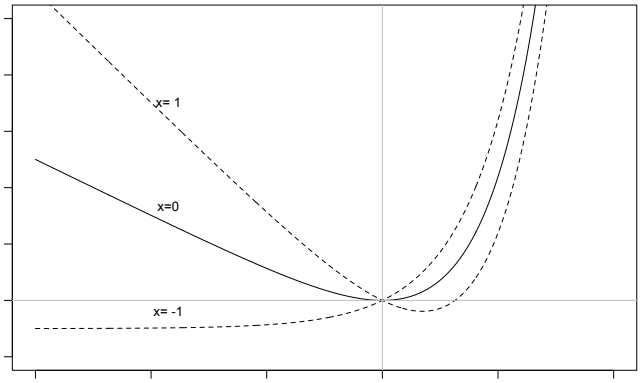
Similarly, $\exp(\lambda(X-x)) \geq 1$ if $X \leq x$ and $\lambda \leq 0$, so that

<2> $$\mathbb{P}\{X \leq x\} \leq \inf_{\lambda \leq 0} \mathbb{P} e^{\lambda(X-x)} = \exp\left(\inf_{\lambda \leq 0}\Big(L_X(\lambda) - \lambda x\Big)\right).$$

Remark. This inequality can also be derived from <1> applied to bound $\mathbb{P}\{-X \geq -x\}$ by $\inf_{\lambda \geq 0} \mathbb{P} e^{\lambda(-X+x)} = \inf_{\lambda \leq 0} e^{\lambda(X-x)}$.

The analysis is simplified if $\mathbb{P}X = 0$ and L_X is finite in a neighborhood of the origin, for then $\dot{L}_X(0) = \mathbb{P}X = 0$ and the convex function L_X is minimized at the origin. Equivalently, we can just work with $X - \mathbb{P}X$ instead of X.

$\lambda \mapsto L_{X-\mathbb{P}X}(\lambda) - \lambda x$



Remark. The picture actually shows the case where $X \sim \text{POISSON}(1)$.

The multiplication of $M_{X-\mathbb{P}X}$ by $e^{-\lambda x}$ tilts the convex function $L_{X-\mathbb{P}X}$ by $-\lambda x$, which ensures that $L_{X-\mathbb{P}X}(\lambda) - \lambda x$ achieves its global minimium on the half line $[0, \infty)$ if $x > 0$ and on the half line $(-\infty, 0]$ if $x < 0$. For the purposes of <1> and <2> we no longer have to consciously think about the sign of x; the infimum in both cases reduces to minimization over the whole real line and everything can be done without regard for the sign of λ In short,

<3> $$\mathbb{P}\{X - \mathbb{P}X \geq x\} \leq \exp\left(-\Lambda(x)\right) \qquad \text{for } x \geq 0,$$

<4> $$\mathbb{P}\{X - \mathbb{P}X \leq x\} \leq \exp\left(-\Lambda(x)\right) \qquad \text{for } x \leq 0,$$

where, for all $x \in \mathbb{R}$,

<5> $$\Lambda(x) := -\inf_{\lambda \in \mathbb{R}} \left(L_{X-\mathbb{P}X}(\lambda) - \lambda x\right) = \sup_{\lambda \in \mathbb{R}} \left(\lambda x - L_{X-\mathbb{P}X}(\lambda)\right).$$

We need only search for the global solution to $\dot{L}_{X-\mathbb{P}X}(\lambda) = x$ to determine Λ, except in those pesky cases where the infimum of the convex function $\lambda \mapsto L_{X-\mathbb{P}X}(\lambda) - \lambda x$ is approached as λ tends to $\pm\infty$.

Remark. The nonnegativity of Λ comes from the zero contribution at $\lambda = 0$. The second expression in <5> identifies the convex function $\Lambda(x)$ as the FENCHEL-LEGENDRE conjugate of the convex function $L_{X-\mathbb{P}X}$. Exciting as the recognition of this conjugate in a probability bound might be, it does not seem to help much in the actual calculation for a given X. Everything comes down to an exercise in Calculus and convexity, which could be completed without any knowledge of the material in §2.4.

*3.2 Behavior of the tail bound near the origin

Even though the tail bounds are not particularly useful for small t it is illuminating to see how the moments of $X - \mathbb{P}X$ affect Λ when its MGF is finite in a neighborhood of the origin.

The MGF gets its name from the coefficients in its power series expansion

$$M_{X-\mathbb{P}X}(\lambda) = \mathbb{P}e^{\lambda(X-\mathbb{P}X)} = 1+\sum\nolimits_{k\in\mathbb{N}} \mu_k\lambda^k/k! \qquad \text{where } \mu_k = \mathbb{P}\left(X-\mathbb{P}X\right)^k.$$

The quantities μ_k is often called the ***k*th central moment** to distinguish it from $\mathbb{P}X^k$. Note that $\mu_1 = 0$ and $\mu_2 = \operatorname{var}(X)$. The quantity $\mu_3/\mu_2^{3/2} = \mathbb{P}(X-\mathbb{P}X)^3/\operatorname{var}(X)^{3/2}$ is called the ***skewness*** of the distribution.

The function $L_{X-\mathbb{P}X} = \log M_{X-\mathbb{P}X}$ also has a power series expansion,

$$L_{X-\mathbb{P}X}(\lambda) = \sum\nolimits_{k\in\mathbb{N}} \kappa_k\lambda^k/k! \qquad .$$

The coefficients κ_k are called the (central?) ***cumulants*** and $L_{X-\mathbb{P}X}$, not surprisingly, is the ***cumulant generating function***. Note that there is no κ_0, because $L_{X-\mathbb{P}X}(0) = \log M_{X-\mathbb{P}X}(0) = 0$.

The cumulants can be related to the moments by equating coefficients in power series expansions. Here is how it works for the first three cumulants.

$$\begin{aligned}
1 + \mu_2\lambda^2/2! + \mu_3\lambda^3/3! + O(\lambda^4) &= \exp\left(\sum\nolimits_{k\in\mathbb{N}} \kappa_k\lambda^k/k!\right) \\
&= 1 + \left(\kappa_1\lambda + \kappa_2\lambda^2/2! + \kappa_3\lambda^3/3! + O(\lambda^4)\right) \\
&\quad + \left(\kappa_1\lambda + \kappa_2\lambda^2/2! + \kappa_3\lambda^3/3! + O(\lambda^4)\right)^2 \\
&\quad + \left(\kappa_1\lambda + \kappa_2\lambda^2/2! + \kappa_3\lambda^3/3! + O(\lambda^4)\right)^3 + O(\lambda^4) \\
&= 1 + \kappa_1\lambda + \kappa_2\lambda^2/2! + \kappa_3\lambda^3/3! + \left(\kappa_1^2\lambda^2 + \kappa_1\kappa_2\lambda^3\right) + \left(\kappa_1^3\lambda^3\right) + O(\lambda^4) \\
&= 1 + \lambda\left(\kappa_1\right) + \frac{\lambda^2}{2!}\left(\kappa_2 + 2\kappa_1^2\right) + \frac{\lambda^3}{3!}\left(\kappa_3 + 6\kappa_1\kappa_2\right) + O(\lambda^4).
\end{aligned}$$

It follows that $\kappa_1 = 0$ and $\kappa_2 = \mu_2$ and $\kappa_3 = \mu_3$.

Remark. Don't get too excited and leap to the conclusion that cumulants are the same as moments. If I hadn't centered the distribution to zero expected value then κ_1 would not be zero and μ_k would be a nasty-looking polynomial in $\kappa_1, \ldots, \kappa_k$. By repeated substitutions we could then write κ_k as another nasty-looking polynomial in the non-central moments $\mathbb{P}X, \ldots, \mathbb{P}X^k$. Even with the centering the remaining cumulants get messier: $\kappa_4 = X^2 - (\mathbb{P}X^2)^2$ and the expression for κ_{10} is a sum of 12 terms.

Back to tail probabilities. Remember that $\Lambda(x)$ is usually obtained by maximizing $\lambda x - L_{X-\mathbb{P}X}(\lambda)$ with respect to λ, with the task coming down to solving

$$t = L'_{X-\mathbb{P}X}(\lambda) = \kappa_2\lambda + \kappa_3\lambda^2/2! + \kappa_4\lambda^3/3! + \ldots \qquad .$$

This regularity suggests that the maximizing value λ_t be expressible as a power series $\sum_{k\in\mathbb{N}} a_k t^k/k!$. The a_k coefficients can be determined by another exercise in coefficient matching. First note that

$$\lambda_t = a_1 t + a_2\frac{t^2}{2!} + a_3\frac{t^3}{3!} + O(t^4), \quad \lambda_t^2 = a_1^2t^2 + a_1a_2t^3 + O(t^4), \quad \lambda_t^3 = a_1^3t^3 + O(t^4).$$

Thus

$$\begin{aligned} t &= \kappa_2\left(a_1t + a_2\frac{t^2}{2!} + a_3\frac{t^3}{3!}\right) + \frac{\kappa_3}{2!}\left(a_1^2t^2 + a_1a_2t^3\right) + \frac{\kappa_4}{3!}(a_1^3t^3) + O(t^4) \\ &= t\left(\kappa_2a_1\right) + \frac{t^2}{2!}\left(\kappa_2a_2 + \kappa_3a_1^2\right) + \frac{t^3}{3!}\left(\kappa_2a_3 + 3\kappa_3a_1a_2 + \kappa_4a_1^3\right) + O(t^4), \end{aligned}$$

implying

$$a_1 = 1/\kappa_2, \qquad a_2 = -\kappa_3/\kappa_2^3, \qquad a_3 = \text{something}.$$

It now follows that

$$\begin{aligned} \Lambda(t) &= t\left(a_1t + a_2t^2/2\right) - \frac{\kappa_2}{2!}\left(a_1^2t^2 + a_1a_2t^3\right) - \frac{\kappa_3}{3!}\left(a_1^3t^3\right) + O(t^4) \\ &= \frac{t^2}{2\kappa_2} - \frac{\kappa_3t^3}{6\kappa_2^3} + O(t^4) \qquad \text{near the origin.} \end{aligned}$$

The contributions from a_3 and κ_4 get absorbed into the $O(t^4)$.

Remark. You might be wondering why I bothered expanding λ_t as a cubic once I had realized the a_3 term would be absorbed into the $O(t^4)$. As a wise friend once advised me, it is always a good idea to expand an approximating series out to at least one term more than one thinks is necessary. A classic example of a failure to expand far enough can be found in the famous paper of Pearson (1900, page 165), which led to many years of argument between him and R. A. Fisher over the number of degrees of freedom for a χ^2 goodness of fit test. See Cochran (1952) for a very clear discussion of Pearson's error.

<6> **Example.** Suppose $X \sim \text{BIN}(n, p)$. Then

$$\mathbb{P}X = np, \quad \kappa_2 = \mathbb{P}(X - np)^2 = npq, \quad \kappa_3 = \mathbb{P}(X - np)^3 = npq(q - p)$$

so that

$$\Lambda(t) = \frac{t^2}{2npq} - \frac{t^3(q-p)}{6(npq)^2} + O(t^4) \qquad \text{near the origin.}$$

This approximation shows that $\Lambda(t) \geq t^2/(2npq)$ if $p \geq 1/2$ and $0 \leq t \approx 0$, in which case

$$\mathbb{P}\{X \geq np + t\} \leq e^{-\Lambda(t)} \leq \exp\left(-\frac{t^2}{2npq}\right) \qquad 0 \leq t \approx 0.$$

In fact, as will be shown in Section 3.7, the inequality holds for all $t \geq 0$ if $p \geq 1/2$. The local property implied by negative skewness suggests a subgaussian upper tail; the convexity of the ψ_{benn} function will transform the local suggestion into a global inequality. □

3.3 Normal

The MGF method is cleanest for the normal distribution. As this Section will show, the method leads to bounds comparable to very sharp inequalities that can be derived using special properties of the normal.

<7> **Example.** If $X \sim N(\mu, \sigma^2)$ then $M_X(\lambda) = \exp(\lambda\mu + \sigma^2\lambda^2/2)$ is finite for all real λ. For $x \geq 0$ inequality <1> gives

$$\begin{aligned}\mathbb{P}\{X \geq \mu + \sigma x\} &\leq \inf_{\lambda \geq 0} \exp(-\lambda(\mu + \sigma x) + \lambda\mu + \lambda^2\sigma^2/2) \\ &= \exp(-x^2/2) \qquad \text{for all } x \geq 0,\end{aligned}$$

the minimum being achieved by $\lambda = x/\sigma$. Analogous arguments, with $X - \mu$ replaced by $\mu - X$, give an analogous bound for the lower tail,

$$\mathbb{P}\{X \leq \mu - \sigma x\} \leq \exp(-x^2/2) \qquad \text{for all } x \geq 0,$$

leading to the inequality $\mathbb{P}\{|X - \mu| \geq \sigma x\} \leq 2e^{-x^2/2}$, which shows that the distribution of X is concentrated near μ. □

Of course the algebra in the Example would have been a tad simpler if I had worked with the standardized variable $(X - \mu)/\sigma$. I did things the messier way in order make the point that if Y is any random variable (not necessarily normally distributed) for which there exist constants ν and τ (not necessarily the mean and variance) for which

$$M_Y(\lambda) = \mathbb{P}e^{\lambda Y} \leq e^{\nu\lambda + \lambda^2\tau^2/2} \qquad \text{for all } \lambda \geq 0$$

then

<8> $$\mathbb{P}\{Y \geq \nu + \tau x\} \leq e^{-x^2/2} \qquad \text{for } x \geq 0.$$

This inequality <8> is usually called a ***subgaussian bound for the upper tail***. However, as explained in Chapter 7, the term 'subgaussian' is also often used in several looser senses.

How good is the upper bound

<9> $$\bar{\Phi}(x) := \mathbb{P}\{Z \geq x\} \leq B(x) := e^{-x^2/2} \qquad \text{for } x \geq 0?$$

Clearly, not so good for x close to 0, because $\bar{\Phi}(0) = 1/2$ whereas $B(0) = 1$. As will be shown by Example <18>, it does capture the most important tail properties if x is very large, but how useful is it for intermediate values of x such as those that appear in typical statistical applications?

<10> **Example.** Anyone who has taken an introductory Statistics course knows that if $T \sim N(\theta, 1)$ under a $\mathbb{P}_\theta$ model then, to two decimal places accuracy,

$$\mathbb{P}_\theta\{T - 1.96 \leq \theta \leq T + 1.96\} = 2\bar{\Phi}(1.96) \approx 0.95.$$

That is, the range $T \pm 1.96$ is a 95% confidence interval for θ.

Consider the effect of using the upper bound $B(x)$ instead of $\bar{\Phi}(x)$ when constructing the confidence interval. It is certainly true that

$$\mathbb{P}_\theta\{T \pm c \text{ fails to contain } \theta\} \leq 2B(c) = 2\exp(-c^2/2),$$

so that the $\mathbb{P}_\theta$ probability that the interval $T \pm c$ fails to contain θ is smaller than $2B(c)$; replacement of $\bar{\Phi}$ by B leads to a "conservative confidence" assertion. Here are some examples (with values rounded to two decimal places):

x	1.64	1.96	2.45	2.58	2.72	3.26
$\bar{\Phi}(x)$	5%	2.5%	0.72%	0.5%	0.33%	0.06 %
$\exp(-x^2/2)$	25.85%	14.65%	5%	3.62%	2.5%	0.5 %

As judged by B, the range $T \pm 1.96$ will contain θ with probability at least 70.7%, which is not very comforting given that the nominal value is 95%. For the 90% interval, the conservative value, $1 - 2 \times 25.85\% = 48.3\%$, is even worse. It takes a great stretching of the imagination to describe either conclusion as 'not bad'.

There is another way to use the upper bound. Instead of stretching imagination we could stretch the interval, from $T \pm 1.96$ to $T \pm c$ with $c = 2.72$. For this interval, the B bound assures coverage of at least 95%. Looking on the bright side, I think the increase from 1.96 to 2.72 is not too

high a price to pay for an appreciable relaxation of the modeling assumptions from normal to subgaussian.

The compromise looks even better when the failure probability is smaller. For example, under the strict $N(\theta,1)$ assumption $T \pm 2.58$ is a 99% confidence interval for θ and $T \pm 3.26$ has, according to B, probability at least 99% of containing θ. □

Sharper tail bounds are possible if we exploit further properties of the normal density, $\phi(x) := (2\pi)^{-1/2}\exp(-x^2/2)$, properties that are not shared by all subgaussian distributions. In fact there is a literature going back over two centuries that contains numerous facts about $\bar{\Phi}$, including several upper and lower bounds. For example,

$$\left(x^{-1}-x^{-3}\right)\phi(x) < \bar{\Phi}(x) <^{-1} \phi(x) \qquad \text{for all } x>0, \tag{11}$$

which can be derived by integrating $\int_x^\infty$ through the pointwise bounds (for $r>0$)

$$\begin{aligned} -\frac{d}{dr}\left[(r^{-1}-r^{-3})\phi(r)\right] &= (1-3r^{-4})\phi(r) < \phi(r) \\ &< (1+r^{-2})\phi(r) = -\frac{d}{dr}\left[r^{-1}\phi(r)\right]. \end{aligned}$$

Inequality <11> is the first of a sequence of approximations essentially due to Laplace (see Notes): for $k=0,2,4,\ldots,$

$$p_{k+1}(1/x) < \mathcal{R}(x) := \bar{\Phi}(x)/\phi(x) < p_k(1/x) \qquad \text{for } x>0, \tag{12}$$

where p_k is a polynomial of degree $2k+1$: $p_0(r) := r$, $p_1(r) := p_0(r) - r^3$, $p_2(r) = p_1(r) + 3r^5$, $p_3(r) = p_2(r) - 15r^7$, ... See Problem [1] for the proof.

The function $\mathcal{R}(x) := \bar{\Phi}(x)/\phi(x)$ is often called the "Mills ratio", for not particularly compelling reasons (see the Notes). I find it more convenient to work with the reciprocal of $\mathcal{R}$,

$$\rho(x) := 1/\mathcal{R}(x) = \phi(x)/\bar{\Phi}(x), \tag{13}$$

which is sometimes called the ***hazard rate*** for the $N(0,1)$ distribution because, for small positive δ,

$$\mathbb{P}\{x \le Z \le x+\delta \mid Z \ge x\} \approx \delta\rho(x) \qquad \text{if } Z \sim N(0,1).$$

Of course all the inequalities in <12> are rather useless for x near 0; the lower bounds $p_k(1/x)$ for odd k are all negative for $0 < x \le 1$. They are most useful for values of x much larger than 1, where they can be inverted to give upper and lower bounds for ρ. For example, the case $k=0$

gives $1 - x^{-2} < x/\rho(x) < 1$, which inverts for small values of $y = 1/x$ to give

<14> $$x < \rho(x) < x\left(1-y^2\right)^{-1} = x\left(1+y^2+y^4+\dots\right) = x + x^{-1} + O(x^{-3}).$$

Similarly for $k = 2$ we have $1 - y^2 + 3y^4 - 15y^6 < x/\rho(x) < 1 - y^2 + 3y^4$, which inverts to give

<15> $$\rho(x) = x + x^{-1} - 2x^{-3} + O(x^{-5}) \qquad \text{as } x \to \infty.$$

We'll be needing both <14> and <15> in Section 3.4.

The literature also contains several bounds that do not blow up as x decreases to 0. For example,

<16> $$(3x + \sqrt{x^2+8})/4 < \rho(x) \le \left(x + \sqrt{x^2+4}\right)/2 \qquad \text{for all } x \in \mathbb{R},$$

the lower bound coming from Birnbaum (1942) and the upper bound from Sampford (1953). See Problems [4] and [3] for proofs.

Remark. The inequalities <16> are not accurate enough for moderate x to serve as a basis for numerical calculation of normal tail probabilities. These days, such results are of little interest once one accords ρ the status of a useful member of the standard repetoire of functions for which highly accurate approximations can be obtained in every decent mathematical or statistical computing package.

For $x > 1$ inequality <11> can be rewritten as

<17> $$\log \bar{\Phi}(x) = -x^2/2 - \log(x) - c_0 - \eta(x)$$

where $c_0 := \log\left(\sqrt{2\pi}\right)$ and $0 \le \eta(x) \le -\log(1-x^{-2})$, which is $\le 2x^{-2}$ for $x \ge \sqrt{2}$. The bound $B(x)$ has captured the $-x^2/2$, which is much more important than the other terms when x becomes large, a fact illustrated by the next Example.

<18> **Example.** Suppose $Z_1, \dots, Z_n$ are random variables, each distributed $N(0,1)$ but, for the moment, not necessarily independent. Define $M_n := \max_{i \le n} Z_i$. A union bound gives some control for the tail:

$$\mathbb{P}\{M_n > x\} \le \sum\nolimits_i \mathbb{P}\{Z_i > x\} = n\bar{\Phi}(x) \le n\exp(-x^2/2) \qquad \text{for } x > 0.$$

In particular, $\mathbb{P}\{M_n > \sqrt{2\log(n) + 2r}\,\} \le e^{-r}$ for each $r \ge 0$. Roughly speaking, with high probability M_n should be not much bigger than $a_n := \sqrt{2\log n}$. This bound has the advantage of being unaffected by possible dependence between the Z_i's. It also has the disadvantage that it is sometimes excessively large. For example, in the extreme case where $Z_i = Z_1$ for all i any bound that involves n would be superfluous.

Remark. The union bound $\mathbb{P}(\cup_i A_i) \le \sum_i \mathbb{P}A_i$ is quite good if the events A_i are independent and $\sum_i \mathbb{P}A_i$ is small. See Problem [8]. It is part of the folklore that if the A_i's are 'almost' independent then the union bound is 'almost' quite good, with the meaning of 'almost' being problem specific.

If the Z_i's are actually independent the union bound can be refined:

$$\begin{aligned}\mathbb{P}\{M_n \le x\} &= \prod\nolimits_i \mathbb{P}\{Z_i \le x\} = \left(1-\bar{\Phi}(x)\right)^n \\ &= \exp\left(n\log(1-\bar{\Phi}(x)\right) \\ &= \exp\left(-n\bar{\Phi}(x) - n\sum\nolimits_{j\ge 2}\left(\bar{\Phi}(x)\right)^j/j\right) \\ &= \exp\left(-n\bar{\Phi}(x) - R_n(x)\right) \qquad \text{where } 0 \le R_n(x) \le \frac{n\bar{\Phi}(x)^2}{2(1-\bar{\Phi}(x))}.\end{aligned} \tag{19}$$

If we choose x_n so that $n\bar{\Phi}(x_n)$ is suitably large then $\mathbb{P}\{M_n > x_n\}$ is small.

To get sharper bounds we need to take care of the $\log(x_n) + c_0$ contributions. It will turn out that most of the distribution of M_n concentrates in a very short interval centered at a point smaller than a_n. To that end consider x_n of the form $a_n - w_n/a_n$ with $|w_n|$ much smaller than a_n. Approximation <17> then gives

$$\begin{aligned}\log\left(n\bar{\Phi}(x_n)\right) &= \log n - (a_n^2 - 2w_n + w_n^2/a_n^2)/2 - \log a_n - c_0 \\ &\qquad - \log(1 - w_n/a_n^2) - \eta(x_n) \\ &= w_n - \log(a_n) - c_0 - \text{small remainder}\end{aligned}$$

where the remainder term is of order $O\left((1+w_n^2)/a_n^2\right)$. Thus, for values $w_n = r + c_0 + \log a_n$ with, say, $|r| = O(\log a_n)$ we have

$$\mathbb{P}\{M_n \le a_n - (r + c_0 + \log a_n)/a_n\} = \exp\left(-B_n(r) - O(B_n(r)^2/n)\right).$$

where

$$B_n(r) := n\bar{\Phi}(x_n) = \exp\left(r + O(\epsilon_n^2)\right) \qquad \text{where } \epsilon_n := a_n^{-1}\log(a_n).$$

For example, when n is large enough, M_n has probability over 0.95 of lying in the range $a_n - (c_0 + \log a_n)/a_n \pm 4/a_n$. □

Remark. Inequalities <11> and <19> are the basis for the classical fact that $a_n(M_n - b_n)$, with $b_n = a_n - (\log\log n + \log(4\pi))/(2a_n)$, converges in distribution. See Leadbetter, Lindgren, and Rootzén (1983, Theorem 1.5.3).

In a digression from MGF methods, the next Section derives inequalities that are much more informative, for theoretical purposes, than <11> and <16>.

*3.4 Global behavior of the normal hazard rate

Perhaps surprisingly, it pays to understand how ρ behaves over the whole real line, and not just for large x or for x near 0. For example, as a short discussion (at the end of the Section) of Stein's method for normal approximation will show, it is more convenient to deal with a single well-behaved function than having to argue differently for different ranges of x values. The proof of the following Theorem, which lists a few basic facts about ρ, also introduces some useful general tricks.

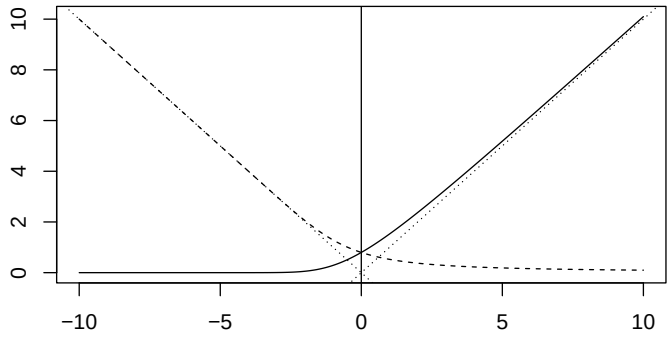
<20> **Theorem.** *The function $\rho(x) := \phi(x)/\bar{\Phi}(x)$ is infinitely differentiable with:*

(i) The function $\log \rho$ is strictly concave on $\mathbb{R}$ and the function ρ is strictly {positive & increasing & convex} on $\mathbb{R}$, with $\rho(x) \to 0$ as $x \to -\infty$ and $\rho(x)/x \to 1$ as $x \to \infty$.

(ii) The function $r(x) := \rho(x) - x$ is strictly {positive & decreasing & convex} on $\mathbb{R}$ with $r(x) \to +\infty$ as $x \to -\infty$ and $xr(x) \to 1$ as $x \to \infty$.

□ *(iii) $\dot{\rho}(x) = r(x)\rho(x) < 1$ for all real x with $r(0) = \rho(0) = \sqrt{2/\pi} \approx 0.798$.*

The functions ρ (solid line), r (dashed line), and $\pm x$ (dotted line)



Proof. The most delicate part of the argument—establishing convexity of ρ—will be left to last. The rest is mostly calculus with just a few probability tricks.

First note that

<21> $$1/\rho(x) = \int_0^\infty \phi(x+t)/\phi(x)\, dt = \int_0^\infty e^{-xt-t^2/2} dt = \sqrt{\pi/2}\,\gamma^z e^{-x|z|},$$

which is strictly decreasing, strictly positive, and

<22> $$\mathcal{L}(x) := \log(1/\rho(x)) = x^2/2 + \log\sqrt{2\pi} + \log\bar{\Phi}(x)$$

is strictly convex on $\mathbb{R}$ by§2.3. Consequently, $\ddot{\mathcal{L}}(x) > 0$ for all real x so that

<23> $$\dot{\mathcal{L}}(x) = x - \phi(x)/\bar{\Phi}(x) = x - \rho(x) = -r(x)$$

is strictly increasing.

For (i), except convexity of ρ: It follows that ρ is strictly {increasing & positive} with $\lim_{x\to-\infty}\rho(x) = 0$ and $\lim_{x\to-\infty}\rho(x) = \infty$, and $\log\rho(x)$ is strictly concave. Approximation <14> show that $\rho(x)/x = 1 + O(x^{-2})$ as $x \to \infty$.

For (ii): The function $r(x) = -\dot{\mathcal{L}}(x)$ is strictly decreasing on the whole real line. Approximation <15> gives $r(x) = x^{-1} + O(x^{-3})$ as $x \to \infty$. Thus $r(x) > 0$ for all x. And $r(x) = \rho(x) - x \to \infty$ as $x \to -\infty$.

For (iii): Equalities <22> and <23> give $-\dot{\rho}(x)/\rho(x) = \dot{\mathcal{L}}(x) = -r(x)$, that is, $\dot{\rho}(x) = \rho(x)r(x)$ for all x. Equality <15> gives $\rho(x) = x + x^{-1} + -2x^{-3} + O(x^{-5})$ and $r(x) = x^{-1} + -2x^{-3} + O(x^{-5})$ as $x \to \infty$, implying

<24> $$\dot{\rho}(x) = \rho(x)r(x) = 1 - x^{-2} + O(x^{-4}) \qquad \text{as } x \to \infty.$$

Of course the limiting behavior does not establish the global property (iii) of the Theorem. A little probability trick comes to the rescue. Direct calculation shows that $\dot{\phi}(r) = -r\phi(r)$ and $\ddot{\phi}(r) = (r^2 - 1)\phi(r)$. In integrated form: for all real x,

$$\int_x^\infty r\phi(r)\,dr/\bar{\Phi}(x) = \phi(x)/\bar{\Phi}(x) = \rho(x),$$
$$\int_x^\infty r^2\phi(r)\,dr/\bar{\Phi}(x) = \int_x^\infty \ddot{\phi}(r) + \phi(r)\,dr/\bar{\Phi}(x) = x\rho(x) + 1.$$

If P_x denotes the probability measure that has density $\phi(r)\{r \ge x\}/\bar{\Phi}(x)$ with respect to LEBESGUE measure on the real line then its expected value is $\rho(x)$ and its variance is $1 + x\rho(x) - \rho(x)^2 = 1 - r(x)\rho(x)$. As P_x is not degenerate its variance is nonzero, whence the asserted inequality for $r(x)\rho(x)$.

Strict convexity of ρ: It suffices to show that $\ddot{\rho}(x) > 0$ for all x.

$$\begin{aligned}\ddot{\rho}(x) &= r(x)\dot{\rho}(x) + \dot{r}(x)\rho(x) \qquad \text{because } \dot{\rho}(x) = \rho(x)r(x)\\ &= r^2(x)\rho(x) + [\rho(x)r(x) - 1]\,\rho(x)\\ &= [\psi(x) - 1]\,\rho(x) \qquad \text{where } \psi(x) := r(x)\,[r(x) + \rho(x)]\,.\end{aligned} \tag{<25>}$$

It suffices to show $\psi(x) > 1$ for all x. Differentiate again.

$$\begin{aligned}\dot{\psi}(x) &= \dot{r}(x)\,[r(x) + \rho(x)] + r(x)\,[\dot{r}(x) + \dot{\rho}(x)] = 2\dot{r}(x)r(x) + \ddot{\rho}(x)\\ &= 2\dot{r}(x)r(x) + [\psi(x) - 1]\,\rho(x) \qquad \text{by } <25>.\end{aligned}$$

In the last line the $2\dot{r}(x)r(x)$ is strictly negative because it equals the derivative of the strictly decreasing function $r(x)^2$. Thus we have the strict inequality

<26> $$\dot{\psi}(x) < [\psi(x) - 1]\,\rho(x) \qquad \text{for all } x \text{ in } \mathbb{R}.$$

To show $\psi(x) > 1$ for all x argue by contradiction (my least favorite method). Suppose there were an x_0 for which $\psi(x_0) \le 1$. By <26>, we would have $\dot{\psi}(x_0) < 0$, implying existence of an x_1 slightly larger than x_0 at which $\psi(x_1) < 1$. As $\lim_{x\to\infty} \psi(x) = 1$, it would follow that the differentiable function ψ must achieve its minimum on $[x_1, \infty)$ at some x_2. Thus $\dot{\psi}(x_2) = 0$ but $\psi(x_2) \le \psi(x_1) < 1$, contradicting <26>. □

I claim that the results listed in Theorem <20> can simplify the discussion of other theory involving the normal distribution. The following discussion illustrates my point.

First a little background. Charles Stein (Stein, 1972, 1986) invented an amazingly powerful method—a jewel of modern probability theory—for obtaining normal approximations. For a variety of random variables X his method provides bounds for $|\mathbb{P}h(X) - \gamma h|$ for a large family of bounded, measurable functions h on the real line. Typically h will be continuous at all but a finite number of points.

The method works by defining a linear transformation $f(\cdot) := \kappa(\cdot, h)$ of h, with κ defined by

$$\begin{aligned} H(r) &:= h(r) - \gamma h = \gamma^s\,[h(r) - h(s)]\,, \\ \phi(x)\kappa(x,h) &:= \gamma^r\{r < x\}H(r) = -\gamma^r\{r > x\}H(r) \qquad \text{because } \gamma H = 0 \\ &= \gamma^r\{r < x\}h(r) - \Phi(x)\gamma h = -\gamma^r\{r > x\}h(r) + \bar{\Phi}(x)\gamma h. \end{aligned} \tag{27}$$

Clearly f is everywhere continuous. At each x where h is continuous the function f is differentiable with $\dot{\phi}(x)f(x) + \phi(x)\dot{f}(x) = \phi(x)H(x)$. By virtue of the identity $\dot{\phi}(x) = -x\phi(x)$, this equality simplifies to

<28> $$\dot{f}(x) - xf(x) = H(x) := h(x) - \gamma h \qquad \text{if } h \text{ is continuous at } x.$$

If X is a random variable then $\mathbb{P}\left(\dot{f}(X) - Xf(X)\right) = \mathbb{P}h(X) - \gamma h$. If we can show, using various calculus and probability tricks, that the expression on the left-hand side of this equality is small then we can deduce $\mathbb{P}h(X) \approx \gamma h$, that is, we will have a kind of normal approximation for X. In particular, if the left-hand side is zero for a large enough collection of h functions then X has a $N(0,1)$ distribution.

Stein's method depends on a small collection of facts that are summarized in the next Lemma.

<29> **Lemma.** *(based on Stein, 1986, pp 25–28) Suppose $f(x) = \kappa(x,h)$ for a bounded measurable function h, as in <27>. Then:*

(i) *For g defined by $g(r) := -h(-r)$ we have $f(x) = \kappa(-x, g)$. Thus it suffices to concentrate the analysis on a general $\kappa(x, h)$ for $x \geq 0$.*

(ii) *$|f(x)| \leq \|H\|_\infty \min(1/\rho(x), 1/\rho(-x)) \leq \sqrt{\pi/2}\|H\|_\infty$ for each $x \in \mathbb{R}$. Consequently, $f(x) \to 0$ as $|x| \to \infty$.*

(iii) *If h is continuous at a particular x then f is differentiable at that x with $|\dot{f}(x)| \leq 2\|H\|_\infty$.*

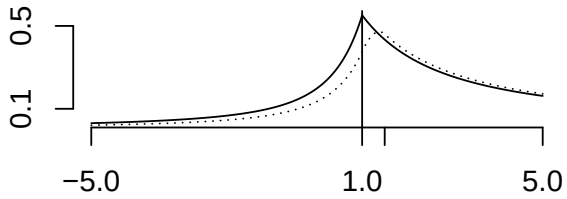
(iv) *If h is* LIPSCHITZ *then so is $\dot{f}$ and $\|\dot{f}\|_{Lip} \leq 2\|h\|_{Lip}$. Consequently, $|f(x+z) - f(x) - z\dot{f}(x)| \leq z^2\|h\|_{Lip}/2$.* □

You might find it helpful to keep two particular cases in mind while reading the proofs of (i) through (iv). For h_1 equal to the indicator function of an interval $(-\infty, x_0]$ we get $\phi(x)f_1(x) = \Phi(x \wedge x_0) - \Phi(x)\Phi(x_0)$. The function f_1 is smooth except for a discontinuity in the first derivative at x_0. It can be used to prove the Berry-Esseen extension of the central limit theorem. See Ho and Chen (1978, §1) for a proof in the case of identically distributed summands, which they attributed to Stein himself. See also Hall and Barbour (1984) for related results with non-identically distributed summands.

For a smoother approximation to $(-\infty, x_0]$ we could use $h_2(x) := h_1(x) + h_\epsilon(x)$ (which appeared in a proof by Stein (1986, pp. 35–36)) where $h_\epsilon(r) := \{x_0 < r < x_1\}(1 + (x_0 - r)/\epsilon)$ for $x_1 := x_0 + \epsilon$ with $\epsilon > 0$. The function h_2 is piecewise linear and LIPSCHITZ continuous, with $0 \leq h_2 \leq 1$ and $\|h_2\|_{\text{Lip}} = 1/\epsilon$. The corresponding $\kappa(x, h_2)$ is given by $f_1(x) + f_\epsilon(x)$ where $\phi(x)f_\epsilon(x) := \lambda_\epsilon(x) - \Phi(x)\lambda_\epsilon(x_1)$ with

$$\lambda_\epsilon(x) = (1 + x_0/\epsilon)\big[\Phi(x \wedge x_1) - \Phi(x \wedge x_0)\big] + (1/\epsilon)\big[\phi(x \wedge x_1) - \phi(x \wedge x_0)\big].$$

Picture of f_1 (solid line), f_2 (dotdash line), for $x_0 = 1$ and $\epsilon = 0.5$:



Notice that f_1 is not differentiable at x_0, although it does have left- and right-derivatives at that point. The function f_2 is actually everywhere differentiable and it has a second derivative everywhere except at x_0 and $x_0 + \epsilon$.

Proof (of Lemma <29>). Assertion (iv) will require the most work; the other assertions are more straightforward.

Proof of (i). For $x = -y$ the replacement of h by the function $g(r) := -h(-r)$ and the symmetry of the $N(0,1)$ distribution give

$$\begin{aligned}\phi(-y)\kappa(-y,h) &= \gamma^r\gamma^s\{-r < -y\}\gamma^s\left[h(-r) - h(-s)\right] \\ &= -\gamma^r\gamma^s\{r > y\}\gamma^s\left[g(r) - g(s)\right] = \phi(y)\kappa(y,g).\end{aligned}$$

□ Thus $\kappa(-y,h) = \kappa(y,g)$.

Proof of (ii). Define $C_0 := \|H\|_\infty$. Then <27> gives two upper bounds for $|f(x)|$:

$$\begin{aligned}|f(x)| &\le \gamma^r\{r > x\}|H(r)|/\phi(x) \le C_0\bar{\Phi}(x)/\phi(x) = C_0/\rho(x), \\ |f(x)| &\le \gamma^r\{r < x\}|H(r)|/\phi(x) \le C_0\Phi(x)/\phi(x) = C_0/\rho(-x).\end{aligned}$$

Thus $|f(x)| \le C_0 \min\left(1/\rho(x), 1/\rho(-x)\right) \le C_0/\rho(0)$, because ρ is an increasing function and $\max(x,-x) \ge 0$. If $|x| \to \infty$ then the smaller
□ of $1/\rho(x)$ and $1/\rho(-x)$ tends to zero.

Proof of (iii). By <28> and (ii),

$$|\dot{f}(x)| \le |H(x)| + |xf(x)| \le C_0 + C_0 \min\left(|x|/\rho(x), |x|/\rho(-x)\right).$$

□ For $x > 0$ use $x \le \rho(x)$; for $x < 0$ use $-x \le \rho(-x)$.

Proof of (iv). Write C_1 for $\|h\|_{\text{Lip}} = \|H\|_{\text{Lip}}$. From <28>, $\dot{f}(x) = G(x) + H(x)$ where $G(x) := xf(x)$ and $\dot{G}(x) = (1+x^2)f(x) + xH(x)$. It suffices to prove $\|G\|_{\text{Lip}} \le C_1$ by showing $|\dot{G}(x)| \le C_1$ for all x. The argument will reveal a surprising (and most fortunate) cancellation by symmetry of γ.

The LIPSCHITZ continuity of h implies that it is also absolutely continuous, in the sense described by UGMTP §3.4, a property that implies existence of a measurable function ψ with $\sup_{x\in\mathbb{R}}|\psi(x)| \le C_1$ for which:

- At $\mathfrak{m}$-almost all x, the function h is differentiable with $\dot{h}(x) = \psi(x)$.
- $h(b) - h(a) = \mathfrak{m}^t\{a < t < b\}\psi(t)$ for each bounded interval $[a,b]$.

As usual, $\mathfrak{m}$ denotes LEBESGUE measure on $\mathcal{B}(\mathbb{R})$.

The main idea is to write $\dot{G}$ as an integral involving ψ, based on similar representations for H and f, then use the bound on $|\psi|$ together with facts about ρ to bound $|\dot{G}(x)|$. Start with H:

$$\begin{aligned}H(r) &= \gamma^s\left(h(r) - h(s)\right) \\ &= \gamma^s\mathfrak{m}^t\psi(t)\left(\{s < t < r\} - \{r < t < s\}\right) \\ &= \mathfrak{m}^t\psi(t)\gamma^s\left(\{s < t\}\{t < r\} - \{r < t\}\{t < s\}\right) \qquad \text{by FUBINI} \\ &= \mathfrak{m}^t\psi(t)\left[\{t < r\}\Phi(t) - \{r < t\}\bar{\Phi}(t)\right]. \qquad <30>\end{aligned}$$

For f, use the second representation in <27>, then <30> and FUBINI:

$$\begin{aligned}
&\phi(x)f(x) = -\gamma^r\{x<r\}H(r)\\
&= \mathfrak{m}^t\psi(t)\gamma^r\{x<r\}\left[-\{t<r\}\Phi(t)+\{r<t\}\bar\Phi(t)\right]\left[\{t<x\}+\{x<t\}\right]\\
<31>\qquad &= -\mathfrak{m}^t\psi(t)\left[\{t<x\}\Phi(t)\bar\Phi(x)+\{x<t\}\bar\Phi(t)\Phi(x)\right]
\end{aligned}$$

because

$$\begin{aligned}
-\Phi(t)\gamma^r\{x<r\}\{t<r\}\{t<x\} &= -\{t<x\}\Phi(t)\bar\Phi(x),\\
\{x<r\}\{r<t\}\{t<x\} &= 0,\\
-\Phi(t)\gamma^r\{x<r\}\{t<r\}\{x<t\} &= -\{x<t\}\Phi(t)\bar\Phi(t),\\
\bar\Phi(t)\gamma^r\{x<r\}\{r<t\}\{x<t\} &= \{x<t\}\bar\Phi(t)\left[\Phi(t)-\Phi(x)\right].
\end{aligned}$$

The terms involving $\Phi(t)\bar\Phi(t)$ cancel.

Combine <30> and <31> to get the representation

$$\begin{aligned}
\dot G(x) = \;& x\dot f(x) + f(x) = (1+x^2)f(x) + xH(x)\\
= \;& \mathfrak{m}^t\psi(t)\Phi(t)\{t<x\}\left[-(1+x^2)\bar\Phi(x)/\phi(x)+x\right]+\\
<32>\qquad & +\mathfrak{m}^t\psi(t)\bar\Phi(t)\{x<t\}\left[-(1+x^2)\Phi(x)/\phi(x)-x\right]
\end{aligned}$$

The two functions of x can again be simplified by fortunate cancellations.

$$\ell(x) := (1+x^2)\bar\Phi(x)/\phi(x) - x = \frac{1+x^2-x(x+r(x))}{\rho(x)} = \frac{1-xr(x)}{\rho(x)},$$

which is strictly positive over the whole real line: trivially for $x \le 0$ because $r(x) > 0$ and by Theorem <20>(iii) because $1 - xr(x) > 1 - \rho(x)r(x) > 0$ for $x > 0$. The other function is just a reflection of ℓ:

$$(1+x^2)\Phi(x)/\phi(x) + x = \ell(-x) = \frac{1+xr(-x)}{\rho(-x)}.$$

Using the assumption that $|\psi(t)| \le C_1$ we now get a very clean inequality from <32>:

$$|\dot G(x)|/C_1 \le \ell(x)\int_{-\infty}^{x}\Phi(t)\,dt + \ell(-x)\int_x^{\infty}\bar\Phi(t)\,dt$$

Elementary calculus gives

$$\begin{aligned}
\int_{-\infty}^{x}\Phi(t)\,dt &= \phi(x)+x\Phi(x) = \phi(x)\left[1+x/\rho(-x)\right] = \phi(x)r(-x)/\rho(-x),\\
\int_{x}^{\infty}\bar\Phi(t)\,dt &= \phi(x)-x\bar\Phi(x) = \phi(x)\left[1-x/\rho(x)\right] = \phi(x)r(x)/\rho(x),
\end{aligned}$$

Now for the triumphant conclusion:

$$\begin{aligned}|\dot{G}(x)|/C_1 &\le \frac{[1-xr(x)]\,\phi(x)r(-x)+[1+xr(-x)]\,\phi(x)r(x)}{\rho(x)\rho(-x)}\\ &= \frac{\phi(x)\,[\rho(-x)+x+\rho(x)-x]}{\rho(x)\rho(-x)}\\ &= \phi(x)\,(1/\rho(x)+1/\rho(-x)) = \bar{\Phi}(x)+\Phi(x) = 1.\end{aligned}$$

Yet another fortunate cancelation.

To bound $|f(x+z)-f(x)-z\dot{f}(x)|$ write it as an integral:

$$|z\int_0^1 \dot{f}(x+tz)-\dot{f}(x)\,dt| \le |z|\int_0^1 \|\dot{f}\|_{\text{Lip}}\,t|z|\,dt.$$ □

Remark. I suspect the fortunate cancellations have something to do with assertion (i) in Lemma <29>, which comes from the symmetry of the standard normal distribution. Perhaps the proof could be simplified by explicit appeals to some sort of invariance property.

3.5 Poisson

The normal distribution represents the prototype for the class of subgaussian distributions. In a similar way the POISSON provides the prototype for a class of distributions that might be called "subPoisson". These distributions behave like subgaussians for moderately large deviations from the mean but decrease only a little faster than the exponential further out in the tails. The BENNETT inequalities in Chapter 8 will provide further examples.

A random variable X has a POISSON(θ) distribution, for some $\theta > 0$, if

$$\mathbb{P}\{X=k\} = e^{-\theta}\theta^k/k! \qquad \text{for } k = 0,1,\ldots.$$

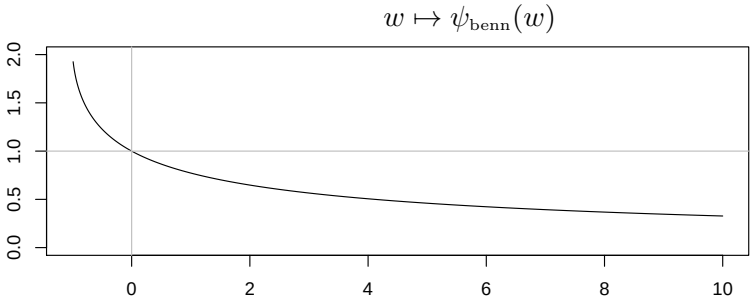
Both $\mathbb{P}X$ and $\text{var}(X)$ equal θ and $M_X(\lambda) = \exp\left(\theta e^\lambda - \theta\right)$. Thus

$$L_{X-\theta}(\lambda) = \theta(e^\lambda - 1 - \lambda) = \theta\mathbb{f}(\lambda) \qquad \text{for all } \lambda \in \mathbb{R}.$$

and $\Lambda(x) = \theta \sup(\lambda x/\theta - \mathbb{f}(\lambda)) = \theta\mathbb{h}(x/\theta)$, where

$$\mathbb{h}(w) := (1+w)\log(1+w) - w = \tfrac{1}{2}w^2\psi_{\text{benn}}(w) \qquad \text{for } w \ge -1,$$

as in §2.2. The function $\psi_{\text{benn}}(\cdot)$ is convex, and decreasing on $[-1,\infty)$ with $\psi_{\text{benn}}(0) = 1$. For large w it decreases like $2w^{-1}\log(w)$.



Thus

$$\mathbb{P}\{X \ge \theta + x\} \le \exp\left(-\frac{x^2}{2\theta}\psi_{\text{benn}}(x/\theta)\right) \qquad \text{for } x \ge 0$$

$$\mathbb{P}\{X \le \theta - x\} \le \exp\left(-\frac{x^2}{2\theta}\psi_{\text{benn}}(-x/\theta)\right) \qquad \text{for } 0 \le x \le \theta.$$

The first inequality shows that the upper tail decreases like a subgaussian in the range $0 \le x \ll \theta$, because $\psi_{\text{benn}}(x/\theta) \approx 1$ for x/θ near 0, but that the tail decay becomes more like $\exp(-x\log(x/\theta))$ further out into the tail. The inequality for the lower tail is more interesting, because $\psi_{\text{benn}}(w) > 1$ for $-1 \le w < 0$. The lower tails drop off even faster than one might expect from the $N(\theta,\theta)$ approximation to the POISSON(θ). This can be interpreted as a skewness effect: $\mathbb{P}(X-\theta)^3$ is the coefficient of $\lambda^3/3!$ in the power series expansion of

$$\begin{aligned}\mathbb{P}e^{\lambda(X-\theta)} &= \exp\Big(\theta(e^\lambda - 1 - \lambda)\Big) \\ &= 1 + \theta\left(\frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots\right) + \frac{\theta^2}{2!}\left(\frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots\right)^2 + \dots,\end{aligned}$$

which is positive. The distribution of $X - \theta$ puts more mass to the right of the origin than the $N(0,\theta)$. That fact slows down the decay in the upper tails but improves the rate of decay in the lower tail.

The $\exp(-\theta\hbar(x/\theta))$ tail bound compares favorably with the probability calculated by means of the STIRLING formula (§2.5) for $k \in \mathbb{N}$:

$$k! = \sqrt{2\pi k}\,(k/e)^k e^{-k+r_k} \qquad \text{where } \frac{1}{12k+1} < r_k < \frac{1}{12k}.$$

If $k = \theta + x$ then

$$\begin{aligned}\log\Big(\sqrt{2\pi k}\,\mathbb{P}\{Y = k\}\Big) &= -\theta + k\log(\theta) - k\log(k) + k - r_k \\ &= -\theta\hbar(x/\theta) - r_k.\end{aligned}$$

The MGF method has captured the most important term, $-\theta\hbar(x/\theta)$, in the exponent.

3.6 Gamma and chi-squared

Suppose X has a GAMMA(α) distribution, the probability measure on $\mathbb{R}^+$ that has density $f_\alpha(r) = r^{\alpha-1}e^{-r}/\Gamma(\alpha)$ with respect to LEBESGUE measure. The positive constant α is often called the *shape parameter*. The expected value and variance of X both equal α and

$$M_X(\lambda) = \int_0^\infty \frac{r^{\alpha-1}e^{-r(1-\lambda)}}{\Gamma(\alpha)}\,dr = (1-\lambda)^{-\alpha}\{\lambda < 1\} + \infty\{\lambda \ge 1\}.$$

Thus $L_{X-\alpha}(\lambda) = -\alpha\lambda - \alpha\log(1-\lambda)$ for $\lambda < 1$, so that

<33> $$\mathbb{P}\{X \ge \alpha + x\} \le \exp\Bigl(-\Lambda(x)\Bigr) \qquad \text{for } x \ge 0$$

<34> $$\mathbb{P}\{X \le \alpha - x\} \le \exp\Bigl(-\Lambda(-t)\Bigr) \qquad \text{for } \alpha > x \ge 0$$

where, for $\alpha + x > 0$,

$$\Lambda(x) = \sup_{\lambda<1} \lambda(\alpha+x) + \alpha\log(1-\lambda) = x - \alpha\log(1+x/\alpha).$$

For $x \ge 0$ we have

$$\Lambda(x) \approx \begin{cases} x^2/(2\alpha) & \text{if } x \text{ is near } 0 \\ x & \text{if } x \text{ is large} \end{cases}$$

and for $0 \le x < \alpha$ it gives a subgaussian lower tail: $\Lambda(x) \ge x^2/(2\alpha)$.

Boucheron, Lugosi, and Massart (2013, page 28) pointed out that the tail could also be controlled by first replacing the logMGF of $X - \alpha$ by a larger function: if $0 \le \lambda < 1$ then

<35> $$\alpha^{-1}L_{X-\alpha}(\lambda) = -\log(1-\lambda) - \lambda = \sum\nolimits_{i\ge2} \frac{\lambda^i}{i} \le \sum\nolimits_{i\ge2} \frac{\lambda^i}{2} = \frac{\lambda^2}{2(1-\lambda)}.$$

This little trick has a desirable consequence: the resulting tail bounds work for a collection of distributions that contains the centered GAMMA(α) as a special case. BL&M referred to this collection as having a "subgamma right tail". As I'll only discuss this tail I'll drop the "right" qualifier.

<36> **Definition.** *For constants $\alpha > 0$ interpret $W \in$ SUBGAMMA(α) to mean*

□ $$\mathbb{P}e^{\lambda W} \le \exp\left(\frac{\alpha\lambda^2/2}{1-\lambda}\right) \qquad \text{for } 0 \le \lambda < 1.$$

Remark. If $W \in \text{subGAMMA}(\alpha)$ then $M_W(\lambda) \le 1 + +O(\lambda^2)$ as $\lambda \searrow 0$, which implies $\mathbb{P}W \le 0$.

Problem [11] shows that $f(\lambda) := \frac{1}{2}\lambda^2/(1-\lambda)$ is convex for $0 \le \lambda < 1$, with FENCHEL-LEGENDRE conjugate

$$\begin{aligned} f^*(y) &:= \sup_{0\le\lambda<1} (\lambda y - f(\lambda)) = 1 + y - (1+2y)^{1/2} \\ &= \frac{y^2}{1+y+(1+2y)^{1/2}} \ge \frac{y^2}{2+2y} \qquad \text{for } y \ge 0. \end{aligned}$$

If $W \in \text{subGAMMA}(\alpha)$ we have $\mathbb{P}\{W \ge x\} \le e^{-\Lambda(x)}$ for $x \ge 0$, where

$$\begin{aligned} \Lambda(x) &= \sup_{0\le\lambda<1} (\lambda x - \alpha f(\lambda)) = \alpha \sup_{0\le\lambda<1} (\lambda x/\alpha - f(\lambda)) \\ &= \alpha f^*(x/\alpha) = \alpha + x - \sqrt{\alpha^2 + 2\alpha x} \qquad \text{<37>} \\ &= x^2 / \left(x + \alpha + (\alpha^2 + 2x\alpha)^{1/2} \right) \\ &\ge x^2/(2\alpha + 2x). \qquad \text{<38>} \end{aligned}$$

The next Example shows how inequality <38> leads to the traditional form of the BERNSTEIN inequality.

<39> **Example.** Suppose $S = W_1 + \cdots + W_n$, a sum of independent random variables with

$$\mathbb{P}W_i = 0, \quad \mathbb{P}W_i^2 = \sigma_i^2, \quad \mathbb{P}|W_i|^k \le k!\sigma_i^2/2 \qquad \text{for } k \ge 3, \text{ for each } i.$$

for finite σ_i^2's. Then, for each i and $\lambda \ge 0$,

$$\begin{aligned} \mathbb{P}e^{\lambda W_i} &\le 1 + \lambda\mathbb{P}W_i + \sum_{k\ge2} \lambda^k \mathbb{P}|W_i|^k/k! \le 1 + \tfrac{1}{2}\sigma_i^2\lambda^2 \sum\nolimits_{k\ge2} \lambda^{k-2} \\ &\le \exp\left(\tfrac{1}{2}\sigma_i^2\lambda^2/(1-\lambda)\right) \qquad \text{for } 0 \le \lambda < 1. \end{aligned}$$

That is, $W_i \in \text{subGAMMA}(\sigma_i^2)$ and $S \in \text{subGAMMA}(\mathcal{V}^2)$ for $\mathcal{V} := \sum_i \sigma_i^2 = \text{var}(S)$. Inequality <38> then gives

$$\mathbb{P}\{S \ge x\} \le \exp\left(-x^2/(2\mathcal{V} + 2x)\right) \qquad \text{for } x \ge 0. \qquad \text{<40>}$$

The assumptions on the W_i's are often relaxed slightly by the introduction of a scaling parameter: $\mathbb{P}|W_i|^k \le \frac{1}{2}k!\mathbb{P}W_i^2 B^{k-2}$ for $k \ge 3$. Equivalently, $\mathbb{P}|W_i/B|^k \le \frac{1}{2}k!\mathbb{P}|W_i/B|^2$ for $k \ge 3$. Inequality <40>, with $\mathcal{V}$ still equal to $\text{var}(S)$, then becomes $\mathbb{P}\{S/B \ge x\} \le \exp\left(-x^2/(2\mathcal{V}/B^2 + 2x)\right)$ or

$$\mathbb{P}\{S \ge z\} \le \exp\left(-z^2/(2\mathcal{V} + 2Bz)\right) \qquad \text{for } z \ge 0.$$

□ after the substitution $z = xB$.

Remark. As BL&M noted, even if the moment assumption is weakened to $\sum_i \mathbb{P}(X_i^+)^k \le \frac{1}{2}k!\mathcal{V}B^{K-2}$ we still get $\log \mathbb{P}\exp(\lambda S - \lambda \mathbb{P}S) \le \frac{1}{2}\mathcal{V}\lambda^2/(1-B\lambda)$ for $0 \le B\lambda < 1$, a refinement that they attributed to Emmanuel Rio. Even though I didn't know any specific cases where the use of X_i^+ instead of $|X_i|$ is the key to a significant improvement, I gladly added Rio's trick to my toolbox. It seemed natural that upper bounds involving $\sum_i X_i$ should only involve upper bounds for the X_i's. See Problem [12]

A more delicate substitution produces another neat tail bound.

<41> **Example.** If $W \in \text{subGamma}(\alpha)$ then $\log \mathbb{P}\{W \ge x\} \le -\Lambda(x)$ for $x \ge 0$. Massart (2006, §2.2.3) realized that <37> could be rearranged into a neater form by solving the quadratic equation $(\alpha + x - z)^2 = \alpha + 2\alpha x$ to show that $\Lambda(x) = z$ for $x := z + \sqrt{2\alpha z}$, so that

<42> $$\mathbb{P}\{W \ge z + \sqrt{2\alpha z}\} \le e^{-z} \qquad \text{for } z \ge 0.$$

In particular, if $X \sim \text{Gamma}(\alpha)$ then

$$\begin{aligned}\mathbb{P}\{\sqrt{X} \ge \sqrt{\alpha} + \sqrt{z}\} &= \mathbb{P}\{X \ge \alpha + 2\sqrt{\alpha z} + z\} \\ &= \mathbb{P}\{X - \alpha \ge z + \sqrt{2\alpha z}\} \le e^{-z} \qquad \text{for } z \ge 0.\end{aligned} \tag{43}$$

Substituting r for $\sqrt{z}$ we get $\mathbb{P}\{\sqrt{X} \ge \sqrt{\alpha} + r\} \le e^{-r^2}$, an example of a "subgaussian-like bound" for an upper tail beyond a point strictly larger than the mean: the JENSEN inequality gives $\mathbb{P}\sqrt{X} < \sqrt{\alpha}$.

Even more specifically, if $X \sim \chi^2_k$ then $X/2 \sim \text{Gamma}(k/2)$ so that inequality <43> implies

$$\mathbb{P}\{\sqrt{X} \ge \sqrt{k} + r\} \le e^{-r^2/2} \qquad \text{for } r \ge 0,$$

giving the convenient bound $\mathbb{P}\{\sqrt{X} \ge 2\sqrt{k}\} \le e^{-k/2}$.

As an application, suppose $Z_1, \ldots, Z_n$ are independent $N(0, I_d)$ random vectors. Then $(\mathbb{P}|Z_i|_2)^2 \le \mathbb{P}|Z_i|_2^2 = d$ and $\sum_{i\le n}|Z_i|_2^2 \sim \chi^2_{nd}$. From the inequality $\left(n^{-1}\sum_{i\le n}|Z_i|_2\right)^2 \le \sum_{i\le n}|Z_i|_2^2/n$ it follows that

$$\mathbb{P}\{n^{-1}\sum\nolimits_{i\le n}|Z_i|_2 \ge 2\sqrt{d}\} \le e^{-nd/2},$$

a neat little bound that is useful in high-dimensional statistical theory. See Wu and Zhou (2021, Section 9), for example. □

Remark. Bennett (1962, p39) obtained a bound similar to <37> under assumptions like those in Example <39>.

The final Example illustrates an advantage of working with subGamma even for random variables with Gamma distributions.

<44> **Example.** Consider the weighted sum $W := \sum_{j=1}^{k} b_j(Z_j^2 - 1)$ where $b_j > 0$ for all j and $Z = (Z_1, \ldots, Z_k) \sim N(0, I_k)$. Laurent and Massart (2000, Lemma 1) showed that

<45> $$\mathbb{P}\{W \geq 2|b|_\infty r + 2|b|_2\sqrt{r}\,\} \leq e^{-r} \qquad \text{for } r \geq 0,$$

where $|b|_\infty := \max_j |b_j|$ and $|b|_2 := (\sum_j b_j^2)^{1/2}$.

Direct integration shows that $\mathbb{P}\exp(\lambda Z_j^2) = (1-2\lambda)^{-1/2}$ for $\lambda < 1/2$. Thus, as in <35>,

$$\log \mathbb{P}\exp(\lambda b_j(Z_j^2 - 1)) = -\lambda b_j - \tfrac{1}{2}\log(1 - 2b_j\lambda) \leq \tfrac{1}{2}b_j^2\lambda^2(1 - 2b_j\lambda)$$

provided $2b_j\lambda < 1$. Consequently, for $0 \leq \lambda|b|_\infty < 1$,

$$\log \mathbb{P}e^{\lambda W} \leq \tfrac{1}{2}\sum_j b_j^2\lambda^2/(1 - 2|b|_\infty\lambda) = \tfrac{1}{2}|b|_2^2\lambda^2/(1 - 2|b|_\infty\lambda),$$

implying $W/(2|b|_\infty) \in \text{subGamma}(\alpha)$ for $\alpha = \frac{1}{2}(|b|_2/|b|_\infty)^2$. Thus inequality <45> follows from <42>. □

3.7 Binomial

The binomial distribution behaves a little like the poisson. It is also a prototype for other inequalities involving sums of bounded random variables.

Remember that X has a $\text{Bin}(n, p)$ distribution if

$$\mathbb{P}\{X = k\} = \binom{n}{k} p^k q^{n-k} \qquad \text{for } k = 0, 1, \ldots, n.$$

Here and subsequently I write q for $1 - p$. The distribution has expected value np, variance npq, and $M_X(\lambda) = (q + pe^\lambda)^n$. The difference $n - X$ has a $\text{Bin}(n, q)$ distribution. Thus

<46> $$\mathbb{P}\{X \leq np - x\} = \mathbb{P}\{n - X \geq nq + x\}.$$

That is, the lower tail for the $\text{Bin}(n, p)$ corresponds exactly to the upper tail for the $\text{Bin}(n, q)$.

Here is the main result:

<47> **Theorem.** *If* $X \sim \text{Bin}(n, p)$ *then*

$$\mathbb{P}\{X \geq np + x\} = \exp\left(-\frac{x^2}{npq} g_p(x)\right) \qquad \text{for } 0 \leq x \leq nq$$

$$\text{where } g_p(x) := q\psi_{\text{benn}}\left(\frac{x}{np}\right) + p\psi_{\text{benn}}\left(\frac{-x}{nq}\right).$$

Remark. From <46>, the companion inequality for the lower tail is

$$\mathbb{P}\{X \le np - x\} \le \exp\left(-\frac{x^2}{2npq} g_q(x)\right) \qquad \text{for } 0 \le x \le np.$$

It is merely a matter of swapping the roles of p and q.

Proof. The quantities $A := 1 + x/(np)$ and $B := 1 - x/(nq)$ play a central role in the argument, which is little more than a routine Calculus exercise (backed up by convexity simplifications and the equality $1 = pA + qB$).

From <3>, $\log \mathbb{P}\{X \ge np + x\} \le -\Lambda(x) = -n \sup_\lambda \mathcal{L}(\lambda)$ where

$$\mathcal{L}(\lambda) := pA\lambda - \log(q + pe^\lambda) \qquad \text{for } \lambda \in \mathbb{R},$$

a concave function with derivative $\dot{\mathcal{L}}(\lambda) = pA - pe^\lambda/(q + pe^\lambda)$.

If $x = nq$ then $A = 1/p$ and $\dot{\mathcal{L}}(\lambda) > 0$ for all λ. The supremum equals $\log(1/p)$ and $\exp(-\Lambda(nq)) = p^n = \mathbb{P}\{X = n\}$.

If $0 \le x < nq$ the supremum is achieved when $A(q + pe^\lambda) = e^\lambda$, that is, $e^\lambda = A/B$. Thus

$$\begin{aligned}
\Lambda(x)/n &= pA \log(A/B) - \log(qB + pA) + \log B \\
&= pA \log A + qB \log B \\
&= p\mathbb{h}(x/np) + p(x/np) + q\mathbb{h}(-x/nq) + q(-x/nq) = g_p(x),
\end{aligned}$$

□ the desired equality.

Remark. The $g_p(x)$ is the same function as the $g(x, n, p)$ in §2.5, which derived sharp approximations for $\mathbb{P}\{X = k\}$ by means of the STIRLING approximation: for $k = np + x$,

$$\mathbb{P}\{X = k\} = \frac{\exp\left[-x^2 g_p(t)/(2npq) + O\left(k^{-1} + (n-k)^{-1}\right)\right]}{\sqrt{2\pi n(p + x/n)(q - x/n)}}.$$

As happened with the $N(0, 1)$, the MGF method captures the main term in the exponent but misses the square root term in the denominator.

As also explained in that Section, the convexity of ψ_{benn} gives

$$g_p(x) \ge \psi_{\text{benn}}\left(\frac{qx}{np} - \frac{px}{nq}\right) = \psi_{\text{benn}}\left(\frac{x(q-p)}{npq}\right) \ge 1 \qquad \text{if } p \ge 1/2.$$

Thus, if $p \ge 1/2$, the upper tail is less than $\exp\left(-x^2/(2npq)\right)$, a clean subgaussian bound with scale parameter $\sqrt{npq}$. If $p < 1/2$ the upper tail is still subgaussian (because $\psi_{\text{benn}}(-x/(nq)) \ge 1$) but with a larger scale parameter $\sqrt{nq}$.

Now let me move a little beyond the BINOMIAL to show that there are several other distributions, with the same expected value as the $\text{BIN}(n, p)$,

that share the tail bounds for the BINOMIAL. In this Chapter these bounds are derived by means of pointwise inequalities for MGFs. As such they leave open the question of whether analogous inequalities would also hold for the exact tail probabilities, not just their MGF-derived upper bounds. Chapter 4 will return to this question

<48> **Example.** Suppose $S \sim \text{PBIN}(p_1, \ldots, p_n)$, that is, S is a sum of independent random variables $Y_1 + \cdots + Y_n$, with $Y_i \sim \text{BER}(p_i)$ for possibly different p_i's. Define $\overline{p} = n^{-1} \sum_i p_i$. Then

$$M_S(\lambda) = \prod\nolimits_{i \le n} \left(q_i + p_i e^{\lambda} \right) = \exp\left(\sum\nolimits_{i \le n} \log(q_i + p_i e^{\lambda}) \right).$$

Concavity of the log() function shows that

<49> $$n^{-1} \sum\nolimits_{i \le n} \log(q_i + p_i e^{\lambda}) \le \log\left(n^{-1} \sum\nolimits_{i \le n} (q_i + p_i e^{\theta}) \right) = \log\left(\overline{q} + \overline{p} e^{\theta} \right).$$

Thus $M_S(\theta) \le M_W(\theta)$ where $W \sim \text{BIN}(n, \overline{p})$ and $\overline{q} = 1 - \overline{p}$. It follows that

$$\mathbb{P}\{S \ge n\overline{p} + x\} \le \exp\left(-\frac{x^2}{2n\overline{p}\overline{q}} g_{\overline{p}}(x) \right) \qquad \text{for } 0 \le x \le n\overline{q},$$

□ with a similar bound for the lower tail.

The convexity idea from the previous Example can be pushed even further.

<50> **Example.** Suppose $T = Y_1 + \cdots + Y_n$, a sum of independent random variables Y_i with $0 \le Y_i \le 1$ and $\mathbb{P}Y_i = p_i$ for each i, and $n\overline{p} = \sum_{i=1}^n p_i$. By convexity of the exponential function,

$$e^{\lambda Y_i} \le (1 - Y_i) + Y_i e^{\lambda} \qquad \text{for each real } \lambda.$$

The inequality holds for all possible realizations of Y_i. Equality is achieved at $Y_i \in \{0, 1\}$. In particular, equality holds when $Y_i \sim \text{BER}(p_i)$, as in the previous Example. Take expectations.

$$\mathbb{P}e^{\lambda Y_i} \le (1 - p_i) + p_i e^{\lambda} = q_i + p_i e^{\lambda} \qquad \text{for each real } \lambda.$$

By independence,

$$M_T(\lambda) = \prod\nolimits_{i \le n} \mathbb{P}e^{\lambda Y_i} \le \prod\nolimits_{i \le n} \left(q_i + p_i e^{\lambda} \right) = M_S(\lambda) \le M_W(\lambda),$$

with $S \sim \text{PBIN}(p_1, \ldots, p_n)$ and $W \sim \text{BIN}(n, \overline{p})$, as in the previous Example. Thus

$$\mathbb{P}\{T \ge n\overline{p} + x\} \le \exp\left(-\frac{x^2}{2n\overline{p}\overline{q}} g_{\overline{p}}(x) \right) \qquad \text{for } 0 \le x \le n\overline{q},$$

□ a result due to Hoeffding (1963, Theorem 1).

3.8 Sampling and the hypergeometric

Both Example <48> and Example <50> involved sums of independent random variables. The MGF approach can also work when there is dependence between the summands, although the argument becomes a little more delicate.

<51> **Example.** Suppose $U = \{u_1, \ldots, u_N\}$ is a finite set, an urn if you like to think that way. In that interpretation the u_i's are the balls. Suppose exactly R of the balls are colored red and the other $B = N - R$ are colored black. If n balls are sampled without replacement then each subset of U with size n has probability $1/\binom{N}{n}$ of being selected and the number of red balls T_n in the sample has a hypergeometric distribution, HYPER(n, R, B), meaning that

$$\mathbb{P}\{T_n = k\} = \binom{R}{k}\binom{B}{n-k}\Big/\binom{N}{n}$$

for each nonnegative integer k such that $k \le R$ and $n - k \le B$.

If the sampling is carried out with replacement then the number of red balls in the sample, S_n, has a BIN(n, p) distribution, where $p = R/N$.

Elementary calculations (Pitman, 1993, §3.6) show that

$$\begin{aligned} \mathbb{P}T_n &= \mathbb{P}S_n = np \\ \operatorname{var}(T_n) &= np(1-p)(N-n)/(N-1) < \operatorname{var}(S_n) = np(1-p). \end{aligned}$$

If n is much smaller than N then there is actually not much difference between HYPER(n, R, B) and BIN(n, p): if a ball is selected then returned to the urn, it is unlikely to be selected again if n/N is very small. If n/N is not so small then, judging by the variances, HYPER(n, R, B) is more concentrated around np than BIN(n, p). A beautiful result by Hoeffding (1963, §6) adds some precision to this intuition. He showed that for each convex function f on the real line,

<52> $$\mathbb{P}f(T_n) \le \mathbb{P}f(S_n).$$

In particular, the choice $f(x) = e^{\lambda x}$ shows that $M_{T_n}(\lambda) \le M_{S_n}(\lambda)$ for all real λ. Any tail bound for the hypergeometric obtained via the MGF argument must therefore be smaller than the corresponding MGF tail bound for the BINOMIAL.

To be more precise, Hoeffding's result didn't involve red balls and black balls. It worked for every function $g : U \to \mathbb{R}$. (The special case where $g(u_i) = 1$ for a red ball an $g(u_i) = 0$ for a black ball get us back to the hypergeometric.) That is, we can take $X_1, \ldots, X_n$ to be a sample from U without replacement and $Y_1, \ldots, Y_n$ to be a sample with replacement. If we define $T_n := \sum_{i \le n} g(X_i)$ and $S_n := \sum_{i \le n} g(Y_i)$ then inequality <52> will still hold for every convex f.

I had some trouble digesting Hoeffding's proof. Even after working through the details I could not have explained to anyone why the method worked. Subsequently I stumbled on a proof by Le Cam (1986a, p534), which involved a much more intuitive explanation, reducing everything to the JENSEN inequality. Unfortunately I again had some trouble convincing myself that all the "obvious facts" were actually true, so I wrote out the following rather more pedantic account based on Le Cam's idea. For technical details see Problem [13].

Here is the key idea. Suppose $Y = (Y_1, Y_2, \dots)$ is obtained by sampling repeatedly with replacement from U. With probability one each member of U appears infinitely often in the Y sequence. If we discard all except the first appearance of each u in U from the Y sequence then we are left with a random permutation, $(X_1, \dots, X_N)$ of U; and $X_1, \dots, X_n$ forms a sample of size n taken without replacement from U.

The sequence $(Y_1, Y_2, \dots)$ will contain repeats, which can be represented as a sequence $\mathcal{C}(Y) = (\mathcal{C}_1(Y), \mathcal{C}_2(Y), \dots)$ of symbols from a set of 'codewords' $\mathbb{B} = \{\boxed{j} : 1 \le j \le N\}$, by the following procedure. Think of $\mathbb{B}$ as ordered: $\boxed{1} < \boxed{2} < \cdots < \boxed{N}$. The code $\mathcal{C}(Y)$ always starts with $\boxed{1}$. If $Y_2 = Y_1$ then $\mathcal{C}_2(Y) = \boxed{1}$, otherwise $\mathcal{C}_2(Y) = \boxed{2}$. And so on. In general, if a Y_i repeats an earlier Y_j then $\mathcal{C}_i(Y) = \mathcal{C}_j(Y)$; if Y_i is different from all previous Y_j's then it receives the smallest unused code symbol. For example, here is how it works for a typical Y:

Y:	u_7	u_3	u_9	u_7	u_2	u_3	u_3	u_{185}	$\dots$
X:	u_7	u_3	u_9		u_2			u_{185}	$\dots$
$\mathcal{C}(Y)$:	$\boxed{1}$	$\boxed{2}$	$\boxed{3}$	$\boxed{1}$	$\boxed{4}$	$\boxed{2}$	$\boxed{2}$	$\boxed{5}$	$\dots$

You should ignore the gaps in the X-vector; I inserted them just to align each X_j with its first appearance in the Y sequence. The corresponding positions in $\mathcal{C}(Y)$ contain a repetition of an earlier code symbol. For example, the second u_7 in the Y sequence has a gap in X and is coded as $\boxed{1}$ because $Y_1 = u_7$.

Together, X and $\mathcal{C}(Y)$ allow us to reconstruct Y: for $i = 1, \dots, N$ replace each $\boxed{i}$ in $\mathcal{C}(Y)$ by X_i, the ith element of X (ignoring the $\sqcup$ characters). More concisely, $Y_j = X_{\mathcal{C}(Y_j)}$, provided we ignore the little box around the code symbol.

I claim that X and $\mathcal{C}(Y)$ are independent.

Remark. Initially I thought the independence was obvious: knowledge of the pattern tells us nothing about the order in which the elements of U are first observed. For example, if $\mathcal{C}(Y) = (\boxed{1}, \boxed{2}, \boxed{3}, \boxed{1}, \dots)$ then we know that Y_1, Y_2, Y_3 are different elements of U and $Y_4 = Y_1$ but we have no information about which three elements of U were involved. Then I began to worry that this assertion was a bit too handwaving. It took me a while to come up with the more rigorous argument given in Problem [13].

Now back to <52>. Remember that $S_n = g(Y_1) + \dots g(Y_n)$ and $T_n = g(X_1) + \dots g(X_n)$. The sum S_n can be re-expressed using the counts

$$N_n(j) = \text{number of times } \boxed{j} \text{ appears amongst } \mathcal{C}(Y_1), \dots, \mathcal{C}(Y_n).$$

For example, if $n = 6$ and $Y = (u_7, u_3, u_9, u_7, u_2, u_3, \dots)$ then

$$(X_1, \dots, X_4) = (u_7, u_3, u_9, u_2) \quad \text{AND} \quad \mathcal{C}(Y) = (\boxed{1}, \boxed{2}, \boxed{3}, \boxed{1}, \boxed{4}, \boxed{2} \dots),$$

so that $N_6(1) = N_6(2) = 2$ and $N_6(j) = 1$ for $j = 3, 4$, which gives

$$g(Y_1) + \dots + g(Y_6) = 2g(X_1) + 2g(X_2) + g(X_3) + g(X_4).$$

Notice that we only need the counts up to $j = 6$, at most, because $(Y_1, \dots, Y_6)$ can involve at most 6 different elements $X_1, \dots, X_6$ of U. In general,

$$S_n = g(Y_1) + \dots + g(Y_n) = \sum\nolimits_{j=1}^{n} N_n(j) g(X_j)$$

and

<53> $$\mathbb{P} f\Big(S_n\Big) = \mathbb{P} f\left(\sum\nolimits_{j=1}^{n} N_n(j) X_j\right)$$

Unfortunately, the final expression is not symmetric in $X_1, \dots, X_n$; it is hard to see how it is related to $\mathbb{P} f(T_n)$. My method for determining patterns broke the symmetry but it can be restored using a sneaky trick. As the $N_n(j)$'s depend only on $\mathcal{C}(Y)$ they are independent of X. We could replace $(X_1, \dots, X_n)$ by any other random sequence $(\widetilde{X}_1, \dots, \widetilde{X}_n)$ that is independent of $\mathcal{C}(Y)$ and has the same distribution as $(X_1, \dots, X_n)$. For example, for any fixed permutation σ of $[\![n]\!] := \{j \in \mathbb{N} : j \le n\}$ we could use $(X_{\sigma(1)}, \dots, X_{\sigma(n)})$:

$$\mathbb{P} f\Big(S_n\Big) = \mathbb{P} f\left(\sum\nolimits_{j=1}^{n} N_n(j) X_{\sigma(j)}\right) \qquad \text{for each permutations } \sigma \text{ of } [\![n]\!].$$

We can even average out over the uniform distribution $\mathbb{Q}$ on the set of all permutations of $[\![n]\!]$ then use the JENSEN inequality to take the $\mathbb{Q}$ integral inside the convex function:

$$\mathbb{P} f\Big(S_n\Big) = \mathbb{Q}^\sigma \mathbb{P} f\left(\sum\nolimits_{j=1}^{n} N_n(j) X_{\sigma(j)}\right) \ge \mathbb{P} f\left(\mathbb{Q}^\sigma \sum\nolimits_{j=1}^{n} N_n(j) X_{\sigma(j)}\right)$$

From the facts that $\sum_{j=1}^{n} N_n(j) = n$ and

$$\mathbb{Q}^\sigma g(X_{\sigma(j)}) = n^{-1} \sum\nolimits_{i=1}^{n} g(X_i) = n^{-1} T_n \qquad \text{for } 1 \le j \le n$$

□ it now follows that $\mathbb{P} f(S_n) \ge \mathbb{P} f(T_n)$, as asserted.

3.9 Problems

For Problems [1] through [5], the function $\phi(x)$ denotes the $N(0,1)$ density and $\bar{\Phi}(x) := \int_x^\infty \phi(t)\,dt$; the functions $\mathcal{R}(\cdot)$ and $\rho(\cdot)$ and $r(\cdot)$ are defined on $\mathbb{R}$ by $1/\rho(x) := \mathcal{R}(x) = \bar{\Phi}(x)/\phi(x)$ and $r(x) := \rho(x) - x$.

[1] Inequality <11> is just the initial part of a sequence of upper and lower bounds for $\mathcal{R}(x)$, which are apparently due to Laplace (see Notes). Each bound is of the form $p(1/x)$ with p a polynomial.

(i) Show that $p(1/x) > \mathcal{R}(x)$ for all $x > 0$ if

$$<54> \qquad -\frac{d}{dt}\left(p(1/t)\phi(t)\right) > \phi(t) \qquad \text{for all } t > 0$$

and $p(1/x) < \mathcal{R}$ for all $x > 0$ if

$$<55> \qquad -\frac{d}{dt}\left(p(1/t)\phi(t)\right) < \phi(t) \qquad \text{for all } t > 0$$

Hint: $\int_x^\infty$.

(ii) Show that<54> holds if and only if $p(s) + s^3\dot{p}(s) > s$ for all $s > 0$. Characterize <55> by the reverse inequality.

(iii) Define a sequence of monomials by $\Delta_0(s) = s$ and $\Delta_k(s) = -s^3\dot{\Delta}_{k-1}(s)$ for $k \geq 1$. Define $a_k := 1 \times 3 \times \cdots \times (2k-1)$. Show that

$$\Delta_k(s) = (-1)^k a_k s^{2k+1} = -(2k-1)s^2\Delta_{k-1}(s) \qquad \text{for } k = 1, 2, \ldots.$$

(iv) Define $p_k(s) = \sum_{i=0}^k \Delta_i(s)$. Show that $p_k(s) + s^3\dot{p}_k(s) = s - \Delta_{k+1}(s)$.

(v) Conclude that $p_k(1/x) > \mathcal{R}(x) > p_{k+1}(1/x)$ for each even k.

[2] Show that $\bar{\Phi}(x) \leq \frac{1}{2}e^{-x^2/2}$ for $x \geq 0$. Hint: From Theorem <20>(i) we have $\mathcal{R}(0) > \mathcal{R}(x)$.

[3] (Birnbaum, 1942). Use the CAUCHY-SCHWARZ inequality and some results from Theorem <20> to show that

$$\phi(x)^2 = \left(\int_x^\infty t\sqrt{\phi(t)}\sqrt{\phi(t)}\,dt\right)^2 \leq \left(x\phi(x) + \bar{\Phi}(x)\right)\bar{\Phi}(x).$$

Deduce that $\rho(x)^2 \leq x\rho(x) + 1$, so that $|\rho(x) - x/2| \leq \sqrt{1 + x^2/4}$, which implies $\rho(x) \leq \left(x + \sqrt{x^2+4}\right)/2$.

[4] (Sampford, 1953) Define $\psi(x) := 2r(x)^2 + xr(x)$, as in the proof of Theorem <20>, which showed that $\psi(x) > 1$ for all $x \in \mathbb{R}$. Argue that $r(x)$ cannot belong to the closed interval $I_x := \{t \in \mathbb{R} : 2t^2 + xt - 1 \leq 0\}$, which has endpoints $\left(-x \pm \sqrt{x^2+8}\right)/4$. Deduce that $r(x) > (-x +$

$\sqrt{x^2+8})/4 = 2/(x+\sqrt{x^2+8}\,)$ and $\rho(x) > (3x+\sqrt{x^2+8})/4$. Note: $r(x) > 0$.

[5] Suppose $Z_1, \ldots, Z_n$ are random variables, each distributed $N(0,1)$ but, for the moment, not necessarily independent. Define $M_n = \max_{i\le n} Z_i$.

(i) Even without independence the MGF approach also gives an upper bound a_n for the expected value of M_n, via Jensen's inequality: for each $\lambda > 0$,

$$\exp(\lambda \mathbb{P} M_n) \le \mathbb{P} e^{\lambda M_n} = \mathbb{P} \max_i e^{\theta Z_i} \le \sum\nolimits_i \mathbb{P} e^{\lambda Z_i} = n e^{\lambda^2/2}.$$

Deduce that $\mathbb{P} M_n \le \inf_{\lambda>0} \left(\log n + \lambda^2/2\right)/\lambda = a_n = \sqrt{2\log(n)}$. The case where $Z_i = Z_1$ for all i shows that the bound is not sharp in general.

(ii) If the Z_i's are independent, show that $\mathbb{P} M_n \ge a_n - c\log(a_n)/a_n$ for some constant c, if n is large enough. First show that

$$M_n \ge \max_{i\le n} Z_i^+ - \left(\sum\nolimits_{i\le n} |Z_i|\right) \{M_n \le 0\}$$

so that $\mathbb{P} M_n \ge \mathbb{P} \max_{i\le n} Z_i^+ - n\mathbb{P}|Z_1|/2^{n-1}$. Then argue that

$$O(n/2^{n-1}) + \mathbb{P} M_n \ge \mathbb{P} \max_{i\le n} Z_i^+ = \int_0^\infty \mathbb{P}\{M_n > t\}\, dt \ge x_n \mathbb{P}\{M_n > x_n\}.$$

Look at Example <18> for a way to choose x_n.

[6] Suppose $Z_1, \ldots, Z_n$ are independent random variables, each distributed $N(0,1)$.

(i) Show that $\mathbb{P}\{\max_{i\le n} |Z_i| \le x_n\} = \left(1 - 2\bar{\Phi}(x_n)\right)^n$.

(ii) Mimic the argument from Example <18> to deduce that $\max_{i\le n} |Z_i|$ concentrates near a_n.

[7] Suppose X has a standard exponential distribution.

(i) Show that $\mathbb{P}\{X \ge x\} = e^{-x}$ for all $x \ge 0$ and $\mathbb{P} X = 1$.

(ii) Show that the method from Section 3.1 gives $\mathbb{P}\{X \ge x\} \le (ex)e^{-x}$ for $x \ge 1$.

(iii) What bound does the method give for $0 \le x < 1$?

Remark. Maybe we need a little bit more than tails decaying like e^{-x} for the MGFmethod to deliver interesting bounds.

[8] Suppose $A_1, \ldots, A_n$ are independent events with $\sum_i \mathbb{P} A_i = \epsilon$, for a small ϵ. Show that

$$\mathbb{P} \cup_i A_i = 1 - \exp\left[\sum\nolimits_i \log(1 - \mathbb{P} A_i)\right] \ge 1 - e^{-\epsilon} = \epsilon - O(\epsilon^2).$$

[9] Use notation from Lemma <29>. Suppose W has distribution P. Suppose we can construct a new random variable $\widetilde{W}$ such that the joint distribution of $(W, \widetilde{W})$ is the same as the joint distribution of $(\widetilde{W}, W)$ (an "exchangeable pair"). Write Δ for $\widetilde{W} - W$ and define

$$\mu_1(w) := \mathbb{P}(\Delta \mid W = w) \quad \text{AND} \quad \mu_2(w) := \mathbb{P}\left(\Delta^2 \mid W = w\right).$$

Assume that $\|h\|_{\text{Lip}}$ is finite.

(i) For real x and $z := y - x$ define an antisymmetric function

$$F(x, y) := (y - x)\,[f(x) + f(y)] = z\left[2f(x) + z\dot{f}(x)\right] + R(x, z),$$

where $R(x, z) := f(x + z) - f(x) - z\dot{f}(x)$. Show that

$$0 = \mathbb{P}F(W, \widetilde{W}) = P^w\left(2\mu_1(w)f(w) + \mu_2(w)\dot{f}(w)\right) + \mathbb{P}R(W, \Delta).$$

(ii) Suppose $\mu_1(w) = -\lambda w + R_1(w)$ and $\mu_2(w) = 2\lambda + R_2(w)$ for some positive constant λ, with R_1 and R_2 small. Deduce from (i) that

$$2\lambda|\mathbb{P}H(W)| \le 2\|h\|_{\text{Lip}}\mathbb{P}|\Delta|^3 + \sqrt{\pi/2}\|H\|_\infty\mathbb{P}|R_1(W)| + 2\|H\|_\infty\mathbb{P}|R_2(W)|.$$

Compare with Stein (1986, pp. 33–35).

[10] Let $\{\xi_{n,i} : i \in [\![n]\!],\ n \in \mathbb{N}\}$ be a triangular array of random variables, independent within each row. It is a classical result (see Petrov, 1975, §IV.4 and Le Cam, 1986) that if $\max_i \mathbb{P}\{|\xi_{n,i}| > \epsilon\} \to 0$ for each $\epsilon > 0$ then $\sum_i \xi_{n,i} \rightsquigarrow N(0, 1)$ iff for each $\epsilon > 0$ we have: $\mathbb{P}\{\max_i |\xi_{n,i}| > \epsilon\} \to 0$ and $\sum_i \text{var}\,(\xi_{n,i}\{|\xi_{n,i}| \le \epsilon\}) \to 1$ and $\sum_i \text{var}\,(\xi_{n,i}\{|\xi_{n,i}| \le \epsilon\}) \to 1$. Thus, for the purpose of developing normal approximations for sums of independent random variables it suffices to consider the case where $\xi_1, \ldots, \xi_n$ are independent random variables with: $|\xi_i| \le \epsilon$ for a fixed $\epsilon > 0$; $\mathbb{P}\xi_i = 0$; and $\sum_i \text{var}(\xi_i) = 1$. This Problem will show how the bounds from Problem [9] can handle such a situation.

Let $\xi_1, \ldots, n, \eta_1, \ldots, \eta_n, J$ be independent random variables, with η_i having the same distribution as ξ_i and J being uniformly distributed on $[\![n]\!]$. Define $W := \sum_i \xi_i$ and $\widetilde{W} = \sum_i \{J \ne i\}\xi_i + \{J = i\}\eta_i$, so that $\Delta := \widetilde{W} - W = \sum_i \{J = i\}(\eta_i - \xi_i)$.

(i) Show that $\mu_1(w) = -w/n$ and $\mu_2(w) = 2n^{-1} + R_2(w)$ where $R_2(w) = n^{-1}P_w\sum_i\left(\xi_i^2 - \sigma_i^2\right)$, where $\sigma_i^2 := \mathbb{P}\xi_i^2$ and P_w is shorthand for the conditional distribution of $\widetilde{W}$ given $W = w$.

(ii) Show that

$$(n\mathbb{P}|R_2(W)|)^2 \le \mathbb{P}\left(\sum\nolimits_i \xi_i^2 - \sigma_i^2\right)^2 \le \sum\nolimits_i \mathbb{P}\xi_i^4 \le \epsilon^2,$$
$$n\mathbb{P}|\Delta|^3 \le |\eta_i - \xi_i|^3 \le 4\epsilon.$$

(iii) Deduce that $|\mathbb{P}H(W)| \le C\max(\|h\|_{\text{Lip}}, \|H\|_\infty)\epsilon$ for some universal constant C.

Remark. The previous Problem is just a sanity check for Problem [9]. The real strength of Stein's approach is its ability to handle problems caused by dependence, which defy classical methods.

[11] Suppose $f(\lambda) := \frac{1}{2}\lambda^2/(1-\lambda)$ for $0 \le \lambda < 1$. Show that f is convex with conjugate $f^*(y) = 1 + y - (1+2y)^{1/2}$ for $y \ge 0$ by the following steps.

(i) Show that $f(1-s) = 1/(2s) - 1 + s/2$ for $0 < s \le 1$. Deduce that f is convex.

(ii) Show that $f^*(y) = 1 + y - \inf_{0<s\le 1}\big(ys + 1/(2s) + s/2\big)$, with the infimum achieved at $s = (1+2y)^{-1/2}$.

(iii) Define $w := (1+2y)^{1/2}$. Show that

$$y/w + w/2 + 1/(2w) = \frac{w^2 - 1 + w^2 + 1}{2w} = w.$$

(iv) Show that

$$f^*(y) = \frac{y^2}{1 + y + (1+2y)^{1/2}} \ge \frac{y^2}{2+2y} \qquad \text{for } y \ge 0.$$

[12] (compare with Boucheron et al., 2013, p37) Suppose $S = X_1 + \cdots + X_n$ is a sum of independent random variables for which $\tau_i^2 := \mathbb{P}X_i^2 < \infty$ for each i and there exists a finite $\mathcal{V}$ with $\sum_i \tau_i^2 \le \mathcal{V}$ and

<56> $$\sum\nolimits_i \mathbb{P}(X_i^+)^k \le k!\mathcal{V}/2 \qquad \text{for } k \ge 3.$$

Show that $S - \mathbb{P}S \in \text{subGamma}(\mathcal{V})$ by the following steps.

(i) Recall that $\mathbb{f}(x) := e^x - 1 - x = \frac{1}{2}x^2\mathbb{a}(x)$ with $\mathbb{a}(x) \le 1$ for $x \le 0$. Deduce that $e^x \le 1 + x + x^2/2 + \sum_{k\ge 3}(x^+)^k/k!$ for all real x. Hint: The $x^2/2$ takes care of $\mathbb{f}(x)$ for $x \le 0$.

(ii) Deduce that, for $0 \le \lambda < 1$,

$$\mathbb{P}e^{\lambda X_i} \le 1 + \lambda\mathbb{P}X_i + \tfrac{1}{2}\sigma_i^2\lambda^2 + \sum\nolimits_{k\ge 3}\lambda^k\mathbb{P}(X_i^+)^k/k!.$$

(iii) Sum over i then invoke <56> to deduce that

$$\log \mathbb{P}e^{\lambda(S-\mathbb{P}S)} \le \mathcal{V}\lambda^2/2 + \sum\nolimits_{k\ge 3}\lambda^k\mathcal{V}/2 = (\mathcal{V}/2)\lambda^2/(1-\lambda)$$

if $0 \le \lambda < 1$. That is, $S - \mathbb{P}S \in \text{subGamma}(\mathcal{V})$.

(iv) Suppose <56> is modified to include a positive scaling factor B, that is, $\sum_i \mathbb{P}(X_i^+)^k \le k!\mathcal{V}B^{k-2}/2$ for $k \ge 3$. Deduce that

$$\log \mathbb{P}e^{\lambda(S-\mathbb{P}S)} \le \tfrac{1}{2}\mathcal{V}\lambda^2/(1-\lambda B)$$

for $0 \le \lambda B < 1$. Hint: Consider S/B.

[13] Here is a rigorous way to establish independence of X and $\mathcal{C}(Y)$ in Example <51>. Notation (derived from Section 3.8):

(a) Regard Y as the identity map (that is, $Y(y) = y$) on $U^{\mathbb{N}}$ equipped with its product sigma-field and product measure $\mathbb{P} = \nu^{\mathbb{N}}$, where ν denotes the uniform distribution on U.

(b) $\mathbb{B} = \{\boxed{i} : i = 1, \ldots, N\}$, the code symbols. Regard $\mathcal{C}$ as a measurable map from $U^{\mathbb{N}}$ into the product space $\mathbb{B}^{\mathbb{N}}$ (equipped with its product sigma-field).

(c) $\mathcal{W}$ = the set of all permutations of U. (Thus $|\mathcal{W}| = N!$.) If $x = (x_1, \ldots, x_N) \in \mathcal{W}$ and $\boxed{i} \in \mathbb{B}$, interpret $x_{\boxed{i}}$ to mean x_i.

(d) Treat $X = X(y)$ as the map from $U^{\mathbb{N}}$ into the set $\mathcal{W}$ that is defined by discarding repetitions of each y_i after its first appearance in y.

(e) For $y = (y_1, y_2, \ldots) \in U^{\mathbb{N}}$ and $B = (b_1, b_2, \ldots) \in \mathbb{B}^{\mathbb{N}}$ define $\mathcal{C}(y) \,[\!]\, n = (\mathcal{C}_1(y), \ldots, \mathcal{C}_n(y))$ and $B \,[\!]\, n = (b_1, \ldots, b_n)$.

The argument:

(i) Show that the distribution of X is uniform on $\mathcal{W}$. Deduce that

$$\mathbb{P}\{X_j = x_j, \ldots, X_N = x_N\} = \frac{1}{N} \times \frac{1}{N-1} \times \frac{1}{N-j+1} = \frac{1}{(N)_j}$$

for each $x \in \mathcal{W}$ and $1 \le j \le N$. (Exchangeability helps.)

(ii) Show that the distribution of $\mathcal{C}(Y)$ concentrates on the set $\mathcal{P}$ of all *feasible* B's in $\mathbb{B}^{\mathbb{N}}$, that is, those B that start with $\boxed{1}$ and for all i, j with $1 \le j < i \le N$ the codeword $\boxed{j}$ first appears in B before $\boxed{i}$.

(iii) Suppose $B \in \mathcal{P}$ and $B \,[\!]\, n$ uses only the k code symbols $\boxed{1}, \ldots, \boxed{k}$. Show that

$$\mathbb{P}\{\mathcal{C}(Y) \,[\!]\, n = B \,[\!]\, n\} = \frac{N(N-1)\ldots(N-k+1)}{N^n} = \frac{(N)_k}{N^n}.$$

Hint: Each repeat of a codeword corresponds to an event that has probability $1/N$ and the first appearance of codeword $\boxed{j}$ indicates a selection from a set $N - j + 1$ elements from U.

(iv) Suppose $x = (x_1, \ldots, x_N) \in \mathcal{W}$ and $B \in \mathcal{P}$, with $B \,[\!]\, n$ using only the k code symbols $\boxed{1}, \ldots, \boxed{k}$. Justify the following assertions.

Define

$$A = \{y : \mathcal{C}(y) \,[\!]\, n = B \,[\!]\, n, \quad X_i = x_i \text{ for } i = 1, \ldots, k\ \}.$$

The assumption about B means that $y_1, \ldots, y_n$ select only k distinct elements of U, namely $U_1 := \{x_i : i = 1, \ldots, k\}$, with first selections occurring in the order $(x_1, \ldots, x_k)$. Thus

$$\mathbb{P}A = \mathbb{P}\{y : y_i = x_{b_i} \text{ for } i = 1, \ldots, n\ \} = (1/N)^n.$$

Conditional on the occurrence of A, the remaining observations $y_{n+1}, y_{n+2}, \ldots$ are left to select each element of $U_2 := \{x_i : n+1 \le i \le N\}$. If we also require $X_i = x_i$ for $k+1 \le i \le N$ then we have specified the order of first selections of the elements of U_2, namely $(x_{k+1}, \ldots, x_N)$. Thus

$$\mathbb{P}\{X_i = x_i \text{ for } k+1 \le i \le N \mid A\} = \frac{1}{N-k} \times \frac{1}{N-k-1} \times \cdots \times \frac{1}{1} = \frac{1}{(N-k)!}.$$

Combining these two results we get

$$\begin{aligned}
&\mathbb{P}\{y : \mathcal{C}(y) \,[\![\, n = B \,[\![\, n,\, X = (x_1, \ldots, x_N)\} \\
&\quad = \mathbb{P}\left(A \cap \{X_i = x_i \text{ for } k+1 \le i \le N\ \}\right) \\
&\quad = (1/N)^n \times \frac{1}{(N-k)!} = \frac{N(N-1)\ldots(N-k+1)}{N^n} \times \frac{1}{N!} \\
&= \mathbb{P}\{\mathcal{C}(Y) \,[\![\, n = B \,[\![\, n\} \times \mathbb{P}\{X = x\}.
\end{aligned}$$

It follows that $\mathcal{C}(Y)$ and X are independent.

3.10 Notes

Bennett (1962) and Hoeffding (1963) are good sources for a host of exponential inequalities. Massart (2003, Chapter 2) and Boucheron, Lugosi, and Massart (2013, Chapter 2) persuaded me that it is better to have the relevant results collected together in one place, rather than deriving them on an ad hoc basis.

Many authors seem to credit Chernoff (1952) with the moment generating trick in <1>, even though the idea is obviously much older. In the 1952 paper Chernoff first noted that Cramér (1938) had already established excellent results for sums of independent random variables using the MGF method. Then he proceeded (page 495) to list a few basic facts:

> Since the results of Cramér are extremely more powerful tha[n] we require here and the (finite) existence of third order moments is not necessary for the results that we desire, we shall state and briefly outline a proof of Theorem 1. Before doing this we shall first formally state some notation and lemmas which we shall use throughout this paper. These lemmas state known results which are rather obvious, depending mainly on Lebesgue's Theorem on integration of monotone sequences [reference to the Saks book].

Remark. Cramér's 1938 paper summarized asymptotic approximations to the tail probabilities for a sum of independent random variables, rather than bounds on those tail probabilities; details appeared in Chapter 7 of Cramér (1937). See Cramér (1976, §4.9) for comments about the conference where he presented the 1938 paper.

Amongst the known results listed by Chernoff was a special case of the 'extended Tchebycheff inequality' (near the start of the proof of his Theorem 1) for which he cited the German original of Kolmogorov (1933). (The inequality, which corresponds to my <1>, appeared in §IV.3 of Kolmogorov's book.) In a subsequent biographical article Chernoff (2004, page 3) acknowledged that while working on his 1952 paper he had originally been ignorant of the 1938 Cramér paper.

In the 1933 book Kolmogorov gave no source for the 'extended Tchebycheff' inequality, although the 1927 edition of Sergei Bernstein's probability book was listed in his Bibliography. Moreover, in a paper celebrating Bernstein's eightieth birthday, Kolmogorov and Sarmanov (1960) noted:

> 3. Beginning in 1921 Sergei Natanovich published a number of papers dealing with various special problems in the application of probability theory ... and in 1927 appeared the first edition of the fundamental text "The Theory of Probability", which was reprinted with large supplements in 1934 and 1946. At the mathematical congresses in Moscow (1927) and Zürich (1932) Sergei Natanovich delivered long survey reports on the problems of probability theory. We ... emphasize that at this time such a wide range of work on all the fundamental theoretical and applied problems of probability theory was a totally new thing. ... It is natural that the theoretical and applied works of Sergei Natanovich and his text in probability theory have determined to a considerable degree the development of research in probability theory in the USSR.

And then

> 4. A whole series of papers by Sergei Natanovich are connected with the strengthening of Chebyshev's inequality [citing papers from 1918, 1924, and 1937] and the calculation of the error in the Laplace formula ...

I took the quotes from the SIAM translation of Volume V number 2 of the Russian original.

Now for the view from the West. The proof of the Bernstein inequality given by Uspensky (1937, pp. 204–206) used the MGF method. He prefaced the "Indication of the Proof" by the remark "S. Bernstein has shown that Tchebycheff's inequality can be considerably improved". I thank Elena Khusainova, who translated four pages from Bernstein's probability book for me, clearly showing that Uspensky was following that book, which appeared in his list of references (page 207).

Hoeffding (1963, page 14) gave Bernstein credit for the MGF approach:

> The method employed to derive the inequalities, which has often been used (apparently first by S. N, Bernstein),...

Bennett (1962) had also cited (page 34) Bernstein's 1927 book (and an earier paper) with the comment (page 35):

> Bernstein's original work was published in Russian, and appears to be unobtainable. It is reported—indirectly—by [Craig, 1933] ... and by [Uspensky (1937, pages 204–206)] who indicates the proof in a series of exercises. The inequality is mentioned or quoted without proof by ... Apart from these brief references, Bernstein's inequality seems to have escaped notice in the English-speaking world.

Craig had commented (his page 94):

> Another interesting and important attempt in this direction due to S. N. Bernstein seems to have generally escaped attention in the English-speaking world, at least, since it has been published only in Russian.

with the footnote

> "Bernstein, S., Theory of Probability, (Moscow, 1927), pp. 159-165. The present account of this work of Bernstein is taken from a lecture of Professor J. V. Uspensky."

Craig also mentioned that his paper was written while he was at Stanford University, where Uspensky was a mathematics faculty member, until his death in 1947.

I think it abundantly clear that credit for the MGF method should go to Bernstein, not Chernoff.

Inequality <11> for $\mathcal{R}(x) := \bar{\Phi}(x)/\phi(x)$ is classical. It corresponds to the first two terms in the asymptotic expansion derived in Problem [1]. See Baricz (2008) or Gasull and Utzet (2014) for recent examples, which include some history. Problem [1] itself corresponds to an analogous result for the error function (that is, $\int_0^x e^{-t^2} dt$) presented by Laplace in his Celestial Mechanics, reprinted 1805 in Volume IV, Book X, Chapter 1, §5 of his collected works (pages 489–492 in the Bowditch translation). He also developed a continued fraction expansion.

The ratio $\mathcal{R}(x)$ is often called the "Mills ratio", because it was discussed by Mills (1926). Actually Mills was just constructing a table of $\mathcal{R}$, using earlier tables for the normal distribution function and numerical methods proposed by other authors. There has been a long history of authors deriving bounds for $\mathcal{R}$, such as the upper bound from Problem [3] and the lower bound from Problem [4].

I first came to appreciate the value of having global control over ρ while working out the theory for Carter and Pollard (2004) and UGMTP Appendix D. Most of the proof for Theorem <20> comes from those two sources, with corrections, plus ideas adapted from Sampford (1953).

The subGamma idea is clearly present in the derivation of Bernstein's inequality from the bernstein moment assumption (see Example <39>) but, to my knowledge, Boucheron, Lugosi, and Massart (2013, page 28) were the first to anoint it as a general concept.

Okamoto (1959) stated the Binomial tail bounds <47> and its analog for the lower tail in a slightly different form, with the comment that "We shall state two Lemmas the first of which is a corollary of a theorem given by [Chernoff 1952, Theorem 1]". He then derived several more attractive bounds that could be derived from the mgf bound. He omitted the calculus (which I provided in Section 3.7) for the mgf bound; he only gave the details for the weaker upper bounds. He also commented that his proof simplified a "tedious, although elementary" calculation by Uspensky (1937, page 102). It seems strange to me that Okamoto did not also cite Uspensky (1937, page 204).

References

Baricz, Á. (2008). Mills' ratio: monotonicity patterns and functional inequalities. *Journal of Mathematical Analysis and Applications 340*(2), 1362–1370.

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association 57*, 33–45.

Birgé, L. and P. Massart (1998). Minimum contrast estimation on sieves: exponential bounds and rates of convergence. *Bernouilli 4*(3), 329–375.

Birnbaum, Z. W. (1942). An inequality for Mills' ratio. *Ann. Math. Statist. 13*, 245–246.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Carter, A. and D. Pollard (2004). Tusnády's inequality revisited. *Annals of Statistics 32*(6), 2731–2741.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations. *Annals of Mathematical Statistics 23*(4), 493–507.

Chernoff, H. (2004). Some reminiscences of my friendship with Herman Rubin. In A. DasGupta (Ed.), *A Festschrift for Herman Rubin*, Volume 45 of *Lecture Note-Monograph Series*, pp. 1–4. Institute of Mathematical Statistics.

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics 23*, 315–3455.

Craig, C. C. (1933). On the Tchebychef inequality of Bernstein. *The Annals of Mathematical Statistics 4*(2), 94–102.

Cramér, H. (1937). *Random Variables and Probability Distributions*. Cambridge University Press.

Cramér, H. (1938). Sur un nouveau théorème-limite de la théorie des probabilités. In *Colloque consacré à la théorie des probabilités, Actualités scientifiques et industrielles 736*, pp. 2–23. Hermann & Cie, Paris. Available at arXiv:1802.05988v3 in the form of a TeXed version of the original paper in French paired with an English translation by Hugo Touchette. Reprinted in: H. Cramér, Collected Works, A. Martin-Löf (Ed.), Vol. II, Springer, Berlin, 1994, p. 895-913.

Cramér, H. (1976). Half a century with probability theory: some personal recollections. *Annals of Probability 4*, 509–546.

Gasull, A. and F. Utzet (2014). Approximating Mills ratio. *Journal of Mathematical Analysis and Applications 420*(2), 1832–1853.

Hall, P. and A. D. Barbour (1984). Reversing the Berry-Esseen inequality. *Proceedings of the American Mathematical Society 90*(1), 107–110.

Ho, S.-T. and L. H. Y. Chen (1978). An L_p bound for the remainder in a combinatorial central limit theorem. *Annals of Probability 6*(2), 231 – 249.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*, 13–30.

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer-Verlag. Second English Edition, *Foundations of Probability* 1950, published by Chelsea, New York.

Kolmogorov, A. N. and O. V. Sarmanov (1960). The work of S. N. Bernshtein on the theory of probability (on his eightieth birthday). *Theory of Probability and Its Applications 5*, 197–203.

Laurent, B. and P. Massart (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics 28*(5), 1302–1338.

Le Cam, L. (1986a). *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag.

Le Cam, L. (1986b). The central limit theorem around 1935. *Statistical Science 1*, 78–96.

Leadbetter, M. R., G. Lindgren, and H. Rootzén (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag.

Massart, P. (2003). *Concentration Inequalities and Model Selection*, Volume 1896 of *Lecture Notes in Mathematics*. Springer Verlag. Lectures given at the 33rd Probability Summer School in Saint-Flour.

Massart, P. (2006). *Concentration Inequalities and Model Selection*, Volume 1896 of *Lecture Notes in Mathematics*. Springer Verlag. Lectures given at the 33rd Probability Summer School in Saint-Flour, July 2003.

Mills, J. P. (1926). Table of the ratio: area to bounding ordinate, for any portion of normal curve. *Biometrika 18*, 395–400.

Okamoto, M. (1959). Some inequalities relating to the partial sum of binomial probabilities. *Annals of the institute of Statistical Mathematics 10*(1), 29–35.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50*(302), 157–175.

Petrov, V. V. (1975). *Sums of Independent Random Variables*. Springer-Verlag. Enlish translation from 1972 Russian edition.

Pitman, J. (1993). *Probability*. Springer.

Sampford, M. R. (1953). Some inequalities on Mill's ratio and related functions. *The Annals of Mathematical Statistics 24*(1), 130–132.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. Sixth Berkeley Symp. Math. Stat. Prob.*, Volume 2, pp. 583–602.

Stein, C. (1986). *Approximate Computation of Expectations*, Volume 7 of *Lecture Notes–Monograph series*. Institute of Mathematical Statistics.

Uspensky, J. V. (1937). *Introduction to Mathematical Probability*. McGraw-Hill.

Wu, Y. and H. H. Zhou (2021). Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations. *Mathematical Statistics and Learning 4*(3), 143–220.

Chapter 4
Poisson-Binomial

SECTION 4.1 *introduces the idea that the Binomial distribution is more spread out than each* POISSON-BINOMIAL *distribution with the same number of trials and the same expected value.*

SECTION 4.2 *identifiies the* POISSON-BINOMIAL *with the set of distributions whose probability generating functions are polynomials with all their roots real.*

SECTION 4.3 *presents some classical facts about polynomials with real roots.*

SECTION 4.4 *shows that all the members of a large family of distributions defined by sampling without replacement, which includes the hypergeometric, have* POISSON-BINOMIAL *distributions.*

SECTION 4.5 *shows that the* BINOMIAL *distribution is almost unimodal, with mode(s) close to the mean.*

SECTION 4.6 *shows that the* POISSON-BINOMIAL *distribution is almost unimodal, with mode(s) close to the mean.*

SECTION 4.7 *shows that* $\mathbb{P}\{\text{PBin} \geq \mu + x\} \leq \mathbb{P}\{\text{Bin} \geq \mu + x\}$ *for* x *slightly larger than zero, when both distributions have mean* μ*.*

SECTION 4.8 *shows that the median of a* POISSON-BINOMIAL *is very close to its expected value.*

4.1 The extreme Binomial

Suppose $S := X_1 + \cdots + X_n$, a sum of independent random variables that take values in $\{0, 1\}$. If $X_i \sim \text{BER}(p)$ for each i then S has a $\text{BIN}(n, p)$ distribution: $\mathbb{P}\{S = k\} = \binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, \ldots, n$. The symbol $\text{BIN}(n, p)$ identifies a particular probabilty measure that concentrates on $[\![0, n]\!] := \{k \in \mathbb{Z} : 0 \leq k \leq n\}$. If $X_i \sim \text{BER}(p_i)$, with the p_i's allowed to be different for each i, then S has a POISSON-BINOMIAL distribution, $\text{PBIN}(\mathbf{p})$ with $\mathbf{p} := (p_1, \ldots, p_n)$.The symbol $\text{PBIN}(\mathbf{p})$ identifies another probabilty measure that concentrates on $[\![0, n]\!]$. The expression for the mass placed at k has a slightly messy form:

$$\mathbb{P}\{S = k\} = A_{k,n}(\mathbf{p}) := \sum\nolimits_{|J|=k} \Big(\prod\nolimits_{j \in J} p_j \Big) \Big(\prod\nolimits_{j \notin J} (1 - p_j) \Big), \tag{1}$$

where $\sum_{|J|=k}$ denotes summation over all $\binom{n}{k}$ subsets J of $[\![n]\!]$ of size k. Each $A_{k,n}$ is a symmetric function of $p_1, \ldots, p_n$, which could be expanded into a sum over all subsets J of $[\![n]\!]$ of terms $C_J \prod_{j \in J} p_j$ for a complicated

set of constants $\{C_J : J \subseteq [\![n]\!]\}$. There is also a recursive formula based on the fact that $T := X_2 + \cdots + X_n \sim \text{PBIN}(p_2, \ldots, p_n)$: for $0 \le k \le n$,

$$\begin{aligned} A_{k,n}(\mathbf{p}) &= \mathbb{P}\{X_1 + T = k\} = p_1 \mathbb{P}\{T = k-1\} + (1-p_1)\mathbb{P}\{T = k\} \\ &= p_1 A_{k-1,n-1}(p_2, \ldots, p_n) + (1-p_1) A_{k,n-1}(p_2, \ldots, p_n). \end{aligned} \tag{<2>}$$

Then $A_{\bullet,n-1}$ could be expanded using $A_{\bullet,n-2}$, and so on.

Remark. When the value of n is clear from context it can safely be omitted, making the formulas slightly easier to decipher. For example, <2> becomes

$$A_k(\mathbf{p}) = p_1 A_{k-1}(\mathbf{p}[-1]) + (1-p_1) A_k(\mathbf{p}[-1])$$

where, **R**-style, $\mathbf{p}[-1] := (p_2, \ldots, p_n)$.

Sections 4.7 and 4.8 will also be dealing with the tail probabilities

$$\mathbb{P}\{S \ge k\} = E_k(\mathbf{p}) = \sum_{j=k}^{n} A_j(\mathbf{p}) \qquad \text{for } 0 \le k \le n. \tag{<3>}$$

Despite the awkward form of <1>, a lot is known about PBIN($\mathbf{p}$). This Chapter will present a few choice examples, including some surprising facts about modes and medians.

You should keep in mind that a random variable W can have a PBIN distribution even if it is not defined as a sum of independent BER(p_i)'s. The random variable W lives on some $(\Omega, \mathcal{F}, \mathbb{P})$, a probability space that might be quite complicated; the distribution of W is a probability measure on $[\![0, n]\!]$ defined by the $\mathbb{P}\{W = k\}$ values. Example <13> will describe in detail such a situation, involving samples of size 2 taken without replacement from a finite population, where PBIN makes an unexpected appearance.

The previous Chapter (§§3.7, 3.8) presented several examples of distributions (including the PBIN and hypergeometric) whose MGF's are smaller than the MGF of a BINOMIAL distribution with the same expected value. Application of the MGF method then gave tail bounds that were smaller than the corresponding tail bounds for the BINOMIAL, inequalities that raised the question of whether the actual tail probabilities really are smaller than the actual tail probabilities for the BINOMIAL. This Chapter will show that they are: in a strong sense, the BINOMIAL is an extreme member of the set of PBIN distributions. In addition, Section 4.4 will show that the hypergeometric is a special case of the PBIN, an assertion that I initially found hard to believe. And in Section 4.8 you will learn a surprising fact about the mode and median of the BINOMIAL: both lie within a distance one of their expected values.

4.2 The PGF of the Poisson-Binomial

As before, suppose $S := X_1 + \cdots + X_n$, a sum of independent random variables with $X_i \sim \text{BER}(p_i)$. Some of the most interesting properties of the distribution for S, the $\text{PBIN}(p_1, \ldots, p_n)$, lie encoded within its probability generating function (PGF),

$$\begin{aligned} g(z) := \mathbb{P}z^S &= a_0 + a_1 z + \cdots + a_n z^n \qquad \text{where } a_k := \mathbb{P}\{S = k\} \\ &= \prod_{i=1}^{n} (1 - p_i + p_i z) \qquad \text{by independence of the } X_i\text{'s.} \end{aligned}$$

The factorization shows that the roots of the polynomial $g(z)$ are all nonnegative real numbers: $-(1-p_i)/p_i$ for each nonzero p_i. More precisely, if $n_0 := |\{i \in [\![n]\!] : p_i = 0\}|$ and $n_1 := |\{i \in [\![n]\!] : p_i = 1\}|$ then $g(z) = \sum_{n_1 \le k \le n-n_0} a_k z^k$ with $a_k > 0$ for $n_1 \le k \le n - n_0$. The n_0 zeros have no effect on S and the ones just contribute a constant n_1 to S. To avoid the trivial case where S is constant we might assume that $0 < p_i < 1$ for at least one i (or even all i). However, there are sometimes good reasons to allow 0's and 1's, as you will see in Sections 4.7 and 4.8.

The all-real-roots property characterizes the POISSON-BINOMIAL.

<4> **Lemma.** *Suppose a random variable W takes values only in $\{0, 1, \ldots, n\}$, with $a_n := \mathbb{P}\{W = n\} \neq 0$ and $h(z) := \mathbb{P}z^W$. If all n roots $r_1, \ldots, r_n$ of h are real then $\max_i r_i \le 0$ and W has a $\text{PBIN}(p_1, \ldots, p_n)$ distribution with $p_i = (1 - r_i)^{-1}$.*

Proof. For $s \ge 0$ the polynomial $h(s) = \sum_k s^k \mathbb{P}\{W = k\}$ is strictly increasing with $h(s) > h(0) = \mathbb{P}\{W = 0\} \ge 0$ for $s > 0$. Thus $r_i \le 0$ and $0 < p_i := 1/(1 - r_i) \le 1$ for each i. Replace $-r_i$ by $1/p_i - 1$ to get

$$h(z) = a_n \prod_i (z + (1 - p_i)/p_i) = \frac{a_n}{\prod_i p_i} \prod_i (1 - p_i + p_i z).$$

□ The equality $h(1) = 1$ forces $a_n = \prod_i p_i$.

4.3 Polynomials with real roots

There is an extensive literature, going back several centuries, regarding the roots of polynomials. See Pólya and Szegö (1976, Part 5, Chap 1) or, for a more exhaustive treatment, Prasolov (2004). Some of the results are useful for POISSON-BINOMIAL purposes.

To avoid constant repetition of the clumsy "all-real-roots", for $n \ge 1$ define $\mathcal{P}_n$ to be the set of all polynomials of degree n with nonnegative, real coefficients and all n of its roots real. (By an argument similar to the one used for the Proof of Lemma <4>, these assumptions force all roots to be ≤ 0.) Write $\mathcal{P}$ for $\cup_{n \in \mathbb{N}} \mathcal{P}_n$. If $f \in \mathcal{P}_n$ then

<5> $$f(z) = a_0 + a_1 z + \cdots + a_n z^n = a_n \prod_{i=1}^{n} (z - r_i) = a_n \prod_{j=1}^{K} (z - R_j)^{N_j}$$

with $a_n \neq 0$ and $\min_i a_i \geq 0 \geq \max_i r_i$, where $\{R_j : j \in [\![K]\!]\}$ are the distinct roots of f with multiplicities $N_j := |\{i \in [\![n]\!] : r_i = R_j\}|$.

<6> **Lemma.** *(i) If $f \in \mathcal{P}_n$ and $g \in \mathcal{P}_m$ then $fg \in \mathcal{P}_{n+m}$.*

(ii) If $f \in \mathcal{P}_n$ with $n \geq 2$ then its derivative, Df, belongs to $\mathcal{P}_{n-1}$.

(iii) If $f \in \mathcal{P}_n$ and $f(z) = \sum_{L \leq k \leq n} a_k z^k = a_n \prod_j (z - R_j)^{N_j}$ with $a_L > 0$, distinct roots $R_1 < R_2 < \cdots < R_K = 0$, and $N_K = L \geq 0$ then the polynomial

$$(\tau f)(z) := z^n f(1/z) = a_n + a_{n-1} z + \cdots + a_L z^{n-L} \in \mathcal{P}_{n-L},$$

has roots $1/R_j$ of multiplicity N_j for $1 \leq j < K$.

□ *(iv) If $f \in \mathcal{P}_n$ then $f(\alpha + \beta z) \in \mathcal{P}_n$ for all real constants α and $\beta \neq 0$.*

Proof. For (i): $f(z)g(z) = C \prod_{i \in [\![n]\!]} (z - r_i) \prod_{j \in [\![m]\!]} (z - s_j)$ for some positive constant C if $g(z)$ has roots $s_1, \ldots, s_m$.

For (ii): Without loss of generality suppose $R_1 < R_2 < \cdots < R_K \leq 0$ in <5>. By the mean value theorem, Df is a polynomial of degree $n-1$ with at least one zero in each of the $K-1$ open intervals (R_j, R_{j+1}). For each j, it also has a root of multiplicity $N_j - 1$ at R_j: if we factorize $f(z)$ as $(z - R_j)^{N_j} h(z)$ then

$$Df(z) = (z - R_j)^{N_j - 1} \left(N_j h(z) + (z - R_j) Dh(z) \right).$$

That gives $(K-1) + \sum_{j=1}^{K} (N_j - 1) = n - 1$ nonpositive roots, counting multiplicities.

For (iii): If $z \neq 0$,

$$\begin{aligned}(\tau f)(z) &= z^{N_1 + \cdots + N_K} \times a_n \left(\prod\nolimits_{1 \leq j < K} (z^{-1} - R_j)^{N_j} \right) z^{-N_K} \\ &= a_n \prod\nolimits_{1 \leq j < K} z^{N_j} (z^{-1} - R_j)^{N_j} = a_n \prod\nolimits_{1 \leq j < K} R_j^{N_j} (R_j^{-1} - z)^{N_j}.\end{aligned}$$

□ For (iv): $\prod_j (\alpha + \beta z - r_j)$ has roots $z = (r_j - \alpha)/\beta$.

The next two Examples make use of the ***falling factorial*** function, which is defined for real x as $(x)_0 = 1$ and

$$(x)_\ell = \prod\nolimits_{j=1}^{\ell} (x - j + 1) = x(x-1)\ldots(x-\ell+1) \qquad \text{for } \ell \in \mathbb{N}.$$

If x is a positive integer then $(x)_x = x!$ and $(x)_\ell = 0$ for $\ell \geq x + 1$. Falling factorials appears naturally if we differentiate z^k repeatedly: $Dz^k = (k)_1 z^{k-1}$, $D^2 z^k = (k)_2 z^{k-2}$,..., $D^\ell z^k = (k)_\ell z^{k-\ell}$ for $\ell \leq k$.

<7> **Example.** Let me show you one way to use Lemma <6> to transform a $\mathcal{P}_n$ polynomial $h(z) := a_L z^L + \cdots + a_n z^n$, with $L \geq 0$ and $\min_j a_j > 0$, into an $f(z) := \sum_{j=0}^m b_j z^j$, with $2 \leq m < n$, by modifying the coefficients. Assume $n - L \geq 3$ to ensure that $m \geq 2$, a condition needed for the next Example. This trick will be most useful in Section 4.4.

Let ℓ and N be positive integers for which $N \geq n$ and $\ell + 2L \leq n - 2$. Define $b_j := a_k(N-k)_\ell$ for $0 \leq k \leq n$ but note that $(N-k)_\ell > 0$ only if $N - k \geq \ell$, or $k \leq m$. The polynomial $f(z) = b_0 + \cdots + b_m z^m$ belongs to $\mathcal{P}_m$ because a sequence of operations from Lemma <6> transforms h into f:

$$\begin{aligned}
h(z) &:= a_L z^L + \cdots + a_n z^n, \\
h_1(z) &:= \tau h(z) = a_n + \cdots + a_L z^{n-L}, \\
h_2(z) &:= z^R h_1(z) = a_n z^R + \cdots + a_L z^N \qquad \text{where } R := L + N - n, \\
h_3(z) &:= D^\ell h_2(z) = (R)_\ell a_n z^{R-\ell} + \ldots\ (N)_\ell a_L z^{N-\ell}, \\
h_4(z) &:= \tau h_3(z) = a_L (N)_\ell + \cdots + a_n (L+R)_\ell z^{N-\ell-(L+R)} \qquad \text{if } R \geq \ell, \\
&= b_0 + b_1 z + \cdots + b_m z^m \qquad \text{where } m := N - \ell - (L+R) = n - (\ell + 2L).
\end{aligned}$$

□ Quite subtle, don't you think?

<8> **Example.** Let me show how to modify coefficients using Lemma <6> to transform a $\mathcal{P}_n$ polynomial $h(z) := \sum_{k=0}^n a_k z^k$, with $n \geq 3$ and $a_k \neq 0$ for all k, into an $f(z) := \sum_{j=0}^m b_j z^j \in \mathcal{P}_m$ with $2 \leq m < n$. This trick will be most useful in Section 4.4.

Let ℓ and N be positive integers for which $2 \leq m := N - \ell < n \leq N$. Define $b_k := (N-k)_\ell a_k$ for $0 \leq k \leq n$ but note that $(N-k)_\ell > 0$ only if $N - k \geq \ell$, or $k \leq m$. The polynomial f belongs to $\mathcal{P}_m$ because a sequence of operations from Lemma <6> transforms h into f:

$$\begin{aligned}
h(z) &:= a_0 + a_1 z + \cdots + a_n z^n, \\
h_1(z) &:= z^{N-n} \tau h(z) = a_n z^{N-n} + \cdots + a_k z^{N-k} + \cdots + a_0 z^N, \\
h_2(z) &= D^\ell h_1(z) = \sum_{k=0}^m (N-k)_\ell\, a_k z^{N-k-\ell}, \\
\tau h_2(z) &= \sum_{k=0}^m b_k z^k = f(z).
\end{aligned}$$

□ Quite subtle, don't you think?

Remark. Actually it would have sufficed to assume only that $h \in \mathcal{P}_n$ with $\min(a_0, a_n) > 0$, as an expansion of

$$a_n \prod_j \left(\sum_{i=0}^{N_j} \binom{N_j}{i} z^i (-R_j)^{N_j - i} \right)$$

would have shown.

The knowledge that a polynomial in $\mathcal{P}$ has only real roots tells us something about the coefficients of the polynomial, information that can be extracted by repeated appeals to Lemma <6>. The short story is that the next result is due to Isaac Newton; see the Notes for the long story.

<9> **Example.** Suppose $g(z) = a_0 + \cdots + a_n z^n \in \mathcal{P}_n$ with $a_0 \neq 0$. Here's how to prove

$$<10> \qquad a_k^2 \geq \frac{(n-k+1)(k+1)}{(n-k)k} a_{k-1}a_{k+1} > a_{k-1}a_{k+1} \qquad \text{for } 1 \leq k \leq n-1.$$

using Lemma <6>.

To slightly simplify notation write K for $k-1$. The K-fold derivative of g is a polynomial of degree $n-K$,

$$D^K g(z) = b_0 + b_1 z + \cdots + b_{n-K} z^{n-K} \qquad \text{where } b_j := a_{K+j}(K+j)_K,$$

which belongs to $\mathcal{P}_{n-K}$. The operator τ transforms $D^K g$ to

$$b_0 z^{n-K} + b_1 z^{n-K-1} + b_2 z^{n-K-2} + \cdots + b_{n-K} \in \mathcal{P}_{n-K}.$$

Another $\ell := n - K - 2$ differentiations leaves $D^\ell \tau D^K g(z)$ as a quadratic in $\mathcal{P}_2$:

$$\begin{aligned} &b_0 z^2 (\ell+2)_\ell + b_1 z (\ell+1)_\ell + b_2 (\ell)_\ell \\ &\quad = (\ell+2)_\ell (K)_k a_K z^2 + (\ell+1)_\ell (K+1)_k a_{K+1} z + +(\ell)_\ell (K+2)_K a_{K+2} \\ &\quad = \ell! K! (Az^2 + Bz + C) \end{aligned}$$

where

$$\begin{aligned} A &= (\ell+1)(\ell+2) \times a_k/(1\times 2) = \tfrac{1}{2}(n-k-1)(n-k) \times a_k \\ B &= (\ell+1)(K+1) \times a_{K+1} = (n-K-1)(K+1) \times a_{K+1} \\ C &= (K+1)(K+2) \times a_{K+1}/(1\times 2). \end{aligned}$$

By Lemma <6>, this quadratic has real roots, which happens if and only if $B^2 \geq 4AC$, which is equivalent to the first inequality in <10>. □

For a more direct coupling proof of $a_k^2 > a_{k-1}a_{k+1}$ see §13.1.

4.4 Sampling and the Poisson-Binomial

The traditional setting for the hypergeometric consists of an urn containing N balls, identical except for color, with R of those balls red and $B = N - R$ of them blue. One takes a sample of size n without replacement. Then W, the number of red balls in the sample, has a $\text{HYPER}(n, R, B)$ distribution:

$$<11> \qquad a_k := \mathbb{P}\{W = k\} = \binom{R}{k}\binom{B}{n-k} \Big/ \binom{N}{n} \qquad \text{for } k = 0, 1, \ldots, n.$$

Of course $a_k = 0$ if $k > R$ or $n - k > B$.

For the purposes of this Section the "urn" will be the subset $[\![N]\!]$ of $\mathbb{N}$ with $[\![R]\!]$ denoting the set of red balls. The following Theorem is a tiny variation on an amazing result due to Vatutin and Mikhailov (1983).

<12> **Theorem.** *Suppose F and G are independent random subsets of $[\![N]\!]$, with $|F|$, the size of F, having a* POISSON-BINOMIAL *distribution and, for a fixed n, G is a randomly chosen (without replacement) subset of size n:*

$$\mathbb{P}\{G = J\} = 1/\binom{N}{n} \qquad \text{for each subset } J \subset [\![N]\!] \text{ of size } n.$$

□ *Then the size $W := |F \cap G|$ also has a* POISSON-BINOMIAL *distribution.*

As a special case we could choose F equal to $[\![R]\!]$, whose size is a degenerate $\text{PBIN}(p_1, \ldots, p_R)$ with each p_i equal to 1. Then W would have a $\text{HYPER}(n, R, B)$ distribution, which thereby emerges as a special case of the POISSON-BINOMIAL. The Theorem could also be iterated for independent subsets $F_1, \ldots, F_k$ with $|F_1|$ distributed PBIN and F_i a random subset of size n_i for $2 \le i \le k$. We could then argue that $|F_1 \cap F_2| \sim \text{PBIN}$, and $|(F_1 \cap F_2) \cap F_3| \sim \text{PBIN}$, and so on.

Remark. V&M actually assumed (in my notation) $F_1, \ldots, F_k$ to be independent random subsets of $[\![N]\!]$ of sizes $n_1, \ldots, n_k$ sampled without replacement. Effectively they showed $|\cap_j F_j|$ has a PBIN distribution, from which they deduced several interesting limit theorems.

I found it helpful to work through the $\text{HYPER}(2, R, B)$ case, using only elementary arguments, before launching into the Proof of the Theorem. The special case gave me some insights into the subtlety of the result.

<13> **Example.** For a sample $\{s_1, s_2\}$ taken without replacement from $[\![N]\!]$ define $\xi_i = \{s_i \in [\![R]\!]\}$, so that the sample contains $W := \xi_1 + \xi_2$ members of $[\![R]\!]$ and $W \sim \text{HYPER}(2, R, B)$ for $B := N - R$.

Marginally speaking, each ξ_i has a $\text{BER}(p)$ distribution, with $p = R/N$, but they are not independent:

$$\mathbb{P}\{\xi_2 = 1 \mid \xi_1 = y_1\} = \begin{cases} (R-1)/(N-1) & \text{if } y_1 = 1 \\ R/(N-1) & \text{if } y_1 = 0 \end{cases}.$$

If both R and N were large then ξ_1 and ξ_2 would be close to independent, in some sense, but not so for an extreme case like $R = 1$ and $N = 3$.

The distribution of W is determined by the joint distribution of ξ_1 and ξ_2: for $y_1, y_2 \in \{0, 1\}$,

$$\mathbb{P}\{\xi_1 = y_1, \xi_2 = y_2\} = \frac{1}{N(N-1)} \begin{cases} R(R-1) & \text{if } y_1 = y_2 = 1 \\ B(B-1) & \text{if } y_1 = y_2 = 0 \\ RB & \text{if } y_1 + y_2 = 1 \end{cases}.$$

It follows that $h(z) := \mathbb{P}z^W = a_0 + a_1 z + a_2 z^2$, where

$$a_0 = \tfrac{B(B-1)}{N(N-1)}, \qquad a_1 = \tfrac{2RB}{N(N-1)}, \qquad a_2 = \tfrac{R(R-1)}{N(N-1)}.$$

The quadratic h has two real roots, $\{r_1, r_2\}$ given by

$$\tfrac{-RB \pm \sqrt{R^2B^2 - BR(B-1)(R-1)}}{R(R-1)} = \tfrac{-RB \pm \sqrt{RB(N-1)}}{R(R-1)},$$

both of them negative. Thus $W \sim \text{PBin}(p_1, p_2)$ for some values p_1, p_2 in $[0,1]$. We could calculate suitable p_1 and p_2 values by following the recipe given by Lemma <4>, but that seems to involve a lot of unenlightening algebra with square roots. Alternatively, we can just note that for $h(z)$ to factorize as $(1 - p_1 + p_1 z)(1 - p_2 + p_2 z)$ we must have

$$p_1 p_2 = a_2, \tag{<14>}$$

$$p_1 + p_2 - 2p_1p_2 = p_1(1-p_2) + (1-p_1)p_2 = a_1, \tag{<15>}$$

$$1 - (p_1 + p_2) + p_1 p_2 = (1-p_1)(1-p_2) = a_0, \tag{<16>}$$

with $p_1, p_2 \in [0,1]$. If there were such a solution then, by <14> and <15>, we would have

$$2\theta := \mathbb{P}W = p_1 + p_2 = a_1 + 2a_2 = \tfrac{2R(N-R) + 2R^2 - 2R}{N(N-1)} = \tfrac{2R}{N},$$

which would force $p_1 = \theta - t$ and $p_2 = \theta + t$ and $a_2 = p_1 p_2 = \theta^2 - t^2$, for some real t. Then we would get

$$t^2 = \theta^2 - a_2 = \tfrac{R^2}{N^2} - \tfrac{R(R-1)}{N(N-1)} = \tfrac{RB}{N^2(N-1)}$$

and p_1 and p_2 take the values $\tfrac{R}{N} \pm \sqrt{\tfrac{(R/N)(B/N)}{N-1}}$.

Remark. You might want to check whether these p_i values agree with the $\{(1-r_1)^{-1}, (1-r_2)^{-1}\}$ promised by Lemma <4>.

Even if my algebra is not quite correct, these calculations do support the idea that $\text{HYPER}(2, R, B)$, as a probability measure on $\{0,1,2\}$, coincides with $\text{PBin}(p_1, p_2)$ for some choice of p_i's. □

Proof (of Theorem <12>). If the PGF $h_0(z) = \mathbb{P}z^{|F|}$ is a polynomial in $\mathcal{P}_m$ then the shifted PGF,

$$h(z) := \mathbb{P}(1+z)^{|F|} = \mathbb{P}\sum_{k=0}^m \binom{|F|}{k} z^k = \sum_{k=0}^m a_k z^k \qquad \text{for } a_k := \mathbb{P}\tbinom{|F|}{k},$$

belongs to $\mathcal{P}_m$: the $1+z$ instead of z just subtracts 1 from each h_0 root.

Similarly, to prove that W has a PBin distribution it suffices to show that the shifted PGF

$$f(z) := \mathbb{P}(1+z)^W = \sum_{k=0}^m \mathbb{P}\binom{W}{k} z^k$$

has only real roots. (Here the sum must stop at, or before, m because $W \le \min(|F|, |G|)$.) To that end, note that

$$\binom{W}{k} = \sum\nolimits_{|J|=k} \{J \subseteq F \cap G\} = \sum\nolimits_{|J|=k} \{J \subseteq F\}\{J \subseteq G\},$$

where the sum runs over all the subsets of $[\![N]\!]$ of size k. (There are $\binom{N}{k}$ of them.) Take expected values then invoke independence of F and G:

<17> $$\mathbb{P}\binom{W}{k} = \sum\nolimits_{|J|=k} \mathbb{P}\{J \subseteq F\}\mathbb{P}\{J \subseteq G\}.$$

This equality is the key to the proof because we can now treat F and G separately.

For G: Consider any J of size k. By assumption, G^c is distributed uniformly over subsets of $[\![N]\!]$ of size $\ell := N - n$. Thus

$$\mathbb{P}\{J \subseteq G\} = \mathbb{P}\{G^c \subseteq J^c\} = \binom{N-k}{\ell} \Big/ \binom{N}{\ell} = C_n (N-k)_\ell$$

where $1/C_n := \ell! \binom{N}{\ell}$. This factor can be taken outside $\sum_{|J|=k}$ because it is the same for every J.

For F: The summation over J gives

$$\begin{aligned}\sum\nolimits_{|J|=k} \mathbb{P}\{J \subseteq F\} &= \mathbb{P} \sum\nolimits_{|J|=k} \{J \subseteq F\} \\ &= \mathbb{P}\,(\text{number of subsets of } F \text{ of size } k) = \mathbb{P}\binom{|F|}{k} = a_k.\end{aligned}$$

Thus

$$f(z) = \sum\nolimits_{k=0}^{m} \mathbb{P}\binom{W}{k} z^k = C_n \sum\nolimits_{k=0}^{m} a_k (N-k)_\ell z^k.$$

Example <8> showed that such an f belongs to $\mathcal{P}_m$. The random variable $W := |F \cap G|$ has a POISSON-BINOMIAL distribution. □

4.5 Mode(s) of the Binomial

You could regard this Section as mainly a warm-up for an analogous, but more delicate, argument for the general POISSON-BINOMIAL.

For a random variable Y taking only integer values,

$$\text{MODE}(Y) := \{k : \mathbb{P}\{Y = k\} = \max\nolimits_i \mathbb{P}\{Y = i\}\}.$$

If $\text{MODE}(Y) = \{\nu\}$, a singleton, then ν is called 'the mode' of the distribution; otherwise the members of $\text{MODE}(Y)$ are called modes.

<18> **Example.** If $S \sim \text{BIN}(n,p)$ with $0 < p < 1$ and $q := 1-p$ then

$$a_j := \mathbb{P}\{S = j\} = \binom{n}{j} p^j q^{n-j} \qquad \text{for } 0 \le j \le n.$$

The ratio of consecutive a_j's is a strictly decreasing function of j:

$$\frac{a_j}{a_{j-1}} = \frac{n! p^j q^{n-j}}{j!(n-j)!} \frac{(n-j+1)!(j-1)!}{n! p^{j-1} q^{n-j+1}} = \frac{(n-j+1)p}{jq}.$$

It follows that

$$a_j/a_{j-1} \qquad \begin{cases} > 1 & \text{if } j < (n+1)p \\ = 1 & \text{if } j = (n+1)p \\ < 1 & \text{if } j > (n+1)p \end{cases} .$$

If $p > n/(n+1)$ then $a_0 < a_1 < \cdots < a_n$, which makes n the unique mode. If $p \le n/(n+1)$, define $L = \lfloor (n+1)p \rfloor$ and $M := \lceil (n+1)p \rceil$, the unique integers satisfying the inequalities

$$L \le (n+1)p < L+1 \quad \text{AND} \quad M - 1 < (n+1)p \le M.$$

If $(n+1)p \in \mathbb{N}$ then $L = M = (n+1)p$ and

$$a_0 < a_1 < \cdots < a_{L-1} = a_{M-1} = a_L = a_M > a_{M+1} > \cdots > a_n.$$

If $(n+1)p \notin \mathbb{N}$ then $L < (n+1)p < M$ and

$$a_0 < a_1 < \cdots < a_{L-1} < a_L = a_{M-1} > a_M > \cdots > a_n.$$

That is, $\text{MODE}(S) = \{L-1, L\} = \{M-1, M\}$ if $(n+1)p$ is an integer, and otherwise $\text{MODE}(S) = \{L\} = \{M-1\}$. In particular, for $p = 1/2$ and n even $n/2$ is the unique mode; and for n odd the modes are $(n-1)/2$ and $(n+1)/2$. □

4.6 Mode(s) of the Poisson-Binomial

Suppose $S := X_1 + \cdots + X_n \sim \text{PBIN}(\mathbf{p})$ for $\mathbf{p} := (p_1, \ldots, p_n)$. As before, define $a_k := \mathbb{P}\{S = k\} = A_k(\mathbf{p})$. From Example <9>, if $0 < p_i < 1$ for each i then $a_k^2 > a_{k-1} a_{k+1}$ for each $1 \le k \le n-1$. Thus

$$\rho_1 := a_1/a_0 > \rho_2 := a_2/a_1 > \cdots > \rho_n := a_n/a_{n-1}.$$

Consequently,

(i) If $\rho_1 < 1$ then $a_0 > a_1 > a_2 > \cdots > a_n$.

(ii) If $\rho_1 = 1$ then $a_0 = a_1 > a_2 > \cdots > a_n$.

(iii) If $\rho_1 > 1$ then there is a largest K in $[\![n]\!]$ for which $\rho_K > 1$. If $K = 1$ then $a_0 > \cdots > a_n$. If $K > 1$ then $a_0 > \cdots > a_K$. If $K < n$ then either $\rho_{K+1} = 1$, which gives, $a_K = a_{K+1} > \ldots$ (meaning that $A_{K+1} > \cdots > a_n$ if $K \le n-2$) or $a_K > \cdots > a_n$.

Remark. As you can see, there are tediously many different cases reflecting slightly different sets of inequalities. Here I have tried to spell out all the gory details because, on early drafts of the book manuscript I had glossed over them, only to later see some beautiful Theorems destroyed by one of the tedious cases. Apparently I am not the only author to have suffered this fate: "We shall always assume $0 < p_i < 1$; the generalizations to $p_i = 0$ or 1 being immediate though sometimes ... burdensome to formulate" (Samuels, 1965, p1272). Indeed. Not surprisingly, some information posted on the WWW is not completely accurate for similar reasons.

When we wish to focus on A_k as a function of a particular p_α, the recursive formula

$$A_k(\mathbf{p}) = p_\alpha A_{k-1}(\mathbf{p}[-\alpha]) + (1-p_\alpha)A_k(\mathbf{p}[-\alpha])$$

becomes particularly useful. For example, an elementary piece of Calculus (Problem [5]) shows: at least for $\mathbf{p} \in (0,1)^n$ and $1 \le k \le n$,

<19> $A_k(\mathbf{p})/A_{k-1}(\mathbf{p})$ is a strictly increasing function of p_α

for fixed value of the other p_i's. This fact plays a central role in the following Theorem, which will handle the part of the range where the a_k's are increasing. When applied to $\sum_i (1-X_i)$ in Corollary <26> it will also take care of the rest of the range.

<20> **Theorem.** *If* $\mathbf{p} := (p_1, \ldots, p_n) \in (0,1)^n$ *with* $n \ge 2$ *and*

$$p_{min} := \min\{p_i : i \in [\![n]\!]\} < p_{max} := \max\{p_i : i \in [\![n]\!]\}$$

□ *then* $A_{k-1}(\mathbf{p}) < A_k(\mathbf{p})$ *for* $1 \le k \le p_{min} + \mu$ *where* $\mu := \sum_{i=1}^n p_i$.

The Proof for Theorem <20> could be written using only the A_k functions, without any mention of random variables. However, the arguments become more transparent when expressed using independent variables $X_0, X_1, \ldots, X_n$ with $X_i \sim \text{BER}(p_i)$ and $p_0 := p_{min}$. In the next Section those arguments will involve many different PBIN's built from the same X_i's. That is, we will need to work with many different probability measures $\mathbb{P}_{\mathbf{p}^*}$ on $\Omega := \{0,1\}^{n+1}$, defined by different $\mathbf{p}^* := (p_0, \ldots, p_n)$ values, with the X_i's as its coordinate maps. In that setting we have:

$$S(\omega) := \sum_{i=1}^n X_i(\omega) \quad \text{AND} \quad S^*(\omega) := \sum_{i=0}^n X_i(\omega),$$

<21> $$T_i(\omega) := S^*(\omega) - X_i(\omega) \qquad \text{for } 0 \le i \le n,$$

$$\mathbb{P}_{\mathbf{p}^*}\{\omega\} := \prod_{i=0}^n \Big(p_i\{\omega_i = 1\} + (1-p_i)\{\omega_i = 0\} \Big) \qquad \text{for } \omega \in \Omega,$$

Actually, $\mathbb{P}_{\mathbf{p}^*}\{S = k\} = A_k(\mathbf{p})$ and $\mathbb{P}_{\mathbf{p}^*}\{S \ge k\} =: E_k(\mathbf{p})$ for $\mathbf{p} := (p_1, \ldots, p_n)$, as before. Similarly

<22> $$\tau_i(k, \mathbf{p}^*) := \mathbb{P}_{\mathbf{p}^*}\{T_i = k\} = A_k(\mathbf{p}^*[-i]) \qquad \text{for } 0 \le k \le n,$$

where $\mathbf{p}^*[-i] := (p_0, p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_n)$. Roughly speaking, T_i behaves like S if p_i is replaced by p_0.

Remark. In this Section the underlying $\mathbb{P}$ will always be $\mathbb{P}_{\mathbf{p}^*}$; there will be no harm if the $\mathbf{p}^*$'s are omitted. The changing subscripts will drive the whole argument in the next two Sections.

Proof (of Theorem <20>). Start from a pair of pointwise equalities:

$$\{S^* = k\} = X_i\{T_i = k-1\} + (1-X_i)\{T_i = k\} \qquad \text{for } 0 \le i \le n,$$
$$k\{S^* = k\} = \sum\nolimits_{j=0}^{n} X_j\{T_j = k-1\}.$$

The first follows directly from $S^* = X_i + T_i$; the second from

$$\sum\nolimits_j X_j\{T_j = k-1\} = \sum\nolimits_j X_j\{S^* = k\} = S^*\{S^* = k\} = k\{S^* = k\}.$$

Take expected values (under $\mathbb{P}_{\mathbf{p}^*}$), using independence of X_i and T_i and $\mathbb{P}\{X_i = 0\} = q_i := 1 - p_i$, to deduce that

<23> $$\mathbb{P}\{S^* = k\} = p_i\tau_i(k-1) + q_i\tau_i(k) \qquad \text{for } 0 \le i \le n,$$

<24> $$k\mathbb{P}\{S^* = k\} = \sum\nolimits_{j=0}^{n} p_j\tau_j(k-1).$$

For $1 \le k \le \mathbb{P}S + p_{min} = \sum_{j=0}^{n} p_j$ we have

$$\begin{aligned}\sum\nolimits_{j=0}^{n} p_j\tau_j(k-1) = k\mathbb{P}\{S^* = k\} &\le \Big(\sum\nolimits_{j=0}^{n} p_j\Big)\,\mathbb{P}\{S^* = k\} \\ &= \sum\nolimits_{j=0}^{n} p_j\Big(p_j\tau_j(k-1) + q_j\tau_j(k)\Big) \qquad \text{by <23>.}\end{aligned}$$

Subtract the first expression from the last.

$$\begin{aligned}0 &\le \sum\nolimits_{j=0}^{n} p_j\Big(p_j\tau_j(k-1) + q_j\tau_j(k) - (p_j+q_j)\tau_j(k-1)\Big) \\ &= \sum\nolimits_{i=0}^{n} p_jq_j\Big(\tau_j(k) - \tau_j(k-1)\Big) \\ &= \sum\nolimits_{j=0}^{n} p_jq_j\tau_j(k-1)\Big(\tau_j(k)/\tau_j(k-1) - 1\Big). \qquad \text{<25>}\end{aligned}$$

The monotonicity property <19> gives

$$\begin{aligned}\tau_j(k)/\tau_j(k-1) &= A_k(p_0, \mathbf{p}^*[-j])/A_{k-1}(p_0, \mathbf{p}^*[-j]) \\ &\le A_k(p_j, \mathbf{p}^*[-j])/A_{k-1}(p_j, \mathbf{p}^*[-j]) \qquad \text{strict if } p_0 < p_j \\ &= A_k(\mathbf{p})/A_{k-1}(\mathbf{p}) = \mathbb{P}\{S = k\}/\mathbb{P}\{S = k-1\}.\end{aligned}$$

The assumption that $p_0 = p_{min} < p_{max}$ together with <25> implies

$$0 < \Big(\mathbb{P}\{S = k\}/\mathbb{P}\{S = k-1\} - 1\Big)\sum\nolimits_{j=0}^{n} p_jq_jA_{k-1}\left(\mathbf{p}^*[-j]\right).$$

□ As $0 < \sum_{j=0}^{n} \ldots$, it follows that $\mathbb{P}\{S = k\} > \mathbb{P}\{S = k-1\}$.

<26> **Corollary.** *Suppose* $\mathbf{p} \in (0,1)^n$, *for some* $n \geq 2$, *with* $\mu := \mathbb{P}_{\mathbf{p}} S := \sum_i^n p_i$ *and* $p_0 := p_{min} < p_{max}$. *Define* $a_i := A_i(\mathbf{p})$. *Then*

(i) $n \geq L := \lfloor p_{min}+\mu \rfloor$ *and* $n+1 \geq M := \lceil p_{max}+\mu \rceil$ *and either* $M = L+1$ *or* $M = L+2$.

(ii) $a_{k-1} < a_k$ *for* $1 \leq k \leq L$.

(iii) $a_{k-1} > a_k$ *for* $M \leq k \leq n$.

(iv) If $M = L+1$ *then* $a_0 < \cdots < a_{M-1} > \cdots > a_n$.

(v) If $M = L+2$ *then* $a_0 < \cdots < a_{M-2}$ *and* $a_{M-1} > \cdots > a_n$, *with no assertion regarding the comparison between* a_{M-2} *and* a_{M-1}. □

Remark. If $M = n+1$ then $a_{M-1} = a_n$; the pattern of inequalities in both (iv) and (v) should then end with a_{M-1}.

Compare with the result from Example <18> for $S \sim \text{BIN}(n, \delta)$. There we have $\mu = n\delta$ with $L := \lfloor \delta + \mu \rfloor$ and $M := \lceil \delta + \mu \rceil$. The patterns are:

<27> $a_0 < \cdots < a_{M-1} = a_M > \cdots > a_n$ and $M = L$ if $\delta + \mu \in \mathbb{N}$,

<28> $a_0 < \cdots < a_{M-1} > a_M > \cdots > a_n$ and $M = L+1$ if $\delta + \mu \notin \mathbb{N}$.

Proof. Start with (i), noting that $p_{min}+\mu < p_{max}+\mu < n+1$ but $p_{max}+\mu$ might be $> n$. By definition of $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$,

$$L \leq p_{min} + \mu < L+1 \quad \text{AND} \quad M - 1 < p_{max} + \mu \leq M.$$

We must have $L < M$ because $p_{min} < p_{max}$. Also

$$L + 2 > 1 + p_{min} + \mu > p_{max} + \mu > M - 1.$$

Thus $0 < M - L < 3$ or, equivalently, $1 + L \leq M \leq L + 2$.

Assertion (ii) just restates the result from Theorem <20>. For (iii) apply the same Theorem to $n - S \sim \text{PBIN}(n, q_1, \ldots, q_n)$ for $q_i := 1 - p_i$. For $1 \leq j \leq q_{min} + (n - \mu) = n + 1 - p_{max} - \mu$,

$$a_{n-j+1} := \mathbb{P}_{\mathbf{p}}\{n - S = j - 1\} < \mathbb{P}_{\mathbf{p}}\{n - S = j\} =: a_{n-j}$$

Set j equal to $n - k + 1$ to get $a_k < a_{k-1}$ for $p_{max} + \mu \leq k \leq n$.

From (i) and (ii) we have

$$a_0 < a_1 < \cdots < a_L \quad \text{AND} \quad a_{M-1} > a_M > \cdots > a_n,$$

□ implying (iv) and (v).

Behavior of MODE(S) if $\mathbf{p}$ has some 0 or 1 elements
A general $\mathbf{p}$ in $[0,1]^n$ requires some extra bookkeeping. Define

<29> $$n_0 := \sum_{i=1}^n \{p_i = 0\},\ n_1 := \sum_{i=1}^n \{p_i = 1\},\ I_{\mathbf{p}} := \{i \in [[n]] : 0 < p_i < 1\}.$$

If $I_{\mathbf{p}} = \emptyset$ then $\mathbb{P}_{\mathbf{p}}\{S = n_1\} = 1$, end of story. If $|I_{\mathbf{p}}| = K \geq 1$ suppose $\mathbf{p} = (\gamma, 1, \ldots, 1, 0, \ldots, 0)$ where $\gamma := (\gamma_1, \ldots, \gamma_K)$ with $0 < \gamma_i < 1$ and $S_K := \sum_{i \in I_{\mathbf{p}}} X_i \sim \text{PBIN}(\gamma)$. Of course, if $n_1 = 0$ the 1's should be omitted and if $n_0 = 0$ the 0's should be omitted. We then have $\mathbb{P}_{\mathbf{p}}\{S = n_1 + i\} = \mathbb{P}_\gamma\{S_K = i\}$ for $0 \leq i \leq K$ and $\text{MODE}(S) = \{n_1 + i : i \in \text{MODE}(S_K)\}$.

4.7 Binomial tails versus Poisson-Binomial tails

Hoeffding (1956, §4) proved the following result.

<30> **Theorem.** *Suppose $Y \sim \text{PBIN}(\mathbf{p})$ and $W \sim \text{BIN}(n, \delta)$ for $0 < \delta < 1$, where $\mathbb{P}Y = \mathbb{P}W = \mu$. Then for $k \in \mathbb{N}$ with $1 + \mu \leq k \leq n$,*

(i) $\mathbb{P}\{Y \geq k\} \leq \mathbb{P}\{W \geq k\}$ for $1 + \mu \leq k \leq n$.

□ *(ii) The inequality in (i) is strict if $p_i \neq \delta$ for at least one i.*

Put another way: for $1 + \mu \leq k \leq n$ the continuous function $E_k(\cdot)$ (see Problem [6]) has a unique maximum over the compact, convex set

$$\mathbb{B}_n(\delta) := \{\mathbf{p} \in [0,1]^n : \sum\nolimits_i p_i = n\delta\}$$

at $\mathbf{u}_\delta := \delta \mathbb{1}_n$, which lies in the relative interior of $\mathbb{B}_n(\delta)$. That is,

<31> $$E_k(\mathbf{u}_\delta) > E_k(\mathbf{p}) \qquad \text{for all } \mathbf{p} \text{ in } \mathbb{B}_n(\delta) \backslash \{\mathbf{u}_\delta\}.$$

The ***relative interior***, rel-int $\mathbb{B}_n(\delta)$, of the parameter set $\mathbb{B}_n(\delta)$ equals $\{\mathbf{p} \in \mathbb{B}_n(\delta) : 0 < p_i < 1 \text{ for all } i\}$. Points in the ***relative boundary***, rel-bnd $\mathbb{B}_n(\delta)$, have $p_i \in \{0, 1\}$ for at least one i. The ***extreme points*** have $p_i \in \{0, 1\}$ for at least $n - 1$ of the coordinates. See Barvinok (2002, Chap II) for an explanation of these polyhedral facts.

Again the arguments become more transparent if written in the notation of <21>, with the role of Y taken over by $S := X_1 + \cdots + X_n$ under $\mathbb{P}_{\mathbf{p}}$ and the role of W taken over by S under $\mathbb{P}_{\mathbf{u}}$. Suppose

<32> $$\mathbf{p} = (p_1, \ldots, p_\kappa, \mathbb{1}_{n_1}, \mathbf{0}_{n_0}) \qquad \text{with } 0 < p_i < 1 \text{ for } 1 \leq i \leq \kappa,$$
$$n_1 \text{ ones, and } n_0 \text{ zeros.}$$

The Theorem is trivial if $\kappa \leq 1$:

(i) If $\kappa = 0$ then $\mu = \mathbb{P}_{\mathbf{p}} S = n_1$ and $\mathbb{P}_{\mathbf{p}}\{S = n_1\} = 1$, which implies $\mathbb{P}_{\mathbf{p}}\{S \geq k\} = 0$ if $k \geq 1 + \mu$.

(ii) If $\kappa = 1$ then $S = n_1 + X_1$ with $\mu = n_1 + p_1$ and $\mathbb{P}_{\mathbf{p}}\{S \le n_1 + 1\} = 1$. If $k \ge 1 + \mu = 1 + n_1 + p_1$ then $k \ge n_1 + 2$, because $p_1 > 0$. Thus $\mathbb{P}_{\mathbf{p}}\{S \ge k\} = 0$.

For $\kappa \ge 2$ the Theorem can be proved by repeated application of a simple procedure that extends an averaging method due to Chebyshev (1846). The idea is that $\mathbf{p}$ should be nudged towards $\mathbf{u}_\delta$ by a sequence of small changes that do not decrease the tail probabilities, with at least one increase. The changes will be defined by functions $\sigma_{\alpha,\beta,t}$, for $\alpha, \beta \in [\![n]\!]$ with $p_\alpha \ne p_\beta$, that map $[0,1]^n$ back into itself. Without loss of generality suppose $p_\beta - p_\alpha = 2r > 0$. For $0 < t \le r$, define $\sigma_{\alpha,\beta,t}(\mathbf{p}) := \boldsymbol{\theta}$ where

<33> $$\theta_i = \begin{cases} p_\alpha + t & \text{if } i = \alpha \\ p_\beta - t & \text{if } i = \beta \\ p_i & \text{otherwise} \end{cases} .$$

That is, $\sigma_{\alpha,\beta,t}$ moves both the coordinates p_a and p_β towards their midpoint, $\lambda := (p_\beta + p_\alpha)/2$, by an amount t. The following Lemma collects together several useful properties of these maps.

<34> **Lemma.** *Suppose* $\mathbf{p} := (p_1, \ldots, p_n) \in [0,1]^n$ *and* $S \sim \text{PBIN}(\mathbf{p})$, *with* $n \ge 2$. *Suppose also that* $p_\beta - p_\alpha = 2r > 0$ *for some* $\alpha, \beta \in [\![n]\!]$. *Define* $\mu := \mathbb{P}S := \sum_i p_i$ *and* $U := S - X_\alpha - X_\beta$ *and* $\boldsymbol{\theta} := \sigma_{\alpha,\beta,t}(\mathbf{p})$ *for some* t *with* $0 < t \le r$. *Then*

(i) $U \sim \text{PBIN}(\boldsymbol{\zeta})$, *where* $\boldsymbol{\zeta} := (p_i : i \in [\![n]\!] \backslash \{\alpha, \beta\}) \in [0,1]^{n-2}$.

(ii) If $\boldsymbol{\zeta} \in (0,1)^{n-2}$ *then*

$$A_{j-1}(\boldsymbol{\zeta}) > A_j(\boldsymbol{\zeta}) \text{ for } n - 2 \ge j \ge M := \lceil \boldsymbol{\zeta}_{max} + \textstyle\sum_j \zeta_j \rceil.$$

(iii) $|\mathbf{p} - \mathbf{u}_\delta|_2^2 = |\boldsymbol{\theta} - \mathbf{u}_\delta|_2^2 + \Delta^2$ *where* $\delta := \mathbb{P}_{\mathbf{p}} S/n$ *and* $\Delta^2 := t(2r - t)$.

(iv) For each k *in* $[\![0{:}n]\!]$,

$$\begin{aligned} E_k(\boldsymbol{\theta}) - E_k(\mathbf{p}) &= \Delta^2 \Big(\mathbb{P}_{\boldsymbol{\zeta}}\{U = k-2\} - \mathbb{P}_{\boldsymbol{\zeta}}\{U = k-1\} \Big). \\ &= \Delta^2 \left(A_{k-2}(\boldsymbol{\zeta}) - A_{k-1}(\boldsymbol{\zeta}) \right). \end{aligned}$$

(v) If $\boldsymbol{\zeta} \in (0,1)^{n-2}$ *and* $n \ge k \ge 1 + \mu$ *then*

$$A_{k-2}(\boldsymbol{\zeta}) > A_{k-1}(\boldsymbol{\zeta}) \qquad \text{if } p_\alpha + p_\beta \ge \boldsymbol{\zeta}_{max}.$$

Proof. Assertion (i) is true by definition. Assertion (ii) follows from Corollary <26> (for $\boldsymbol{\zeta}_{min} < \boldsymbol{\zeta}_{max}$) and by <28> (for $\boldsymbol{\zeta}_{min} = \boldsymbol{\zeta}_{max}$).

For (iii): The contributions from $j \notin \{\alpha, \beta\}$ cancel, leaving

$$\begin{aligned}(p_\alpha - \delta)^2 + (p_\beta - \delta)^2 - (p_\alpha + t - \delta)^2 - (p_\beta - t - \delta)^2 \\ = 2t(p_\beta - \delta - p_\alpha + \delta) - 2t^2 = \Delta^2.\end{aligned}$$

For (iv): Note that $\theta_\alpha + \theta_\beta = p_\alpha + p_\beta = 2r$ and $\theta_\alpha\theta_\beta = p_\alpha p_\beta + \Delta^2$, so that

$$\begin{aligned}\psi_1(\theta_\alpha, \theta_\beta) := \theta_\alpha(1-\theta_\beta) + (1-\theta_\alpha)\theta_\beta = \psi_1(p_\alpha, p_\beta) - 2\Delta^2, \\ \psi_0(\theta_\alpha, \theta_\beta) := (1-\theta_\alpha)(1-\theta_\beta) = \psi_0(p_\alpha, p_\beta) + \Delta^2.\end{aligned}$$

Define $B_j := \mathbb{P}_{\boldsymbol{\zeta}}\{U \ge j\} := E_j(\boldsymbol{\zeta})$ and $\tau_j := \mathbb{P}_{\boldsymbol{\zeta}}\{U = j\} := A_j(\boldsymbol{\zeta})$ so that $B_{k-2} + B_k - 2B_{k-1} = \tau_{k-2} - \tau_{k-1}$. By independence,

$$\begin{aligned}E_k(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}\{S \ge k\} = \sum\nolimits_{j=0}^{2} \mathbb{P}_{\boldsymbol{\theta}}\{X_\alpha + X_\beta = j\} B_{k-j} \\ = \theta_\alpha\theta_\beta B_{k-2} + \psi_1(\theta_\alpha, \theta_\beta) B_{k-1} + \psi_0(\theta_\alpha, \theta_\beta) B_k \\ = (p_\alpha p_\beta + \Delta^2) B_{k-2} + \big(\psi_1(p_\alpha, p_\beta) - 2\Delta^2\big) B_{k-1} + \big(\psi_0(p_\alpha, p_\beta) + \Delta^2\big) B_k \\ = E_k(\mathbf{p}) + \Delta^2 \left(-B_{k-2} + 2B_{k-1} - B_k\right) = E_k(\mathbf{p}) + \Delta^2(\tau_{k-2} - \tau_{k-1}).\end{aligned}$$

For (v): Corollary <26> gives $A_{M-1}(\boldsymbol{\zeta}) > A_M(\boldsymbol{\zeta}) > \cdots > A_{n-2}(\boldsymbol{\zeta})$ for $M := \lceil \zeta_{max} + \mathbb{P}_{\boldsymbol{\zeta}} U \rceil$. If $k \ge 1 + \mathbb{P}_{\mathbf{p}} S$ then

$$k \ge 1 + p_\alpha + p_\beta + \mathbb{P}_{\boldsymbol{\zeta}} U \ge 1 + \zeta_{max} + \mathbb{P}_{\boldsymbol{\zeta}} U.$$

□ Equivalently, $k - 1 \ge M := \lceil \zeta_{max} + \mathbb{P}_{\boldsymbol{\zeta}} U \rceil$, so that $A_{k-2}(\boldsymbol{\zeta}) > A_{k-1}(\boldsymbol{\zeta})$.

If $n = 2$ then $U = 0$, so that the Lemma boils down to a well-known fact: $(\lambda + r)(\lambda - r) = \lambda^2 - r^2$. However, for $n = 3$ it tells us something slightly less obvious.

<35> **Example.** Suppose $\mathbf{p} = (p_1, p_2, p_3)$ with $p_1 = \lambda - r \le p_2 \le p_3 = \lambda + r$. If $p_2 = 0$ or $p_2 = 1$ the sum $S = X_1 + X_2 + X_3$ is not very interesting, so let me assume $0 < p_2 < 1$. Consider the case $\boldsymbol{\theta} = \sigma_{1,2,r}(\mathbf{p}) = (\lambda, p_2, \lambda)$, for which $\Delta = r^2$ and $U = X_2 \sim \text{BER}(p_2)$.

Write d_k for $\mathbb{P}_{\boldsymbol{\theta}}\{S \ge k\} - \mathbb{P}_{\mathbf{p}}\{S \ge k\}$. Then the Lemma gives $d_0 = 0$ and $d_1 = -r^2 p_2$ (not interesting). However $d_2 = r^2(1 - 2p_2)$, which tells us that only for $p_2 < 1/2$ do we learn something interesting. Was that
□ obvious?

The Proof of Theorem <30> (in the form <31>) will consist of repeated appeals to Lemma <34>. There are two distinct cases to consider:

(a) $\mathbf{p}$ lies in rel-int $\mathbb{B}_n(\delta)$

(b) $\mathbf{p}$ lies in rel-bnd $\mathbb{B}_n(\delta)$

In both cases we need to construct a sequence $\mathbf{p}^{(0)} = \mathbf{p}, \mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots$, with $\mathbf{p}^{(\ell)} \to \mathbf{u}_\delta$, by $\mathbf{p}^{(\ell+1)} := \sigma_{\alpha_\ell, \beta_\ell, t_\ell}(\mathbf{p}^{(\ell)})$ for well-chosen $(\alpha_\ell, \beta_\ell, t_\ell)$.

Proof (of Theorem <30> under (a): $\mathbf{p} \in$ rel-int $\mathbb{B}_n(\delta)$).
Of course we may assume $\mathbf{p} \neq \mathbf{u}_\delta$.

The construction for $\ell = 0$ shows the general pattern. Remember that

$$0 < p_{min} = p_1 \le p_2 \le \cdots \le p_{max} = p_n < 1 \qquad \text{with } p_n - p_1 = 2r > 0.$$

As $\sum_i p_i = n\delta$, we must have $p_1 < \delta < p_n$. Define $\mathbf{p}^{(1)} := \boldsymbol{\theta} := \sigma_{1,n,t}(\mathbf{p})$ with $t := \min(\delta - p_1, p_n - \delta)$. Then at least one of $\{\theta_1, \theta_n\}$ must equal δ, with $p_1 + p_n > p_{max} \ge \zeta_{max}$ and $E_k(\boldsymbol{\theta}) > E_k(\mathbf{p})$. Lemma <34> then gives (iv) and (v).

For $\mathbf{p}^{(2)}$ it simplifies notation if we first sort $\mathbf{p}^{(1)}$ into increasing order then apply the same strategy: define $\mathbf{p}^{(2)} := \sigma_{1,n,t_2}(\mathbf{p}^{(1)})$ for $t_2 := \min\bigl(\delta - \mathbf{p}^{(1)}[1], \mathbf{p}^{(1)}[n] - \delta\bigr)$. And so on. Each step adds at least one new coordinate equal to δ. In fewer than n steps we reach a $\mathbf{p}^{(\ell)}$ equal to $\mathbf{u}_\delta$ and $E_k(\mathbf{p}^{(\ell)}) > E_k(\mathbf{p})$.

□ Thus Theorem <30> holds for $\mathbf{p}$ in the relative interior of $\mathbb{B}_n(\delta)$.

Proof (of Theorem <30> under (b): $\mathbf{p} \in$ rel-bnd $\mathbb{B}_n(\delta)$).
We need some preliminary steps to push $\mathbf{p}$ into the relative interior. It will help to write S_d when referring to $X_1 + \cdots + X_d$ because the argument will jump around between different dimensions.

For the sake of notation let me initially assume that $\mathbf{p}$ has at least one coordinate equal to 1. Define

$$\begin{aligned}
\mathbf{p} &= (p_1, \ldots, p_\kappa, \mathbb{1}_{n_1}, \mathbf{0}_{n_0}) \qquad \text{with } 0 \le p_1 \le \cdots \le p_\kappa < 1 \text{ and } n_0 + n_1 \ge 1,\\
\boldsymbol{\zeta} &:= \mathbf{p}[1{:}\kappa - 1] = (p_1, \ldots, p_{\kappa-1}) \in (0,1)^{\kappa-1},\\
\boldsymbol{\gamma} &:= \mathbf{p}[1{:}\kappa + 1] = (p_1, \ldots, p_{\kappa-1}, p_\kappa, 1) = (\boldsymbol{\zeta}, p_\kappa, 1) \in [0,1]^{\kappa+1},\\
\boldsymbol{\gamma}^* &:= \sigma_{\kappa,\kappa+1,r}(\boldsymbol{\gamma}) = (\boldsymbol{\zeta}, r, r) \in (0,1)^{\kappa+1} \qquad \text{for } r := (1 - p_\kappa)/2,\\
\mathbf{p}^* &:= (\boldsymbol{\gamma}^*, \mathbb{1}_{n_1-1}, \mathbf{0}_{n_0}) = (\boldsymbol{\zeta}, r, r, \mathbb{1}_{n_1-1}, \mathbf{0}_{n_0}).
\end{aligned}$$

We then have

$$\begin{aligned}
E_k(\mathbf{p}) &= \mathbb{P}_{\mathbf{p}}\{S_n \ge k\} = \mathbb{P}_{\mathbf{p}}\{S_{\kappa+1} + n_1 - 1 \ge k\}\\
&= \mathbb{P}_{\boldsymbol{\gamma}}\{S_{\kappa+1} \ge \ell\} = E_\ell(\boldsymbol{\gamma}) \qquad \text{where } \ell := k - (n_1 - 1)\\
&= E_\ell(\boldsymbol{\gamma}^*) - r^2\Bigl(A_{\ell-2}(\boldsymbol{\zeta}) - A_{\ell-1}(\boldsymbol{\zeta})\Bigr) \qquad \text{by Lemma <34>(iv)}\\
&< E_\ell(\boldsymbol{\gamma}^*) \qquad \text{by Lemma <34>(v) because } p_\kappa + 1 \ge p_{\kappa-1}\\
&= \mathbb{P}_{\boldsymbol{\gamma}^*}\{S_{\kappa+1} \ge \ell\} = \mathbb{P}_{\mathbf{p}^*}\{S_n \ge k\} = E_k(\mathbf{p}^*).
\end{aligned}$$

We have killed off a 1 and increased $\sum_{i \le n}\{0 < p_i^* < 1\}$ to $\kappa + 1$. Repeated applications of this trick eventually leave us with $E_k(\mathbf{p}) < E_k(\boldsymbol{\lambda})$ where $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_{\kappa+n_1}, \mathbf{0}_{n_0})$ with $0 < \lambda_i < 1$ for $i \le \kappa + n_1$.

To avoid cluttering the notation with lots more symbols, let me just reduce the previous paragraph to an assertion that "without loss of generality we may assume $n_1 = 0$" then recycle notation by assuming

$$\mathbf{p} = (\boldsymbol{\zeta}, \mathbf{0}_{n_0}) \qquad \text{where } 0 < \zeta_1 \le \cdots \le \zeta_\kappa < 1 \text{ with } \kappa = n - n_0.$$

Remark. Assuming that we are starting from $\mathbf{p} = (p_1, \ldots, p_\kappa, \mathbb{1}_{n_1}, \mathbf{0}_{n_0})$ with $0 < p_i < 1$ for $i \in [\![\kappa]\!]$ and $\sum_i p_i = n\delta$ and $n_1 > 0$, the **R** pseudocode would be:

```
1:  function KILL.ONES(p,κ,n1,δ)
2:      while n1 > 0 do
3:          p[1:κ] ← sort(p[1:κ])          % now p1 ≤ ··· ≤ pκ
4:          r ← (1 − pκ)/2
5:          p ← σκ,κ+1,r(p)
6:          n1 ← n1 − 1; κ ← κ + 1
7:      end while
8:      Return p
9:  end function
```

Maybe it would be better to write all algorithmic arguments using a language designed for such things.

We need a slightly different trick to kill off the 0's. This time start from $\mathbf{p} = (\boldsymbol{\zeta}_\kappa, \mathbf{0}_{n_0})$, with $\boldsymbol{\zeta}_\kappa := (p_1, \ldots, p_\kappa) \in (0,1)^\kappa$, and $n_0 > 0$. As $\boldsymbol{\zeta}_\kappa \in \text{rel-int}\, \mathbb{B}_\kappa(\delta)$ we can apply Theorem <30> case (a), the result established by the previous Proof, to show that

$$E_k(\boldsymbol{\zeta}_\kappa) < E_k(\delta \mathbb{1}_\kappa) = \mathbb{P}\{Y_\kappa \ge k\} \qquad \text{if } Y_\kappa \sim \text{BIN}(\kappa, \delta).$$

Problem [7] shows (using Lemma <34>) that

$$\begin{aligned} \mathbb{P}\{Y_\kappa \ge k\} &< \mathbb{P}\{Y_{\kappa+1} \ge k\} \qquad \text{where } Y_{\kappa+1} \sim \text{BIN}(\kappa+1, \mu/(\kappa+1)) \\ &< \mathbb{P}\{Y_{\kappa+2} \ge k\} \qquad \text{where } Y_{\kappa+2} \sim \text{BIN}(\kappa+2, \mu/(\kappa+2)) \\ &\ldots \\ &< \mathbb{P}\{Y_n \ge k\} \qquad \text{where } Y_n \sim \text{BIN}(n, \mu/n). \end{aligned}$$

As $\mu/n = \delta$ we now have Theorem <30> holding when $\mathbf{p}$ lies in the relative boundary of $\mathbb{B}_n(\delta)$. □

Remark. Think of $E_k(\mathbf{p})$ as the height at $\mathbf{p}$ for an $(n+1)$-dimensional mountain sitting on the base $\mathbb{B}_n(\delta)$. Then the $(\mathbf{p}^{(\ell)}, E_k(\mathbf{p}^{(\ell)}))$'s plot the path of an $(n+1)$-dimensional bunny rabbit hopping up to the top of the mountain. If you replaced the discrete changes by something smoother—say $\boldsymbol{\theta}(t) = \sigma_{1,n,t}(\mathbf{p})$ for $0 \le t \le \min(\delta - p_1, p_n - \delta)$, and so on—the strategy would give a zigzag path in $\mathbb{B}_n(\delta)$ along which $E_k(\cdot)$ increased from $E_k(\mathbf{p})$ to $E_k(\mathbf{u}_\delta)$. Maybe there exists an even better smooth path connecting $\mathbf{p}$ and $\mathbf{u}_\delta$, with the increasing property corresponding to some differential inequality. Similar ideas will appear in §6.2, where two multivariate normal distributions $N(0, V_0)$ and $N(0, V_1)$ are connected by a smooth path $\{V_t : 0 \le t \le 1\}$ through the parameter space of positive semi-definite matrices. In that setting the choice of a good path appears to be a matter of some inspired guesswork.

4.8 The median of the Poisson-Binomial

The following argument repackages a clever idea of Jogdeo and Samuels (1968), who made clever use of the inequality from Theorem <30>.

Recall that a median of a random variable S is any real number R for which $\mathbb{P}\{S \geq R\} \geq 1/2$ and $\mathbb{P}\{S \leq R\} \geq 1/2$.

<36> **Theorem.** *Suppose* $S_n \sim \text{PBIN}(\mathbf{p})$ *with* $\mathbf{p} \in [0,1]^n$ *and* $\mu := \mathbb{P}S_n \in (0,n)$. *Then* $\lfloor \mu \rfloor \leq R \leq \lceil \mu \rceil$ *for every median of* S_n.

Proof. Define $\delta_j := (\mu + j)/(n + 2j)$ for $j = 0, 1, \ldots$ and $k := \lceil \mu \rceil$. To show that $R \leq k$ it suffices to shown that $\mathbb{P}_{\mathbf{p}}\{S_n \geq k+1\} < 1/2$. As a first step, add in a pair of innocuous new BER variables by introducing $\mathbf{p}^* := (\mathbf{p}, 1, 0) \in [0,1]^{n+2}$. Clearly

$$\begin{aligned} \mathbb{P}_{\mathbf{p}}\{S_n \geq k+1\} = \mathbb{P}_{\mathbf{p}^*}\{S_{n+2} \geq k+2\} & \qquad \text{with } \mathbb{P}_{\mathbf{p}^*} S_{n+2} = \mu + 1 \\ < \mathbb{P}\{W_1 \geq k+2\} & \qquad \text{where } W_1 \sim \text{BIN}(n+2, \delta_1), \end{aligned}$$

the inequality coming from $k + 2 \geq 1 + \mathbb{P}_{\mathbf{p}^*} S_{n+2}$ and Theorem <30>.

Repeated applications of the same idea then give

$$\mathbb{P}\{W_1 \geq k+2\} < \mathbb{P}\{W_2 \geq k+3\} \cdots < \mathbb{P}\{W_N \geq k+N\} < \ldots$$

where $W_j \sim \text{BIN}(n+2j, \delta_j)$, with $\mathbb{P}W_j = \mu + j$ and $\sigma_j^2 := \text{var}(W_j) = (\mu+j)(1-\delta_j)$. A classical CLT (see UGMTP 7.29, for example) ensures that $Z_N := (W_N - (\mu + N))/\sigma_j \rightsquigarrow N(0,1)$ as $N \to \infty$. Thus

$$\mathbb{P}\{W_N \geq k+N\} = \mathbb{P}\{Z_N \geq (k-\mu)/\sigma_N\} \nearrow \bar{\Phi}(0) = 1/2.$$

It follows that $\mathbb{P}_{\mathbf{p}}\{S_n \geq k+1\} < 1/2$.

An analogous argument with $n - S$, which is also PBIN distributed, shows that $R \geq \lfloor \mu \rfloor$. □

4.9 Problems

[1] Suppose f is a convex function on $[0,1]$ and $\theta \in (0,1)$. Amongst probability measures P on $[0,1]$ with mean θ, show that Pf is minimized by the $\text{BER}(\theta)$ distribution.

[2] (Hoeffding, 1956, last paragraph of page 713) As in Section 4.7, think of S as the sum of coordinate maps on the set $\{0,1\}^n$ equipped with $\mathbb{P}_{\boldsymbol{p}}$, a product of $\text{BER}(p_i)$ distributions. For a given real valued function g on the set $\{0, 1, \ldots, n\}$ define a map $G : [0,1]^n \to \mathbb{R}$ by

$$G(\mathbf{p}) = \mathbb{P}_{\mathbf{p}} g(S) = \sum\nolimits_{k=0}^{n} g(k) \mathbb{P}_{\mathbf{p}}\{S = k\}.$$

(i) Show that $G(\mathbf{p}) = G(p_{\sigma(1)}, \ldots, p_{\sigma(n)})$ for each permutation σ of $[n]$ and that $p_j \mapsto G(\boldsymbol{p})$ is linear for each j (with the other coordinates held fixed). Hint: Look at equations <1> and <23>.

(ii) Conversely, suppose a map $H : [0,1]^n \to \mathbb{R}$ has the two properties (symmetry and linearity in each coordinate) identified in part (i). For $0 \le k \le n$ define $h(k) := H(\mathbf{1}_k, \mathbf{0}_{n-k})$, where the argument denotes a vector with 1 repeated k times followed by 0 repeated $n-k$ times. Show that $H(\mathbf{p}) = \mathbb{P}_{\mathbf{p}} h(S)$. Hint:

$$\begin{aligned} H(\mathbf{p}) &= H(p_1 \times 1 + (1-p_1) \times 0, p_2 \ldots, p_n) \\ &= p_1 H(1, p_2, \ldots, p_n) + (1-p_1) H(0, p_2, \ldots, p_n) \\ &= p_1 \Big[p_2 H(1,1,p_3 \ldots, p_n) + (1-p_2) H(1,0,p_3 \ldots, p_n) \Big] \\ &\quad + (1-p_1) \Big[p_2 H(0,1,p_3 \ldots, p_n) + (1-p_2) H(0,0,p_3 \ldots, p_n) \Big], \end{aligned}$$

and so on. Remember that the order of the arguments of H is irrelevant.

[3] (Chebyshev 1846; Hoeffding 1956, Corollary 2.1) Suppose $G(\mathbf{p}) = \mathbb{P}_{\mathbf{p}} g(S)$, as in the previous Problem. As a continuous function on the compact set $[0,1]^n$ it must achieve its maximum at some point $\boldsymbol{a}$, which need not be unique. Amongst the maximizers choose $\boldsymbol{a}$ of the form $(a_1, \ldots, a_m, \mathbf{0}_{n_0}, \mathbf{1}_{n_1})$ with $0 < a_1 \le \cdots \le a_m < 1$ and $n_0 + n_1$ maximal.

(i) Define $b := (a_m + a_1)/2$ and $t_0 = (a_m - a_1)/2$ and $\epsilon := \min(b, 1-b)$. Show that $a_1 = b - t_0$ and $a_m = b + t_0$ and $0 \le t_0 < \epsilon$.

(ii) Define $T = S - X_1 - X_m$. Show that $G(\boldsymbol{a}) = a_1 a_m A + 2bB + C$ where

$$A = \mathbb{P}_{\boldsymbol{a}} \Big[g(2+T) - 2g(1+T) + g(T) \Big],$$
$$B = \mathbb{P}_{\boldsymbol{a}} \Big[g(1+T) - g(T) \Big], \quad \text{AND} \quad C = \mathbb{P}_{\boldsymbol{a}} g(T)$$

do not depend on a_1 or a_m.

(iii) Define $\boldsymbol{b} \in [0,1]^n$ by $b_1 = b - t$ and $b_m = b + t$ and $b_j = a_j$ otherwise, for a value of t with $|t| \le \epsilon$.

(iv) Show that $0 \le G(\boldsymbol{a}) - G(\boldsymbol{b}) = (t^2 - t_0^2)A$.

(v) Show that $A < 0$ by considering the choice $t = \epsilon$: if $A \ge 0$ then $G(\boldsymbol{b}) \ge G(\boldsymbol{a})$, which would mean that $\boldsymbol{b}$ is another maximizer with

$$\sum\nolimits_j (\{b_j = 0\} + \{b_j = 1\}) > n_0 + n_1.$$

(vi) Consider $t = 0$. Show that $G(\boldsymbol{b}) > G(\boldsymbol{a})$ if $t_0 \ne 0$.

(vii) Conclude that $a_1 = \cdots = a_m = b$. That is, $S - n_1$ has a $\text{BIN}(m, b)$ distribution under $\mathbb{P}_{\boldsymbol{a}}$.

[4] Suppose $h_1(t) := a_1 + b_1 t$ and $h_2(t) := a_2 + b_2 t$ for $0 < t < 1$ with $h_2(t) \neq 0$ for all t. Show that $h_2(t)^2 D\Big(h_1(t)/h_2(t)\Big) = b_1 a_2 - a_1 b_2$, where D denotes differentiation with respect to t. Deduce that $h_1(t)/h_2(t)$ is strictly increasing if $b_1 a_2 > a_1 b_2$.

[5] From equation <2>: $A_k(\mathbf{p}) = p_1 A_{k-1}(\mathbf{p}[-1]) + (1 - p_1) A_k(\mathbf{p}[-1])$ for $\mathbf{p} := (p_1, \ldots, p_n)$. Define $h_1(p_1) := A_k(\mathbf{p})$ and $h_2(p_1) =: A_k(\mathbf{p})$ for fixed $\mathbf{p}[-1]$. Use Problem [4] to show that $h_1(p_1)/h_2(p_1)$ is strictly increasing if $p_i > 0$ for $i \geq 2$. Hint: Example <9> tells you something useful about the coefficients.

[6] Show that the function $E_k(\mathbf{p}) := \mathbb{P}_{\mathbf{p}}\{S \geq k\}$, for $k \in [\![n]\!]$, is a LIPSCHITZ continuous function on $[0,1]^n$. Hint: $A_k(\mathbf{p}) = \sum_{J \subseteq [\![n]\!]} C_J \prod_{j \in J} p_j$ and $|a_i \ldots a_n - b_1 \ldots b_n| \leq |\mathbf{a} - \mathbf{b}|_1$ for $\mathbf{a}, \mathbf{b} \in [\![0,1]^n$.

[7] Suppose $Y \sim \text{BIN}(n, \delta)$ with $0 < \delta < 1$ and $W \sim \text{BIN}(n+1, \theta)$ with $\mu = n\delta = (n+1)\theta$. With $\boldsymbol{\theta} := \theta \mathbb{1}_{n+1}$, show that

$$\mathbb{P}\{Y \geq k\} = E_k(\delta \mathbb{1}_n) < E_k(\boldsymbol{\theta}) = \mathbb{P}\{W \geq k\} \qquad \text{for } 1 + \mu \leq k \leq n.$$

Argue as follows.

(i) With $\boldsymbol{\zeta} := \delta \mathbb{1}_{n-1} \in \text{rel-int}\, \mathbb{B}_{n-1}(\delta)$ and $\mathbf{p} := (\boldsymbol{\zeta}, \delta, 0) \in \mathbb{B}_{n+1}(\theta)$, define

$$\mathbf{p}^* := \sigma_{n,n+1,\delta/2}(\mathbf{p}) = (\boldsymbol{\zeta}, \delta/2, \delta/2).$$

By Lemma <34> (iv) and (v), $E_k(\mathbf{p}) < E_k(\mathbf{p}^*)$ for $1 + \mu \leq k \leq n$ because $p_n + p_{n+1} \geq \boldsymbol{\zeta}_{max}$.

(ii) Complete the argument by invoking Theorem <30> for $\mathbf{p}^*$, which lies in rel-int $\mathbb{B}_{n+1}(\theta)$.

4.10 Notes

Pitman (1997) gave an excellent survey of facts about polynomials with only real roots, with a host of probabilistic and combinatorial applications.

The inequality <10>, in Section 4.3, was attributed to Isaac Newton by Hardy, Littlewood, and Pólya (1989, Sections 2.22, 4.3), who cited page 173 of Newton's *Arithmetica universalis: sive de compositione et resolutione arithmetica liber [Opera, I.]*. Unfortunately I was unable to find the target (in Latin) of this citation, or an English translation thereof. At first I found it easier to follow the sketch proof by Lévy (1937, page 88). I was puzzled why HLP needed to replace the polynomial $p(x)$ by $y^n p(x/y)$ until I saw an argument by Vatutin and Mikhailov (1983), which revealed that the HLP trick was merely a device to change the order of the coefficients in the same way as operator τ in Lemma <6> does. I decided to

include my more pedantic treatment of τ in that Lemma because I was worried about the role of roots of $p(z)$ that equal zero.

Fortunately, my problems with Newton's inequality were solved when I looked at an unpublished paper by Stein (1990, page 5) that pointed me at Whiteside (2008, Volume 1, page 519), which contained a translation of some of Newton's work on roots of equations. Footnote (38) on page 523 described Newton's method in modern algebraic notation. As explained in the General Introduction to Whiteside (2008, Volume 5), the *Arithmetica Universalis* evolved from a set of notes submitted to a Cambridge library by Newton, in fulfillment of an obligation attached to the chair he held. They were later published by another professor, much to Newton's displeasure. Interesting reading.

I first learned the surprising fact about the median of the Binomial from Le Cam (1983, §3), who cited Jogdeo and Samuels (1968), which became my main source for Section 4.8. Section 4.6 is my reworking of material from a short (but most impressive) paper by Samuels (1965), with much help from Hoeffding (1956). As the reader might suspect, I had a lot of trouble with strict inequalities; counterexamples to my theorems kept popping up.

Samuels attributed the results that appear in my Section 4.7 to Hoeffding, although he also noted that similar results were obtained by Chebyshev (1846), who used the pairwise averaging trick to show that the tail probability $E_k(p_1\{S \geq m\}$ is maximized when there is a value θ such that every p_j belongs to $\{0, \theta, 1\}$. Apparently unaware of Chebyshev's paper, Hoeffding proved a similar fact for a general $\mathbb{P}g(S)$. See Problem [3]. See Seneta (1998, pages 206–207) and Maistrov (2014, pages 191,196) for very interesting discussions of Chebyshev (1846).

References

Barvinok, A. (2002). *A Course in Convexity*, Volume 54 of *Graduate Studies in Mathematics*. American Mathematical Society.

Chebyshev, P. L. (1846). Démonstration élémentaire d'une proposition générale de la théorie des probabilités. *Journal für die reine und angewandte Mathematik (Crelle's journal) 33*, 259–267. (The author's surname was spelled 'Tchebichef' in this paper.).

Hardy, G., J. Littlewood, and G. Pólya (1989). *Inequalities*. Cambridge University Press. (Reprint of the the 1952 second edition).

Hoeffding, W. (1956). On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics 27*(3), 713–721.

Jogdeo, K. and S. M. Samuels (1968). Monotone convergence of binomial probabilities and a generalization of Ramanujan's theorem. *Annals of Mathematical Statistics 39*, 1191–1195.

Le Cam, L. (1983). A remark on empirical measures. In P. J. Bickel, K. Doksum, and J. L. Hodges (Eds.), *Festschrift for Erich Lehmann*, pp. 305–327. Belmont, California: Wadsworth.

Lévy, P. (1937). *Théorie de l'addition des variables aléatoires*. Paris: Gauthier-Villars. References from the 1954 second edition.

Maistrov, L. E. (2014). *Probability Theory: A Historical Sketch*. Academic Press. Translated from the 1967 Russian original by Samual Kotz.

Pitman, J. (1997). Probabilistic bounds on the coefficients of polynomials with only real zeros. *Journal of Combinatorial Theory, Series A 77*(2), 279–303.

Pollard, D. (2001). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.

Pólya, G. and G. Szegö (1976). *Problems and Theorems in Analysis II: Theory of Functions. Zeros. Polynomials. Determinants. Number Theory. Geometry*. Springer-Verlag.

Prasolov, V. V. (2004). *Polynomials*. Springer. (Translated from the 2001 Russian edition).

Samuels, S. M. (1965). On the number of successes in independent trials. *Annals of Mathematical Statistics 36*, 1272–1278.

Seneta, E. (1998). Early influences on probability and statistics in the Russian empire. *Archive for History of Exact Sciences 53*(3-4), 201–213.

Stein, C. (1990). Application of Newton's identities to a generalized birthday problem and to the Poisson binomial distribution. Technical Report 354, Stanford Department of Statistics. Available as report EFS NSF 354 at https://statistics.stanford.edu/resources/technical-reports.

Vatutin, V. and V. Mikhailov (1983). Limit theorems for the number of empty cells in an equiprobable scheme for group allocation of particles. *Theory of Probability and Its Applications 27*(4), 734–743.

Whiteside, D. T. (Ed.) (2008). *The mathematical papers of Isaac Newton*. Cambridge University Press. (A paperback reprint of the 8 volumes of an earlier (1967–?) hardcover printing. Described by CUP as "an annotated and critical edition, of all the known mathematical papers of Isaac Newton ... Translations of papers in Latin face the original text

and notes are printed on the page-openings to which they refer, so far as possible. ").

Chapter 5
Orlicz spaces

SECTION 5.1 *introduces* $\mathcal{L}^\Psi$ *spaces, generalizations of the usual* $\mathcal{L}^p$ *spaces, which have proved useful in the study of sample paths of stochastic processes and in the theory of empirical processes.*

SECTION 5.2 *describes some useful properties of* ORLICZ *functions* Ψ *on* $\mathbb{R}^+$*, the convex functions that are used to define the* $\mathcal{L}^\Psi$ *spaces. They are shown to have left- and right-derivatives everywhere, which leads to useful integral representation.*

SECTION 5.3 *describes a set of* ORLICZ *functions that grow much faster than any pth power, a property that leads to several useful inequalities.*

SECTION 5.4 *presents a few simple methods for proving maximal inequalities using* ORLICZ *norms. These inequalities are the basis for some chaining bounds that were popular in empirical process theory.*

SECTION 5.5 *shows how the conjugate of* Φ *of an* ORLICZ *function* Ψ *is related to the* YOUNG *inequality, an analog of the* HÖLDER *inequality. The* ORLICZ *space* $\mathcal{L}^\Phi$ *is then identified with a subset of the space of continuous linear functionals on* $\mathcal{L}^\Psi$*.*

*SECTION 5.6 *describes some of the differences and similarities between* $\mathcal{L}^p$ *and* $\mathcal{L}^\Psi$ *if* Ψ *grows rapidly enough.*

5.1 What is an ORLICZ space?

During the 1980's, bounds involving norms on ORLICZ spaces of functions became very popular in the empirical process literature, in part I suspect because chaining arguments (see Chapter 10) are cleaner with norms than with tail probabilities. More recently, the literature seems to have turned back to working with tail probabilities.

An ORLICZ space is a most useful generalization of the concept of an $\mathcal{L}^p$ space. It replaces the pth power function by a general ORLICZ function, that is, a convex, increasing function $\Psi : \mathbb{R}^+ \to \mathbb{R}^+$ with $\Psi(0) = 0$ and $\Psi(x) \to \infty$ as $x \to \infty$. The concept applies with any measure but I have used it mostly for probability measures. The special case where $\Psi_2(x) = \exp(x^2) - 1$ is particularly useful in the study of subgaussianity.

Remark. Here and subsequently in this Chapter "increasing" is treated as a synonym for the clumsier "nondecreasing". It allows an ORLICZ function Ψ to be zero on an interval $[0, x_0]$ but then convexity forces Ψ to be strictly increasing on $[x_0, \infty)$.

Henceforth $\mathcal{Y}$ will denote the set of all ORLICZ functions on $\mathbb{R}^+$, with $\mathcal{Y}_\infty$ the subset where $\Psi(x)/x \to \infty$ as $x \to \infty$.orlicz

<1> **Definition.** *For a measure space $(\mathcal{X}, \mathcal{A}, \mu)$ and an* ORLICZ *function Ψ define $\mathcal{L}^\Psi(\mathcal{X}, \mathcal{A}, \mu)$ to be the set of all $\mathcal{A}$-measurable real-valued functions f on $\mathcal{X}$ for which $\mu\Psi(|f|/c_0) < \infty$ for at least one positive constant c_0. For each such f the* **orlicz *norm*** $\|f\|_\Psi$ *(or Ψ-norm) is defined as*

$$\|f\|_\Psi := \inf\{c > 0 : \mu\Psi(|f(x)|/c) \le 1\}.$$

□

□

Remark. Some authors define $\|f\|_\Psi$ for every measurable f, interpreting the infimum as $+\infty$ when $\mu\Psi(|f(x)|/c) = \infty$ for each $c > 0$. The space $\mathcal{L}^\Psi$ then consists of all f for which $\|f\|_\Psi < \infty$. Actually, it was Luxemburg (1955) who first introduced the norm $\|f\|_\Psi$; Orlicz (1932, 1936) had worked with another norm, which I will temporarily denote by $\|\cdot\|_\bullet$ to avoid confusion. However the name Orlicz seems to be firmly attached to $\|\cdot\|_\Psi$ in a lot of the literature. As shown in Problem [17], the two norms are equivalent, in the sense that $\|f\|_\Psi \le \|f\|_\bullet \le 2\|f\|_\Psi$ for all f in $\mathcal{L}^\Psi$. I like to use the letter $\mathcal{Y}$ as an indirect nod to the work of William Henry Young. See Section 5.5 for more about that remarkable mathematician.

If $\mu\Psi(|f|/c_0) < \infty$, a dominated convergence argument shows that $\mu\Psi(|f|/(c_0+n)) \to 0$ as $n \to \infty$. The set $\{c > 0 : \mu(|f|/c) \le 1\}$ is either $(0, \infty)$, in which case $\|f\|_\Psi = 0$ and $f = 0$ ae$[\mu]$, or an interval of the form $[c_1, \infty)$ in which case $\|f\|_\Psi = c_1$ (Problem [1]).

For the special case where $\Psi(x) = x^p$, the inequality $\mu\Psi(|f|/c)) \le 1$ is equivalent to $\mu|f|^p \le c^p$, which ensures that the ORLICZ norm agrees with the usual $\|f\|_p$ norm (for $p \ge 1$) and the ORLICZ space is the same as the usual $\mathcal{L}^p(\mathcal{X}, \mathcal{A}, \mu)$ space.

For the general case it is important to keep the constant c inside that Ψ function. For example, here is how to prove that $\|f+g\|_\Psi \le \|f\|_\psi + \|g\|_\Psi$ for $f, g \in \mathcal{L}^\Psi$: for each $c > \|f\|_\Psi$ and $d > \|g\|_\Psi$, convexity and monotonicity of Ψ give

$$\Psi\left(\frac{|f+g|}{c+d}\right) \le \Psi\left(\frac{c|f|/c}{c+d} + \frac{d|f|/d}{c+d}\right) \le \frac{c}{c+d}\Psi(|f|/c) + \frac{d}{c+d}\Psi(|g|/d).$$

Integrate both sides with respect to μ, using the fact that $\mu\Psi(|f|/c) \le 1$ and $\mu\Psi(|g|/d) \le 1$ to deduce that

$$\mu\Psi\left(\frac{|f+g|}{c+d}\right) \le 1 \qquad \text{if } c > \|f\|_\Psi \text{ and } d > \|g\|_\Psi.$$

It follows that $\|f+g\|_\Psi \le c + d$. Complete the argument by letting c decrease to $\|f\|_\Psi$ and d decrease to $\|g\|_\Psi$. (According to Problem [1], you could also take $c = \|f\|_\Psi$ and $d = \|g\|_\Psi$, except in some trivial cases.)

Remark. As usual $\|\cdot\|_\Psi$ fails to be a norm on $\mathcal{L}^\Psi$ only because $\|f\|_\Psi = 0$ if and only if $\mu\{x : f(x) \neq 0\} = 0$. The usual trick of working with the space $L^\Psi(\mathcal{X}, \mathcal{A}, \mu)$ of μ-equivalence classes makes $\|\cdot\|_\Psi$ a true norm and L^Ψ a BANACH space.

The most important ORLICZ functions for my purposes are the power functions, $\Psi(x) = x^p$ for a fixed $p \geq 1$, and the exponential power functions $\Psi_\alpha(x) := \exp(x^\alpha) - 1$ for $\alpha \geq 1$. The function Ψ_2 corresponds to subgaussian, and Ψ_1 to subexponential, tail behaviour.

Remark. It is also possible (Problem [9]) to define an ORLICZ function Ψ_p, for $0 < p < 1$, that behaves like $\exp(x^p)$ far enough from the origin.

When seeking to bound $\|f\|_\Psi$ for some f in $\mathcal{L}^\Psi$ I often find it easier to first obtain a constant c_0 for which $\mu\Psi(|f|/c_0) \leq K_0$, where K_0 is a constant larger than 1. As the next Lemma shows, a simple convexity argument turns such an inequality into a bound on $\|f\|_\Psi$.

<2> **Lemma.** *Suppose Ψ is an* ORLICZ *function. If $\mu\Psi(|f|/c_0) \leq K_0$ for some finite constants c_0 and $K_0 > 1$ then $\|f\|_\Psi \leq c_0 K_0$.*

Proof. For each θ in $[0, 1]$ convexity of Ψ gives

$$\mu\Psi\left(\frac{\theta|f|}{c_0}\right) \leq \theta\mu\Psi\left(\frac{|f|}{c_0}\right) + (1-\theta)\Psi(0) \leq \theta K_0.$$

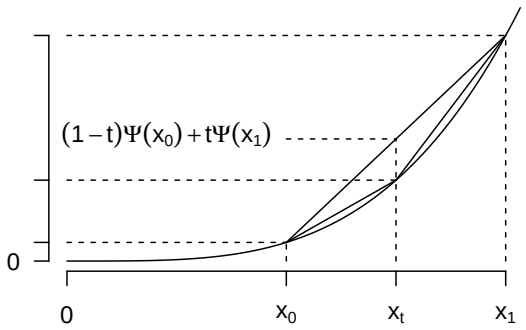
□ The choice $\theta = 1/K_0$ makes the last bound equal to 1.

5.2 Facts about ORLICZ functions

Suppose $\Psi \in \mathcal{Y}$. Define the ***slope function***

$$S(x, y) := \frac{\Psi(y) - \Psi(x)}{y - x} \qquad \text{for } 0 \leq x < y.$$

The slope is always nonnegative because Ψ is increasing.



Most of the useful properties of ORLICZ functions follow from a pair of inequalities for the slope function. For given values $0 \le x_0 < x_1 = x_0 + D$ and $0 < t < 1$ define $x_t := (1-t)x_0 + tx_1$. Note that $x_t - x_0 = tD$ and $x_1 - x_t = (1-t)D$. Convexity of Ψ gives

$$\Psi(x_t) \le (1-t)\Psi(x_0) + t\Psi(x_1),$$

which implies

$$S(x_0, x_t) = \frac{\Psi(x_t) - \Psi(x_0)}{x_t - x_0} \le \frac{t\Psi(x_1) - t\Psi(x_0)}{tD} = S(x_0, x_1),$$
$$S(x_t, x_1) \ge \frac{(1-t)\Psi(x_1) - (1-t)\Psi(x_0)}{(1-t)D} = S(x_0, x_1).$$

Re-expressed more suggestively, (first with $x = x_0$ and $y = x_t$ and $y' = x_1$ then with $x = x_0$ and $x' = x_t$ and $y = x_1$) the inequalities become

$$S(x,y) \le S(x,y') \qquad \text{if } 0 \le x < y < y'$$
$$S(x,y) \le S(x',y) \qquad \text{if } 0 \le x < x' < y.$$

That is,

<3> the function $y \mapsto S(x,y)$ is increasing for each fixed x;
the function $x \mapsto S(x,y)$ is increasing for each fixed y.

Remark. The preceding argument used only the convexity of Ψ. Properties <3> are general facts about convex functions on $\mathbb{R}$ or a subinterval thereof. If Ψ is also increasing then $S \ge 0$.

<4> **Theorem.** *If $\Psi \in \mathcal{Y}$ then*

(i) At each $x \ge 0$ the right-derivative

$$\dot{\Psi}_R(x) = \lim_{t \searrow 0} \left(\Psi(x+t) - \Psi(x)\right)/t$$

exists. The function $\dot{\Psi}_R$ is nonnegative, increasing, and continuous from the right.

(ii) At each $x > 0$ the left-derivative

$$\dot{\Psi}_L(x) = \lim_{w \nearrow 0} \left(\Psi(x) - \Psi(w)\right)/(x-w)$$

exists. The function $\dot{\Psi}_L$ is nonnegative, increasing, and continuous from the left.

(iii) The function Ψ is continuous on $\mathbb{R}^+$.

(iv) If $0 \le x < y$ *then* $\dot\Psi_R(x) \le \dot\Psi_L(y) \le \dot\Psi_R(y)$ *and*

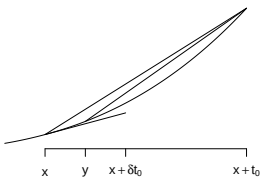
$$(y-x)\dot\Psi_R(x) \le \Psi(y) - \Psi(x) \le (y-x)\dot\Psi_L(y).$$

(v) $\Psi(x) = \int_0^x \dot\Psi_R(t)\,dt = \int_0^x \dot\Psi_L(t)\,dt$ *for each* $x > 0$. *Consequently, the set* $\{t > 0 : \dot\Psi_R(t) > \dot\Psi_L(t)\}$ *has zero* LEBESGUE *measure.* □

Proof. For (i): The function $t \mapsto S(x, x+t)$ is increasing and nonnegative. Thus $S(x, x+t)$ decreases to a finite limit $\dot\Psi_R(x) \ge 0$ as $t \searrow 0$.

If $x < x'$ and $t > 0$ then $S(x, x+t) \le S(x, x'+t) \le S(x', x'+t)$ Let t decrease to 0 to deduce that $\dot\Psi_R(x) \le \dot\Psi_R(x')$.

For right-continuity at a given x, we need to show that $\dot\Psi_R(y) \approx \dot\Psi_R(x)$ if y lies in a small enough interval to the right of x. Given a small $\epsilon > 0$ first choose a $t_0 > 0$ for which $\dot\Psi_R(x) \le S(x, x+t_0) \le \dot\Psi_R(x) + \epsilon$ then consider $\delta > 0$ and y for which $x < y < x + \delta t_0 < x + t_0$, so that



$$\begin{aligned}
\dot\Psi_R(x) \le \dot\Psi_R(y) &= \lim_{t\searrow 0} S(y, y+t) \qquad \text{decreasing limit} \\
&\le S(y, x+t_0) = \frac{\Psi(x+t_0) - \Psi(y)}{x + t_0 - y} \\
&\le \frac{\Psi(x+t_0) - \Psi(x)}{x + t_0 - (x + \delta t_0)} = \frac{S(x, x+t_0)}{1-\delta} \\
&\le \frac{(1-\delta)\dot\Psi_R(x) + \delta\dot\Psi_R(x) + \epsilon}{1-\delta} \\
&\le \dot\Psi_R(x) + 2\epsilon \qquad \text{if } \delta \text{ is small enough.}
\end{aligned}$$

For (ii): For $0 \le w < x$ the function $\mapsto S(w, x)$ is increasing and nonnegative. Thus $S(w, x)$ increases to a limit $\dot\Psi_L(x) \ge 0$ as $w \nearrow 0$. The limit is finite because

$$\begin{aligned}
S(w, x) &\le S(w, x+t) \qquad \text{for each } t > 0 \\
&\le S(x, x+t) \searrow \dot\Psi_R(x) \qquad \text{as } t \searrow 0,
\end{aligned}$$

showing that $\dot\Psi_L(x) \le \dot\Psi_R(x) < \infty$. The rest of the argument is analogous to the argument for (i), except that approximations are made from below instead of from above.

For (iii): Existence of left and right derivatives implies continuity.

For (iv): The inequality $\dot\Psi_L(y) \le \dot\Psi_R(y)$ was already established in the proof of (ii). If $x < y$ then

$$\dot\Psi_R(x) = \lim_{w\searrow x} S(x, w) \le S(x, y) \le \lim_{w\nearrow y} S(w, y) = \dot\Psi_L(y).$$

Multiply through by $y - x$ to get the other inequality.

For (v): For a positive integer n define $\delta = x/n$ and $x_i = i\delta$ for $i = 0, 1, 2, \ldots$. Then we have

$$\begin{aligned}
\Psi(x) &= \sum\nolimits_{i=1}^{n} \Big(\Psi(x_i) - \Psi(x_{i-1}) \Big) \\
&\geq \sum\nolimits_{i=1}^{n} \delta \dot{\Psi}_R(x_{i-1}) \qquad \text{by (iv)} \\
&\geq \delta \dot{\Psi}_R(0) + \sum\nolimits_{i=2}^{n} \int_{x_{i-2}}^{x_{i-1}} \dot{\Psi}_R(t)\, dt = \delta \dot{\Psi}_R(0) + \int_0^{x_{n-1}} \dot{\Psi}_R(t)\, dt \\
&\to \int_0^x \dot{\Psi}_R(t)\, dt \qquad \text{as } n \to \infty, \text{ by dominated donvergence.}
\end{aligned}$$

Similarly,

$$\Psi(x) \leq \sum\nolimits_{i=1}^{n} \delta \dot{\Psi}_R(x_i) \leq \sum\nolimits_{i=1}^{n} \int_{x_i}^{x_{i+1}} \dot{\Psi}_R(t)\, dt = \int_{x_2}^{x_{n+1}} \dot{\Psi}_R(t)\, dt,$$

which also converges to $\int_0^x \dot{\Psi}_R(t)\, dt$ as $n \to \infty$. Together the upper and lower bounds establish the first asserted integral representation. The argument for $\dot{\Psi}_L$ is similar.

By subtraction, $0 = \int_0^x \Big(\dot{\Psi}_R(t) - \dot{\Psi}_L(t) \Big)\, dt = 0$. By (iv), the integrand is nonegative. Hence the integrand must equal 0 except on a set of zero LEBESGUE measure. □

Remarks.

(i) Problem [4] strengthens part (v) by showing that the both the nondecreasing functions $\dot{\Psi}_L$ and $\dot{\Psi}_R$ can have, at worst, only countably many jumps. Thus the set $\{t > 0 : \dot{\Psi}_R(t) > \dot{\Psi}_L(t)\}$ is not just LEBESGUE negligible but is also, at worst, countably infinite.

(ii) Problem [5] show that every Ψ on $\mathbb{R}^+$ defined by $\Psi(x) = \int_0^x \psi(t)\, dt$, with ψ nonnegative and nondecreasing, is an ORLICZ function. Moreover, such a ψ must satisfy the inequalities

$$\dot{\Psi}_L(t) \leq \psi(t) \leq \dot{\Psi}_R(t) \qquad \text{for all } t > 0.$$

Thus ψ can differ from $\dot{\Psi}_L$ or $\dot{\Psi}_R$ only at the (at worst countable) set of t's at which $\dot{\Psi}_R(t) \neq \dot{\Psi}_L(t)$. In this sense, the function Ψ essentially determines ψ uniquely.

(iii) The condition $\Psi(x)/x \to \infty$ as $x \to \infty$ that is also required for ORLICZ functions in $\mathcal{Y}_\infty$ can also be written as $\psi(t) \to \infty$ as $t \to \infty$ because

$$\psi(x) \geq x^{-1}\Psi(x) \geq x^{-1} \int_{x_0}^{x} \psi(t)\, dt \geq x^{-1}(x - x_0)\psi(x_0)$$

for $x \geq x_0$, for arbitrarily large x_0.

5.3 Orlicz functions that grow exponentially fast

To avoid a lot of nitpicking, each Orlicz function Ψ in this Section should be assumed to have $\{\Psi = 0\} = \{0\}$.

For some purposes, it helps to have $\Psi(x)/x^p \to \infty$ as $x \to \infty$, for each finite p. For example, Ledoux and Talagrand (1991, page 310) required

$$\Psi^{-1}(ab) \le C_0 \left(\Psi^{-1}(a) + \Psi^{-1}(b)\right) \qquad \text{if } \min(a,b) \ge a_0, \tag{5}$$

for constants C_0 and $a_0 > 0$. Equivalently,

$$\Psi(x)\Psi(y) \le \Psi\left(C_0(x+y)\right) \qquad \text{if } \min(x,y) \ge x_0. \tag{6}$$

Write $\mathcal{Y}_{\text{LT}}$ for the set of all Orlicz functions that satisfy this condition for each strictly positive x_0, with C_0 allowed to depend on x_0. (Often such a bound also holds for x_0 equal to 0.)

The main result in this Section provides a simple sufficient condition for <6>.

Van der Vaart and Wellner (1996, Lemma 2.2.2) imposed a similar-looking condition to control the Ψ-norm of a maximum:

$$\Psi(x)\Psi(y) \le \Psi\left(C_1 xy\right) \qquad \text{if } \min(x,y) \ge x_1. \tag{7}$$

for constants C_1 and $x_1 > 0$. Write $\mathcal{Y}_{\text{VW}}$ for the set of all Orlicz functions that satisfy this condition for each strictly positive x_1, with C_1 depending on x_1.

It is easy to show for the Orlicz function $\Psi(x) = x^p$, with $p \ge 1$, that $\Psi \in \mathcal{Y}_{\text{VW}}$ but $\Psi \notin \mathcal{Y}_{\text{LT}}$: consider the case where $x = y$. Thus <7> does not require exponential growth, although its most interesting uses do involve rapidly growing Orlicz functions—see Example <12>. However, the $\mathcal{Y}_{\text{LT}}$ condition implies $\Psi(x)^2 \le \Psi(2C_0 x)$ for $x \ge x_0 > 0$, an exponential growth rate. (It also forces $2C_0 \ge 1$ because $\Psi(x) \le \Psi(x)^2$ when $\Psi(x) \ge 1$.)

For the purposes of deriving sufficent conditions for rapid growth it is cleaner to work with $g(x) := \log\left(1 + \Psi(x)\right)$, a nonnegative, increasing function with $g(0) = 0$. The g function inherits continuity and existence of left and right derivatives from Ψ. Conversely, a quick way to generate Orlicz functions is to start from a twice-differentiable, increasing function g with $g(0) = 0$. If we define $\Psi(x) = e^{g(x)} - 1$ then $\ddot{\Psi}(x) = \left(\dot{g}(x)^2 + \ddot{g}(x)\right)\Psi(x)$, which shows that $\Psi \in \mathcal{Y}$ if $\dot{g}(x)^2 + \ddot{g}(x) \ge 0$ for all $x > 0$.

<8> **Definition.** *Define $\mathcal{Y}_{\exp}$ to be the set of functions $\Psi = e^g - 1$ in $\mathcal{Y}$ for which there are constants K_0 and x_1 such that $2g(x) \le g\left(K_0 x\right)$ if $x \ge x_1$.* □

As you will see in later Chapters, the $\mathcal{Y}_{\exp}$ property turns out to be a very natural assumption for a lot of theory developed in the past few decades. foobar

<9> **Lemma.** *Suppose $\Psi = e^g - 1$ is an* ORLICZ *function with $\{\Psi = 0\} = \{0\}$.*

(i) If $\Psi \in \mathcal{Y}_{\text{exp}}$ then there exist a constant C_g such that $2g(x) \le g(C_g x)$ for all $x \in \mathbb{R}^+$. Consequently, $g(x) + g(y) \le g\left(C_g \max(x,y)\right)$ for all x, y in $\mathbb{R}^+$.

(ii) If $\Psi \in \mathcal{Y}_{\text{exp}}$ then $\Psi(x)\Psi(y) \le \Psi(C_g \max(x,y))$ for all x, y in $\mathbb{R}^+$ and (equivalently) $\Psi^{-1}(ab) \le C_g \max\left(\Psi^{-1}(a), \Psi^{-1}(b)\right)$ for all a, b in $\mathbb{R}^+$. Consequently, $\Psi \in \mathcal{Y}_{\text{LT}}$.

(iii) $\mathcal{Y}_{\text{exp}} \subset \mathcal{Y}_{\text{VW}}$

(iv) If $\Psi \in \mathcal{Y}_{\text{exp}}$ then there is a positive x_0 such that $g(x) \ge (x/C_g x_0)^\alpha g(x_0)$ for $x \ge x_0$, where $\alpha := 1/\log_2 C_g$.

Proof. For (i), first extend the upper bound on $2g(x)$ to cover a neighborhood of the origin. Define $x_0 := \Psi^{-1}(1)/4$. Then, from the facts that $y \mapsto \Psi(y)/y$ is an increasing function on $(0,\infty)$ and $y/(1+y) \le \log(1+y) \le y$ for $y > 0$, we have

$$\begin{aligned} g(4x) &= \log\left(1 + \Psi(4x)\right) \\ &\ge \Psi(4x)/\left(1 + \Psi(4x)\right) \\ &\ge 4\Psi(x)\left(1 + \Psi(4x_1)\right) \qquad \text{because } \Psi(4x)/(4x) \ge \Psi(x)/x \\ &= 2\Psi(x) \ge 2g(x) \qquad \text{for } 0 \le x \le x_0. \end{aligned}$$

Extend to cover the range $[x_0, x_1]$. Define K_1 by $2g(x_1) = g(K_1 x_0)$, so that

$$\max_{x_0 \le x \le x_1} 2g(x)/g(K_1 x) \le 2g(x_1)/g(K_1 x_0) = 1.$$

With $C_g := \max(K_0, 4, K_1)$ we then have $2g(x) \le g(C_g x)$ for all x in $\mathbb{R}^+$. For the consequence note that $g(x) + g(y) \le 2g(\max(x,y))$.

For (ii) and (iii), first note that

$$\begin{aligned} \Psi(x)\Psi(y) &= e^{g(x)+g(y)} - e^{g(x)} - e^{g(y)} + 1 \\ &\le e^{g(x)+g(y)} - 1 \le e^{g(C_g \max(x,y))} - 1 = \Psi\left(C_g \max(x,y)\right). \end{aligned}$$

For Ψ in $\mathcal{Y}_{\text{LT}}$ use $\max(x,y) \le x + y$ and for Ψ in $\mathcal{Y}_{\text{VW}}$ use $\max(x,y) \le xy/x_1$ when $\min(x,y) \ge x_1$.

For (iv), suppose $C_g^{k-1} x_0 \le x < C_\Psi^k x_0$ with $k \in \mathbb{N}$. By induction, $2^p g(x_0) \le g(C_g^p x_0)$ for each p in $\mathbb{N}_0$. Thus

$$g(x) \ge g(C_g^{k-1} x_0) \ge 2^{k-1} g(x_0) = \left(C_g^\alpha\right)^{(k-1)} g(x_0).$$

□ By definition of k we have $C_g^{k-1} > x/(C_\Psi x_0)$.

<10> **Example.** The constant C_g calculated for the previous Lemma might be unnecessarily large. Consider the case where $g_\alpha(x) := x^\alpha$ for some $\alpha \geq 1$. By direct calculation, $\dot{g}(x)^2 + \ddot{g}(x) = \alpha x^{2\alpha-1}(\alpha x + (\alpha - 1)) \geq 0$ and $2x^\alpha \leq (Kx)^\alpha$ if $K = 2^{1/\alpha}$. Thus $\Psi_\alpha(x) := \exp(x^\alpha) - 1$ is an ORLICZ function in $\mathcal{Y}_{\exp}$ if $1 \leq \alpha < \infty$.

The simple form of the g_α function allows us to derive a result slightly sharper than Lemma <9>. As before, for all x, y in $\mathbb{R}^+$,

$$\Psi_\alpha(x)\Psi_\alpha(y) \leq \exp\left(g_\alpha(x) + g_\alpha(y)\right) - 1.$$

If $x + y > 0$ and $x/(x+y) = r$ then

$$\frac{g_\alpha(x) + g_\alpha(y)}{g_\alpha(x+y)} = R_\alpha(r) := r^\alpha + (1-r)^\alpha \qquad \text{for } 0 \leq r \leq 1.$$

The function R_α is convex on $[0,1]$, with maximum value 1 achieved at the endpoints $\{0,1\}$ and minimum value $(1/2)^{\alpha-1}$ at $r = 1/2$, which implies

$$g_\alpha(x) + g_\alpha(y) \leq g_a(x+y) \qquad \text{for all } x, y \text{ in } \mathbb{R}^+.$$

Consequently $\Psi_\alpha \in \mathcal{Y}_{\text{LT}}$. We also have

$$\Psi_\alpha(x)\Psi_\alpha(y) \leq e^{g_\alpha(x+y)} - 1 = \Psi_\alpha(x+y) \qquad \text{for all } x, y \in \mathbb{R}^+$$

and

$$\Psi_\alpha^{-1}(ab) \leq \Psi_\alpha^{-1}(a) + \Psi_\alpha^{-1}(b) \qquad \text{for all } a, b \in \mathbb{R}^+,$$

□ which is much cleaner than <5>.

Problem [9] extends the Example to values of α in $(0,1)$: there is a Ψ_α in $\mathcal{Y}_{\exp}$ for which $\Psi_\alpha(x) = \exp(x^\alpha) - 1$ when x is large enough.

5.4 Norms of maxima of finitely many variables

As noted near the start of the Chapter, my interest in ORLICZ norms was first aroused by their use with the chaining method for constructing simple uniform approximations to stochastic processes $\{X_t : t \in T\}$. That method requires some sort of probabilistic control over the increments $X_t - X_s$, expressible in terms of some distance function on T. When that distance takes the form of an ORLICZ norm $\|X_t - X_s\|_\Psi$ and when the probabilistic approximations are expressed using some other ORLICZ norm, the chaining method takes a particularly simple form. The key idea is to control the maximum of finite sets of random variables. The following Theorem collects together a few techniques that I have extracted from various chaining proofs in the literature. The results are stated for probability measures.

<11> **Theorem.** *Let $X_1, \ldots, X_N$ be nonnegative random variables,* ***not necessarily independent****, but all defined on the same probability space. Define $M := \max_{i\le N} X_i$. Suppose $\max_{i\le N} \|X_i\|_\Psi \le 1$ for some* ORLICZ *function Ψ.*

(i) $\mathbb{P}M \le \Psi^{-1}(N)$.

(ii) $\mathbb{P}_B M \le \Psi^{-1}(N/\mathbb{P}B)$ *where $\mathbb{P}_B$ denotes conditional expectation given an event B with $\mathbb{P}B > 0$.*

(iii) If $\Psi \in \mathcal{Y}_{\exp}$ then $\|M\|_\Psi \le 2C\Psi^{-1}(N)$ for some constant C.

Proof. The assumption that $\max_{i\le N} \|X_i\|_\Psi \le 1$ implies $\mathbb{P}\Psi(X_i) \le 1$ for each i.

For (i): Even though the assertion is a special case of assertion (ii), I think it helps build intuition to see this proof first. Jensen's inequality then non-negativity and monotonicity of Ψ imply

$$\Psi(\mathbb{P}M) \le \mathbb{P}\max_{i\le N}\Psi(X_i) \le \sum\nolimits_{i\le N} \mathbb{P}\Psi(X_i) \le N.$$

For (ii): Partition B into disjoint subsets B_i such that $M = X_i$ on B_i. Without loss of generality assume $\mathbb{P}B_i > 0$ for each i. (Alternatively, just discard those B_i with zero probability.) By definition,

$$\mathbb{P}_B M = \mathbb{P}_B\left(\sum\nolimits_i B_i X_i\right) = \sum\nolimits_i \frac{\mathbb{P}B_i}{\mathbb{P}B}\mathbb{P}_{B_i}X_i.$$

By Jensen's inequality,

$$\Psi(\mathbb{P}_{B_i}X_i) \le \mathbb{P}_{B_i}\Psi(X_i) = \frac{\mathbb{P}B_i\Psi(X_i)}{\mathbb{P}B_i} \le \frac{\mathbb{P}\Psi(X_i)}{\mathbb{P}B_i} \le \frac{1}{\mathbb{P}B_i}.$$

Thus

$$\begin{aligned}\Psi(\mathbb{P}_B M) &= \Psi\left(\sum\nolimits_i \frac{\mathbb{P}B_i}{\mathbb{P}B}\mathbb{P}_{B_i}X_i\right)\\ &\le \sum\nolimits_i \frac{\mathbb{P}B_i}{\mathbb{P}B}\Psi(\mathbb{P}_{B_i}X_i) \qquad \text{by convexity of } \Psi\\ &\le \sum\nolimits_i \frac{\mathbb{P}B_i}{\mathbb{P}B}\frac{1}{\mathbb{P}B_i} = \frac{N}{\mathbb{P}B}.\end{aligned}$$

Perhaps it would be better to write the last equality as an inequality, to cover the case where some $\mathbb{P}B_i$ are zero.

For (iii): Choose $x_1 = \min(1, \Psi^{-1}(1))$, so that $\beta := \Psi^{-1}(N) \ge x_1$. On the set $\{M \ge \beta C_1\}$ we then have

$$N\Psi(M/\beta C_1) = \Psi(\beta)\Psi(M/\beta C_1) \le \Psi(M) \le \sum\nolimits_i \Psi(X_i).$$

Thus $\Psi(M/\beta C_1) \le \Psi(1) + N^{-1}\sum_i \Psi(X_i)$ for all values of M. Take expected values to deduce that $\mathbb{P}\Psi(M/\beta C_1) \le \Psi(1)+1$, which, by Lemma <2>, implies $\|M\|_\Psi \le C_1(2\Psi(1)+1)\beta$. □

See §10.6 for an example that demonstrates the surprising power of inequality (ii) from the Theorem.

<12> **Example.** Suppose $\{Z(t) : t \in T\}$ is a stochastic process indexed by a set equipped with a metric for which $\|Z(s) - Z(t)\|_{\Psi_2} \le d(s,t)$ for $s, t \in T$.

Lemma <9> shows that each of the ORLICZ functions $\Psi_p(x) = \exp(x^p) - 1$, with $p \ge 1$ satisfy the assumption of inequality (iii). In particular, it holds for Ψ_2. If $p \ge 1$ and $k = \lceil p/2 \rceil$ then, for a constant C_p depending on p,

$$\|X\|_p \le \|X\|_{2k} \le C_p \|X\|_{\Psi_2} \qquad \text{for each random variable } X$$

by virtue of the fact that $\mathbb{P}|X/c|^{2k} \le k! \mathbb{P}\Psi_2(|X|/c)$. The constant $(k!)^{1/2k}$ is of order $O(\sqrt{p})$.

In consequence, if s_i, t_i are pairs of points with $d(s_i, t_i) > 0$ then

$$\left\| \max_{i \le N} \frac{|X(s_i) - X(t_i)|}{d(s_i, t_i)} \right\|_p \le K_p \sqrt{1 + \log N}.$$

This bound can be used to derive the inequalities from Pollard (1989, Section 3) that became the main technical tool for the cube-root asymptotic theory developed by Kim and Pollard (1990). □

5.5 Conjugacy and the YOUNG inequality

This Section explains a connection between ORLICZ functions and the conjugates from Chapter 2. Before writing this Chapter, I regarded conjugacy as a slightly exotic topic that would not not really be needed in the rest of the book. Subsequently I came to think that this convexity idea is actually closely related to several constructions that appear in later Chapters. For example, the splitting idea of Fernique for deriving maximal inequalities, which is described and extended in §11.5, seems to be closely related to the YOUNG inequality.

As noted in §2.4, it is convenient to define the conjugate of an ORLICZ function Ψ by first extending its domain of definition to the whole real line: define $\Psi(x) = +\infty$ for $x < 0$. The conjugate Ψ^* is defined by

$$\Psi^*(y) := \sup_{x \in \mathbb{R}} F(x,y) \qquad \text{where } F(x,y) := yx - \Psi(x) \text{ for } x, y \in \mathbb{R}.$$

Write Λ for the restriction of Ψ^* to $\mathbb{R}^+$. The main task in this Section is to find conditions to ensure that Λ is also an ORLICZ function and then to show how Ψ and Λ fit together to produce a result due to Young (1912).

The choice $x = 0$ gives the value $F(0,y) = 0$, and the choice $x < 0$ gives the value $F(x,y) = -\infty$. It follows that

<13> $$\Lambda(y) = \sup_{x \ge 0} \Big(yx - \Psi(x) \Big) \qquad \text{for } y \ge 0.$$

Temporarily write $\psi(t)$ for $\dot\Psi_R(t)$. By Theorem <4> we have $\Psi(x) = \int_0^x \psi(t)\,dt$ for all $x \geq 0$. If $c := \sup_t \psi(t)$ were finite then we would have $\Psi(x) \leq cx$ for all x, implying $\Lambda(y) = +\infty$ for $y > c$. As we want Λ to be an ORLICZ function we need $\psi(t) \to \infty$ as $t \to \infty$. Equivalently, we need $\Psi(x)/x \to \infty$ as $x \to \infty$, that is, $\Psi \in \mathcal{Y}_\infty$.

<14> **Theorem.** *If $\Psi \in \mathcal{Y}_\infty$ then $\Lambda \in \mathcal{Y}_\infty$ with $\Lambda(y) = \int_0^\infty (y - \dot\Psi_R(t))^+\,dt$.*

Proof. As above, write $\psi(t)$ for $\dot\Psi_R(t)$ and define $q(y) := \inf\{t \geq 0 : \psi(t) \geq y\}$, with $q(y) = 0$ if $\psi(0) > y$. The assumption that Ψ belongs to $\mathcal{Y}_\infty$ ensures that $q(y) \to \infty$ as $y \to \infty$.

For $x \geq 0$ and $y \geq 0$ we have

$$F(x,y) = xy - \Psi(x) = \int_0^x y - \psi(t)\,dt.$$

By definition, $y - \psi(t) > 0$ if $t < q(y)$ and $y - \psi(t) \leq 0$ if $t > q(y)$. Thus $x \mapsto F(x,y)$ is increasing while $0 \leq x < q(y)$ and nonincreasing for $x > q(y)$. (There is actually be a flat spot if $\{t : \psi(t) = y\}$ is a nondegenerate interval.) It follows that

$$\Lambda(y) = \sup_x F(x,y) = \int_0^{q(y)} y - \psi(t)\,dt = \int_0^\infty (y - \psi(t))^+\,dt.$$

Also, for $y > 0$ we have

$$\Lambda(y)/y = \sup_x F(x,y) = \int_0^\infty (1 - \psi(t)/y)^+\,dt.$$

At each fixed t the integrand increases monotonely to 1 as $y \uparrow \infty$. By monotone convergence it follows that $\Lambda(y)/y \uparrow \infty$ as $y \uparrow \infty$. □

<15> **Example.** Everything is much simpler when the ORLICZ function Ψ is everywhere differentiable. For example, if $\Psi(x) = x^p/p$ with $1 < p < \infty$ and $p^{-1} + q^{-1} = 1$ then the maximizing x in the definition of $\Lambda(y)$ equals $y^{q/p}$, so that $\Lambda(y) = y^q/q$. Without the $1/p$ in the definition of Ψ the function would be a messier multiple of y^q.

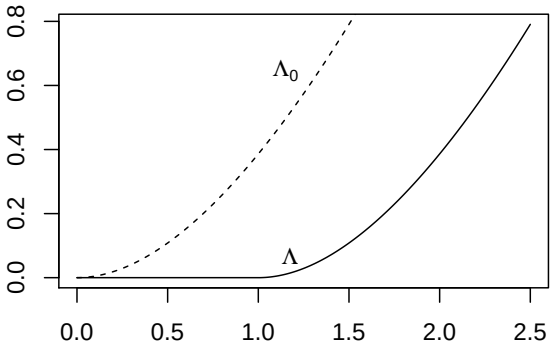
A tad more interesting is the case where $\Psi(x) = e^x - 1$. For $y \leq 1 = \dot\Psi(0)$ the maximum is achieved at $x = 0$, so that $\Lambda(y) = 0$. If $y > 1$ the maximum is now achieved at $x = \log y$ so that $\Lambda(y) = 1 + y\log y - y$. That is

$$\Lambda(y) = 0\{0 \leq y \leq 1\} + (1 + y\log y - y)\,\{y > 1\}$$

The flat spot on $[0,1]$ comes from the fact that $\Psi'(0) > 0$.

The flat spot can be removed by subtracting off a linear function to get a zero right-hand derivative at the origin: if $\Psi_0(x) := e^x - 1 - x$ then

$\dot{\Psi}_0(x) = e^x - 1$. Now the maximum is achieved at $x = \log(1+y)$ so that $\Lambda_0(y) = \Lambda(1+y) = (1+y)\log(1+y) - y$.



Problem [11] shows that $\mathcal{L}^{\Psi_0}(\mathcal{X}, \mathcal{A}, \mathbb{P}) = \mathcal{L}^{\Psi}(\mathcal{X}, \mathcal{A}, \mathbb{P})$, for probability measures $\mathbb{P}$, with

$$K_\Psi \|X\|_{\Psi_0} \geq \|X\|_\Psi \geq \|X\|_{\Psi_0} \qquad \text{for all } X \in \mathcal{L}^\Psi$$

and $\mathcal{L}^\Lambda = \mathcal{L}^{\Lambda_0}$ with

$$\|Y\|_\Lambda \leq \|Y\|_{\Lambda_0} \leq K_\Lambda \|Y\|_\Lambda \qquad \text{for all } Y \in \mathcal{L}^\Lambda.$$

□ where $K_\Psi = 1 + \Psi^{-1}(1) \approx 1.7$ and $K_\Lambda = 1 + \Lambda(2) \approx 2.3$.

Theorem <14> has another, more impressive, consequence. The conclusion from the Theorem,

$$\Lambda(y) = \sup_{x \geq 0} xy - \Psi(x) \qquad \text{with equality when } x = q(y) := \inf\{t \geq 0 : \dot{\Psi}_R(t) \geq y\},$$

can be rewritten as

<16> $$\Lambda(y) + \Psi(x) \geq xy \qquad \text{for all } x, y \geq 0,$$

with equality if $x = q(y)$. This result is often called "Young's inequality". A very similar result was derived by William Henry Young, whom I cannot resist quoting:

> The method [...] has the great advantage as regards the details of the work that it is based on an extremely simple inequality. From this point of view it is far superior to the other methods of proof, almost intuitive as they are, when the necessary preliminary theorems have been proved, that I have indicated elsewhere. Instead of requiring the generalisation of Schwarz's well-known inequality, it suffices to be acquainted with the generalisation of the relation
>
> $2ab \leq a^2 + b^2$, namely, that $(p+1)ab \leq a^{p+1} + pb^{1+1/p}$.
>
> In the prosecution of research one is often hampered by the difficulty of generalising a known formula, owing to the very simplicity and obviousness of its statement. The inequality just written down possesses, however, the great advantage over Schwarz's inequality that its generalisation is almost immediate.
>
> [Young, 1912, pp 225-226]

Why don't mathematicians write like that any more?

Remark. The equality $\Psi^{**} = \Psi$ ensures that there is a symmetry in the roles of Λ and Ψ. That is, $\Lambda^* = \Psi$. There is also equality in <16> if $y = \Psi'_R(x)$.

The YOUNG inequality tells us something about linear functionals on ORLICZ spaces. Remember that a linear map $T : \mathcal{Z} \to \mathbb{R}$ from a vector space $\mathcal{Z}$ equipped with a norm (or semi-norm) $\|\cdot\|$ is continuous if and only if

$$\|T\| := \sup\{|Tz| : \|z\| \le 1\} \qquad \text{is finite.}$$

For ancient reasons, such a T is also called a continuous linear ***functional***. (Once upon a time, functions whose domains were sets of functions were considered scary objects.)

For example, if $1 < p < \infty$ and $p^{-1} + q^{-1} = 1$ then each g in $\mathcal{L}^q$ defines a continuous linear functional on $\mathcal{L}^p$ by

$$T_g(f) := \mu(fg) \qquad \text{for each } f \in \mathcal{L}^p.$$

The continuity comes directly from the HÖLDER inequality:

$$|T_g(f)| = |\mu(fg)| \le \|f\|_p \|g\|_q ,$$

which implies $\|T_g\| \le \|g\|_q$. In fact more is true: $\|T_g\| = \|g\|_q$ and every continuous linear functional T on $\mathcal{L}^p(\mathcal{X}, \mathcal{A}, \mu)$ can be represented (μ-almost uniquely) by means of a function $g \in \mathcal{L}^q(\mathcal{X}, \mathcal{A}, \mu)$ (Folland, 1999, Theorem 6.15).

The analogous results for general ORLICZ spaces take a slightly less pleasing form. The role played by the HÖLDER inequality is taken over by the YOUNG inequality

<17> **Example.** Suppose Ψ and Λ are conjugate ORLICZ functions, both in $\mathcal{Y}_\infty$. Suppose also that $f \in \mathcal{L}^\Psi$ and $g \in \mathcal{L}^\Lambda$. For finite constants $c > \|f\|_\Psi$ and $d > \|g\|_\Lambda$, the YOUNG inequality gives

$$|fg|/(cd) \le \Psi(|f|/c) + \Lambda(|g|/d),$$

which integrates to give

$$\mu|fg|/(cd) \le \mu\Psi(|f|/c) + \mu\Lambda(|g|/d) \le 2.$$

Rearrange then take infima over c and d to deduce that

$$\mu|fg| \le 2\|f\|_\Psi \|g\|_\Lambda .$$

This inequality tells us that $T_g(f) := \mu(fg)$ is a well defined linear functional on $\mathcal{L}^\Psi$ for which satsifies

$$\|T_g\| := \sup\{|T_g(f)| : \|f\|_\Psi \le 1\} \le 2\|g\|_\Lambda .$$

That is, g defines a continuous linear functional on $\mathcal{L}^\Psi$.

Note the extra factor of 2. Problem [16] provides the corresponding lower bound, $\|g\|_\Lambda \le \|T_g\|$.

The gap between the upper and lower bounds for $\|T_g\|$ might seem a small price to pay for the generalization from $\mathcal{L}^p$ to $\mathcal{L}^\Psi$. Unfortunately there is worse news. These bounds leave open the possibility that there might be other continuous linear functionals on $\mathcal{L}^\Psi$ that cannot be represented as $f \mapsto \mu(fg)$ for some function g. Indeed, as shown in the next □ Section, if Ψ increases rapidly enough then such functionals do exist.

*5.6 Comparison of $\mathcal{L}^p$ and $\mathcal{L}^\Psi$

As you saw from the Young quotation, the original motivation for the YOUNG inequality was generalization of the theory of $\mathcal{L}^p$ spaces. In some ways $\mathcal{L}^\Psi$ is better behaved than $\mathcal{L}^p$; in other ways worse. To keep the story as simple as possible, in this Section I consider only the case where the underlying measure is a probability measure, even though much of the discussion could be extended to other measures.

The use of the norms to bound tail probabilities is an example where $\mathcal{L}^\Psi$ can be better behaved than $\mathcal{L}^p$.

<18> **Example.** If $f \in \mathcal{L}^p$ with $0 < \sigma = \|f\|_p$ then

$$\mathbb{P}\{|f| \ge t\} \le \mathbb{P}|f/t|^p = \sigma^p t^{-p} \qquad \text{for } t > 0.$$

The implication does not go the other way. If $\mathbb{P}\{|f| \ge t\} \le Ct^{-p}$ for all $t > 0$ then, by the TONELLI inequality,

$$\mathbb{P}|f|^p = \int_0^\infty pt^{p-1}\mathbb{P}\{|f| > t\}\,dt \le Cp\int_0^\infty t^{-1}dt = \infty.$$

The problem occurs as t gets large; for t near zero the tail bound can be improved to $\min(1, Ct^{-p})$. Of course an infinite upper bound does not make $\mathbb{P}|f|^p$ infinite. The example where $U \sim \text{UNIF}(0,1)$ and $X = (1/U)^{1/p}$ □ provides a specific example where $\mathbb{P}\{X \ge t\} = \min(1, t^{-p})$ but $\mathbb{P}X^p = \infty$.

Remark. If $p \ge 1 + \delta$ with $\delta > 0$ and $\mathbb{P}\{|f| \ge t\} \le Ct^{-p}$ for $t > 0$ then $\mathbb{P}|f|^{p-\delta} < \infty$.

The story for $\mathcal{L}^\Psi$ is slightly more encouraging. Suppose $\Psi(x) = \int_0^x \psi(t)\,dt$ for each $x \ge 0$ (as in Section 5.2) and Ψ increases rapidly enough to ensure that

<19> $$\int_0^\infty \frac{\psi(t)}{\max(1, \Psi(At))}\,dt < \infty \qquad \text{for some constant } A > 0.$$

Then the next Example shows that there is an equivalence between the tail bound and finiteness of the $\mathcal{L}^\Psi$ norm. It is not hard to show that each Ψ_p, for $p \ge 1$, has property <19>.

<20> **Example.** If $f \in \mathcal{L}^\Psi$ with $0 < \sigma = \|f\|_\Psi$ then

$$\mathbb{P}\{|f| \geq t\} \leq \min\left(1, \mathbb{P}\frac{\Psi\left(|f|/\sigma\right)}{\Psi(t/\sigma)}\right) \leq \min\left(1, \frac{1}{\Psi(t/\sigma)}\right) \qquad \text{for } t > 0.$$

Conversely, suppose Ψ has property <19> and there is a positive constant a for which

$$\mathbb{P}\{|f| \geq t\} \leq \min\left(1, 1/\Psi(at)\right) = 1/\max(1, \Psi(at)) \qquad \text{for each } t > 0.$$

Then the TONELLI inequality gives

$$\begin{aligned}\mathbb{P}\Psi(|f|/c) &= \mathbb{P}\int_0^\infty \psi(t)\{|f|/c > t\}\, dt \\ &= \int_0^\infty \psi(t)\mathbb{P}\{|f|/c > t\}\, dt \\ &\leq \int_0^\infty \frac{\psi(t)}{\max(1, \Psi(act))}\, dt \quad .\end{aligned}$$

□ The final integral is finite if $ac \geq A$.

As Problem [14] shows, the characterization of $\mathcal{L}^\Psi$ in Example <20> via tail probabilities comes at a cost: under the rapid-growth assumption <19> there always exist continuous linear functionals that are not of the form $f \mapsto \mathbb{P}(fg)$ for some g.

In trying to understand what goes wrong, I found it instructive to try to adapt the $\mathcal{L}^p$ proof for $1 < p < \infty$ (Folland, 1999, Section 6.2) to a general ORLICZ function Ψ and a probability space $(\mathcal{X}, \mathcal{A}, \mathbb{P})$. We start with a continuous linear map T from $\mathcal{L}^\Psi(\mathcal{X}, \mathcal{A}, \mathbb{P})$ into the real line. There is no loss of generality (Bourbaki, 2004, Section II.2.2) in assuming that $T(f) \geq 0$ whenever $f \geq 0$.

Here is an outline the steps of that proof. Notice the switch from $\mathcal{L}^\Psi$ to $\mathcal{L}^p$ at step (iii).

(i) The indicator function $\mathbb{1}_A$ of a set $A \in \mathcal{A}$ has $\mathbb{P}\Psi(\mathbb{1}_A/c) = \mathbb{P}A\Psi(1/c)$. It follows that $\|\mathbb{1}_A\|_\Psi = 0$ if $\mathbb{P}A = 0$ and $1/\Psi^{-1}(1/\mathbb{P}A)$ otherwise. The set function $\nu A := T(\mathbb{1}_A)$ inherits finite additivity from the linearity of T and countable additivity from the continuity of T: If $A_n \downarrow \emptyset$ then

$$|\nu(A_n)| = |T(\mathbb{1}_{A_n})| \leq \|T\|\, \|\mathbb{1}_{A_n}\|_\Psi \to 0 \qquad \text{because } \mathbb{P}A_n \to 0.$$

Thus ν is a countably additive measure with $\nu\mathcal{X} = T(\mathbb{1}) < \infty$ and for which $\nu A = 0$ whenever $\mathbb{P}A = 0$. By the RADON-NIKODYM theorem (Folland, 1999, Section 3.2), there exists a nonnegative, $\mathbb{P}$-integrable function Δ for which $\nu A = \mathbb{P}(A\Delta)$ for each $A \in \mathcal{A}$.

(ii) If $h(x) = \sum_i \alpha_i \{x \in A_i\}$ is a simple function then, by linearity of T and of the integrals with respect to ν and $\mathbb{P}$,

$$Th = \sum\nolimits_i \alpha_i T(\mathbb{1}_{A_i}) = \sum\nolimits_i \alpha_i \nu(A_i) = \nu f = \mathbb{P}(f\Delta).$$

(iii) The set of all simple functions is dense in $\mathcal{L}^p$. That is, for each f in $\mathcal{L}^p(\mathcal{X}, \mathcal{A}, \mathbb{P})$ and each $\epsilon > 0$ there exists a simple function h_ϵ for which $\|f - h_\epsilon\|_p < \epsilon$. Continuity of T extends the equalities from step (ii) from simple functions to all of $\mathcal{L}^p$.

(iv) We then play around with various choices for f in $\mathcal{L}^p$ to show that $\mathbb{P}\Delta^q < \infty$, where $p^{-1} + q^{-1} = 1$ and $\|\Delta\|_q = \|T\|$.

It is step (iii) that fails for an ORLICZ function Ψ satisfying the rapid-growth condition. As shown in Problem [14], under that condition there are always simple probability spaces $(\mathcal{X}, \mathcal{A}, \mathbb{P})$ with some function in $f_0 \in \mathcal{L}^\Psi(\mathcal{X}, \mathcal{A}, \mathbb{P})$ that cannot be closely approximated in $\mathcal{L}^\Psi$ norm by any bounded measurable function, let alone by a simple functions.

The main difficulty is well illustrated by the following Example for the ORLICZ function $\Psi_2(x) = \exp(x^2) - 1$.

<21> **Example.** Let $\mathbb{P}$ be the $N(0,1)$ distribution on $\mathcal{B}(\mathbb{R})$. Consider $f_0(x) = x$, which belongs to $\mathcal{L}^{\Psi_2}$ because

$$\begin{aligned} 1 + \mathbb{P}\Psi_2(f_0^2/c^2) &= (2\pi)^{-1/2} \int_{\mathbb{R}} \exp(-x^2/2 + x^2/c^2)\, dx \\ &= \begin{cases} c/\sqrt{c^2 - 2} & \text{if } c^2 > 2 \\ +\infty & \text{if } 0 < c^2 \le 2 \end{cases}. \end{aligned}$$

Consequently, $\|f_0\|_{\Psi_2} = \sqrt{8/3}$ and $f_0 \in \mathcal{L}^{\Psi_2}$.

Surprisingly, no bounded function (let alone a simple function) can approximate f_0 very closely. Suppose $|h(x)| \le K < \infty$. The fact that $x - K \ge x/2$ when $x \ge 2K$ implies

$$1 + \mu\Psi_2(|f_0 - h|/c) \ge (2\pi)^{-1/2} \int_{2K}^{\infty} \exp(-x^2/2 + x^2/4c^2)\, dx,$$

which is infinite if $0 < c \le 1/\sqrt{2}$. Thus $\|f_0 - h\|_\Psi > 1/\sqrt{2}$.

It follows that f_0 does not belong to the smallest closed subspace $\mathcal{H}$ of $\mathcal{L}^{\Psi_2}$ that contains all the bounded, measurable functions. By the HAHN-BANACH theorem Folland (1999, Theorem 5.8) there exists a continuous linear functional T on $\mathcal{L}^{\Psi_2}$ with $T(h) = 0$ for all h in $\mathcal{H}$ and $T(f_0) \ne 0$. There can be no measurable function Δ for which $f\Delta \in \mathcal{L}^1$ and $T(f) = \mathbb{P}(f\Delta)$ for all $f \in \mathcal{L}^{\Psi_2}$: the choice $h = \operatorname{sgn}(\Delta)$ would force $\mathbb{P}|\Delta| = 0$ and consequently $\mathbb{P}(f_0\Delta) = 0 \ne T(f_0)$. □

5.7 Problems

[1] Suppose $f \in \mathcal{L}^\Psi(\mathcal{X}, \mathcal{A}, \mu)$ and $c_1 := \|f\|_\Psi > 0$. Show that $\mu\Psi(|f|/c_1) \le 1$. Hint: Use dominated convergence and $\mu\Psi\left(|f|/(c_1 + n^{-1})\right) \le 1$ for $n \in \mathbb{N}$.

[2] Suppose $\{f_n : n \in \mathbb{N}\} \subset \mathcal{L}^\Psi = \mathcal{L}^\Psi(\mathcal{X}, \mathcal{A}, \mu)$ and $0 \le f_n \uparrow f$ pointwise. If $c_0 := \sup_n \|f\|_\Psi < \infty$ show that $f \in \mathcal{L}^\Psi$ and $\|f\|_\Psi = c_0$. Hint: Use monotone convergence for $\mu\Psi(f_n/c)$ with $c > c_0$.

[3] Suppose $h(x) = e^{g(x)}$, where g is a twice differentiable real-valued function on $\mathbb{R}^+$. Show that h is convex on any interval J for which $(g'(x))^2 + g''(x) \ge 0$ for $x \in \text{int}(J)$.

[4] Let $\dot\Psi_L$ and $\dot\Psi_R$ be the left- and right-derivative functions for an ORLICZ function Ψ. For $x > 0$ let $\dot\Psi(x-)$ denote the limit of $\dot\Psi_R(y)$ as $y \nearrow x$, which exists because $\dot\Psi_R$ is an increasing function.

(i) For an $x > 0$, let $\{y_i\}$ be a strictly increasing sequence in $\mathbb{R}^+$ for which $0 \le y_1 < y_2 < \cdots \nearrow x$. Explain why $\dot\Psi_R(y_i) \le \dot\Psi_L(y_{i+1})$ for each i. Deduce that $\dot\Psi_R(x-) = \dot\Psi_L(x)$, so that Ψ is differentiable at x if and only if the size $\dot\Psi(x) - \dot\Psi_L(x)$ of the jump at x is zero.

(ii) For each $\epsilon > 0$ and $A > 0$ show that there are at most finitely many x in $(0, A]$ for which $\dot\Psi_R(x) - \dot\Psi_L(x) > \epsilon$. Deduce that $\dot\Psi_R$ has at most countably many jumps in $\mathbb{R}^+$ and hence $\dot\Psi_R(x) \ne \dot\Psi_L(x)$ for, at worst, a countable set of x in $\mathbb{R}^+$.

[5] Suppose $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ is nondecreasing. For x in $\mathbb{R}^+$ define $\Psi(x) = \int_0^x \psi(r)\,dr$. Show that $\Psi \in \mathcal{Y}$. Hint: Suppose $x_\theta = \overline\theta x_0 + \theta x_1$, where $\overline\theta = 1 - \theta$ with $x_0 < x_1$ and $0 < \theta < 1$. Show that

$$\overline\theta\Psi(x_0) + \theta\Psi(x_1) - \Psi(x_\theta) = -\overline\theta \int_{x_0}^{x_\theta} \psi(r)\,dr + \theta \int_{x_\theta}^{x_1} \psi(r)\,dr.$$

Use the bounds $\psi(r_0) \le \psi(x_\theta) \le \psi(r_1)$ for $r_0 \le x_\theta \le r_1$.

[6] Suppose $\Psi \in \mathcal{Y}$ has a representation $\Psi(x) = \int_0^x \psi(t)\,dt$ with ψ an increasing function on $\mathbb{R}^+$. Show that $\dot\Psi_R(t) \ge \psi(t)$ for all $t \ge 0$ and $\dot\Psi_L(t) \le \psi(t)$ for all $t > 0$. Hint: For $\delta > 0$ we have

$$\delta\dot\Psi_R(t_0 + \delta) \ge \int_{t_0}^{t_0+\delta} \dot\Psi_R(t)\,dt = \Psi(t_0 + \delta) - \Psi(t_0) = \int_{t_0}^{t_0+\delta} \psi(t)\,dt.$$

[7] Suppose $\Psi(x) = e^{g(x)} - 1$, where $g(x) = (x \log x)^+$.

(i) Show that g is convex and $g(x) + g(y) \le g(x + y)$ for all $x, y \ge 0$. Deduce that $\Psi \in \mathcal{Y}_{\text{exp}}$.

(ii) Show that $\|X\|_\Psi < \infty$ if X has a POISSON(θ) distribution.

[8] Suppose Ψ_0 is a convex, nonnegative, increasing function defined on an interval $[a, \infty)$, with $a > 0$. Suppose also that there exists an $x_0 \in [a, \infty)$ for which $\Psi_0(x_0)/x_0 = \gamma := \inf_{x \ge a} \Psi_0(x)/x$. Show that the extension $\Psi(x) := \gamma x\{0 \le x < x_0\} + \Psi_0(x)\{x \ge x_0\}$ is an ORLICZ function.

[9] For a fixed α in $(0, 1)$ define $f_\alpha(x) = e^{x^\alpha}$ for $x \ge 0$.

(i) Show that $x \mapsto f_\alpha(x)$ is convex for $x \ge z_\alpha := (\alpha^{-1} - 1)^{1/\alpha}$ and concave for $0 \le x \le z_\alpha$.

(ii) Define $x_\alpha = (1/\alpha)^{1/\alpha}$ and $\tau_\alpha = f_\alpha(x_\alpha)/x_\alpha = (\alpha e)^{1/\alpha}$. Show that

$$\Psi_\alpha(x) := \tau x\{x < x_\alpha\} + f_\alpha(x)\{x \ge x_\alpha\}$$

is an ORLICZ function.

(iii) Show that there exists a constant K_α for which

$$\Psi_\alpha(x) \le \exp(x^\alpha) \le K_\alpha + \Psi_\alpha(x) \qquad \text{for all } x \ge 0.$$

(iv) Show that $\Psi_\alpha(x)\Psi_\alpha(y) \le \Psi_\alpha(2^{1/\alpha}xy)$ for $x \wedge y \ge x_\alpha$.

[10] Suppose $\Psi \in \mathcal{Y}$ and $X \in \mathcal{L}^\Psi(\mathcal{X}, \mathcal{A}, \mathbb{P})$, for a probability measure $\mathbb{P}$. Show that $\mathbb{P}|X| \le \Psi^{-1}(1)\,\|X\|_\Psi$. Hint: the JENSEN inequality.

[11] Suppose $\Psi \in \mathcal{Y}_\infty$ with $\dot{\Psi}_R(0) = \tau > 0$. Define $\Psi_0(x) = \Psi(x) - \tau x$ for $x \ge 0$.

(i) Show that $\Psi_0 \in \mathcal{Y}$. Suppose $X \in \mathcal{L}^{\Psi_0}(\mathcal{X}, \mathcal{A}, \mathbb{P})$, for a probability measure $\mathbb{P}$. For $K_\Psi := 1 + \tau\Psi_0^{-1}(1)$ show that $K_\Psi\,\|X\|_{\Psi_0} \ge \|X\|_\Psi \ge \|X\|_{\Psi_0}$. Deduce that $\mathcal{L}^\Psi = \mathcal{L}^{\Psi_0}$. Hint: previous Problem.

(ii) Suppose Ψ has conjugate Λ and that Ψ_0 has conjugate Λ_0. Show that $\Lambda_0(y) = \Lambda(y + \tau)$ for $y \ge 0$.

(iii) Define $K_\Lambda := 1 + \Lambda(2\tau)$. For each $Y \in \mathcal{L}^\Lambda$ show that

$$\|Y\|_\Lambda \le \|Y\|_{\Lambda_0} \le K_\Lambda\,\|Y\|_\Lambda$$

Deduce that $\mathcal{L}^{\Lambda_0} = \mathcal{L}^\Lambda$. Hint: The first inequality follows from $\Lambda_0 \ge \Lambda$. For the second inequality first show that $\Lambda_0(|Y|/c) \le \Lambda(2\tau) + \Lambda(2|Y|/c)$ then take expected values with $c > \|Y\|_\Lambda$.

[12] Suppose $n_0 = 1 \le n_1 < n_2 < \ldots$ is an increasing sequence of positive integers with $\sup_{k\in\mathbb{N}} n_k/n_{k-1} = B < \infty$ and $f : [1, \infty) \to \mathbb{R}^+$ is an increasing function with $\sup_{t\ge1} f(Bt)/f(t) = A < \infty$. For each random variable X show that $\sup_{r\ge1} \|X\|_r/f(r) \le A \sup_{k\in\mathbb{N}} \|X\|_{n_k}/f(n_k)$.

[13] For $p \ge 1$ prove the existence of positive constants c_p an C_p for which

$$c_p\,\|X\|_{\Psi_p} \le \sup_{r\ge1} \frac{\|X\|_r}{r^{1/p}} \le C_p\,\|X\|_{\Psi_p}$$

for every random variable X. Hint: previous Problem.

[14] Suppose $\mathbb{P}$ is LEBESGUE measure on the $\mathcal{B}(0,1)$ and $\Psi \in \mathcal{Y}_\infty$ satisfies the rapid-growth condition <19>:

$$\int_0^\infty \frac{\psi(t)}{\max(1, \Psi(At))}\, dt < \infty \qquad \text{for some constant } A > 0.$$

Define $f_0(x) = \Psi^{-1}(1/x)$.

(i) Show that $\mathbb{P}\{x : f_0(x) \geq t\} = \mathbb{P}\{x \leq 1/\Psi(t)\} = \min(1, 1/\Psi(t))$. Deduce from Example <20> that $f \in \mathcal{L}^\Psi$.

(ii) Suppose h is a $\mathcal{B}(0,1)$-measurable function that is bounded in absolute value by a constant K. Define $A_k = \{x \leq 1/\Psi(2K)\}$. Argue that $f_0(x) \geq 2K$ on A_k, so that $f_0(x)/2 \geq |h(x)|$

$$\mathbb{P}\Psi(2|f_0 - h|) \geq \mathbb{P}\Psi(f_0)\{x \in A_k\} = \mathbb{P}\{x \in A_k\}x^{-1} = \infty.$$

Deduce that $\|f_0 - h\|_\Psi \geq 1/2$.

(iii) Let $\mathcal{H}$ denote the closure (in $\mathcal{L}^\Psi$) of the set of all bounded functions in $\mathcal{L}^\Psi$. Explain why $f_0 \notin \mathcal{H}$. Use the HAHN-BANACH theorem (Folland, 1999, Theorem 5.8) to prove the existence of a continuous linear functional T on $\mathcal{L}^\Psi$ for which $T(h) = 0$ for $h \in \mathcal{H}$ and $T(f) \neq 0$. Explain why there cannot exist a function g in $\mathcal{L}^\Lambda$ for which $T(f) = \mathbb{P}(fg)$ for each f in $\mathcal{L}^\Psi$. Hint: Consider $h = \operatorname{sgn}(g)$.

[15] For $f \in \mathcal{L}^\Psi(\mathcal{X}, \mathcal{A}, \mu)$ show that $\|f\|_\Psi \leq 1 + \mu\Psi(|f|)$. Hint: If $c = \|f\|_\Psi > 1$ then convexity gives $\Psi(t/c) \leq \Psi(t)/c$ for $t > 0$.

[16] (cf. Garling, 2007, Theorem 6.3.2) Suppose $\Psi \in \mathcal{Y}_\infty$ and $(\mathcal{X}, \mathcal{A}, \mu)$ is a σ-finite measure space. Suppose a function $g : \mathcal{X} \to \mathbb{R}$ is $\mathcal{A}$-measurable and $fg \in \mathcal{L}^1(\mathcal{X}, \mathcal{A}, \mu)$ whenever $f \in \mathcal{L}^\Psi$. Suppose also that the linear functional $T(f) := \mu(fg)$ on $\mathcal{L}^\Psi$ is continuous. Let Λ denote the conjugate of Ψ, as in Section 5.5. Show that $g \in \mathcal{L}^\Lambda$ and

$$\|g\|_\Lambda \leq \|T\| := \sup\{|T(f)| : \|f\|_\Psi \leq 1\}.$$

Follow these steps.

(i) Replace $f \in \mathcal{L}^\Psi$ by $|f|\operatorname{sgn}(g)$ to show that there is no loss of generality in assuming that $g \geq 0$ and $\|T\| = \sup\{\mu(fg) : 0 \leq f \text{ and } \|f\|_\Psi \leq 1\}$.

(ii) By σ-finiteness there exist sets B_n in $\mathcal{A}$ for which $B_n \uparrow \mathcal{X}$ and $\mu B_n < \infty$ for each n. Define $g_n(x) = \{x \in B_n\}\,(n \wedge g(x))$, so that $d_n := \|g_n\|_\Lambda \leq \Lambda(n)\mu B_n < \infty$. Also define $f_n(x) = \Lambda'_R(g_n(x)/d_n)$, which is also bounded by a constant multiple of $\{x \in B_n\}$, so that $c_n := \|f_n\|_\Psi < \infty$. (Here I am tacitly assuming that $d_n \neq 0$ and, later, $c_n > 0$. I leave it to you to handle

the trivial cases where either $d_n = 0$ or $c_n = 0$.) Use the YOUNG inequality (for the case where there is equality) and Problem [15] to show that

$$\|T\|/d_n \geq \mu\left(f_n g_n/(c_n d_n)\right) = \Big[\mu\Psi(f_n) + \mu\Lambda(g_n/d_n)\Big]/c_n \geq 1.$$

(iii) Invoke Problem [2] to deduce that $\|T\| \geq \lim_{n\to\infty} c_n = \|g\|_\Lambda$.

[17] Let Ψ and Λ be conjugate members of $\mathcal{Y}_\infty$. For $f \in \mathcal{L}^\Psi(\mathcal{X}, \mathcal{A}, \mu)$ define

$$\|f\|_\bullet = \sup\{\|fg\|_1 : \|g\|_\Lambda \leq 1\}.$$

If μ is σ-finite, appeal to Example <17> and Problem [16], with the roles of Ψ and Λ interchanged, to deduce $\|f\|_\Psi \leq \|f\|_\bullet \leq 2\|f\|_\Psi$ for all f in $\mathcal{L}^\Psi$.

5.8 Notes

If I remember rightly, my introduction to the virtues of using expected values instead of tail probabilities in empirical process theory came from Giné and Zinn (1984, Lemmas 2.8 and 2.9). That paper (in its preprint form) and advice from its authors led me to the work of Pisier (1980, 1983), who used ORLICZ norms to establish sufficient conditions for existence of versions of stochastic processes with continuous sample paths. It was then just a short step to the amazing work of Ledoux and Talagrand (1991, Chapter 11).

My subsequent ORLICZ education came from Dudley (1999, Appendix H) and Garling (2007, Chapter 6), whose brief accounts led me to the beautiful PhD thesis of Luxemburg (1955), where the $\|\cdot\|_\Psi$ is defined (page 43). My remarks about the difference between the norms defined by Orlicz and by Luxemburg are based on a rather quick reading of Orlicz (1932, 1936). There appears to be some controversy about the assignment of priority in this area. I do not claim to have made any careful study of the early contributions; my historical remarks should be treated with caution.

There also seems to be some disagreement in the literature about the precise definition of an ORLICZ function. Garling (2007, page 73) used the name *Young function* for the members of my $\mathcal{Y}_\infty$. Dudley (1999, Appendix H) used the name *Young-Orlicz modulus* for a my $\mathcal{Y}$ and *Orlicz modulus* for those Ψ in my $\mathcal{Y}_\infty$ for which $\dot{\Psi}_R(0) = 0$. The latter also correspond to the N-functions of Krasnosel'skiĭ and Rutickiĭ (1961).

The special ORLICZ functions $\Psi(x) = x^p$ for $p = 1, 2$ and $\Psi(x) = \Psi_2(x) := e^{x^2} - 1$ were behind the inequalities from Pollard (1989, Section 3). I was (perhaps overly) pleased with myself for being able to replace the $\mathcal{L}^p$ norms by the $\mathcal{L}^{\Psi_2}$ norm in Pollard (1990, Section 3). Subsequently van der Vaart and Wellner (1996, Lemma 2.2.2) extended the inequality to a large collection of ORLICZ functions. It now seems that the inequality in part (ii) of Theorem <11>, which I learned from Ledoux

and Talagrand (1991, Chapter 11), is even better at delivering maximal inequalities. In the notes to that chapter, L&T gave credit to Fernique (1983), by noting that their "Theorem 11.2 is equivalent to the (perhaps somewhat unorthodox) formulation of [Fernique]". I think they were referring to Fernique's Lemme 2.2.1 and Corollaire 3.1.

More recently, van de Geer and Lederer (2013) used the ORLICZ function

$$\Psi(x) = \exp\left((\sqrt{1+2Lx}-1)^2/L\right),$$

for positive constants L, and Wellner (2017) used the ORLICZ function

$$\Psi(x) = \exp\left(2L^{-2}\hbar(Lx)\right) - 1 \qquad \text{where } \hbar(x) := (1+x)\log(1+x) - x,$$

to derive empirical process maximal inequalities of the BERNSTEIN\BENNETT variety (see Chapter 8).

My enthusiasm for things ORLICZ grew after I finally understood the role played by the YOUNG inequality as an extension of the HÖLDER inequality, a powerful tool for generalizing from $\mathcal{L}^p$ to $\mathcal{L}^\Psi$. It was with some disappointment that I eventually learned that the dual space of $\mathcal{L}^\Psi$ could not always be identified with $\mathcal{L}^\Lambda$, for the ORLICZ function Λ that is conjugate to Ψ. Whence Section 5.6. I have no idea whether the counterexamples in that Section are well known.

Remark. It would be better to call the YOUNG inequality "one of the inequalities to which the name of W. H. Young has been attached". FOURIER theorists think of the YOUNG inequality as another result involving convolutions (Folland, 1999, p 247).

By the way, Dudley (2003, Notes to Section 5.1) made a good case for adding the name Leonard James Rogers to that of Hölder for the HÖLDER inequality.

References

Bourbaki, N. (2004). *Integration I: Chapters 1–6*. Elements of Mathematics. Springer-Verlag. English translation by S. K. Berberian of *Intégration* 1965, 1967, 1959.

Dudley, R. M. (1999). *Uniform Central Limit Theorems*. Cambridge University Press.

Dudley, R. M. (2003). *Real Analysis and Probability* (2nd ed.), Volume 74 of *Cambridge studies in advanced mathematics*. Cambridge University Press.

Fernique, X. (1983). Regularité de fonctions aléatoires non gaussiennes. *Springer Lecture Notes in Mathematics 976*, 1–74. École d'Été de Probabilités de St-Flour XI, 1981.

Folland, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications*. Wiley.

Garling, D. J. H. (2007). *Inequalities: A Journey into Linear Analysis*. Cambridge University Press.

Giné, E. and J. Zinn (1984). Some limit theorems for empirical processes. *Annals of Probability 12*, 929–989. (plus 9 pages of contributed discussion).

Kim, J. and D. Pollard (1990). Cube root asymptotics. *Annals of Statistics 18*, 191–219.

Krasnosel'skiĭ, M. A. and Y. B. Rutickiĭ (1961). *Convex Functions and Orlicz Spaces*. Noordhoff. Translated from the first Russian edition by Leo F. Boron.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. New York: Springer.

Luxemburg, W. A. J. (1955). *Banach function spaces*. Ph. D. thesis, TU Delft. http://resolver.tudelft.nl/uuid:252868f8-d63f-42e4-934c-20956b86783f.

Orlicz, W. (1932). Über eine gewisse Klasse von Räumen vom Typus B. *Bulletin International de l'Académie Polonaise des Sciences et des Lettres, Classe de Mathématique et Naturelles, série A 8*(9), 207–220.

Orlicz, W. (1936). Über Räume (L^M). *Bulletin International de l'Académie Polonaise des Sciences et des Lettres, Classe de Mathématique et Naturelles, série A, 93*, 107.

Pisier, G. (1980). Conditions d'entropie assurant la continuité de certains processus et applications à l'analyse harmonique. In *Séminaire d'analyse fonctionnelle, 1979-80*, pp. 1–41. École Polytechnique Palaiseau. Available from http://archive.numdam.org/.

Pisier, G. (1983). Some applications of the metric entropy condition to harmonic analysis. *Springer Lecture Notes in Mathematics 995*, 123–154. (A collection of lecture notes from the 1980-81 Special Year in Analysis at the University of Connecticut Department of Mathematics).

Pollard, D. (1989). Asymptotics via empirical processes (with discussion). *Statistical Science 4*, 341–366.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*, Volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Hayward, CA: Institute of Mathematical Statistics.

Pollard, D. (2001). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.

van de Geer, S. and J. Lederer (2013). The Bernstein–Orlicz norm and deviation inequalities. *Probability theory and related fields 157*(1-2), 225–250.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Process: With Applications to Statistics*. Springer-Verlag.

Wellner, J. A. (2017). The Bennett-Orlicz norm. *Sankhya A 79*(2), 355–383.

Young, W. H. (1912). On classes of summable functions and their Fourier series. *Proc. R. Soc. Lond. A 87*(594), 225–229.

Chapter 6
Multivariate normal

SECTION 6.1 *describes some wondrous properties of the (multivariate) normal distribution. By way of preview, it presents two examples of what can be achieved by using the masterly trick of studying relationships between different normals by integrating along a path.*

SECTION 6.2 *describes the path method. It also draws your attention to a simple integration-by-parts trick used by Charles Stein to prove amazing facts about normal approximation.*

SECTION 6.3 *presents three different path methods for slightly different versions of the* LIPSCHITZ *concentration inequality.*

*SECTION 6.4 *describes the gaussian isoperimetric inequality.*

SECTION 6.5 *derives some comparison inequalities (initially due to Slepian, with a generalization due to Gordon) by means of the path method.*

*SECTION 6.6 *uses the* GORDON *inequality to derive a special case of a result by Dvoretzky regarding cross sections of convex bodies in high dimensional Euclidean spaces.*

SECTION 6.7 *presents Chatterjee's proof of Fernique's inequality.*

6.1 Introduction

The normal distribution has many amazing properties. This Chapter presents a few results that give a glimpse of the heights achieved by the those who have worked on the theory of gaussian processes in the last half century or so. Even for those readers who are primarily interested in nongaussian processes, it helps to see what is possible in the cleanest case before plunging into more general theory; often the gaussian results provide the benchmarks against which the nongaussian analogs are judged. In any case, the gaussian results provide an example of beautiful probability theory, which is worthy of attention for aesthetic reasons.

A little bit of notation is needed before I can describe the main focus of the Chapter. As in §3.3, the symbol Φ denotes the $N(0,1)$ distribution function with $\bar{\Phi} = 1 - \Phi$, and ϕ denotes the standard normal density. The symbol γ_n denotes the $N(0, I_n)$, the probability measure on $\mathcal{B}(\mathbb{R}^n)$ with density $\phi_n(x) = \prod_j \phi(x_j) = (2\pi)^{-n/2} \exp(-|x|^2/2)$ with respect to LEBESGUE measure on $\mathbb{R}^n$. More generally, if $\mu \in \mathbb{R}^n$ and V is an $n \times n$ positive semi-definite matrix then $N(\mu, V)$ denotes the probability measure on $\mathcal{B}(\mathbb{R}^n)$ with Fourier transform $\exp(it'\mu - t'Vt/2)$. Such a measure is often called a ***multivariate normal*** (MVN for short) distribution. It

has mean μ and variance matrix V. If μ is zero the distribution is said to be a ***centered* mvn**. Only if V is nonsingular does the $N(\mu, V)$ have a density with respect to LEBESGUE measure. Sometimes it helps to add a subscript, $N_n(\dots)$, if there is any ambiguity about the dimension of the Euclidean space where the measure lives. For example, if $X \sim N_n(\mu, V)$ and F is an $m \times n$ constant matrix then FX has a $N_m(F\mu, FVF')$ distribution. As a special case for m equal to n, if $FF' = V$ and $Z \sim N_n(0, I_n)$ then $\mu + FZ \sim N_n(\mu, V)$. Finally, the general fact that independence implies zero correlation can be run in the other direction for the MVN: if

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(\mu, V) \qquad \text{with } V = \begin{pmatrix} V_x & C_{x,y} \\ C'_{x,y} & V_y \end{pmatrix},$$

then X and Y are independent if and only if $C_{x,y} = 0$.

All the assertions in the previous paragraph are easy to prove using Fourier transforms.

Now back to the main focus of the Chapter. The unifying concept is that of a ***path argument*** (described in Section 6.2) to establish two very useful properties of the MVN:

(i) subgaussian concentration of $f(Z)$ if $Z \sim N(0, I_n)$ and $f : \mathbb{R}^n \to \mathbb{R}$ is LIPSCHITZ (in Section 6.3);

(ii) comparison inequalities relating probabilistic bounds for two centered MVN distributions based on simple comparisons between their variance matrices (in Sections 6.5 and 6.7).

Here is a preview.

6.1.1 Subgaussian concentration

Remember that a function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be LIPSCHITZ if there exists a finite constant κ for which

$$|f(y) - f(z)| \le \kappa |y - z| \qquad \text{for all } y, z \text{ in } \mathbb{R}^n.$$

Here $|\cdot|$ denotes the usual Euclidean (ℓ^2) distance, $|w|^2 := \sum_i w_i^2$. The smallest κ for which the inequality holds is called the LIPSCHITZ constant, which I denote by $\|f\|_{\text{Lip}}$.

The following concentration result is proved in Section 6.3.3. A much simpler argument, which produces a slightly larger subgaussian scale parameter, appears in in Section 6.3.1. Of course I suggest that you start with the easier derivation.

<1> **Theorem.** *If $Z \sim \gamma_n = N(0, I_n)$ and $f : \mathbb{R}^n \to \mathbb{R}$ is a* LIPSCHITZ *function with $\|f\|_{Lip} \le \kappa$ then $f(Z) - \gamma_n f$ has a subgaussian distribution,*

$$\mathbb{P}e^{\lambda(f(Z) - \gamma_n f)} \le e^{\lambda^2 \kappa^2 / 2} \qquad \text{for each } \lambda \in \mathbb{R},$$

□ *which implies $\mathbb{P}\{|f(Z) - \gamma_n f| \ge t\kappa\} \le 2e^{-t^2/2}$ for each $t \ge 0$.*

Remark. Notice that n does not appear explicitly in the upper bound, although it might be hiding inside the constant κ.

The bound is sharp, in the following sense: If u is a unit vector then the linear function $f(x) = u'x$ is LIPSCHITZ with $\|f\|_{\text{Lip}} = 1$ and the function $f(Z)$ has a $N(0,1)$ distribution if $Z \sim \gamma_n$.

<2> **Example.** Let M equal $\sup_i Y_i$ for a gaussian process $\{Y_i : i \in \mathbb{N}\}$. If both $m := \mathbb{P}M$ and $\sigma^2 := \sup_{i\in\mathbb{N}} \text{var}(Y_i)$ are finite then Theorem <1> can be used to show that

<3> $$\mathbb{P}\{|M - m| \geq \sigma x\} \leq 2\exp(-x^2/2) \qquad \text{for all } x \geq 0.$$

In special cases—such as the maximum of independent $N(0,1)$-distributed variables, as discussed in §3.3—there exist much tighter bounds. However, inequality <3> has the great virtues of both being impervious to the effects of possible dependence between the Y_i and of not depending (except through m and σ) on the size of the index set.

All the real work is carried out for the case of a finite index set. Suppose $W := (Y_1, \ldots, Y_n) \sim N_n(\mu, V)$. We may write W as $\mu + LZ$ with $Z \sim \gamma_n$ and an $n\times n$ constant matrix L for which $LL' = V$. If ℓ_i denotes the ith row of L then $Y_i = \mu_i + \ell_i Z$ and $\sigma^2 \geq \text{var}(Y_i) = \text{var}(\mu_i + \ell_i Z) = \ell_i I_n \ell_i' = |\ell_i|^2$.

Define a real-valued function $f_n(z) := \max_{i\leq n} (\mu_i + \ell_i z)$ on $\mathbb{R}^n$. It is LIPSCHITZ with $\|f_n\|_{\text{Lip}} \leq \sigma$: for $z_1, z_2 \in \mathbb{R}^n$,

$$\begin{aligned} |f_n(z_1) - f_n(z_2)| &= |\max_{i\leq n}(\mu_i + \ell_i z_1) - \max_{i\leq n}(\mu_i + \ell_i z_2)| \\ &\leq \max_i |(\mu_i + \ell_i z_1) - (\mu_i + \ell_i z_2)| \\ &\leq \max_i |\ell_i|\,|z_1 - z_2| \qquad \text{by CAUCHY-SCHWARZ} \\ &\leq \sigma |z_1 - z_2|. \end{aligned}$$

Define $M_n := \max_{i\leq n} Y_i = \max_{i\leq n}(\mu_i + \ell_i Z) = f_n(Z)$. Theorem <1> gives $\mathbb{P}\exp(\lambda(M_n - \mathbb{P}M_n)) \leq \exp(\lambda^2\sigma^2/2)$ for all real λ, which implies $\mathbb{P}\{M_n \geq \mathbb{P}M_n + \sigma x\} \leq e^{-x^2/2}$. Put another way,

$$\mathbb{P}\{M_n > m + \sigma(x - \epsilon)\} \leq e^{-(x-\epsilon)^2/2} \qquad \text{for } 0 < \epsilon < x \text{ and each } n.$$

As n tends to ∞ the events $\{M_n > r\}$ increase to $\{M > r\}$ for each r. Consequently, $\mathbb{P}\{M \geq m+\sigma x\} \leq \mathbb{P}\{M > m+\sigma(x-\epsilon)\} \leq e^{-(x-\epsilon)^2/2}$ for $0 < \epsilon < x$. Let ϵ decrease to zero to obtain a one-sided analog of the asserted inequality. A similar argument applied to $-f_n$ followed by a passage to the limit leads to a similar bound for the lower tail of $M - \mathbb{P}M$. □

Remark. You might be puzzled by the sneaky use of a strict inequality near the end of the argument. The trick is made necessary by the annoying fact that if $M(\omega) > r$ then $M_n(\omega) > r$ for all n large enough, but if $M(\omega) \geq r$ then we might have $M_n(\omega) < r$ for all n.

6.1.2 Comparison inequalities

The following inequality is a special case of a result proved in Section 6.7.

<4> **Theorem.** *(Fernique, 1975, Section 2.1) Suppose both $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ have centered* MVN *distributions. If $\mathbb{P}|X_i - X_j|^2 \le \mathbb{P}|Y_i - Y_j|^2$ for all i, j then $\mathbb{P}\max_i X_i \le \mathbb{P}\max_i Y_i$.*

□

<5> **Example.** (SUDAKOV "minoration") If $Y := (Y_1, Y_2, \ldots, Y_n)$ has a centered MVN distribution with $\mathbb{P}|Y_j - Y_k|^2 \ge \delta^2$ for all distinct $j \ne$ and k, then $\mathbb{P}\max_{i\le n} Y_i \ge C_{\text{sud}}\delta\sqrt{\log_2 n}$ with C_{sud} a universal (positive) constant.

I leave to you the trival case where $n = 1$.

The asserted inequality follows directly from Theorem <4>. Consider first the case where $n = 2^k$, for some $k \in \mathbb{N}$, so that the index set can be identified with $\mathcal{S} := \{-1, +1\}^k$. Construct $\{X_\mathfrak{s} : \mathfrak{s} \in \mathcal{S}\}$ from a set $Z_1, \ldots, Z_k$ of independendent $N(0,1)$'s:

$$X_\mathfrak{s} := \tfrac{1}{2}\delta k^{-1/2}\sum_{j=1}^k \mathfrak{s}_j Z_j.$$

For $\mathfrak{s} \ne \mathfrak{s}'$,

$$\mathbb{P}|X_\mathfrak{s} - X_{\mathfrak{s}'}|^2 = \tfrac{1}{4}\delta^2 k^{-1}\sum_{j=1}^k (\mathfrak{s}_j - \mathfrak{s}'_j)^2 \le \delta^2 \le \mathbb{P}|Y_\mathfrak{s} - Y_{\mathfrak{s}'}|^2.$$

The X process is more tractable than Y. Its expected maximum can be calculated exactly.

$$\begin{aligned}\mathbb{P}\max_\mathfrak{s} X_\mathfrak{s} &= \tfrac{1}{2}\delta k^{-1/2}\mathbb{P}\left(\max_\mathfrak{s}\sum_{j=1}^k \mathfrak{s}_j Z_j\right)\\ &= \tfrac{1}{2}\delta k^{-1/2}\mathbb{P}\sum_{j=1}^k |Z_j| = \tfrac{1}{2}\delta k^{1/2}\mathbb{P}|Z_1| = \tfrac{1}{2}\delta k^{1/2}\sqrt{2/\pi}.\end{aligned}$$

For a general $n \ge 2$ define $k = \lfloor \log_2 n \rfloor$

$$\mathbb{P}\max_{i\le n} Y_i \ge \mathbb{P}\max_{i\le 2^k} Y_i \ge \tfrac{1}{2}\delta k^{1/2}\sqrt{2/\pi}.$$

The choice $C_{\text{sud}} = (4\pi)^{-1/2}$ suffices. □

Remark. The lower bound for $\mathbb{P}\max_i Y_i$ is sharp within a constant, in the following sense. Suppose $\mathbb{P}|Y_j - Y_k|^2 \le (A\delta)^2$ for all j, k, with A a positive constant. First note that

$$\mathbb{P}\max_i Y_i = \mathbb{P}Y_1 + \mathbb{P}\max_i(Y_i - Y_1) = \mathbb{P}\max_i(Y_i - Y_1).$$

Then argue as in §3.3. For each $\lambda > 0$,

$$\begin{aligned}\exp(\lambda\mathbb{P}\max_i Y_i) &= \mathbb{P}\exp(\lambda\max_i(Y_i - Y_1)) \qquad \text{by Jensen's inequality}\\ &\le \mathbb{P}\sum\nolimits_i \exp(\lambda(Y_i - Y_1)) \le n e^{\lambda^2 A^2\delta^2/2}.\end{aligned}$$

Take logs of both sides, divide through by λ, then choose $\lambda = \sqrt{2\log n}/(A\delta)$ to minimize, leaving $\mathbb{P}\max_i Y_i \le A\delta\sqrt{2\log n}$.

6.2 The path method

The path method usually involves the construction of a smooth multivariate gaussian process $\{W_t : 0 \le t \le 1\}$ with W_0 and W_1 suggested by the particular problem at hand. For example, to attack Theorem <4> the random vector W_0 could be chosen to have the same distribution as X and the random vector W_1 could be chosen to have the same distribution as Y. For f a suitably smooth function on $\mathbb{R}^n$ we have a representation

$$\begin{aligned} f(W_1) - f(W_0) &= \int_0^1 df(W_t)/dt\, dt \\ &= \int_0^1 \langle \dot{f}(W_t), \dot{W}_t \rangle\, dt = \sum\nolimits_{i \le n} \int_0^1 \dot{f}_i(W_t) W_{t,i}\, dt, \end{aligned} \tag{<6>}$$

where $\dot{W}_t$ denotes the derivative of W_t with respect to t, with ith component $\dot{W}_{t,i}$, and $\dot{f} = \nabla f$ denotes the derivative of the map $x \mapsto f(x)$, with ith component $\dot{f}_i(x) := \partial f(x_1, \ldots, x_n)/\partial x_i$. To learn something about $\mathbb{P}f(W_1) - \mathbb{P}f(W_0)$ we need to control $\mathbb{P}\dot{f}_i(W_t)\dot{W}_{t,i}$ at each point of the path.

Remark. As I have been using the prime symbol $'$ to denote transpose of a vector or matrix I need something different to indicate derivatives. For a real-valued function $D = D(x_1, \ldots, x_n)$ on $\mathbb{R}^n$ the derivative will be denoted by $\dot{D}$ (or sometimes ∇D), with components $\dot{D}_i = \partial D(x)/\partial x_i$. If D also depends on another real argument, t, then $\dot{D}_t$ will denote $\partial D(x,t)/\partial t$. Similarly $\ddot{D}_{i,j}$ will denote the partial derivative $\partial^2 D(x)/\partial x_i \partial x_j$. The function $\sum_i \ddot{D}_{i,i}$ is called the Laplacian. It is often denoted by ΔD, a notation that I need to avoid because I want to reserve Δ to denote an increment.

In general, the choice of a good path is a rather mysterious business, although the examples I know all seem to share a few tricks. Talagrand (2003, Section 1.3) suggested that the choice of starting point of the path is the real subtlety:

> To study a difficult situation one can compare it to a simpler one, by finding a path between them and controlling derivatives along this path. This is an old idea. In practice we are given the difficult situation, and the key to the effectiveness of the method is to find the correct simple situation to which it should be compared. This can be done only after the problem is well understood. To insist upon the fact that the choice of the path is the real issue, we call this method the smart path method. (More precisely, the real issue is in the choice of the "easy end of the path". Once this has been chosen, the choice of the path itself will be rather canonical, except for its "orientation". We make the convention that the "smart path"

moves from the "easy end" to the "hard end") The smart path method, under various forms, will be the main tool throughout the book.

Often the path is constructed in a rather rigid way, starting from independent random vectors $W_0 \sim N(0, V_0)$ and $W_1 \sim N(0, V_1)$. For smooth deterministic functions a_t and b_t with $a_0 = 1 = b_1$ and $a_1 = b_0 = 0$ we define $W_t = a_t W_0 + b_t W_1$. That gives $\dot{W}_t = \dot{a}_t W_0 + \dot{b}_t W_1$, which has a $N\left(0, (\dot{a}_t)^2 V_0 + (\dot{b}_t)^2 V_1\right)$ distribution. More importantly, the joint distribution of W_t and $\dot{W}_t$ is also MVN, with

<7> $$\operatorname{cov}(\dot{W}_t, W_t) = \dot{a}_t a_t V_0 + \dot{b}_t b_t V_1.$$

It often seems to help if we choose a_t and b_t so that $a_t^2 + b_t^2 = 1$ for each t, a constraint that implies

<8> $$0 = \partial(a_t^2 + b_t^2)/\partial t = 2\left(\dot{a}_t a_t + \dot{b}_t b_t\right)$$

so that

<9> $$\operatorname{cov}(\dot{W}_t, W_t) = \dot{b}_t b_t \left(V_1 - V_0\right).$$

This simplification is most convenient when dealing with theorems that impose assumptions on the difference $V_1 - V_0$. See Theorem <33> for an example.

Strangely enough, the argument in Section 6.3.1 involves a path construction with $V_1 = V_0$, so that $\operatorname{cov}(\dot{W}_t, W_t) = 0$, implying that $\dot{W}_t$ and W_t are independent. In that case, the random variables $g(W_0)$ and $g(W_1)$ have the same distribution, so it would be pointless to take expected values in <6>. Instead, as you'll see, a little JENSEN trick gives a bound on the expected value of $\exp\left(\lambda(g(W_1) - g(W_0)\right)$.

If $V_1 \neq V_0$ the expected value $\mathbb{P}\dot{g}_i(W_t)\dot{W}_{t,i}$ can still be simplified by using an integration-by-parts trick that was often used by Charles Stein to perform miracles. See the Notes.

<10> **Theorem.** *[i-b-p trick] Suppose $(Z, X_1, \ldots, X_m)$ has a* MVN *distribution with $\mathbb{P}Z = 0$. (No assumptions are made about the means or covariances for the X_i's.) Suppose also that a continuously differentiable function $G : \mathbb{R}^m \to \mathbb{R}$ has partial derivatives $\dot{G}_j(x_1, \ldots, x_m) := \partial G(x_1, \ldots, x_m)/\partial x_j$ for which $\mathbb{P}|\dot{G}_j(X_1, \ldots, X_m)| < \infty$ for each i. Then*

□ $$\mathbb{P}ZG(X_1, \ldots, X_m) = \sum\nolimits_{j \le m} \tau_j \mathbb{P}\dot{G}_j(X_1, \ldots, X_m) \qquad \text{where } \tau_j := \mathbb{P}(ZX_j).$$

For the proof see Problems [2] (the one dimensional case) and [3], which involves many appeals to the one-dimensional case.

For future reference, here is a result that summarizes these ideas. Rather than imposing explicit assumptions on the f in <6>, I'll stick with the vague term 'suitably regular' to suggest any requirement that would ensure integrability of various random variables and justify taking derivatives insides integrals with respect to $\mathbb{P}$ and integrations-by-parts. Twice continuous differentiability with second partial derivatives $\ddot{f}_{i,j}(x)$ that grow no faster than $\exp(C|x|)$ would suffice in the next Lemma.

<11> **Lemma.** *Suppose $W_t = a_t W_0 + b_t W_1$ for $0 \le t \le 1$, where W_0 and W_1 are independent random vectors with $W_\alpha \sim N_n(0, V_\alpha)$ and a_t, b_t are chosen so that $a_t^2 + b_t^2 = 1$ and $0 = b_0 < b_t \uparrow 1$ as $t \uparrow 1$. If $G(t) := \mathbb{P}g(W_t)$ for a suitably regular real-valued function g on $\mathbb{R}^n$ then*

$$\dot{G}(t) = \sum_{i \le n} \mathbb{P}\dot{W}_{t,i}\dot{g}_i(W_t) = \dot{b}_t b_t \sum_{i,j \in [\![n]\!]} (V_1[i,j] - V_0[i,j])\, \mathbb{P}\ddot{g}_{i,j}(W_t). \qquad \square$$

Remark. The assumption on b_t ensures that $\dot{b}_t b_t = d(b_t^2/2)/dt > 0$, a convenience if we want to show H is an increasing function of t.

The choice $a_t = \sqrt{1-t}$ and $b_t = \sqrt{t}$, or its reparametrization $a_t = \cos(\pi t/2)$ and $b_t = \sin(\pi t/2)$, works well for all of the applications in this Chapter. We could also change the index set to some other subinterval, such as $a_t = \cos(t)$ and $b_t = \sin(t)$ with $0 \le t \le \pi/2$.

6.3 Concentration of LIPSCHITZ functionals

This Section contains three different path arguments, presented in order of increasing subtlety, that show $f(Z) - \gamma_n f$ has a subgaussian distribution if $Z \sim \gamma_n = N(0, I_n)$ and f is LIPSCHITZ. The easiest argument (in Section 6.3.1) yields a subgaussian bound with scale parameter $\pi\|f\|_{\text{Lip}}/2$. Section 6.3.2 improves the scale parameter to $\sqrt{2}\|f\|_{\text{Lip}}$ by using a slightly better path. Finally, Section 6.3.3 uses a path constructed from a BROWNIAN motion to obtain the scale factor $\|f\|_{\text{Lip}}$. A completely rigorous argument in the third case uses the ITÔ formula from stochastic calculus. On the assumption that my readers might not be completely familiar with that formula, I also include a heuristic argument that gives some insight into what is really going on.

The three different proofs helped me when I was first trying to understand what made one path smarter than another.

An analog of Theorem <1>, with centering at the median rather than at the expected value, can also be derived easily from a very deep result known as the ***gaussian isoperimetric inequality***, which is discussed briefly in Section 6.4.

Remark. Unfortunately, most proofs that I know of for that inequality are quite involved. It is worth knowing at least a little about isoperimetry because it lurks behind several interesting facts about the normal distribution. It would, however, be discouraging if the only path to Theorem <1> ran through isoperimetry.

Roughly speaking, when proving mgf results about a generic LIPSCHITZ function f there is no harm in pretending that it is infinitely differentiable, with derivative $\dot{f}(x)$ (sometimes denoted by ∇f) whose euclidean length is everywhere bounded above by $\|f\|_{\text{Lip}}$. Such a simplification is justified by Problem [1], which shows how to use convolution smoothing to construct a family $\{f_\sigma : \sigma > 0\}$ of infinitely differentiable functions with $\|f_\sigma\|_{\text{Lip}} \le \|f\|_{\text{Lip}}$ and $\sup_x |f_\sigma(x) - f(x)| \to 0$ as $\sigma \to \infty$ and

$$<12> \qquad |\dot{f}_\sigma(x)|_2 = \sqrt{\sum\nolimits_i \left(\partial f_\sigma(x)/\partial x_i\right)^2} \le \|f\|_{\text{Lip}} \qquad \text{for all } x.$$

Moreover, $\mathbb{P}\exp(\lambda f_\sigma(W)) \to \mathbb{P}\exp(\lambda f(W))$ as $\sigma \to 0$ for each real λ if $\mathbb{P}\{|W| \ge t\}$ decreases fast enough. Subgaussianity suffices.

6.3.1 The PISIER-MAUREY method

The method in this subsection comes from Theorem 2.2 of Pisier (1986, page 176), who commented that "The proof below is a simplification, due to Maurey, of my original proof which used an expansion in Hermite polynomials". Pisier was primarily interested in a concentration bound for the norm of a BANACH-valued gaussian process X. He also noted that his method would "apply to more general functions of X than the norm of X (and even vector valued functions) provided a suitable bound is known for the gradients of the functions".

The proof starts with a symmetrization trick. (Compare with Chapter 13.) For notational reasons, which should soon be apparent, let me rewrite the mgf for $f(Z) - \mathbb{P}f(Z)$ as an iterated integral, then invoke the JENSEN inequality for the integral appearing within the exponent:

$$\begin{aligned} \mathbb{P}e^{\lambda(f(X)-\mathbb{P}f(X))} &= \gamma_n^x \exp\left(\lambda(f(x) - \gamma_n^y f(y))\right) \\ &\le \gamma_n^x \gamma_n^y \exp\left(\lambda(f(x) - f(y))\right) = \mathbb{P}e^{\lambda(f(X)-f(Y))}, \end{aligned} \qquad <13>$$

where X and Y are independent random vectors that both have distribution γ_n.

Remark. The argument could also be written as a conditional JENSEN inequality using the fact that $\mathbb{P}\left(f(Y) \mid X\right) = \gamma_n f$. I prefer FUBINI to conditioning when I am trying to deal carefully with symmetrization arguments involving independence.

Notice that $\text{var}(f(X) - f(Y)) = 2\text{var}(f(X))$. This symmetrization approach inevitably leads to at least a doubling of the squared scaling factor.

The distribution of $|X - Y|$ depends on n. If we were to bound the difference $f(X) - f(Y)$ in the exponent by $\kappa|X - Y|$ we would introduce an explicit dependence on n in the upper bound. Instead we need to exploit cancellations due to independence along a one-dimensional path from $W_0 = Y$ to $W_1 = X$. For $0 \le t \le 1$ define $a_t = \cos(t\pi/2)$ and $b_t = \sin(t\pi/2)$ and $W_t = a_t W_0 + b_t W_1$. The derivative with respect to t along the path equals

$$\dot{W}_t = \frac{\partial W_t}{\partial t} = \dot{a}_t W_0 + \dot{b}_t W_1 = \frac{\pi}{2}\left(-b_t W_0 + a_t W_1\right) \sim N(0, (\pi/2)^2 I_n).$$

As explained in Sections 6.2, this choice ensures that $\operatorname{cov}(\dot{W}_t, W_t) = 0$, making the random vectors $(2/\pi)\dot{W}_t$ and W_t independent, each with distribution γ_n. Moreover

$$f(X) - f(Y) = f(W_1) - f(W_0) = \int_0^1 \frac{\partial f(W_t)}{\partial t}\, dt = \int_0^1 \langle \dot{W}_t, \dot{f}(W_t)\rangle\, dt.$$

By the JENSEN inequality for LEBESGUE measure on $[0,1]$,

$$\mathbb{P}\exp\left(\lambda\left(f(X) - f(Y)\right)\right) \le \mathbb{P}\int_0^1 \exp\left(\lambda\langle \dot{W}_t, \dot{f}(W_t)\rangle\right)\, dt.$$

For each fixed s in $\mathbb{R}^n$ we have $\mathbb{P}\exp(\langle \dot{W}_t, s\rangle) = \exp(\pi^2|s|_2^2/8)$. Independence of $\dot{W}_t$ and $\dot{f}(W_t)$ lets us treat $\lambda\dot{f}(W_t)$ like a constant s (more conditioning? or Fubini?) to deduce that

$$\begin{aligned}\mathbb{P}\exp\left(\lambda\left(f(X) - f(Y)\right)\right) &\le \int_0^1 \mathbb{P}\exp\left(\lambda^2\pi^2|\dot{f}(W_t)|_2^2/8\right)\, dt \\ &\le \exp\left(\lambda^2\kappa^2\pi^2/8\right) \qquad \text{by } <12>.\end{aligned}$$

We have a subgaussian bound with scale parameter $\pi\kappa/2$. Derive the tail bound as in §3.3.

Remark. The argument does not work with the choice $a_t = \sqrt{1-t}$ and $b_t = \sqrt{t}$, because then $\dot{W}_t \sim N(0, \sigma_t^2 I_n)$ with $1/\sigma_t^2 = 4t(1-t)$. The average over the $\dot{W}_t$ distribution then leaves a σ_t^2 in the exponent. The trigonometric parametrization, which gives a constant variance matrix for $\dot{W}_t$, seems better in this case.

6.3.2 The smart path method

This refinement of the path method from Section 6.3.1 comes from Talagrand (2011, Section 1.3). The method this time is similar to the one in 6.3.1 , except that now the path argument is carried out directly using the function

$$g(w) := g(x,y) := \exp(\lambda f(x) - \lambda f(y)) \qquad \text{for } w = (x,y) \in \mathbb{R}^{2n}.$$

Notice that

$$\begin{aligned}\dot{g}(w) &= \left(\lambda \dot{f}(x) g(w), -\lambda \dot{f}(y) g(w)\right) \\ &= \left(\dot{f}_1(x), \ldots, \dot{f}_n(x), -\dot{f}_1(y), \ldots, -\dot{f}_n(y)\right) \lambda g(w).\end{aligned}$$

To avoid some notational confusion, I'll use Greek letters (α, β) when referring to elements of $[\![2n]\!] = \{1, 2, \ldots, 2n\}$ and Roman letters $(i, j, \ldots)$ when referring to elements of $[\![n]\!] = \{1, 2, \ldots, n\}$.

The smarter path is defined using three independent random vectors X, Y an Z, each distributed $N(0, I_n)$, and $a_t = \sqrt{1-t}$ and $b_t = \sqrt{t}$:

$$\begin{aligned}W_t &= (X_t, Y_t) \qquad \text{for } 0 \le t \le 1, \\ X_t &= (X_{t,1}, \ldots, X_{t,n}) = a_t Z + b_t X, \\ Y_t &= (Y_{t,1}, \ldots, Y_{t,n}) = a_t Z + b_t Y.\end{aligned}$$

Remark. You might find it informative to check that the argument still works with $a_t = \cos(t)$ and $b_t = \sin(t)$ for $0 \le t \le \pi/2$.

For some calculations it helps to think of the $\{X_t\}$ and $\{Y_t\}$ as two n-dimensional gaussian processes, each with the same distribution as the $\{W_t\}$ process from Section 6.3.1. In particular,

$$\operatorname{cov}(\dot{X}_t, X_t) = \left(\dot{a}_t a_t + \dot{b}_t b_t\right) I_n = 0 = \operatorname{cov}(\dot{Y}_t, Y_t).$$

The processes inherit a dependence from the Z that they share. In particular, $W_0 = (Z, Z)$ and $W_1 = (X, Y)$, so that

<14> $$\Delta V := V_1 - V_0 := \operatorname{var}(W_1) - \operatorname{var}(W_0) = I_{2n} - \begin{pmatrix} I_n & I_n \\ I_n & I_n \end{pmatrix} = \begin{pmatrix} 0 & -I_n \\ -I_n & 0 \end{pmatrix}.$$

Consequently

$$\Delta V[\alpha, \beta] = \begin{cases} -1 & \text{if } \alpha = \beta - n = i \in [\![n]\!] \text{ or } \alpha - n = \beta = i \in [\![n]\!] \\ 0 & \text{otherwise} \end{cases}.$$

The sparsity of ΔV greatly simplifies the application of Lemma <11> to the function $G(t) := \mathbb{P} g(W_t)$:

$$\begin{aligned}\dot{G}(t) &= \dot{b}_t b_t \sum\nolimits_{\alpha,\beta \in [\![2n]\!]} \Delta V[\alpha, \beta] \mathbb{P} \ddot{g}_{\alpha,\beta}(W_t) \\ &= \tfrac{1}{2} \sum\nolimits_{i \in [\![n]\!]} \mathbb{P} 2\lambda^2 \dot{f}_i(X_t) \dot{f}_i(Y_t) g(W_t) \\ &\le \lambda^2 \mathbb{P} |\dot{f}(X_t)| \, |\dot{f}(Y_t)| g(W_t) \qquad \text{by CAUCHY-SCHWARZ} \\ &\le \lambda^2 \kappa^2 G(t) \qquad \text{by inequality <12>}.\end{aligned}$$

Remark. It helped me to write the the αth component of $\dot{g}$ as

$$\dot{g}_\alpha(w) = \begin{cases} \lambda \dot{f}_i(x) g(w) & \text{if } \alpha = i \in [\![n]\!] \\ -\lambda \dot{f}_i(y) g(w) & \text{if } \alpha - n = i \in [\![n]\!] \end{cases} .$$

The partial derivative with respect to w_β, with $|\alpha - \beta| = n$, then acts only on the $g(w)$ factor.

That is, $d(\log(G(t))/dt \le \lambda^2\kappa^2$ and

$$\log G(1) \le \log G(0) + \int_0^1 \lambda^2\kappa^2 \, dt = \lambda^2\kappa^2.$$

Take exponentials, using the fact that $G(1) = \mathbb{P}g(X,Y)$ and $G(0) = \mathbb{P}g(Z,Z) = 1$, to deduce that $\mathbb{P}g(X,Y) \le \exp\left(\lambda^2\kappa^2\right)$. Finally, again invoke the JENSEN inequality <13> to conclude that

$$\mathbb{P}\exp\left(\lambda f(X) - \lambda\gamma_n f\right) \le \mathbb{P}\exp\left(\lambda f(X) - \lambda f(Y)\right) = G(1) \le e^{\lambda^2(2\kappa^2)/2}.$$

The centered random variable $f(X) - \gamma_n f$ is subgaussian with scale parameter $\sqrt{2}\kappa$.

6.3.3 The stochastic calculus method

The PISIER-MAUREY argument showed that $f(Z) - \gamma_n f$ is subgaussian with scale factor approximately 1.57κ Talagrand's argument brought the scale factor down to approximately 1.4κ. In this Section, with a lot more technical effort the scale factor is reduced to κ, the best possible. For many purposes, where the size of various constants is not a great concern, the PISIER-MAUREY bound should suffice. If you find yourself in that situation then you might want to skip this Section. However, the stochastic calculus approach (apparently due to Maurey) has proved itself useful in other problems. You might find it helpful, therefore, to at least glance at the heuristic argument just to get a feel for how this approach works.

This very smart proof creates a different sort of path, from $\gamma_n f$ to $f(Z)$, using a stochastic integral with respect to an n-dimensional BROWNIAN motion, $B_t = (B_{t,1}, \ldots, B_{t,n})$ for $0 \le t \le 1$. That is, the $\{B_{t,i} : 0 \le t \le 1\}$, for $i = 1, \ldots, n$, are independent BROWNIAN motions on $[0,1]$.

Remark. I have no satisfying intuitive explanation for why a wiggly BROWNIAN motion path does better than the more rigid paths from the previous two Sections. A miracle, perhaps.

Write $\mathcal{F}_t$ for $\sigma\{B_s : 0 \le s \le t\}$, a sigma-field that corresponds to what we learn by watching $\{B_s : 0 \le s \le t\}$. Given $\mathcal{F}_t$, the post-t process $\{B_s : t \le s \le 1\}$ is BROWNIAN motion started from B_t. In particular, the conditional distribution of $B_1 - B_t$ is $N_n(0, (1-t)I_n)$, so that

<15> $$M_t := \mathbb{P}_{\mathcal{F}_t} f(B_1) = F(B_t, t) \qquad \text{where } F(x,t) := \gamma_n^z f(x + z\sqrt{1-t}\,).$$

The process $\{(M_t, \mathcal{F}_t) : 0 \le t \le 1\}$ is a martingale.

Notice the reappearance of our old friend $a_t = \sqrt{1-t}$. The function F satisfies the boundary conditions $F(x,1) = f(x)$ and $F(0,0) = \gamma_n f$. If we again assume f to be very smooth then the derivative $\dot{F} = \partial F(x,t)/\partial x$, with components $\dot{F}_i(x,t)$, is controlled by the LIPSCHITZ property:

$$\sqrt{\sum\nolimits_i \dot{F}_i(x,t)^2} = |\dot{F}(x,t)| \le \gamma_n^z |\dot{f}(x + a_t z)| \le \kappa. \tag{16}$$

The key to the whole subgaussian proof is a stochastic integral representation of M in a form that allows us to exploit <16>. The rigorous argument uses the ITÔ formula to express M as a stochastic integral:

$$M_t - \gamma_n f = \int_0^t \langle \dot{F}(B_s, s), dB_s \rangle = \sum\nolimits_i \int_0^t \dot{F}_i(B_s, s)\, dB_{s,i}. \tag{17}$$

Instead of starting with the ITÔ formula, I'll first present a non-rigorous calculation that gives a rough idea of what is going on. If you are comfortable with stochastic calculus you can skip the next few pages of heuristics by jumping straight to equation <24>.

The trick with stochastic integration is to carry TAYLOR expansions of functions of B_t out to second order, using the fact that for $\delta > 0$

$$\Delta B_t := B_{t+\delta} - B_t = (\Delta B_{t,1}, \ldots, \Delta B_{t,n})$$

has a $N(0, \delta I_n)$ distribution independent of $\mathcal{F}_t$. In particular,

$$\mathbb{P}_{\mathcal{F}_t} \Delta B_{t,i} = 0 \quad \text{AND} \quad \mathbb{P}_{\mathcal{F}_t} \Delta B_{t,i} B_{t,j} = \begin{cases} \delta & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}. \tag{18}$$

Accordingly, if $0 \le t < 1$ and δ is positive and small enough,

$$\begin{aligned} \Delta M_t := M_{t+\delta} - M_t \approx \dot{F}_t(B_t, t)\delta + \sum\nolimits_{i \in [\![n]\!]} \dot{F}_i(B_t, t) \Delta B_{t,i} \\ + \tfrac{1}{2} \sum\nolimits_{i,j \in [\![n]\!]} \ddot{F}_{i,j}(B_t, t) \Delta B_{t,i} \Delta B_{t,j}. \end{aligned} \tag{19}$$

The approximation sign ($\approx$) is intended to suggest that terms of order smaller than δ (in some probabilistic sense) have been omitted. A rigorous argument would need to include error terms. As before, $\dot{F}_t$ denotes $\partial F(x,t)/\partial t$, and $\dot{F}_i$ denotes $\partial F(x,t)/\partial x_i$, and so on.

We can get a good approximation for $M_1 - M_0$ by summing up the increments ΔM_t over a fine grid $\mathbb{G}(m) = \{\ell\delta : \ell = 0, 1, \ldots, m-1\}$, where $\delta = 1/m$ for a large positive integer m.

The martingale property for M_t and <18> give

$$0 = \mathbb{P}_{\mathcal{F}_t} \Delta M_t \approx \dot{F}_t(B_t, t)\delta + \tfrac{1}{2} \sum\nolimits_i \ddot{F}_{i,i}(B_t, t)\delta,$$

which suggests (correctly) that $\dot{F}_t + \frac{1}{2}\sum_i \ddot{F}_{i,i} = 0$. That is,

<20> $$\partial F(x,t)/\partial t + \tfrac{1}{2}\sum\nolimits_i \partial^2 F(x,t)/\partial x_i^2 = 0 \qquad \text{for } x \in \mathbb{R}^n \text{ and } 0 < t < 1,$$

an equality that is easy to verify by means of an appeal to the i-b-t trick: see Problems [2] and [5].

With the help from <20>, the approximation <19> simplifies to

<21> $$\Delta M_t \approx \sum\nolimits_i \dot{F}_i(B_t,t)\Delta B_{t,i} + R_1(t,\delta) + R_2(t,\delta)$$

where

$$R_1(t,\delta) = \tfrac{1}{2}\sum\nolimits_i \ddot{F}_{i,i}(B_t,t)\left(\Delta B_{t,i}^2 - \delta\right),$$
$$R_2(t,\delta) = \tfrac{1}{2}\sum\nolimits_{i,j}\{i \neq j\}\ddot{F}_{i,j}(B_t,t)\Delta B_{t,i}\Delta B_{t,j}.$$

Both $\mathbb{P}_{\mathcal{F}_t} R_1$ and $\mathbb{P}_{\mathcal{F}_t} R_2$ are zero. If we could keep B_t in a bounded region then the contributions from the $\ddot{F}_{i,j}$'s would stay bounded and a conditioning argument could be used to show that the contributions from R_1 and R_2 would be small in an $\mathcal{L}^2$ sense if m were large.

Such a boundedness effect could be achieved by borrowing a stopping time trick that is often used in proofs of Itô's formula (see Karatzas and Shreve, 1988, page 149, for example): replace B_t by the stopped process $Z_t = B_{t\wedge\tau(r)}$ where $\tau(r) := \inf\{t \in \mathbb{R}^+ : |B_t| > r\}$. Unfortunately, such a substitution would complicate the heuristics based on <18> but, as you'll soon see, it is easy to accommodate in a rigorous stochastic calculus argument.

In the heuristic spirit I'll just ignore the R_1 and R_2 contributions, simplifying <21> to $\Delta M_t \approx \sum_i \dot{F}_i(B_t,t)\Delta B_{t,i}$. Another conditioning arguments (which kills off cross-product terms) then gives

$$\mathbb{P}_{\mathcal{F}_t}(\Delta M_t)^2 \approx \sum\nolimits_i \mathbb{P}_{\mathcal{F}_t}\dot{F}_i(B_t,t)^2\delta = \mathbb{P}_{\mathcal{F}_t}|\dot{F}(B_t,t)|^2\delta \le \kappa^2\delta \qquad \text{by <16>.}$$

Remark. In stochastic calculus jargon, this approximation corresponds to a DOOB-MEYER decomposition of the submartingale M_t^2.

Further approximation then leads to

$$\mathbb{P}_{\mathcal{F}_t} e^{\lambda\Delta M_t} \approx 1 + \lambda\mathbb{P}_{\mathcal{F}_t}\Delta M_t + \tfrac{1}{2}\lambda^2\mathbb{P}_{\mathcal{F}_t}(\Delta M_t)^2 \lessapprox \exp(\lambda^2\kappa^2\delta/2).$$

Successive conditioning steps finally produce (approximately) the conclusion

<22> $$\mathbb{P}e^{\lambda M_1 - \lambda\gamma_n f} \approx \mathbb{P}\prod\nolimits_{t\in\mathbb{G}(m)} e^{\lambda\Delta M_t} \le \exp\left(m\lambda^2\kappa^2\delta/2\right) = e^{\lambda^2\kappa^2/2},$$

which should give the desired subgaussian bound in the limit.

End of heuristics.

To turn the last few pages of approximations into a rigorous argument one would need to engage in exactly the sort of calculations that lead to the ITÔ formula. Instead of providing those details I'll just refer you to the places in that excellent book by C&W = Chung and Williams (2014) where the stochastic calculus is derived.

Start again with the martingale $M_t = F(B_t, t)$ for $0 \le t \le 1$. By C&W Theorem 5.10 (the multidimensional version of the ITÔ formula), the crude approximation <19> is replaced by

$$F(B_t,t) - F(B_0,0) = \int_0^t \dot{F}_t(B_s,s)ds + \sum\nolimits_{i\in[\![n]\!]} \int_0^t \dot{F}_i(B_s,s)\, dB_{s,i} + \tfrac{1}{2}\int_0^t \sum\nolimits_{i,j\in[\![n]\!]} \int_0^t \ddot{F}_{i,j}(B_s,s) d[B_{\cdot,i}, B_{\cdot,j}]_s \quad . \tag{<23>}$$

The quadratic variation-covariation process of B is a deterministic matrix valued-process with (i,j)th component

$$[B_{\cdot,i}, B_{\cdot,j}]_t = \begin{cases} t & \text{for } i = j \\ 0 & \text{for } i \ne j \end{cases} .$$

All the terms with $i \ne j$ in the final double sum in <23> vanish, leaving only the $i = j$ terms:

$$\int_0^t \ddot{F}_{i,i}(B_s,s) d[B_{\cdot,i}, B_{\cdot,i}]_s = \int_0^t \ddot{F}_{i,i}(B_s,s)\, ds.$$

Thus <23> becomes

$$F(B_t,t) - F(B_0,0) = \int_0^t \dot{F}_t(B_s,s) + \tfrac{1}{2}\sum\nolimits_i \ddot{F}_{i,i}(B_s,s)\, ds + \sum\nolimits_i \int_0^t \dot{F}_i(B_s,s)\, dB_{s,i}. \tag{<24>}$$

Once again the identity <20> kills the $\int_0^t \ldots ds$ contribution, leaving

$$M_t - \gamma_n f = \sum\nolimits_i N_{t,i} \qquad \text{where } N_{t,i} := \int_0^t \dot{F}_i(B_s,s)\, dB_{s,i}. \tag{<25>}$$

Remark. In fact we don't even need <20> because $M_t - \sum_i N_{t,i}$ is a continuous local martingale whose sample paths are locally of bounded variation, which forces it to be (for almost all paths) equal to $M_0 = \gamma_n f$. See C&W Corollary 4.5.

The quadratic variation $[M,M]$ process, which is usually abbreviated to $[M]$, corresponds to the limiting form of the heuristic $\sum_{t\in\mathbb{G}(m)} (\Delta M_t)^2$.

It can be derived from the variation-covariation processes of the BROWNIAN motions: by C&W Theorem 5.7,

$$\begin{aligned}[M]_t &= \sum\nolimits_{i,j}[N_{\cdot,i}, N_{\cdot,j}]_t \\ &= \sum\nolimits_{i,j}\int_0^t \dot{F}_i(B_s,s)F_j(B_s,s)\, d[B_{\cdot,i}, B_{\cdot,j}]_s \\ &= \sum\nolimits_i \int_0^t \dot{F}_i(B_s,s)^2 ds = \int_0^t |\dot{F}(B_s,s)|^2 \le \kappa^2 t. \end{aligned} \tag{<26>}$$

Finally, to bound the MGF $\mathbb{P}e^{\lambda(M_t - M_0)}$ we can use another stochastic calculus trick to replace the peeling argument <22>. By C&W Theorem 6.2, the process $Z_t := \exp(\lambda M_t - \frac{1}{2}\lambda^2[M]_t)$ is a local martingale, that is, for some sequence of bounded stopping times τ_j that increases pointwise to ∞, the process $Z(t \wedge \tau_j)$ is a martingale. For each j,

$$\begin{aligned} e^{\lambda M_0} = \mathbb{P}Z_0 = \mathbb{P}Z(1\wedge\tau_j) &= \mathbb{P}\exp\left(\lambda M_{1\wedge\tau_j} - \tfrac{1}{2}\lambda^2[M]_{1\wedge\tau_j}\right) \\ &\ge \mathbb{P}\exp\left(\lambda M_{1\wedge\tau_j} - \tfrac{1}{2}\lambda^2\kappa^2\right). \end{aligned}$$

That is,

$$\mathbb{P}\exp\left(\lambda M_{1\wedge\tau_j}\right) \le \exp\left(\lambda\gamma_n f + \tfrac{1}{2}\lambda^2\kappa^2\right) \qquad \text{for each } j.$$

Complete the argument by letting j tend to infinity. An appeal to the FATOU lemma as $j \to \infty$ then gives

$$\mathbb{P}e^{\lambda f(B_1)} = \mathbb{P}e^{\lambda M_1} \le \liminf_j \mathbb{P}e^{\lambda M_{1\wedge\tau_j}} \le \exp\left(\lambda\gamma_n f + \tfrac{1}{2}\lambda^2\kappa^2\right),$$

the desired subgaussian bound.

*6.4.0 The gaussian isoperimetric inequality

For each subset A of $\mathbb{R}^n$ define $A^\delta := \{z \in \mathbb{R}^n : d(a,A) \le \delta\}$, where $d(z,A) := \inf\{|z-y| : y \in A\}$.

<27> **Gaussian isoperimetric inequality.** *If A is a* BOREL *subset of $\mathbb{R}^n$ with $\gamma_n A = \Phi(\alpha)$ then $\gamma_n A^\delta \ge \Phi(\alpha+\delta)$ for each $\delta \ge 0$. The lower bound is achieved when A is any closed halfspace with gaussian measure $\Phi(\alpha)$.*

It is the reduction from an n-dimensional problem, with n arbitrarily large, to a one-dimensional calculation for the lower bound that makes the isoperimetric inequality so powerful. The next Example, which derives concentration of LIPSCHITZ functionals around a median, illustrates.

Recall that a median of a (real valued) random variable X is any constant m for which $\mathbb{P}\{X \ge m\} \ge 1/2$ and $\mathbb{P}\{X \le m\} \ge 1/2$. Such an m always exists, but it need not be unique.

<28> **Example.** Suppose f is a LIPSCHITZ function on $\mathbb{R}^n$ with $\|f\|_{\text{Lip}} \le \kappa$. Under γ_n, the random variable $f(z)$ has at least one median, a number M for which

$$\gamma_n\{f(z) \le M\} \ge \tfrac{1}{2} \qquad \text{AND} \qquad \gamma_n\{f(z) \ge M\} \ge \tfrac{1}{2}.$$

Define $A = \{z \in \mathbb{R}^n : f(z) \le M\}$ so that $\gamma_n A \ge 1/2 = \Phi(0)$. If $d(x, A) < u$ then there exist a point $z \in A$ with $d(z, x) < u$. From the LIPSCHITZ property and the fact that $f(x) \le M$ we then get

$$f(x) < f(z) + \kappa u < M + \kappa u.$$

Conversely, if $f(x) \ge M + \kappa u$ then $d(x, A) \ge u$. It follows that

$$\gamma_n\{f(x) \ge M + \kappa u\} \le \gamma_n\{d(x, A) \ge u\} \le \bar{\Phi}(0 + u) \le \tfrac{1}{2}\exp(-u^2/2).$$

An analogous argument for deviations from the set $\{z : f(z) \ge M\}$ gives the companion lower bound. Together the two bounds give a concentration property for f,

<29> $$\gamma_n\{z : |f(z) - M| \ge \kappa y\} \le 2\bar{\Phi}(u) \le \exp(-y^2/2),$$

□ where M is a median for f under γ_n.

Remark. Inequality <29> also gives concentration around the mean $\mu = \mathbb{P}f(Z)$ because it also implies that the mean and the median are close to each other:

$$|\mu - M| \le \gamma_n|f - M| \le \kappa \int_0^\infty \gamma_n\{|f(z) - M| \ge \kappa y\}\, dy = C\kappa,$$

with $C = 1/(2\sqrt{2\pi})$. Thus

$$\gamma_n\{|f - \mu| \ge \kappa(C + y)\} \le \gamma_n\{|f - M| \ge \kappa y\} \le \exp(-y^2/2).$$

6.5 Tail bound comparisons

One of the earliest comparison inequalities for the MVN/ is due to Slepian (1962, page 498).

<*30*> **Slepian's comparison inequality.** *For some finite index set B suppose $X = (X_b : b \in B)$ and $Y = (Y_b : b \in B)$ both have centered* MVN *distributions. Suppose also that, for all $(b, b') \in B \times B$,*

<31> $$\Delta(b, b') := \mathbb{P}(Y_b Y_{b'}) - \mathbb{P}(X_b X_{b'}) \le 0 \qquad \text{with equality when } b = b'.$$

□ *Then $\mathbb{P} \cup_b \{X_b \ge r_b\} \le \mathbb{P} \cup_b \{Y_b \ge r_b\}$ for each r in $\mathbb{R}^B$.*

Remarks.

(i) Of course the prime on b' does not denote a derivative. I just need some way of distinguishing between two elements of B. I'll also write $X(b)$ for X_b, and so on, if the subscripts get too messy.

(ii) If you are wondering whether the equality constraint in <31> could be omitted, consider the case where $B = \{1, 2\}$ with $Y \sim N_2(0, I_2)$ and $X_1 = Y_1$ and $X_2 = 100Y_2$. Note that

$$\mathbb{P}\left(\{X_1 \geq 5\} \cup \{X_2 \geq 5\}\right) \approx 1/2,$$
$$\mathbb{P}\left(\{Y_1 \geq 5\} \cup \{Y_2 \geq 5\}\right) \approx 0,$$

which clearly would violate the conclusion of the Theorem.

(iii) I often have trouble remembering which way the inequalities should go. It helps me to rewrite the covariance assumptions as $\mathbb{P}X_b^2 = \mathbb{P}Y_b^2$ and $\mathbb{P}|X_b - X_{b'}|^2 \leq \mathbb{P}|Y_b - Y_{b'}|^2$ for all b, b'. The Y components are more 'spread out' than the X components, which makes $\max_b Y_b$ stochastically larger than $\max_b X_b$, that is,

$$\mathbb{P}\{\max_b X_b \geq r\} \leq \mathbb{P}\{\max_b Y_b \geq r\} \qquad \text{for each real } r.$$

This property ensures (Marshall et al., 2011, Chapter 11) existence of a pair of random variables on some new probability space, M_X with the same distribution as $\max_b X_b$ and M_Y with the same distribution as $\max_b Y_b$, such that $M_X \leq M_Y$ almost surely. For each increasing function h we then have $h(M_X) \leq h(M_Y)$ almost surely and hence

<32> $$\mathbb{P}h(\max_b X_b) = \mathbb{P}h(M_X) \leq \mathbb{P}h(M_Y) = \mathbb{P}h(\max_b Y_b),$$

provided the expected values are well defined (no $\infty - \infty$ difficulties). In particular $\mathbb{P}\max_b Y_b \leq \mathbb{P}\max_b Y_b$. Compare with the version of the Fernique inequality stated in Theorem <4>, which does not require the equality $\mathbb{P}X_b^2 = \mathbb{P}Y_b^2$.

Gordon (1985) extended the SLEPIAN inequality to doubly indexed arrays with a more complicated comparison between tail events. If you are wondering why anyone would be interested in this $\cap\cup$ stuff take a look at Section 6.6.

<33> **Theorem.** *Suppose A and B are finite sets and $X = (X_{a,b} : (a, b) \in A \times B)$ and $Y = (Y_{a,b} : (a, b) \in A \times B)$ both have centered* MVN *distributions. Suppose also that, for all $i = (a, b)$ and $j = (a', b')$ in $A \times B$,*

$$\Delta(i, j) := \mathbb{P}Y_iY_j - \mathbb{P}X_iX_j \qquad \begin{cases} = 0 & \text{if } i = j \\ \leq 0 & \text{if } a = a' \text{ and } b \neq b' \\ \geq 0 & \text{if } a \neq a' \end{cases}.$$

Then

$$\mathbb{P} \cap_{a\in A} \cup_{b\in B} \{X_{a,b} \geq r_{a,b}\} \leq \mathbb{P} \cap_{a\in A} \cup_{b\in B} \{Y_{a,b} \geq r_{a,b}\}$$

□ *for each r in $\mathbb{R}^{A\times B}$.*

<34> **Corollary.** *Under the conditions of the Theorem, by taking $r_{a,b} = r$ for all a, b we get $\mathbb{P}\{\min_a \max_b X_{a,b} \geq r\} \leq \mathbb{P}\{\min_a \max_b Y_{a,b} \geq r\}$ for each real r. That is, $\min_a \max_b Y_{a,b}$ is stochastically larger than $\min_a \max_b X_{a,b}$, implying*

$$\mathbb{P}h\left(\min_a \max_b X_{a,b}\right) \leq \mathbb{P}h\left(\min_a \max_b Y_{a,b}\right)$$

for each increasing function h, provided there are no $\infty - \infty$ difficulties.

Notice that the GORDON inequality reduces to the SLEPIAN inequality if the set A is taken to be a singleton.

Remark. The covariance assumptions of Theorem <33> imply, for all distinct $i = (a, b)$ and $j = (a', b')$ in $A \times B$,

$$\begin{aligned}&\mathbb{P}|Y_i - Y_j|^2 - \mathbb{P}|X_i - X_j|^2 \\ &= \left(\mathbb{P}Y_i^2 + \mathbb{P}Y_j^2 - \mathbb{P}X_i^2 - \mathbb{P}X_j^2\right) - 2\left(\mathbb{P}Y_iY_j - \mathbb{P}X_iX_j\right) \\ &\begin{cases} \geq 0 & \text{if } a = a' \text{ and } b \neq b' \\ \leq 0 & \text{if } a \neq a' \end{cases}.\end{aligned} \tag{35}$$

Gordon (1985, Theorem 1.4) actually established an analog of the FERNIQUE inequality,

$$\mathbb{P}\min_{a\in A}\max_{b\in B} X_{a,b} \leq \mathbb{P}\min_{a\in A}\max_{b\in B} Y_{a,b}, \tag{36}$$

under assumption <35>. That is, he derived the special case of Corollary <34> where h is the identity map without requiring that $\mathbb{P}X_{a,b}^2 = \mathbb{P}Y_{a,b}^2$ for all (a, b) in $A \times B$. Unfortunately, the proof is more complicated than the proof of Theorem <33>, just as the original proof of the FERNIQUE inequality is more complicated than the proof of Theorem <30>.

See Section 6.6 for a Lemma and an Example that show the extra costs incurred by the extra assumption that $\mathbb{P}X_{a,b}^2 = \mathbb{P}Y_{a,b}^2$ for all (a, b).

Kahane (1986) provided a most elegant derivation of both the SLEPIAN and GORDON inequalities, by a single theorem whose proof can be slightly modified to become a path argument with an appeal to i-b-p trick. The theorem involves a twice continuously differentiable function f for which various differentiations inside expectations are legitimate. Once again, rather than hardcoding the necessary regularity requirements into the Theorem I will merely refer to f as having 'suitably integrable partial derivatives'. For example, it would suffice that f be derived from a LIPSCHITZ function by means of smoothing operations like those in Problem [1]. For the derivation of the SLEPIAN and GORDON inequalities, the f is used to approximate sets that are products of intervals of the form $[r, \infty)$.

<37> **Theorem.** *(Kahane) Suppose $X = (X_i : i \in I)$ and $Y = (Y_i : i \in I)$ both have centered* MVN *distributions, for some finite index set I. Suppose also that*

(i) *a function $f : \mathbb{R}^I \to \mathbb{R}$ is twice continuously differentiable and has 'suitably integrable partial derivatives';*

(ii) *the set $I \times I$ can be partitioned into three disjoint subsets $\mathcal{Z}$, $\mathcal{P}$, and $\mathcal{N}$ for which*

$$\Delta(i,j) := \mathbb{P}Y_iY_j - \mathbb{P}X_iX_j \qquad \begin{cases} = 0 & \text{if } (i,j) \in \mathcal{Z} \\ \geq 0 & \text{if } (i,j) \in \mathcal{P} \\ \leq 0 & \text{if } (i,j) \in \mathcal{N} \end{cases}$$

$$\ddot{f}_{i,j}(w) := \frac{\partial^2 f}{\partial w_i \partial w_j} \qquad \begin{cases} \geq 0 & \text{if } (i,j) \in \mathcal{P} \\ \leq 0 & \text{if } (i,j) \in \mathcal{N} \end{cases} \qquad \text{for all } w \in \mathbb{R}^I.$$

□ *Then $\mathbb{P}f(X) \leq \mathbb{P}f(Y)$.*

Proof. Let W_0 and W_1 be independent random vectors, with W_0 having the same distribution as X and W_1 having the same distribution as Y. As before, define $W_t = a_tW_0 + b_tW_1$, this time (just for the sake of variety) with $a_t = \cos(t)$ and $b_t = \sin(t)$ for $0 \leq t \leq \pi/2$. Define $H(t) := \mathbb{P}f(W_t)$. Note that $\dot{a}_t = -b_t$ and $\dot{b}_t = a_t$, so that

$$\mathbb{P}\left(\dot{W}_tW_t'\right) = \dot{a}_ta_t\text{var}(X) + \dot{b}_tb_t\text{var}(Y) = a_tb_t\left(\text{var}(Y) - \text{var}(X)\right).$$

Or, in coordinate form, $\tau_t(i,j) := \mathbb{P}\dot{W}_{t,i}W_{t,j} = a_tb_t\Delta(i,j)$. By assumption and the fact that $a_tb_t \geq 0$, the product $\tau_t(i,j)\ddot{f}_{i,j}(w)$ is nonegative for every w in $\mathbb{R}^I$.

As before, we now have

$$\begin{aligned} \dot{H}(t) &= \mathbb{P}\langle \dot{W}_t, \dot{f}(W_t)\rangle \\ &= \sum\nolimits_{i\in I} \mathbb{P}\dot{W}_{t,i}\dot{f}_i(W_t) \\ &= \sum\nolimits_{i\in I, j\in I} \tau_t(i,j)\mathbb{P}\ddot{f}_{i,j}(W_t) \qquad \text{by Theorem <10>} \\ &\geq 0. \end{aligned}$$

□ It follows that $\mathbb{P}f(X) = F(0) \leq F(1) = \mathbb{P}f(Y)$, as asserted.

The remainder of this Section is devoted to explaining how Kahane's result covers both Theorems <30> (SLEPIAN) and <33> (GORDON). Even though the first Theorem is a special case of the second, I'll prove both results because the first proof gives some insight into the second.

The SLEPIAN inequality involves two centered MVN distributions, represented by random vectors $X = (X_b : b \in B)$ and $Y = (Y_b : b \in B)$. The Theorem gives conditions under which is that $\mathbb{P}g(X - r) \leq \mathbb{P}g(Y - r)$ for each r in $\mathbb{R}^B$, where g is the indicator function of $\{z \in \mathbb{R}^B : z_b \geq 0\}$. In de Finetti notation, g can be reexpressed as

$$g(z) = 1 - \prod\nolimits_{b\in B}\{z_b < 0\},$$

a convenient representation of a union as a complement of an intersection.

To employ Kahane's result we need to approximate the indicator function of $(-\infty, 0)$ by a smooth, decreasing function ψ_σ that equals 1 on $(-\infty, -\sigma)$ and 0 on $[0, \infty)$. (In fact, existence of a bounded, continuous derivative $\dot{\psi}_\sigma$ will suffice, as far as smoothness is concerned.) At some stage the positive parameter σ is sent to zero, which ensures that

$$f(z) := 1 - \prod\nolimits_{b\in B} \psi_\sigma(z_b)$$

converges pointwise to $g(z)$. If we can show that $\mathbb{P}f(X - r) \le \mathbb{P}f(Y - r)$ for each r in $\mathbb{R}^B$ then an appeal to dominated convergence will give the coreesponding inequality with f replaced by g.

The function f has partial derivatives, for distinct b and b' in B,

$$\dot{f}_b(z) := \frac{\partial f(z)}{\partial z_b} = -\dot{\psi}_\sigma(z_b) \prod\nolimits_{\beta\in B\setminus\{b\}} \psi_\sigma(z_\beta),$$
$$\ddot{f}_{b,b'}(z) := \frac{\partial^2 f(z)}{\partial z_b \partial z_{b'}} = -\dot{\psi}_\sigma(z_b)\dot{\psi}_\sigma(z_{b'}) \prod\nolimits_{\beta\in B\setminus\{b,b'\}} \psi_\sigma(z_\beta).$$

The function $\dot{f}_b$ is everywhere ≥ 0 because ψ_σ is decreasing and nonnegative. The function $\ddot{f}_{b,b'}$ is everywhere ≤ 0.

The signs of the partial derivatives are unchanged if $f(z)$ is replaced by $f(z - r)$.

The assumptions of Theorem <37> hold when applied to $f(\cdot - r)$ for a fixed r in $\mathbb{R}^B$ with $I = B$ and $\mathcal{Z} = \{(b, b') \in B \times B : b = b'\}$ and $\mathcal{N} = I\setminus\mathcal{N}$. It follows that $\mathbb{P}f(X - r) \le \mathbb{P}f(Y - r)$, which implies the SLEPIAN inequality in the limit as $\sigma \to 0$.

For the GORDON inequality, $I = A \times B$ and

$$\mathcal{Z} = \{(a, b, a', b') \in I \times I : a = a', b = b'\},$$
$$\mathcal{N} = \{(a, b, a', b') \in I \times I : a = a', b \ne b'\},$$
$$\mathcal{P} = \{(a, b, a', b') \in I \times I : a \ne a'\}.$$

Think of elements of $\mathbb{R}^I$ as $|A|$ by $|B|$ matrices. If $Z \in \mathbb{R}^I$ define $Z_a := (Z[a, b] : b \in B)$, the Ath row of Z. The function g is replaced by

$$G(Z) = \prod\nolimits_{a\in A} g(Z_a) = \cap_{a\in A} \cup_{b\in B} \{Z[a, b] \ge 0\},$$

which is approximated by the smooth function

$$F(Z) := \prod\nolimits_{a\in A} f(Z_a).$$

If $i = (a, b)$ and $j = (a', b')$ then

$$\dot{F}_i(Z) := \frac{\partial F(Z)}{\partial Z[a, b]} = \dot{f}_b(Z_a) \prod\nolimits_{\alpha\in A\setminus\{a\}} f(Z_\alpha),$$

which is everywhere ≥ 0, and

$$\ddot{F}_{i,j}(Z) = \begin{cases} \ddot{f}_{b,b'}(Z_a) \prod_{\alpha \in A \setminus \{a\}} f(Z_\alpha) \leq 0 & \text{if } a = a' \\ \dot{f}_b(Z_a) \dot{f}_{b'}(Z_{a'}) \prod_{\alpha \in A \setminus \{a,a'\}} f(Z_\alpha) \geq 0 & \text{if } a \neq a' \end{cases} .$$

That is, the partial derivative $\ddot{F}_{i,j}$ is everywhere ≥ 0 if $(i,j) \in \mathcal{P}$ and is everywhere ≤ 0 if $(i,j) \in \mathcal{N}$, which agrees with the sign of $\Delta(i,j)$. The partial derivatives for $F(Z-R)$ follow the same pattern, for each $R = (r_{a,b})$ in $\mathbb{R}^{A \times B}$. Theorem <37> gives $\mathbb{P}F(X-R) \leq \mathbb{P}F(Y-R)$. The limit as σ tends to zero then gives the inequality asserted by Theorem <33>.

*6.6 An application of the GORDON inequality

Both inequalities let us control complicated gaussian processes by means of simpler gaussian processes, as illustrated by the following result of Gordon (1985, Theorem 2.1).

<38> **Lemma.** *Let $G = [g_{i,j}]$ be an $N \times k$ matrix of independent standard normals, and let $Z \sim N(0, I_N)$ be independent of $W \sim N(0, I_k)$ If $\mathbb{A}$ is a compact subset of $\mathbb{R}^k$ and $\mathbb{B}$ is a compact subset of $\mathbb{R}^N$, with $0 \in \mathbb{B}$, then*

$$\begin{aligned} &\alpha_{\min} \mathbb{P} \sup_{b \in \mathbb{B}} \langle b, Z \rangle - \beta_{\max} \mathbb{P} \sup_{a \in \mathbb{A}} |\langle a, W \rangle| - \alpha_{\max} \beta_{\max} \\ &\quad \leq \mathbb{P} \inf_{a \in \mathbb{A}} \sup_{b \in \mathbb{B}} \langle b, Ga \rangle \leq \mathbb{P} \sup_{a \in \mathbb{A}} \sup_{b \in \mathbb{B}} \langle b, Ga \rangle \\ &\leq \alpha_{\max} \mathbb{P} \sup_{b \in \mathbb{B}} \langle b, Z \rangle + \beta_{\max} \mathbb{P} \sup_{a \in \mathbb{A}} |\langle a, W \rangle| + \alpha_{\max} \beta_{\max} \end{aligned}$$

where $\alpha_{\min} := \inf_{a \in \mathbb{A}} |a|_2$ and $\alpha_{\max} := \inf_{a \in \mathbb{A}} |a|_2$ and $\beta_{\max} = \sup_{b \in \mathbb{B}} |b|_2$.

Proof. It suffices to prove an analogous set of inequalities with $\mathbb{A}$ replaced by a finite subset A and $\mathbb{B}$ replaced by a finite subset B. (Then invoke dominated convergence as A expands up to a countable dense subset of $\mathbb{A}$ and B expands up to a countable dense subset of $\mathbb{B}$.)

Define two gaussian processes indexed by $\mathbb{A} \times \mathbb{B}$:

$$\begin{aligned} X_{a,b} &= |a|_2 |b|_2 \xi + \langle b, Ga \rangle = |a|_2 |b|_2 \xi + \sum\nolimits_{i,j} b_i g_{i,j} a_j, \\ Y_{a,b} &= |a|_2 \langle b, Z \rangle + |b|_2 \langle a, W \rangle, \end{aligned}$$

where $\xi \sim N(0,1)$ is independent of G. The role of ξ is to ensure that $\mathbb{P}X_{a,b}^2 = \mathbb{P}Y_{a,b}^2$. It comes at the slight $\pm \alpha_{\max}\beta_{\max}$ cost because

$$|X_{a,b} - \langle b, Ga \rangle| \leq \alpha_{\max} \beta_{\max} |\xi| \qquad \text{for all } (a,b) \in \mathbb{A} \times \mathbb{B},$$

which implies

$$\begin{aligned} &|\inf_{a \in \mathbb{A}} \sup_{b \in \mathbb{B}} X_{a,b} - \inf_{a \in \mathbb{A}} \sup_{b \in \mathbb{B}} \langle a, Gb \rangle| \leq \alpha_{\max} \beta_{\max} |\xi|, \\ &|\sup_{a \in \mathbb{A}, b \in \mathbb{B}} X_{a,b} - \sup_{a \in \mathbb{A}, b \in \mathbb{B}} \langle a, Gb \rangle| \leq \alpha_{\max} \beta_{\max} |\xi|. \end{aligned}$$

The GORDON inequality does most of the work. It is easy to check, for (a, b) and (a^*, b^*) in $A \times B$, that

$$\begin{aligned}\mathbb{P}X_{a,b}X_{a^*,b^*} &= |a|_2\,|a^*|_2\,|b|_2\,|b^*|_2 + \langle a, a^*\rangle\langle b, b^*\rangle,\\ \mathbb{P}Y_{a,b}Y_{a^*,b^*} &= |a|_2\,|a^*|_2\,\langle b, b^*\rangle + |b|_2\,|b^*|_2\,\langle a, a^*\rangle,\end{aligned}$$

which gives

$$\mathbb{P}X_{a,b}X_{a^*,b^*} - \mathbb{P}Y_{a,b}Y_{a^*,b^*} = \left(|b|_2\,|b^*|_2 - \langle b, b^*\rangle\right)\left(|a|_2\,|a^*|_2 - \langle a, a^*\rangle\right)$$

<39> $$\begin{cases} = 0 & \text{if } a = a^* \\ \geq 0 & \text{if } a \neq a^* \end{cases}.$$

The inequality

$$\mathbb{P}\min_{a\in A}\max_{b\in B} Y_{a,b} \leq \mathbb{P}\min_{a\in A}\max_{b\in B} X_{a,b}$$

follows from Corollary <34> with the roles of X and Y reversed. The inequality

$$\mathbb{P}\max_{a\in A, b\in B} X_{a,b} \leq \mathbb{P}\max_{a\in A, b\in B} Y_{a,b}$$

follows from the SLEPIAN inequality via <32>.

The calculations now involve only the Y process. Write $\mathcal{Y}_{\mathbb{A}}$ for $\sup_{a\in\mathbb{A}}|\langle a, W\rangle|$ and $\mathcal{Y}_B$ for $\max_{b\in B}\langle b, Z\rangle$. Then, from the definition of $Y_{a,b}$,

$$|a|_2\mathcal{Y}_B - \beta_{\max}\mathcal{Y}_{\mathbb{A}} \leq \max_{b\in B} Y_{a,b} \leq |a|_2\mathcal{Y}_B + \beta_{\max}\mathcal{Y}_{\mathbb{A}},$$

□ Take the maximum or minimum over $a \in A$ then take expectations.

<40> **Example.** A compact, convex subset K of $\mathbb{R}^N$ with non-empty interior is called a ***convex body***. It is called symmetric if $-x \in K$ for each $x \in K$. A famous result of Dvoretzky (1961) asserts that low-dimensional cross sections of symmetric, convex bodies look like balls. More precisely, for each $\epsilon > 0$ there is a small $\eta(\epsilon) > 0$ such that: if K is a symmetric, convex body in $\mathbb{R}^N$ and k is a positive integer with $k \leq \eta(\epsilon)\log N$ then there exists a k-dimensional subspace $\mathcal{H}$ of $\mathbb{R}^N$ and a closed ball B in $\mathcal{H}$ for which

<41> $$B \subset K \cap \mathcal{H} \subset (1+\epsilon)B.$$

In fact it also doesn't much matter if we require B to be a ball or an ellipsoid. See Problem [6].

In this Example I'll derive Dvoretzky's theorem by applying the GORDON inequality, but only for the special case where $K = [-1, +1]^N$.

Remark. The tricks for this special case can be extended to cover more general convex bodies. See Gordon (1985, 1988, 2005), for example.

The set K is the unit ball for the ℓ_∞^N norm, $|x|_\infty := \max_j |x_j|$. It is an easy exercise to prove that $|x|_\infty$ is equal to $\sup\{\langle b, x\rangle : b \in K^o\}$, where $K^o = \{b \in \mathbb{R}^N : |b|_1 \le 1\}$ is the unit ball for the ℓ_1^N norm. The little superscript 'o' means that the convex body K^o is the polar of K, that is, $K^o = \{b \in \mathbb{R}^N : \langle x, b\rangle \le 1 \text{ for all } x \in K\}$.

The subspace $\mathcal{H}$ will be defined as $\operatorname{span}\{z_1, \ldots, z_k\}$, where $\mathfrak{z} = (z_1, \ldots, z_k)$ is a carefully chosen realization of $\mathfrak{Z} = (Z_1, \ldots, Z_k)$, for independent $N(0, I_N)$-distributed random vectors Z_j.

If $a = ru$ with $r \ge 0$ and $|u|_2 = 1$ then $|a|_\infty = r|u|_\infty$. Thus

$$|a|_2 h(\mathfrak{Z}) \le \left|\sum\nolimits_j a_j Z_j\right|_\infty \le |a|_2 H(\mathfrak{Z}) \qquad \text{for each } a \text{ in } \mathbb{R}^k,$$

where

$$h(\mathfrak{Z}) := \inf_{u\in\mathbb{A}}\left|\sum\nolimits_j u_j Z_j\right|_\infty = \inf_{u\in\mathcal{U}} \sup_{b\in K^o} \langle b, \sum\nolimits_j u_j Z_j\rangle$$
$$H(\mathfrak{Z}) := \sup_{u\in\mathbb{A}}\left|\sum\nolimits_j u_j Z_j\right|_\infty = \sup_{u\in\mathcal{U}} \sup_{b\in K^o} \langle b, \sum\nolimits_j u_j Z_j\rangle.$$

We are now in the setting described by Lemma <38> for $\mathbb{B} = K^o$, with $\alpha_{\max} = \alpha_{\min} = 1 = \beta_{\max}$. The Lemma gives

<42> $$\mathbb{P}|Z|_\infty - \mathbb{P}|W|_2 - 1 \le \mathbb{P}h(\mathfrak{Z}) \le \mathbb{P}H(\mathfrak{Z}) \le \mathbb{P}|Z|_\infty + \mathbb{P}|W|_2 - 1,$$

where $Z \sim N(0, I_N)$ and $W \sim N(0, I_k)$.

The quantity $\mathbb{P}|W|_2$ is $\le \sqrt{\mathbb{P}|W|_2^2} = \sqrt{k}$. As explained in §3.4, $\max_{j\le N} Z[j]$ concentrates very tightly around a value only slightly smaller than $\sqrt{2\log N}$. A cruder form of the argument shows that $\mathbb{P}|Z_0|_\infty = \mathbb{P}\max_{j\le N}|Z[j]| \ge (1 - o(1))\sqrt{\log N}$, with $\mathbb{P}|Z_0|_\infty \ge 0.65\sqrt{\log N}$ for $N \ge 2$ suggested numerically. See Problem [8]. It is safe to assert for $N \ge 2$ that, for some positive constant c,

$$c\sqrt{\log N} - \sqrt{k} - 1 \le \mathbb{P}h(\mathfrak{Z}) \le \mathbb{P}H(\mathfrak{Z}) \le c\sqrt{\log N} + \sqrt{k} + 1,$$

and hence, if $\sqrt{k} + 1 \le c\epsilon\sqrt{\log(N)}/3$ and $0 < \epsilon < 1$,

$$\frac{\mathbb{P}H(\mathfrak{Z})}{\mathbb{P}h(\mathfrak{Z})} \le \frac{c\sqrt{\log N} + \sqrt{k} + 1}{c\sqrt{\log N} - \sqrt{k} - 1} \le \frac{1 + \epsilon/3}{1 - \epsilon/3} \le 1 + \epsilon.$$

From Problem [7] there must exist some realization $\mathfrak{z} = (z_1, \ldots, z_k)$ of $\mathfrak{Z}$ for which $0 < h(\mathfrak{z}) \le H(\mathfrak{z}) \le (1 + \epsilon)h(\mathfrak{z})$. If we write R for $H(\mathfrak{z})$ then

$$|a|_2 R/(1 + \epsilon) \le |\sum\nolimits_i a_i z_i|_\infty \le |a|_2 R \qquad \text{for each } a \text{ in } \mathbb{R}^k.$$

Let $\mathcal{H}$ be the k-dimensional subspace of $\mathbb{R}^N$ spanned by $z_1, \ldots, z_k$ and define $T : \mathbb{R}^k \to \mathcal{H}$ by $Te_i = z_i$, where $\{e_1, \ldots, e_k\}$ is the usual orthonormal basis for $\mathbb{R}^k$. Define $B = \{a \in \mathbb{R}^k : |a|_2 \le 1/R\}$. The ellipsoid $D := TB$ has the properties asserted by Dvoretzky's theorem.

□ $$D \subset \mathcal{H} \cap K_\infty = \left\{ \sum_i a_i z_i : |\sum_i a_i z_i|_\infty \le 1 \right\} \subset (1+\epsilon)D.$$

6.7 A generalized FERNIQUE inequality

The main topic of this Section is the proof of a stronger version of Theorem <4> from Section 6.1.2: for random vectors $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ with centered MVN distributions, if

<43> $$\mathbb{P}|X_i - X_j|^2 \le \mathbb{P}|Y_i - Y_j|^2 \qquad \text{for all } i, j.$$

then

<44> $$\mathbb{P} \max_i X_i \le \mathbb{P} \max_i Y_i.$$

Fernique (1975, Section 2.1) deduced <44> from a stronger inequality for the ranges $\max(Y_i - Y_j)$ and $\max(X_i - X_j)$. He pointed out that <43> is weaker than the assumption made for the SLEPIAN inequality, in that it does not require $\mathbb{P}X_i^2 = \mathbb{P}Y_i^2$ for each i.

Fernique's proof combined a path argument, FOURIER inversion, and a decomposition of $\mathbb{R}^n$ into polyhedral subsets corresponding to the regions where the max is achieved by different coordinates. Subsequently, Chatterjee (2005) gave a much simpler path proof based on a smooth approximation to the maximum function. Chatterjee's method also covered a result of Vitale (2000), who had employed an ingenious limit argument to show that <44> also implies result,

<45> $$\mathbb{P} \max_i (X_i + \mu_i) \le \mathbb{P} \max_i (Y_i + \mu_i) \qquad \text{for all } \mu \in \mathbb{R}^n.$$

That is, Vitale effectively removed the assumption that the variables are centered to zero expected values.

Remark. See Pollard (2001, Section 12.3) for a a slightly flawed exposition of Fernique's original method. As a friend pointed out to me, with my usual interpretation of $\{i \ne j\}$ as the indicator function of the set $\{i \in [\![n]\!] : i \ne j\}$ the quantity $\max\{i \ne j\} x_i$ on page 276 would correspond to the function $L_j(x)^+$, not the intended $L_j(x) := \max\{x_i : i \in [\![n]\!] \backslash \{j\}\}$.

Chatterjee's smooth approximation to $m(w) := \max_{i \le n} w_i$ is sometimes called a ***soft maximum***. The relevant facts are summarized in the next Lemma.

<46> **Lemma.** *For w in $\mathbb{R}^n$ and $\lambda > 0$ define $S(w,\lambda) = \sum_{i\le n} \exp(\lambda w_i)$ and $\mathfrak{M}(w,\lambda) := \lambda^{-1}\log S(w,\lambda)$. Then*

(i) $\mathfrak{M}(w,\lambda) = m(w) + R(w,\lambda)$ where $0 \le R(w,\lambda) \le \lambda^{-1}\log n \to 0$ as $\lambda \to \infty$.

(ii) The function $w \mapsto \mathfrak{M}(w,\lambda)$ is infinitely differentiable with

$$p_i(w,\lambda) := \dot{\mathfrak{M}}_i(w,\lambda) := \frac{\partial \mathfrak{M}(w,\lambda)}{\partial w_i} = e^{\lambda w_i}/S(w,\lambda)$$

and

$$\ddot{\mathfrak{M}}_{i,j}(w,\lambda) := \frac{\partial^2 \mathfrak{M}(w,\lambda)}{\partial w_i \partial w_j} = \begin{cases} \lambda p_i(w,\lambda) - \lambda p_i(w,\lambda)^2 & \text{if } j = i \\ -\lambda p_i(w,\lambda) p_j(w,\lambda) & \text{if } j \ne i \end{cases}.$$

Proof. For (i) write $S(w,\lambda)$ as $e^{\lambda m w)} \sum_{i\le n} e^{\lambda(w_i - m(w))}$ then note that $\lambda(w_i - m(w)) \le 0$ for all i with equality at any w_i that achieves the maximum, $m(w)$.

□ The equalities in (ii) are just calculus exercises.

The proof of <45> also depends on a simple real variable fact: for any real numbers $p_1,\ldots,p_n$ with $\sum_i p_i = 1$ and real numbers $z_1,\ldots,z_n$ we have the identity

$$\begin{aligned} \sum_{i,j} z_i z_j \Big(\{i=j\} p_i - p_i p_j \Big) &= \sum_i p_i z_i^2 - \Big(\sum_i p_i z_i \Big)^2 \\ &= \tfrac{1}{2} \sum_{i,j} p_i p_j (z_i - z_j)^2 \end{aligned} \qquad \text{<47>}$$

Define $G(t) := G(t,\mu,\lambda) = \mathbb{P} g(W_t)$ for $g(w) := g(w,\mu,\lambda) := \mathfrak{M}(w+\mu,\lambda)$, with both μ and λ held fixed. Write $p_i(w) = p_i(w,\mu,\lambda)$ for $\dot{g}_i(w,\mu,\lambda) = \dot{\mathfrak{M}}_i(w+\mu,\lambda)$ and let $\widetilde{X}, \widetilde{Y}, W_0, W_1$ be independent random vectors, with both $\widetilde{X}$ and W_0 distributed like X and both $\widetilde{Y}$ and W_1 distributed like Y. Then Lemma <11> and <47> give

$$\begin{aligned} \dot{G}(t) &= \dot{b}_t b_t \sum\nolimits_{i,j\in[\![n]\!]} \mathbb{P}\left(\widetilde{Y}_i \widetilde{Y}_j - \widetilde{X}_i \widetilde{X}_j \right) \mathbb{P} \ddot{\mathfrak{M}}_{i,j}(W_t + \mu, \lambda) \\ &= \lambda \dot{b}_t b_t \mathbb{P} \sum\nolimits_{i,j\in[\![n]\!]} \left(\widetilde{Y}_i \widetilde{Y}_j - \widetilde{X}_i \widetilde{X}_j \right) \Big(\{i=j\} p_i(W_t) - p_i(W_t) p_j(W_t) \Big) \\ &= \tfrac{1}{2} \lambda \dot{b}_t b_t \mathbb{P} \sum\nolimits_{i,j\in[\![n]\!]} \left((\widetilde{Y}_i - \widetilde{Y}_j)^2 - (\widetilde{X}_i - \widetilde{X}_j)^2 \right) p_i(W_t) p_j(W_t) \\ &= \tfrac{1}{2} \lambda \dot{b}_t b_t \sum\nolimits_{i,j\in[\![n]\!]} D_{ij} \mathbb{P} p_i(W_t) p_j(W_t) \end{aligned}$$

where $D_{i,j} := \mathbb{P}|Y_i - Y_j|^2 - \mathbb{P}|X_i - X_j|^2 \ge 0$ for all (i,j).

The function G is increasing, which implies

$$G(0) = \mathbb{P}\mathfrak{M}(X+\mu,\lambda) \le \mathbb{P}\mathfrak{M}(Y+\mu,\lambda) = G(1).$$

The desired inequality <45> then emerges as the limit when λ tends to ∞.

Remark. It seemed equally natural to me that Gordon's extension of Fernique's inequality, inequality <36>, should be derivable by replacing min max by $\mathfrak{m}\mathfrak{M}$, where $\mathfrak{m}$ is a 'soft minimum' function. I tried this approach but did not succeed. Maybe some clever reader can make this idea work, perhaps by using a diferent form of smooth approximation to min max.

6.8 Problems

[1] Suppose f is a real valued function on $\mathbb{R}^n$ with $\|f\|_{\text{Lip}} \le \kappa$. Let ψ be an infinitely differentiable, nonnegative function on $\mathbb{R}^n$ with compact support and $\int \psi(z)\,dz = 1$. For each $\sigma > 0$, define

$$f_\sigma(x) := \int_{\mathbb{R}^n} f(x+\sigma z)\psi(z)\,dz = \sigma^{-n}\int f(w)\psi\left((w-x)/\sigma\right)\,dw.$$

(i) Show that

$$\sup_x |f_\sigma(x) - f(x)| \le \sup_x \int |f(x+\sigma z) - f(x)|\psi(z)\,dz \le \kappa\sigma \int |z|\psi(z)\,dz.$$

Deduce that f_σ converges uniformly to f as σ tends to zero.

(ii) Show that f_σ is infinitely differentiable.

(iii) Use the inequality

$$|f_\sigma(x) - f_\sigma(y)| \le \int |f(x+\sigma z) - f(y+\sigma z)|\psi(z)\,dz \le \kappa|x-y|$$

to show that $\|f_\sigma\|_{\text{Lip}} \le \kappa$.

(iv) Show that all the partial derivatives of f_σ are also LIPSCHITZ functions (with LIPSCHITZ norms depending on σ and ψ).

(v) From (iii), we have $|f_\sigma(x+tu) - f_\sigma(x)| \le t\kappa$ for each unit vector u and each $t > 0$. Use the fact that $\left(f_\sigma(x+tu) - f_\sigma(x)\right)/t \to \langle u, \nabla f_\sigma\rangle$ as $t \searrow 0$ to show that $\sum_i (\partial f_\sigma/\partial x_i)^2 = |\nabla f_\sigma|^2 \le \kappa^2$.

(vi) Suppose W has a subgaussian distribution. Show that $\mathbb{P}e^{\lambda f_\sigma(W)} \to \mathbb{P}e^{\lambda f(W)}$ for each real λ as $\sigma \to 0$. Hint: dominated convergence using $|f_\sigma(x)| \le |f(0)| + \kappa|x| + 1$ for small enough σ.

[2] Suppose Z is $N(0,1)$ distributed and F is an absolutely continuous, real-valued function on the real line with almost sure (LEBESGUE) derivative f. That is f is integrable at least on each bounded interval and $F(b)-F(a) = \int_a^b f(t)\,dt$ for $-\infty < a < b < \infty$. If $\mathbb{P}|f(Z)| < \infty$, show that $\mathbb{P}|ZF(Z)| < \infty$ and $\mathbb{P}ZF(Z) = \mathbb{P}f(Z)$ by these steps. The argument is essentially just a very careful integration-by-parts, making sure there are no hidden $\infty - \infty$ cancellations.

(i) Explain why we may assume $F(0) = 0$ and $f \geq 0$ without loss of generality. Note that these assumptions imply that $xF(x) \geq 0$ for all x. Hint: Split into $f^{\pm}$ contributions.

(ii) Let ϕ denote the $N(0,1)$ density. Use the FUBINI theorem to show that

$$\mathbb{P}\{Z > 0\}ZF(Z) = \int_0^\infty \int_0^\infty z\phi(z)\{0 \leq t \leq z\}f(t)\,dt\,dz = \int_0^\infty f(t)\phi(t)\,dt.$$

(iii) Argue similarly for $\mathbb{P}\{Z < 0\}ZF(Z)$.

[3] Suppose $(Z, X_1, \ldots, X_m)$ has a MVN distribution with $Z \sim N(0, \sigma^2)$ (but no assumptions about the means or covariances for the X_i's). Suppose also that $G : \mathbb{R}^m \to \mathbb{R}$ is continuously differentiable, with partial derivatives $\dot{G}_i(x_1, \ldots, x_m) := \partial G(x_1, \ldots, x_m)/\partial x_i$. If $\mathbb{P}|ZG(X_1, \ldots, X_m)| < \infty$ and $\mathbb{P}|\dot{G}_i(X_1, \ldots, X_m)| < \infty$ for each i, show that

$$\mathbb{P}ZG(X_1, \ldots, X_m) = \sum\nolimits_{i \leq m} \tau_i \mathbb{P}\dot{G}_i(X_1, \ldots, X_m) \qquad \text{where } \tau_i := \mathbb{P}ZX_i.$$

Follow these steps.

(i) Without loss of generality suppose σ equals 1. (Equivalently, divide both sides of the asserted representation by σ.)

(ii) Show that Z is independent of the random variables $Y_i := X_i - \tau_i Z$ for $i \in [\![m]\!]$. [Check that $\operatorname{cov}(Z, Y_i) = 0$.] For fixed y_i values, use Problem [2] to show that

$$\begin{aligned}
&\mathbb{P}\left(ZG(X) \mid Y_i = y_i \text{ for } i \in [\![m]\!]\right) \\
&\quad = \gamma_n^z z G(y_1 + \tau_1 z, \ldots, y_m + \tau_m z) \\
&\quad = \sum\nolimits_i \tau_i \gamma_n^z \dot{G}_i(y_1 + \tau_1 z, \ldots, y_m + \tau_m z).
\end{aligned}$$

(iii) Average out over the y_i's with respect to the joint distribution of $Y_1, \ldots, Y_m$.

[4] (cf. Isserlis 1918; or google "Wick"; or look at Janson 1997, Theorem 1.28) Suppose $X \sim N_n(0, V)$. For each subset A of $[\![n]\!]$ define $M_A := \mathbb{P}\prod_{i \in A} X_i$.

(i) Use Problem [3] to show that $M_{[\![n]\!]} = \sum_{j \geq 2} V_{1,j} M_{[\![n]\!] \setminus \{1,j\}}$.

(ii) Deduce that $M_{[\![n]\!]}$ is either zero (for n odd) or it can be written as a sum of products of off-diagonal elements of V.

(iii) Calculate the representations for $n = 4$ and $n = 5$.

[5] For suitably regular f (such as the f_σ from Problem [1]) and $a_t = \sqrt{1-t}$, show that the function $F(x,t) = \gamma_n^z f(x + z a_t)$ (as in equation <15>) satisfies

$$\begin{aligned}
\partial F(x,t)/\partial t &= \dot{a}_t \sum\nolimits_i \gamma_n^z z_i \dot{f}_i(x + z a_t) \qquad \text{for } 0 < t < 1 \\
&= a_t \dot{a}_t \sum\nolimits_i \gamma_n^z \ddot{f}_{i,i}(x + z a_t) \qquad \text{by Problem [2]} \\
&= -\tfrac{1}{2} \sum\nolimits_i \partial^2 F(x,t)/\partial x_i^2.
\end{aligned}$$

Remark. If the "−" sign in the final line were removed the equation would become $\partial F(x,t)/\partial t = \frac{1}{2}\sum_i \partial^2 F(x,t)/\partial x_i^2$, the heat equation. The negative sign comes from the $1-t$ in the definition of a_t; time is running backwards. Maybe the partial differential equation could be called the 'backwards heat equation'.

[6] A subset D of $\mathbb{R}^N$ is called an ellipsoid in a k-dimensional subspace $\mathcal{H}$ of $\mathbb{R}^N$ if there is a linear map T that is a bijection from $\mathbb{R}^k$ onto $\mathcal{H}$ of $\mathbb{R}^N$ such that such that $D = h + TB$, where $B = \{x \in \mathbb{R}^k : |x|_2 \le 1\}$ and $h \in \mathcal{H}$.

(i) Suppose T has the singular value decomposition $Tv_i = \lambda_i u_i$ for $1 \le i \le k$, for singular values $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_k > 0$ with $\{v_i : i \le k\}$ an onb of $\mathbb{R}^k$ and $\{u_i : i \le k\}$ an onb of $\mathcal{H}$. Show that

$$TB = \Big\{ \sum\nolimits_{i\le k} t_i u_i : \sum\nolimits_{i\le k} (t_i/\lambda_i)^2 \le 1 \Big\}.$$

In particular, TB is a closed ball in $\mathcal{H}$ if $\lambda_i = \lambda_1$ for all i.

(ii) Show that there is a subspace $\mathcal{L}_0$ of $\mathbb{R}^k$ of dimension approximately $k/2$ for which $D \cap (T\mathcal{L}_0)$ is a closed ball. Argue as follows. Suppose k is even, $k = 2\ell$. Choose any value λ with $\lambda_\ell \ge \lambda \ge \lambda_{\ell+1}$. For $i \le \ell$ choose positive numbers c_i and s_i for which $c_i^2 + s_i^2 = 1$ and $c_i^2\lambda_i + s_i^2\lambda_{k-i+1} = \lambda^2$. Define $V_i = c_i v_i + s_i v_{k-i+1}$ and $U_i = \left(c_i\lambda_i u_i + s_i\lambda_{k-i+1}u_{k-i+1}\right)/\lambda$. Show that V_i and U_i are unit vectors with $TV_i = \lambda U_i$ for $i \le \ell$. Define $\mathcal{L}_0 = \operatorname{span}\{V_1, \ldots, V_\ell\}$.

If k is odd, $k = 2\ell + 1$, argue similarly with $\lambda = \lambda_{\ell+1}$ and $V_{\ell+1} = v_{\ell+1}$.

[7] Suppose h and H are random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ for which $0 < h \le H$ almost surely and $\mathbb{P}H \le (1+\epsilon)\mathbb{P}h < \infty$ for some positive ϵ. Argue as follows to show that there exists an ω in Ω for which $1 \le H(\omega)/h(\omega) \le 1+\epsilon$. Define a new probability measure $\mathbb{Q}$ by $d\mathbb{Q}/d\mathbb{P} = h/\mathbb{P}h$. Show that $\mathbb{Q}(H/h) \le 1+\epsilon$, so that $\mathbb{Q}\{\omega : H(\omega)/h(\omega) > 1+\epsilon\} < 1$.

[8] (repeated from Chapter 3) Suppose $\xi_1, \ldots, \xi_N$ are independent random variables, each $N(0,1)$ distributed. Define $M_n = \max_{j\le N} |\xi_j|$ and $x_n = \sqrt{\log n}$.

(i) Show that

$$\mathbb{P}M_n = \int_0^\infty \mathbb{P}\{M_n > t\}\, dt \ge x_n \mathbb{P}\{M_n > x_n\}.$$

(ii) Show that $\mathbb{P}\{M_n > t\} = 1 - (1 - 2\bar{\Phi}(t))^n \ge 1 - \exp\left(-2n\bar{\Phi}(t)\right)\big)$.

(iii) Use Laplace's bound from §3.4, $\bar{\Phi}(t) \ge (t^{-1} - t^{-3})\phi(t)$ for $t > 1$, to show that $\mathbb{P}\{M_n > x_n\} \to 1$ as $n \to \infty$.

(iv) Deduce that $\mathbb{P}M_n \ge x_n(1 - o(1))$.

[9] Suppose $\{X_i : i \in \mathbb{N}\}$ is a gaussian process with $X_i \sim N(0, \sigma_i^2)$. Define $M_n := \max_{i \le n} X_i$ and $M_\infty := \sup_{i \in \mathbb{N}} X_i$. Show that $M_\infty < \infty$ almost surely if and only if $\mathbb{P}M_\infty < \infty$. Argue as follows for the nontrivial implication. Suppose $\mathbb{P}\{M_\infty < \infty\} = 1$.

(i) Show that there exists an $R \in \mathbb{R}$ for which

$$\bar{\Phi}(1) > \mathbb{P}\{M_\infty > R\} \ge \mathbb{P}\{M_n > R\} \ge \mathbb{P}\{X_n > R\} = \bar{\Phi}(R/\sigma_n).$$

Deduce that $\sigma := \sup_n \sigma_n \le R$.

(ii) Choose any constant r_0 for which $1 > 2e^{-r_0^2/2} + \bar{\Phi}(1)$. Write m_n for $\mathbb{P}M_n$. Use the concentration inequality from Section 6.1 to show that

$$\mathbb{P}\{M_n \ge m_n - \sigma r_0\} \ge 1 - 2e^{-r_0^2/2} > \bar{\Phi}(1) > \mathbb{P}\{M_n > R\}.$$

Deduce that $m_n - \sigma r_0 \le R$ for all n.

(iii) Deduce that $\mathbb{P}M_\infty \le R + \sigma r_0 < \infty$. Hint: $0 \le |X_1| + M_n \nearrow |X_1| + M_\infty$.

(iv) Extend the argument to a two-sided equivalence: $\mathbb{P}\{\sup_i |X_i| < \infty\} = 1$ iff $\mathbb{P} \sup_i |X_i| < \infty$. Hint: $\sup_i |X_i| = \max(\sup_i X_i, \sup_i(-X_i))$.

[10] Let $\mathbb{H}$ be a HILBERT space with inner product $\langle \cdot, \cdot \rangle$, and an orthonormal basis $\{e_\alpha : \alpha \in I\}$. Let $\{\eta_\alpha : \alpha \in I\}$ be a set of independent random variables, each distributed $N(0,1)$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. (Kolmogorov's existence theorem would provide such a space.)

(i) For h in $\mathbb{H}$ define $I(h) := \{\alpha \in I : \langle h, e_\alpha \rangle \ne 0\}$, a countable set for which $h = \sum_{\alpha \in I(h)} e_\alpha \langle h, e_\alpha \rangle$, a series that converges in norm. Show that the sum $Z_h := \sum_{\alpha \in I(h)} \eta_\alpha \langle h, e_\alpha \rangle$ converges in the $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ sense. Hint: For each finite subset A of I, show that

$$\mathbb{P}\Big|\sum\nolimits_{\alpha \in A} \eta_\alpha \langle h, e_\alpha \rangle\Big|^2 = \sum\nolimits_{\alpha \in A} |\langle h, e_\alpha \rangle|^2.$$

Note: The limit is defined only up to a P-equivalence. Perhaps it would be more elegant to work with limits in $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

(ii) Show that $\{Z_h : h \in \mathbb{H}\}$ is a centered gaussian process with $\operatorname{cov}(Z_g, Z_h) = \langle g, h \rangle$ for $g, h \in \mathbb{H}$. Remark: This process is sometimes called the ***isonormal process indexed by*** $\mathbb{H}$. See Dudley (2014, §2.2.1) for an insightful discussion.

[11] Let $\{X_t : t \in T\}$ be a centered gaussian process, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Equip T with the semi-metric $d(s,t) := \|X_s - X_t\|_2 := \big(\mathbb{P}|X_s - X_t|^2\big)^{1/2}$.

(i) Show that $X_s = X_t$ ae[$\mathbb{P}$] if $d(s,t) = 0$. Explain why this fact might cause difficulties with sample-path continuity if the inequality could not be strengthened to $X_s(\omega) = X_t(\omega)$ for all ω.

(ii) Partion T into equivalence classes $[t] := \{s \in T : d(s,t) = 0\}$. Let T_0 be a subset of T built up by choosing one member from each equivalence class. Show that $\{X_t : t \in T_0\}$ can be identified with a subset of the HILBERT space $\mathbb{H} := L^2(\Omega, \mathcal{F}, \mathbb{P})$ via the map $X_t \leftrightarrow [X_t]$ in such a way that

$$\operatorname{cov}(X_s, X_t) = \langle [X_s], [X_t] \rangle$$

where $[X_t]$ denotes the equivalence class of X_t in $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Remark: This representation explains why some authors regard all centered gaussian processes as subsets of an isonormal process indexed by a HILBERT space.

6.9 Notes

I know little about the early history of the path idea, except that it appears in the work of Plackett (1954, page 353), Chover (1962), and Slepian (1962, page 483).

The most stunning fact about γ_n—the so-called isoperimetric inequality—was established independently by Borell (1975) and Sudakov and Tsirel'son (1978), a translation from a 1974 paper in Russian. See the concise and informative book by Ledoux (2001) for more about concentration inequalities. For an exposition of a proof due to Ehrhard (1983a,b) see Pollard (2001, Section 12.5). The inequality can be rewritten more compactly as

$$\gamma_k(A^\delta) \geq \Phi\left(\Phi^{-1}(\gamma_n A) + \delta\right),$$

slightly disguising the fact that equality is achieved by halfspaces but leading towards a functional form of the inequality that was developed by Bobkov (1996, 1997). See Boucheron, Lugosi, and Massart (2000, pages 290–298, 303–307) for an insightful discussion of Bobkov's method. For elegant reformulations of the functional approach see Ledoux (1998) and Barthe and Maurey (2000).

The lower bound in Example <5> was proved by Sudakov (1973) using a different method.

The integration by parts formula, $\mathbb{P}ZF(Z) = \mathbb{P}\dot{F}(Z)$ if $Z \sim N(0,1)$, (see Problem [2]) actually characterizes the $N(0,1)$ distribution. This fact lies at the heart of Charles Stein's astounding theory of normal approximation, as exposited in his Lecture notes (Stein, 1986). See page 21 of that volume for a proof of the characterization. Undoubtedly the same trick has been used for many other purposes. See Problem [4], for example.

A huge part of the modern theory of Gaussian process (and stochastic processes in general) grew from the ideas of Xavier Fernique. For example, see Fernique (1975, 1983) or search for his name at the French digital

mathematics library, http://www.numdam.org/. See also Ledoux and Talagrand (1991, page 87–88) and Dudley (2014, Chapter 2 Notes) for more about the history of comparison methods and where credit is due.

My introduction to the stochastic integration method described in Section 6.3.1 came from Adler (1990, Section 2.1), who attributed the method to Maurey and Pisier, citing lectures by Pisier (1986). In those lectures Pisier (page 180) had provided a sketch of Maurey's argument, prefaced by the comment "B. Maurey found a proof of theorem 2.1 with the best constant ... His proof uses stochastic integrals and apparently does not extend to the setting of theorem 2.2." Only later did I notice that Ledoux (2001, page 45) gave credit for the stochastic calculus proof to Cirel'son, Ibragimov, and Sudakov (1976), with the comment that their paper "was unfortunately ignored for a long time". That 1976 paper had attributed the concentration result (via the isoperimetric inequality) to the 1974 Russian version of Sudakov and Tsirel'son (1978), with the remark (page 25) that the 1974 proof was "not a purely probabilistic one ... But sometimes the following assertion, provable in a purely probabilistic way, can replace it". They then gave a stochastic calculus argument that started in the same way as the proof in Section 6.3.1, with the identification of the martingale M, but then they invoked a time-change argument (Chung and Williams, 2014, Section 9.3) to represent M as $\gamma_n f + W([M]_t)$ for a new BROWNIAN motion W. In my notation, the rest of their proof used the fact (inequality <26>) that $[M]_1 \le \kappa^2$ to deduce that

$$\begin{aligned}\mathbb{P}\{M_1 \ge \gamma_n f + \kappa x\} &\le \mathbb{P}\{\sup_{0\le s\le \kappa^2} W_s \ge \kappa x\} \\ &= 2\mathbb{P}\{W_{\kappa^2}/\kappa \ge x\} \qquad \text{reflection principle} \\ &= 2\bar{\Phi}(x).\end{aligned}$$

That is, they actually established a result sharper than the $e^{-x^2/2}$ bound.

See Davidson and Szarek (2001, Section II) and Vershynin (2018) for discussion of ways that concentration and comparison inequalities enter the theory of random matrices.

References

Adler, R. J. (1990). *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*, Volume 12 of *Lecture Notes–Monograph series*. Hayward, CA: Institute of Mathematical Statistics.

Barthe, F. and B. Maurey (2000). Some remarks on isoperimetry of Gaussian type. *Annales de l'Institut Henri Poincaré, Probability and Statistics 36*(4), 419–434.

Bobkov, S. (1996). A functional form of the isoperimetric inequality for the Gaussian measure. *Journal of Functional Analysis 135*, 39–49.

Bobkov, S. G. (1997). An isoperimetric inequality on the discrete cube, and an elementary proof of the isoperimetric inequality in Gauss space. *Annals of Probability 25*(1), 206–214.

Borell, C. (1975). The Brunn-Minkowski inequality in Gauss space. *Inventiones Mathematicae 30*, 207–216.

Boucheron, S., G. Lugosi, and P. Massart (2000). A sharp concentration inequality with applications. *Random Structures and Algorithms 16*, 277–292.

Chatterjee, S. (2005). An error bound in the Sudakov-Fernique inequality. Technical report, arXiv:math/0510424.

Chover, J. (1962, March). Certain convexity conditions on matrices with applications to Gaussian processes. *Duke Mathematical Journal 29*(1), 141–150.

Chung, K. L. and R. J. Williams (2014). *Introduction to Stochastic Integration* (Second ed.). Boston: Birkhäuser.

Cirel'son, B., I. Ibragimov, and V. Sudakov (1976). Norms of Gaussian sample functions. In *Proceedings of the Third Japan-USSR Symposium on Probability Theory*, Volume 550 of *Springer Lecture Notes in Mathematics*, pp. 20–41. Springer.

Davidson, K. R. and S. J. Szarek (2001). Local operator theory, random matrices and Banach spaces. *Handbook of the geometry of Banach spaces 1*(317-366), 131. (Available in preprint form at http://case.edu/artsci/math/szarek/publications.html).

Dudley, R. M. (2014). *Uniform Central Limit Theorems* (2nd ed.), Volume 142 of *Cambridge studies in advanced mathematics*. Cambridge University Press. (First edition, 1999).

Dvoretzky, A. (1961). Some results on convex bodies and Banach spaces. In *Proceedings of the International Symposium on linear spaces (held at Hebrew University of Jerusalem, July 5-12, 1960)*, pp. 123–160. Academic Press; New York, Pergamon Press.

Ehrhard, A. (1983a). Symétrisation dans l'espace de Gauss. *Mathematica Scandinavica 53*, 281–301.

Ehrhard, A. (1983b). Un principe de symétrisation dans les espaces de Gauss. *Springer Lecture Notes in Mathematics 990*, 92–101.

Fernique, X. (1975). Regularité des trajectoires des fonctions aléatoires gaussiennes. *Springer Lecture Notes in Mathematics 480*, 1–97. École d'Été de Probabilités de Saint-Flour IV, 1974.

Fernique, X. (1983). Regularité de fonctions aléatoires non gaussiennes. *Springer Lecture Notes in Mathematics 976*, 1–74. École d'Été de Probabilités de St-Flour XI, 1981.

Gordon, Y. (1985). Some inequalities for Gaussian processes and applications. *Israel Journal of Mathematics 50*(4), 265–289.

Gordon, Y. (1988). Gaussian processes and almost spherical sections of convex bodies. *The Annals of Probability 16*(1), 180–188.

Gordon, Y. (2004-2005). A note on an observation of G. Schechtman. *Springer Lecture Notes in Mathematics 1910*, 127–132. Geometric Aspects of Functional Analysis, Israel Seminar.

Isserlis, L. (1918). On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika 12*(1/2), 134–139.

Janson, S. (1997). *Gaussian Hilbert Spaces*. Cambridge Tracts in Mathematics. Cambridge University Press.

Kahane, J.-P. (1986). Une inégalité du type de Slepian et Gordon sur les processus gaussiens. *Israel Journal of Mathematics 55*(1), 109–110.

Karatzas, I. and S. E. Shreve (1988). *Brownian Motion and Stochastic Calculus* (First ed.). New York: Springer-Verlag.

Ledoux, M. (1998). A short proof of the Gaussian isoperimetric inequality. In *High Dimensional Probability*, Volume 43 of *Progress in Probability*, pp. 229–232. Springer.

Ledoux, M. (2001). *The Concentration of Measure Phenomenon*, Volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. New York: Springer.

Marshall, A. W., I. Olkin, and B. C. Arnold (2011). *Inequalities: Theory of Majorization and its Applications* (Second ed.), Volume 143 of *Series in Statistics*. Springer.

Pisier, G. (1986). Probabilistic methods in the geometry of Banach spaces. In G. Letta and M. Pratelli (Eds.), *Probability and Analysis*, Volume 1206 of *Lecture Notes in Mathematics*, pp. 167–241. Springer Berlin Heidelberg.

Plackett, R. L. (1954). A reduction formula for normal multivariate integrals. *Biometrika 41*(3/4), 351–360.

Pollard, D. (2001). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.

Slepian, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal 41*, 463–501.

Stein, C. (1986). *Approximate Computation of Expectations*, Volume 7 of *Lecture Notes–Monograph series*. Institute of Mathematical Statistics.

Sudakov, V. (1973). A remark on the criterion of continuity of Gaussian sample function. In *Proceedings of the Second Japan-USSR Symposium on Probability Theory, Kyoto, August 2–9, 1972*, Volume 330 of *Springer Lecture Notes in Mathematics*, pp. 444–454.

Sudakov, V. N. and B. S. Tsirel'son (1978). Extremal properties of half-spaces for spherically invariant measures. *Journal of Soviet Mathematics 9*, 419–434. (Translated from Zapiski Nauchnykh Seminarov Leningradskogo Otdeleniya Matematicheskogo instituta im. V. A. Steklova AN SSSR, Vo!. 41, pp. 14-24, 1974.).

Talagrand, M. (2003). *Spin Glasses: a Challenge to Mathematicians*, Volume 46 of *Ergbnisse der Mathematik und ihrer Grenzgebiete*. New York: Springer-Verlag.

Talagrand, M. (2011). *Mean Field Models for Spin Glasses: Volume I: Basic Examples*, Volume 54 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer Science & Business Media.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.

Vitale, R. A. (2000). Some comparisons for Gaussian processes. *Proceedings of the American Mathematical Society 128*(10), 3043–3046.

Chapter 7
Subgaussian distributions

SECTION 7.1 *defines the subgaussian family of distributions, providing two useful examples.*

SECTION 7.2 *shows that the subgaussian property is equivalent to several other distributional properties.*

SECTION 7.3 *discusses the* HOEFFDING *inequality for sums of independent bounded random variables.*

SECTION 7.4 *shows how the* HOEFFDING *inequality can be extended to sums of martingale differences.*

SECTION 7.5 *presents an application of the martingale version of the* HOEFFDING *inequality, although the resulting bound falls far short of what is possible using other methods.*

7.1 Definition of subgaussian

As noted in §3.3, every random variable X for which $\mathbb{P}e^{\lambda X} \le e^{\sigma^2\lambda^2/2}$, with $\sigma > 0$ and all $\lambda \in \mathbb{R}$, shares with the $N(0,\sigma^2)$ a two-sided exponential tail bound for $r \ge 0$,

<1> $$\begin{aligned} \mathbb{P}\{X \ge r\} &\le \inf_{\lambda>0} \mathbb{P}e^{\lambda(X-r)} = \exp(-r^2/2\sigma^2) \\ \mathbb{P}\{X \le -r\} &\le \inf_{\lambda\le 0} \mathbb{P}e^{\lambda(X+r)} = \exp(-r^2/2\sigma^2) \end{aligned}$$

As shown by Problems [1] and [2], the constant σ^2 might need to be strictly greater than $\operatorname{var}(X)$. To avoid any hint of the convention that σ^2 denotes a variance, I feel it is safer to replace σ by another symbol.

<2> **Definition.** *Say that a random variable X has a subgaussian distribution with scale factor $0 < \tau < \infty$, denoted by $X \in \text{SUBG}(\tau^2)$, if $\mathbb{P}e^{\lambda X} \le \exp(\tau^2\lambda^2/2)$ for all real λ. Write $\tau(X)$ for the smallest τ for which such a bound holds.*

Remark. Notice that $\text{SUBG}(\tau^2)$ denotes not a single distribution but rather a whole family of distributions, hence the $\in$ instead of $\sim$. This means that it makes no sense to speak of a maximum likelihood estimator (mle) for the parameter of $\text{SUBG}(\tau^2)$ based on independent observations $X_1,\ldots,X_n$, although it does make sense to speak of the behavior of the mle, $\widehat{\tau}^2 = n^{-1}\sum_i (X_i - \overline{X})^2$, derived for the $N(0,\tau^2)$ model when the distributional assumption is relaxed to subgaussianity. In the same way, one could examine the behavior of the mle under other departures from normality, as the Robustniks like to do.

Some authors call the SUBG(τ^2) the centered SUBGAUSSIAN. More generally, if $\mathbb{P}e^{\lambda X} \le \exp(\lambda\nu + \lambda^2\tau^2/2)$ for all real λ then X has a SUBG(ν, τ^2) distribution, with ν the location and τ the scale parameter. In this case ν must equal $\mathbb{P}X$ and $\mathbb{P}(X-\nu)^2 \le \tau^2$. If the bound on M_X held only for nonnegative λ we could only deduce that $\mathbb{P}X \le \nu$ and $\mathbb{P}(X-\nu)^2 \le \tau^2$. It

The first of the two inequalities <1> is sometimes called a ***subgaussian bound for the upper tail***. Somewhat confusingly, the term 'subgaussian' is also often used in a looser sense as an indication that a tail probability decreases at an $\exp(-Ct^2)$ rate for some positive constant C. See §3.6 for the upper tail of $\sqrt{X}$ when X has a GAMMA(α) distribution, although the centering constant $\sqrt{\alpha}$ is larger than $\mathbb{P}\sqrt{X}$. Even sloppier is the habit (which I have) of referring to a tail bound as 'approximately subgaussian' in particular regions. See, for example, the discussion of the upper tail for the GAMMA(α) distribution. The sloppy justification for such a habit is that some applications for tail bounds involve only deviations in the 'approximately subgaussian' region.

If X has a SUBG(β^2) disribution, inequality <1> gives a pair of tail bounds,

$$\mathbb{P}\{\pm X \ge r\} \le \exp\left(-\frac{r^2}{2\beta^2}\right) \qquad \text{for all } r \ge 0, \tag{<3>}$$

In fact (Theorem <6>) existence of such a tail bound for some $\beta > 0$ is equivalent to subgaussianity.

<4> **Example.** Suppose $Y = (Y_1, \ldots, Y_n)$ has a multivariate normal distribution. Define $M := \max_{i\le n} Y_i$ and $\sigma^2 := \max_{i\le n} \operatorname{var}(Y_i)$. From §6.1 we have
□ That is, $M - \mathbb{P}M \in$ SUBG(σ^2), a most surprising and wonderful fact.

Many useful results that hold for gaussian variables can be extended to subgaussians. In empirical process theory, conditional subgaussianity plays a key role in symmetrization arguments (see Chapter 13). The next example captures the main idea.

<5> **Example.** Suppose $S = \sum_{j\le n} c_j \mathfrak{s}_j$, where the c_j's are constants and the $\mathfrak{s}_j$'s are independent ***sign variables*** (often called **rademacher** variables, for dubious reasons) with $\mathbb{P}\{\mathfrak{s}_j = +1\} = 1/2 = \mathbb{P}\{\mathfrak{s}_j = -1\}$. Via the fact that $k!2^k \le (2k)!$ for each positive integer k we have

$$\mathbb{P}e^{\lambda c_j \mathfrak{s}_j} = \tfrac{1}{2}\left(e^{\lambda c_j} + e^{-\lambda c_j}\right) = \sum\nolimits_{k=0}^{\infty} (\lambda c_j)^{2k}/(2k)! \le \exp(\lambda^2 c_j^2/2)$$

so that $\mathbb{P}e^{\lambda S} = \prod_{j\le n} \mathbb{P}e^{\lambda c_j \mathfrak{s}_j} \le e^{\lambda^2\tau^2/2}$ with $\tau^2 := \sum_{j\le n} c_j^2$. That is, the sum S has a SUBG(τ^2) distribution, so that

$$\mathbb{P}\{\pm S \ge r\} \le \exp\left(-\tfrac{1}{2}r^2/\sum\nolimits_i c_i^2\right) \qquad \text{for } r \ge 0.$$

The $\pm$ is just my lazy way of indicating that the inequality holds for both S
□ and $-S$, that is, it holds for both upper and lower tails.

Remark. In the literature, the $\mathfrak{s}_i$ sign variables are often called RADEMACHER variables, presumably because of a perceived similarity to a particular orthonormal basis used by Rademacher (1922). Notably, Talagrand (2021) preferred the name "Bernoulli random variables". Unfortunately his terminology clashes with my use of $\text{BER}(p)$ for the distribution that puts mass p at 1 and mass $1-p$ at 0.

More generally, if S is a sum of independent subgaussians $X_1, \ldots, X_n$ then the equality $\mathbb{P}e^{\lambda S} = \prod_j \mathbb{P}e^{\lambda X_j}$ shows that $X \in \text{SUBG}(\tau^2)$ for $\tau^2 = \sum_j \tau^2(X_j)$. For sums of independent variables we need consider only the subgaussianity of each summand.

7.2 Characterizations of subgaussianity

The calculations at the end of Section 7.1 show the advantage of defining subgaussianity using moment generating functions. For other purposes equivalent definitions, as identified by the following Theorem, are sometimes more convenient.

<6> **Theorem.** *For each random variable X with expected value 0, the following assertions are equivalent.*

(i) X has a subgaussian disribution (with $\tau(X) < \infty$)

(ii) $L(X) := \sup_{k\in\mathbb{N}} \|X\|_{2k} / \sqrt{k}$ is finite

(iii) $X \in \mathcal{L}^{\Psi_2}$ for the ORLICZ *function $\Psi_2(x) = \exp(x^2) - 1$*

(iv) there exists a positive constant β for which

$$\mathbb{P}\{|X| \ge t\} \le 2\exp\left(-\tfrac{1}{2}t^2/\beta^2\right) \qquad \text{for all } t \ge 0.$$

More precisely,

$$\|X\|_{\Psi_2} / \sqrt{6} \le \beta(X) \le \tau(X) \le \sqrt{8}\, L(X)$$

and $L(X) \le \|X\|_{\Psi_2} \le \sqrt{2e} L(X)$, where $\beta(X)$ denotes the smallest β for which (iv) holds.

Remarks.

(i) Implicit in the sequences of inequalities is the assertion that finiteness of one quantity implies finiteness of another quantity.

(ii) It might be more natural use $\sqrt{2k}$ instead of $\sqrt{k}$ in the definition of $L(X)$. The choice $\sqrt{k}$ ensures that $\|X\|_p \le L(X)\sqrt{p}$ for all real $p \ge 1$: if $k = \lceil p/2 \rceil$ then $\|X\|_p \le \|X\|_{2k}$ and $p \ge k$.

Proof. Simplify notation by writing β for $\beta(X)$ and τ for $\tau(X)$. and L for $L(X)$. Also, assume that X is not a constant, to avoid annoying trivial cases.

For $\beta \le \tau$: Use inequality <1>.

For $\|X\|_{\Psi_2}$ and β:

$$\begin{aligned}\mathbb{P}\Psi_2\left(\frac{|X|}{\beta\sqrt{6}}\right) &= \mathbb{P}\int_0^\infty e^t\{|X|^2 \ge 6\beta^2 t\}\,dt \\ &= \int_0^\infty e^t \mathbb{P}\{|X| \ge \beta\sqrt{6t}\}\,dx \\ &\le 2\int_0^\infty \exp(x-3x)\,dx = 1.\end{aligned}$$

For L versus $\|X\|_{\Psi_2}$: This result is mostly an exercise in applying the inequality $k/e \le (k!)^{1/k} \le k$. It does not need $\mathbb{P}X = 0$. First note that $\mathbb{P}\Psi_2(|X|/D) = \sum_{k\in\mathbb{N}} \|X/D\|_{2k}^{2k}/k!$. If $\infty > D > \|X\|_{\Psi_2}$ then the sum is bounded by 1, which implies $\|X\|_{2k} \le D(k!)^{1/2k} \le D\sqrt{k}$ for each k. Conversely,

$$\mathbb{P}\Psi_2(|X|/D) \le \sum\nolimits_{k\in\mathbb{N}} (L\sqrt{k}/D)^{2k}(e/k)^k = 1 \qquad \text{if } D = \sqrt{2e}L.$$

For $\tau \le 4L$: Let X' be a random variable with the same distribution as X but independent of X. By the triangle inequality, $\|X - X'\|_{2k} \le 2\|X\|_{2k} \le 2L\sqrt{k}$ for all $k \in \mathbb{N}$. By independence, $\mathbb{P}_X X' = \mathbb{P}X' = 0$. Via Jensen's inequality for the conditional expectation operator $\mathbb{P}_X$ we have

$$\begin{aligned}\mathbb{P}e^{\lambda X} = \mathbb{P}e^{\lambda(X - \mathbb{P}_X X')} &\le \mathbb{P}e^{\lambda(X-X')} = 1 + \sum\nolimits_{j\in\mathbb{N}} \lambda^j \mathbb{P}(X-X')^j/j! \\ &= 1 + \sum\nolimits_{k\in\mathbb{N}} \frac{\lambda^{2k}\mathbb{P}(X-X')^{2k}}{(2k)!} \qquad \text{symmetry kills odd moments} \\ &\le 1 + \sum\nolimits_{k\in\mathbb{N}} \frac{(\lambda 2L\sqrt{k})^{2k}}{k^k k!} = 1 + \sum\nolimits_{k\in\mathbb{N}} \frac{(4\lambda^2L^2)^k}{k!} = e^{4L^2\lambda^2},\end{aligned}$$

and so on. If you worry about the legitimacy of taking $\mathbb{P}$ inside $\sum_j$, note that

$$\sum\nolimits_{j\in\mathbb{N}} \mathbb{P}|\lambda^j(X-X')^j|/j! \le \sum\nolimits_{j\in\mathbb{N}} |\lambda|^j(2L\sqrt{j})^j/(j/e)^j < \infty$$

□ then appeal to dominated convergence.

Remark. As mentioned in Chapter 3, the name 'subgaussian' is often applied in a loose sense to any tail bound that decreases like $\exp(-Ct^2)$. For example, §3.7 showed that the upper tail for the $\text{BIN}(n,p)$ is subgaussian in this sense, with even the best constant $C^{-1} = 2np(1-p)$ when $p \geq 1/2$.

I tried without much success to extend Theorem <6> to an analogous result for 'subgaussianity of a single tail'. Part of the difficulty was the possibility that the centering constant need not be equal to the mean. See Problem [1]. Another difficulty was the failure of symmetrization when only one tail was required to decrease in a subgaussian way.

Before giving up completely I did have some partial success with the idea that subgaussian in the upper tail should correspond to subgaussianity of $(X-\nu)^+$ for some constant ν. Unfortunately, the fact that $\mathbb{P}(X-\nu)^+ = 0$ only when $\mathbb{P}\{X > \nu\} = 0$ got in the way of clean Taylor expansions near the origin.

7.3 The HOEFFDING inequality for independent summands

In a very famous paper, Hoeffding (1963) proved several exponential tail bounds for sums of independent random variables. He also extended the results to a number of dependent settings (including martingales—see Section 7.4). The following result is probably the best known from his paper.

<7> **Theorem.** *(Hoeffding, 1963, Theorem 2) Suppose $S_n = \sum_i X_i$, a sum of independent random variables with $a_i \leq X_i \leq b_i$ for each i, for constants a_i and b_i. Then*

$$S_n - \mathbb{P}S_n \in \text{SUBG}(\tau^2) \qquad \text{for } \tau^2 := \textstyle\sum_i c_i^2, \text{ where } c_i := (b_i - a_i)/2.$$

Consequently,

<8> $$\mathbb{P}\{S_n - \mathbb{P}S_n \geq r\} \leq \exp\left(-r^2/2\tau^2\right) \qquad \text{for each } r \geq 0,$$

with an analogous inequality for the lower tail.

Proof. First show that $X_i - \mathbb{P}X_i$ has a $\text{SUBG}(c_i^2)$ distribution, for each i. The simplest approach uses facts about $L_i(\lambda) := \log \mathbb{P}e^{\lambda X_i}$ that were established in §2.3:

$$\dot{L}_i(\lambda) = \mathbb{P}_{\lambda,i} X_i \qquad \text{AND} \qquad \ddot{L}_i(\lambda) = \text{var}_\lambda(X_i) := \mathbb{P}_{\lambda,i}\left(X_i - \mathbb{P}_{\lambda,i}X_i\right)^2,$$

where $\mathbb{P}_{\lambda,i}$ is the probability measure defined by the tilted density $e^{\lambda X_i}/\mathbb{P}e^{\lambda X_i}$ with respect to $\mathbb{P}$. As $\mathbb{P}_{\lambda,i}\{a_i \leq X_i \leq b_i\} = 1$ it follows that

$$\text{var}_\lambda(X_i) \leq \mathbb{P}_{\lambda,i}\left(X_i - (a_i+b_i)/2\right)^2 \leq c_i^2 \qquad \text{for every } \lambda.$$

In particular, for some λ^* between 0 and λ,

$$L_i(\lambda) = L_i(0) + \lambda \dot{L}_i(0) + \tfrac{1}{2}\lambda^2 \ddot{L}_i(\lambda^*) \leq 0 + \lambda \mathbb{P}X_i + \tfrac{1}{2}\lambda^2 c_i^2.$$

It then follows that

$$\mathbb{P}e^{\lambda(S_n-\mathbb{P}S_n)} = \prod_{i\le n} \mathbb{P}e^{\lambda(X_i-\mathbb{P}X_i)} \le e^{\lambda^2\sum_{i\le n} c_i^2/2},$$

□ the desired subgaussian bound.

Remark. Hoeffding (1963, §4) first used convexity of the exponential function to get

<9> $$e^{\lambda X_i} \le \frac{b-X_i}{b-a}e^{\lambda a} + \frac{X_i-a}{b-a}e^{\lambda b}.$$

Then he took expected values:

$$\begin{aligned} \mathbb{P}e^{\lambda X_i} &\le \theta_i e^{\lambda a_i} + (1-\theta_i)e^{\lambda b_i} \qquad \text{where } \theta_i := (b_i - \mathbb{P}X_i)/(b_i-a_i) \\ &\le \exp\left(\lambda\mathbb{P}X_i + c_i^2\lambda^2/2\right) \qquad \text{by brute force calculus.} \end{aligned} \tag{10}$$

The inequality for <10> could also be established by noting that $\theta_i e^{\lambda a_i} + (1-\theta_i)e^{\lambda b_i}$ is the MGF of a random variable W for which $\theta_i = \mathbb{P}\{W = a_i\} = 1 - \mathbb{P}\{W = b_i\}$ and $\mathbb{P}W = \theta_i a_i + (1-\theta_i)b_i = \mathbb{P}X_i$. The brute force calculation is essentially equivalent to the representation of the second derivative of $\log \mathbb{P}e^{\lambda W}$ as a variance under a tilted distribution.

<11> **Example.** The result from Example <5> is a special case of Theorem <7> with $X_j = c_j\mathfrak{s}_j$, a random variable that takes values in the interval $[-c_j, c_j]$, with $\mathbb{P}X_j = 0$ and $\operatorname{var}(X_j) = c_j^2$. In a strong sense this case is extreme, because the distribution of X_j concentrates at the endpoints of the interval.

In general, one quarter of the squared length of the range can be a vast overestimate of the variance. For example, suppose $S \sim \text{BIN}(n,p)$. Then $\mathbb{P}S = np$ and $S - np = \sum_{j\le n} X_j$ where $X_j = \xi_j - p$ with $\xi_j \sim \text{BER}(p)$. The random variable X_j takes values in the interval from $-p$ to $1-p$, which has length 1. From Theorem <7>, $\mathbb{P}\{S - np \ge r\} \le \exp(-2r^2/n)$ for $r \ge 0$. The bound for the lower tail is the same. Compare with the tail bounds from §3.7:

<12> $$\mathbb{P}\{S \ge np + r\} \le \exp\left(-\frac{r^2}{2npq}g_{n,p}(r)\right) \qquad \text{for } 0 \le r \le nq$$

where $q = 1-p$ and, for $r \ge 0$,

$$\begin{aligned} g_{n,p}(r) &:= q\psi_{\text{benn}}\left(\frac{r}{np}\right) + p\psi_{\text{benn}}\left(\frac{-r}{nq}\right) \\ &\ge \psi_{\text{benn}}\left(q\frac{r}{np} + p\frac{-r}{nq}\right) \qquad \text{by convexity of } \psi_{\text{benn}} \\ &= \psi_{\text{benn}}\left(\frac{(q-p)r}{npq}\right) \gtreqless 1 \text{ if } p \gtreqless 1/2. \end{aligned}$$

Inequality <12>, which is derived from a minimization with the true mgf, cannot be worse than the bound from Theorem <7>, which carried out an analogous minimization that started from an upper bound for the same mgf. For p equal to $1/2$ and r much smaller than n the bound from Theorem <7> is comparable to the bound given by <12>; for $p \neq 1/2$ the $2r^2/n$ is markedly inferior to the $r^2/(2npq)$ (particularly so if $p > 1/2$, which ensures that $g_{n,p}(r) \geq 1$). □

The previous Example highlights a weakness of Theorem <7> when the distribution of a bounded random variable concentrates most of its probability a long way from the endpoints of its support. The squared range times $1/4$ can then be a very poor substitute for a variance term. Chapter 8 will show how better control is possible when the actual variances enter a tail bound.

7.4 From independence to martingales

Theorem <7> has a martingale analog, which has proved particularly useful for the development of concentration inequalities. The result is often attributed to Azuma (1967), even though Hoeffding (1963, page 18) had already noted that, if $S_m = X_1 + \cdots + X_m$ for $1 \leq m \leq n$, then

> "..., the inequalities of Theorems 1 and 2 remain true if the assumption that $X_1, X_2, \ldots, X_n$ are independent is replaced by the weaker assumption that the sequence $S'_m = S_m - ES_m$, $m = 1, 2, \ldots, n$, is a martingale, that is,
>
> $$E(S'_m \mid S'_1, \ldots, S'_j) = S'_j, \qquad 1 \leq j \leq m \leq n, \quad \text{(H2.18)}$$
>
> with probability one. ... [H2.18] implies that the conditional mean of X_m for S'_{m-1} fixed is equal to its unconditional mean. A slight modification of the proofs of Theorems 1 and 2 yields the stated result."

He also noted that martingale methods provide stronger maximal inequalities. (I added the 'H' to indicate that (2.18) was Hoeffding's equation number.)

I am not completely sure about what Hoeffding meant regarding the martingale analog of his Theorem 2. I suspect it involved writing $S_n - \mathbb{P}S_n$ as a sum of random variables $D_1 + \cdots + D_n$ with zero conditional means and $a_i \leq D_i \leq b_i$, applying inequality <9> to each D_i then taking conditional expectations.

In spite of this evidence, in the literature the martingale analog of Hoeffding's Theorem 2 is often attributed to Azuma; 'Hoeffding's inequality' usually refers to the result for independent summands described in Section 7.3.

The extension to martingales might appear rather easy. For a martingale $\{(S_i, \mathcal{F}_i) : 0 \le i \le n\}$ we would have $S_0 = \mathbb{P}(S_n \mid \mathcal{F}_0) = \mathbb{P}S_n$ if $\mathcal{F}_0 = \{\emptyset, \Omega\}$, the trivial sigma-field. We could even replace S_i by $S_i - \mathbb{P}S_n$ and then forget about S_0 altogether, working directly with the martingale differences $D_1, \ldots, D_n$ for which $\mathbb{P}(D_i \mid \mathcal{F}_{i-1}) = 0$ almost surely and $S_i = D_1 + \cdots + D_i$ for $1 \le i \le n$. We would also have $\mathcal{F}_i = \sigma\{D_1, \ldots, XD_i\}$ with $\mathcal{F}_0$ again the trivial sigma-field. We would need an assumption like

<13> $$\mathbb{P}\{a_i \le D_i \le b_i \mid \mathcal{F}_{i-1}\} =_{a.s} 1,$$

for possibly random a_i and b_i that depend only on $\{D_j : j \le i-1\}$. One could then hope that the methods from the Proof of Theorem <7> carry over to show, for $\lambda \ge 0$, that

<14> $$\mathbb{P}\left(e^{\lambda D_i} \mid \mathcal{F}_{i-1}\right) \le_{a.s.} \exp\left(\lambda^2 c_i(\omega)^2/2\right) \qquad \text{if } c_i \ge (b_i - a_i)/2.$$

Unfortunately these seemingly straightforward calculations conceal some delicate measurabilty issues if we adopt the usual KOLMOGOROV interpretation for $\mathbb{P}(\cdots \mid \mathcal{F}_{i-1})$.

A message from the measure theory police:
You could ignore this message at the risk of violating the usual laws of measure theoretic probability.

With the KOLMOGOROV*-style interpretations of conditional expectations, assertions <13> and <14> mean*

<15> $$\mathbb{P}g(D_1, \ldots, D_{i-1})\{a_i \le D_i \le b_i\} = \mathbb{P}g(D_1, \ldots, D_{i-1}),$$

<16> $$\mathbb{P}g(D_1, \ldots, D_{i-1})e^{\lambda D_i} \le \mathbb{P}g(D_1, \ldots, D_{i-1})e^{\lambda^2 c_i^2/2},$$

at least for bounded, non-negative, measurable functions g. The a.s. subscripts in <13> and <14> reflect the fact that <15> and <16> determine the random variables $\mathbb{P}\{a_i \le D_i \le b_i \mid \mathcal{F}_{i-1}\}$ and $\mathbb{P}\left(e^{\lambda D_i} \mid \mathcal{F}_{i-1}\right)$ only up to almost sure equivalences. In particular, the second of these conditional expectations could be changed without warning on a negligible set that depends on λ, a possibility that bodes ill for any attempt to differentiate with respect to λ.

Equalities <15> and <16> also seem to require a_i, b_i and c_i to be measurable functions of their arguments, $D_1, \ldots, D_{i-1}$. Unfortunately, such an assumption would create measurability complications for the setting that will be described in Example <19>.

When delicate measurability issues arise it is often wise to follow the example of careful authors (such as Doob and Dudley) and work with an explicit representation for the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In this case, we could choose $\Omega = \mathbb{R}^n$ with generic member $\omega = (\omega_1, \ldots, \omega_n)$ and $D_i(\omega) = \omega_i$, the coordinate map. The probability measure $\mathbb{P}$ corresponds to the joint distribution of $D_1, \ldots, D_n$ and $\mathcal{F}_i$ is the sigma-field

generated by $\omega_1, \ldots, \omega_i$. Also with this notation, the vector $(D_1, \ldots, D_i)$ could be written as $\omega[1{:}i]$ and the vector $(D_{i+1}, \ldots, D_n)$ as $\omega[i+1{:}n]$. The great advantage of such a representation is that $\mathbb{P}$ can be built up from an initial distibution P_1 for D_1 and a set of conditional distributions $P_{\omega[1:i-1]}$ for $\mathbb{P}\{D_i \in \cdot \mid D_1, \ldots, D_{i-1})$. That is, at least for bounded, measurable functions $g : \mathbb{R}^n \to \mathbb{R}$, we have

$$\mathbb{P}g(D_1, \ldots, D_n) = P_1^{\omega_1} P_{\omega_1}^{\omega_2} \ldots P_{\omega[1:n-1]}^{\omega_n} g(\omega),$$

$$\mathbb{P}\left(g(D_1, \ldots, D_n) \mid \mathcal{F}_i\right) =_{a.s.} g_i(\omega[1{:}i]) := \mathbb{Q}_{\omega[1:i]}^{\omega[i+1:n]} g(\omega),$$

where, symbolically, $\mathbb{Q}_{\omega[1:i]} := P_{\omega[1:i]}^{\omega_{i+1}} \ldots P_{\omega[1:n-1]}^{\omega_n}$. The $P_{\omega[\ldots]}$'s are often called ***regular conditional probability distributions*** *(Breiman, 1968, Section 4.3) or sometimes* ***disintegrations*** *(UGMTP §5.3).*

Equalities <15> and <16> can now be rewritten as:

$P_{\omega[1:i-1]}$ concentrates on the interval $[a_i, b_i]$,
possibly with a_i and b_i both functions of $\omega[1{:}i-1]$;

<17> $P_{\omega[1:i-1]} e^{\lambda \omega_i} \le \exp\left(\lambda^2 (b_i - a_i)^2/8\right) \le \exp\left(\lambda^2 c_i^2/2\right)$,
where c_i could depend on $\omega[1{:}i-1]$ and $c_i \ge (b_i - a_i)/2$,

without almost sure caveats. The inequality <17> could be proved by either of the methods described in Section 7.3. Neither equation relies on an assumption that a_i, b_i, c_i depend on $\omega[1{:}i-1]$ in a measurable way, because the averaging is carried out separately for each fixed $\omega[1{:}i-1]$, a small but significant improvement over the KOLMOGOROV*-style interpretation.*

□ *End of message.*

If a_i and b_i are actually constants then the method used to prove Theorem <7> carries over with minimal changes. For martingale differences $\{D_i\}$ with truly random c_i's a new complication arises. We now need c_i (but not a_i and b_i) to be a measurable function of $\omega[1{:}i-1]$. That is, we need c_i to be $\mathcal{F}_{i-1}$-measurable, so that factors can be pulled out one at a time by successive conditioning arguments. Start with the D_n contribution:

$$\mathbb{P}\left(e^{\lambda S_n} \mid \mathcal{F}_{n-1}\right) = e^{\lambda S_{n-1}} \mathbb{P}\left(e^{\lambda D_n} \mid \mathcal{F}_{n-1}\right) \le e^{\lambda S_{n-1}} e^{\lambda^2 c_n^2/2} \qquad \text{a.s.} \quad .$$

The $\exp(\lambda^2 c_n^2/2)$ would present an obstacle if we now tried to condition on $\mathcal{F}_{n-2}$. The obstacle can be removed by an old martingale trick: absorb the $\mathcal{F}_{n-1}$-measurabile factor involving c_n into the left-hand side:

$$\mathbb{P}\left(e^{\lambda S_n - \lambda^2 c_n^2/2} \mid \mathcal{F}_{n-1}\right) \le e^{\lambda S_{n-1}} \qquad \text{a.s.} \quad .$$

A further conditioning on $\mathcal{F}_{n-2}$ would then throw out an $\exp\left(\lambda^2 c_{n-1}^2/2\right)$ factor. Of course this new obstacle could also be removed by also absorbing the c_{n-1} contribution into the left-hand side. And so on. This trick was much used in the literature on martingale central limit theorems. It is the idea behind the Proof of the next Theorem.

<18> **Theorem.** *Let $((D_i, \mathcal{F}_i) : i \in [\![n]\!])$ be a martingale difference sequence, with $\mathbb{P}(D_i \mid \mathcal{F}_{i-1}) =_{a.s.} 0$, with $\mathcal{F}_0$ the trivial sigma-field $\{\emptyset, \Omega\}$. Define $S_i := \sum_{j=1}^{i} D_j$ for $i \in [\![n]\!]$.*

(i) Suppose the conditional distribution of D_i given $\mathcal{F}_{i-1}$ concentrates on an interval $[a_i, b_i]$, with both a_i and b_i being functions of $D_1, \ldots, D_{i-1}$.

(ii) Let c_i be any $\mathcal{F}_{i-1}$-measurable upper bound for $(b_i - a_i)/2$. Define $V_0 = 0$ and $V_i := \sum_{j=1}^{i} c_j^2$ with for $i \in [\![n]\!]$.

Then for each positive constant σ^2,

$$\mathbb{P}\{S_n \geq r, V_n \leq \sigma^2\} \leq e^{-r^2/(2\sigma^2)} \qquad \text{for } r \geq 0.$$

□ *Consequently, $\mathbb{P}\{S_n \geq r\} \leq e^{-\lambda^2 (r/\sigma)^2/2} + \mathbb{P}\{V_n > \sigma^2\}$ for $r \geq 0$.*

Proof. The second inequality comes from adding $\mathbb{P}\{S_n \geq r, V_n > \sigma^2\}$ to both sides of the first inequality.

For the main argument define $W_0 := e^{\lambda^2\sigma^2/2}$ and, for $j \in [\![n]\!]$,

$$W_j := \exp\left(\lambda S_j + \lambda^2(\sigma^2 - V_j)/2\right) = W_{j-1} \exp\left(\lambda D_j - \lambda^2 c_j^2/2\right).$$

Take conditional expectations then invoke inequality <14> to deduce

$$\mathbb{P}\left(W_j \mid \mathcal{F}_{j-1}\right) = W_{j-1} e^{-\lambda^2 c_j^2/2} \mathbb{P}\left(e^{\lambda D_j} \mid \mathcal{F}_{j-1}\right) \leq W_{j-1},$$

which implies $\mathbb{P}W_j \leq \mathbb{P}W_{j-1}$. For each positive λ we have a pointwise inequality,

$$\{S_n \geq r, V_n \leq \sigma^2\} \leq \exp\left(\lambda(S_n - r) + \lambda^2(\sigma^2 - V_n)/2\right) = e^{-\lambda r} W_n.$$

Take expected values to deduce

$$\mathbb{P}\{S_n \geq r, V_n \leq \sigma^2\} \leq e^{-\lambda r} \mathbb{P}W_n \leq e^{-\lambda r} \mathbb{P}W_0 = e^{-\lambda r + \lambda^2\sigma^2/2} \qquad \text{for } \lambda \geq 0.$$

The usual subgaussian minimization argument then delivers the asserted tail bound. □

<19> **Example.** Suppose $\xi := (\xi_1, \ldots, \xi_n)$ is a vector of independent random quantities, each ξ_i taking values in some measurable space $(\mathcal{X}_i, \mathcal{A}_i)$, with ξ_i having distribution P_i. To avoid awkward measurability issues, assume ξ_i to be the ith coordinate map on $\mathcal{X} := \prod_i \mathcal{X}_i$, equipped with product measure $\mathbb{P} := \otimes_i P_i$ on its product sigma-field $\mathcal{F} := \otimes_i \mathcal{A}_i$. Thus $\xi_i(x) := x_i$ for $x := (x_1, \ldots, x_n) \in \mathcal{X}$.

Suppose f is a bounded, product-measurable, real-valued function on $\mathcal{X}$ with the ***bounded difference*** property: for constants c_i,

$$|f(y) - f(z)| \leq c_i \qquad \text{if } y, z \in \mathcal{X} \text{ differ only in the } i\text{th coordinate.}$$

More succinctly,

<20> $$|f(y_1,\ldots,y_n) - f(z_1,\ldots,z_n)| \le \sum\nolimits_{i\le n} c_i\{y_i \ne z_i\} \qquad \text{for all } y,z \in \mathcal{X}.$$

Let $\mathcal{F}_i$ denote the sigma-field generated by $\xi_1,\ldots,\xi_i$, with $\mathcal{F}_0 : \{\emptyset,\mathcal{X}\}$. Define a martingale $\{(S_i,\mathcal{F}_i) : 0 \le i \le n\}$ by

$$S_0 := 0 \quad \text{AND} \quad S_i := \mathbb{P}\left(f(\xi) \mid \mathcal{F}_i\right) - \mathbb{P}f(\xi) \qquad \text{for } i \ge 1.$$

An appeal to Theorem <18> with $\sigma^2 = \sum_{i=1}^n c_i^2$ will give the subgaussian bound

<21> $$\mathbb{P}\{f(\xi) \ge r + \mathbb{P}f(\xi)\} \le \exp\left(-r^2/2\sigma^2\right) \qquad \text{for } r \ge 0.$$

Remark. Inequality <20> is sometimes referred to as a LIPSCHITZ condition (for a weighted HAMMING distance). Inequality <21> is often referred to as the MCDIARMID inequality.

To calculate a clean version of $\mathbb{P}\left(f(\xi) \mid \mathcal{F}_i\right)$ we have only to average out with respect to $\mathbb{Q}_{i+1} = \otimes_{i<j\le n} P_j$:

$$\begin{aligned}\mathbb{P}\left(f(\xi) \mid \mathcal{F}_i\right) = f_i(\xi_1,\ldots,\xi_i) &:= \mathbb{Q}_{i+1}^z f(\xi_1,\ldots,\xi_i,z_{i+1},\ldots,z_n)\\ &= P_{i+1}^{z_{i+1}}\ldots P_n^{z_n} f(\xi_1,\ldots,\xi_i,z_{i+1},\ldots,z_n).\end{aligned}$$

By construction, $f(\xi) - \mathbb{P}f(\xi) = S_n = \sum_{j=1}^n D_j$, a sum of martingale differences $D_i := S_i - S_{i-1}$ for $i \in [\![n]\!]$. The increment can also be written as

$$\begin{aligned}D_i &= f_i(\xi_1,\ldots,\xi_i) - f_{i-1}(\xi_1,\ldots,\xi_{i-1}) = f_i(x_1,\ldots,x_i) - f_{i-1}(x_1,\ldots,x_{i-1})\\ &= \mathbb{Q}_{i+1}^{z[i+1:n]} f(x_1,\ldots,x_i,z_{i+1},\ldots,z_n) - \mathbb{Q}_i^{z[i:n]} f(x_1,\ldots,x_{i-1},z_i,\ldots,z_n)\\ &= \mathbb{Q}_{i+1}P_i^{z_i}\Big[f(x_1,\ldots,x_i,z_{i+1},\ldots,z_n) - f(x_1,\ldots,x_{i-1},z_i,z_{i+1},\ldots,z_n)\Big].\end{aligned}$$

The extra $P_i^{z_i}$ has no effect on the x_i. We could deduce from this equality that $|D_i| \le c_i$, which suggests we could choose $a_i = -c_i$ and $b_i = c_i$. However, as pointed out by McDiarmid (1989, page 159, after Lemma 4.1), this choice would contribute an extra, unwanted factor of $1/4$ in the exponent of <21>. He took a better approach, based on a very clever insight.

First note that f_i inherits a bounded difference property from f:

$$\begin{aligned}&|f_i(y_1,\ldots,y_i) - f_i(w_1,\ldots,w_i)|\\ &\quad\le \mathbb{Q}_i|f(y_1,\ldots,y_i,z_{i+1},\ldots,z_n) - f(w_1,\ldots,w_i,z_{i+1},\ldots,z_n)|\\ &\quad\le \sum\nolimits_{j\le i} c_j\{y_j \ne w_j\} \qquad \text{via } <20>.\end{aligned}$$

This inequality implies $a_i \le D_i \le b_i$ where

$$a_i = \inf_{s\in\mathcal{X}_i} f_i(\xi_1,\ldots,\xi_{i-1},s) - f_{i-1}(\xi_1,\ldots,\xi_{i-1}),$$
$$b_i = \sup_{t\in\mathcal{X}_i} f_i(\xi_1,\ldots,\xi_{i-1},t) - f_{i-1}(\xi_1,\ldots,\xi_{i-1}).$$

As $\xi_1,\ldots,\xi_{i-1}$ are regarded as fixed for current purposes, the f_{i-1} term has no effect on the length of $[a_i,b_i]$:

$$\begin{aligned} 0 \le b_i - a_i &= \sup_{t\in\mathcal{X}_i} f_j(\xi_1,\ldots,\xi_{j-1},t) - \inf_{s\in\mathcal{X}_i} f_j(\xi_1,\ldots,\xi_{j-1},s) \\ &= \sup_{s,t\in\mathcal{X}_i} |f_i(\xi_1,\ldots,\xi_{i-1},t) - f_j(\xi_1,\ldots,\xi_{j-1},s)|, \end{aligned}$$

□ which is $\le c_i$ by the bounded difference inequality for f_i. And so on.

Remark. If you worry about the effect of a taking a supremum over a possibly uncountably infinite set $\mathcal{X}_i$ you should note the lack of any measurabilty assumptions about a_i and b_i in Theorem <18>; it is only c_i that needs to be $\mathcal{F}_{i-1}$-measurable. You might also want to read the message from the measure theory police.

McDiarmid (1989, page 168, proof of (6.10)) was obviously aware of the measurability issues. I am not quite sure exactly what he meant by $\operatorname{ess\,inf}[D_k \mid \mathcal{F}_{k-1}]$, but it is clearly intended to avoid a non-measurable supremum. McDiarmid is a careful author.

At this point you are probably expecting to witness a triumphal procession of applications of Theorems <7> and <18> to problems in combinatorics and graph theory. I will give only two examples. The first (bin packing) is quite straightforward. I include it here because it was one of the first examples that convinced me concentration is a general topic worth careful study. The second example (in Section 7.5) will emphasize the limitations of these methods. For more uplifting applications you should peruse the beautiful books by Steele (1997) and Alon and Spencer (2000), or read McDiarmid (1989, 1998), or even just google "Hoeffding", "Azuma", "bounded difference", or "McDiarmid inequality". See Boucheron, Lugosi, and Massart (2013, Chap 6), especially the Notes to the chapter, for some of the generalizations of the method of bounded differences.

<22> **Example.** Suppose objects of independent random sizes $\xi_1,\ldots,\xi_n$, each with distribution P concentrated on $[0,1]$, are packed into bins of capacity 1. No objects may be split and the contents of no bin is allowed to exceed capacity. Define $f(\xi_1,\ldots,\xi_n)$ to be the smallest number of bins needed to contain all the objects.

The function f has the bounded difference property on $[0,1]^n$, with $c_i = 1$ for each i. To see why, suppose vectors y and z differ only in the ith coordinate. Every packing for y can then be converted to a packing for z with one more bin by discarding y_i and putting z_i into a new bin. It follows that $f(z) \le 1 + f(y)$. Reverse the roles of y and z to get the companion inequality $f(y) \le 1 + f(z)$.

□

Remark. Inequality <21> is an example of a concentration inequality for suitable functions of independent variables. That area of research unwent an amazing expansion in the late 1990's, largely due to the contributions of Michel Talagrand. See Talagrand (1996) and Ledoux (2001) for most readable accounts.

7.5 A cautionary example

The following example was mentioned by Kim and Vu (2000) and Vu (2002) as motivation for developing a more general inequality (see §8.4) that is analogous to a BENNETT inequality for martingales.

<23> **Example.** (random graphs) Consider a graph with vertex set $[[n]]$ whose edges are a subset of the set $\mathcal{E}$ of all $|\mathcal{E}| := \binom{n}{2}$ pairs of distinct vertices. Two edges are said to be adjacent if they share an endpoint. Three distinct edges form a triangle if together they contain only three vertices: that is, the edges are $\{i,j\}$, $\{j,k\}$, and $\{k,i\}$ for distinct vertices i,j,k. Write $\mathcal{T}$ for the set of all triangles, subsets of $\mathcal{E}$ of size 3 that involve exactly 3 vertices. The set $\mathcal{T}$ has size $\binom{n}{3}$.

The random graph $G_n(p)$, which oftens carries the names of Erdös and Rényi, chooses its edges by means of a set of independent random variables $\{\xi_e : e \in \mathcal{E}\}$, with each ξ_e distributed $\text{BER}(p)$. That is, $G_n(p)$ includes e when $\xi_e = 1$, an event with probability p.

Remark. Actually Erdös and Rényi (1960) took the number of edges to be a nonrandom number N. They considered the graph $\Gamma_{n,N}$ whose edge set is a sample of size N selected without replacement from $\mathcal{E}$. As explained by Janson, Łuczak, and Ruciński (2000, Section 1.4), the random graph $G_n(p)$ behaves asymptotically very like $\Gamma_{n,N}$ with $N = \binom{n}{2}p$.

The number of triangles in $G_n(p)$ equals

$$f(\xi) := \sum\nolimits_{\{e_1,e_2,e_3\}\in\mathcal{T}} \xi_{e_1}\xi_{e_2}\xi_{e_3}.$$

The expected number of triangles is $\theta := \mathbb{P}f(\xi) = \binom{n}{3}p^3$.

To fit this example into the bounded difference setting of Example <19> we need to enumerate $\mathcal{E}$ as a sequence $(e_j : 1 \le j \le |\mathcal{E}|)$. I don't know of any particularly clever method for enumerating that can improve the concentration bound for $f(\xi)$.

A change in a single y_e can have a big effect on $f(y)$. For example, if $y_e = 1$ for all $e \in \mathcal{E}$ and just one y_e is changed to zero then $n-2$ triangles disappear. The constants corresponding to the c_i's in inequality <20> must all be equal to $n-2$. The two-sided analog of inequality <21> gives

<24> $$\mathbb{P}\{|f(\xi) - \mathbb{P}f(\xi)| \ge r\} \le 2\exp\left(-\frac{r^2}{2|\mathcal{E}|(n-2)^2}\right) \qquad \text{for } r \ge 0.$$

We would need an $r = r_n$ with $r_n/n^2 \to \infty$ to send the tail bound to zero as $n \to \infty$.

Better bounds are available. For example, arguments based on the CHEN-STEIN method show that $f(\xi)$ is approximately $P_\theta = \text{POISSON}(\theta)$ distributed. More precisely, as shown by Barbour, Holst, and Janson (1992, Section 5.1),

$$\sup_A |\mathbb{P}\{f(\xi) \in A\} - P_\theta A| = O(np^2) \qquad \text{if } np \to \infty.$$

In particular, the choice $A = \{i \in \mathbb{N}_0 : |j - \theta| > r\}$ and the Poisson tail bound from §3.5,

$$P_\theta A \le 2 \exp\left(-\frac{r^2}{2\theta}\psi_{\text{benn}}(r/\theta)\right) = 2\exp(-\theta \mathbb{h}(r/\theta)),$$

gives

$$\mathbb{P}\{|f(\xi) - \theta| > s_n \theta^{1/2}\} \to 0$$

if, for example, $s_n \to \infty$ and $s_n/\theta^{1/2} \to 0$ and $np^2 \to 0$.

Remark. The appeal to CHEN-STEIN has the drawback that it imposes the extraneous requirement that $f(\xi)$ have an approximate POISSON distribution. My only point in invoking CHEN-STEIN is to make sure you realize that <24> is very crude. It is possible to have concentration without an approximate POISSON distribution.

Why has inequality <24> fallen so far short of the CHEN-STEIN result? The blame lies with the pessimistic choice $c_i = n - 2$ forced by the unlikely configuration y with $y_e = 1$ all $e \in \mathcal{E}$. It is much more likely that a possible edge, say $e = \{1, 2\}$, is involved in only about np^2 potential triangles: the expected number of vertices $k \ge 3$ for which $\xi_{\{1,k\}} = 1 = \xi_{\{2,k\}}$ is equal to $(n-2)p^2$. A change in ξ_e probably changes $f(\xi)$ by an amount of order np^2. If by some modification of the argument we could reduce the effective c_i to a term of order np^2 then the exponent in the bound <24> would be greatly improved. Roughly speaking, this is the idea behind the KIM-VU method, which will be discussed in §8.4. To be continued. □

7.6 Problems

[1] Suppose X is a random variable whose moment generating function $M_X(\lambda)$ is finite in a neighborhood of the origin. Suppose also that there exist constants ν and $\tau > 0$ for which $M_X(\lambda) \le \exp(\nu\lambda + \frac{1}{2}\tau^2\lambda^2)$ for all $\lambda \ge 0$, but not necessarily for all real λ.

(i) Show that $\mathbb{P}\{X \ge \nu + \tau t\} \le e^{-t^2/2}$ for all $t \ge 0$.

(ii) Use the Taylor expansions for λ near zero,

$$\begin{aligned} M_X(\lambda) &= 1 + \lambda \mathbb{P}X + o(\lambda), \\ \exp(\nu\lambda + \tfrac{1}{2}\sigma^2\lambda^2) &= 1 + \lambda\nu + o(\lambda), \\ \mathbb{P}e^{\lambda(X-\nu)} &= 1 + \lambda(\mathbb{P}X - \nu) + \tfrac{1}{2}\lambda^2 \mathbb{P}(X-\nu)^2 + o(\lambda^2), \\ \exp(\tfrac{1}{2}\sigma^2\lambda^2) &= 1 + \tfrac{1}{2}(\tau^2)\lambda^2 + o(\lambda^2), \end{aligned}$$

to deduce that $\mathbb{P}X \le \nu$ and $\operatorname{var}(X) \le \mathbb{P}(X-\nu)^2 \le \tau^2$. As Problem [2] shows, τ^2 might need to be strictly larger than the variance.

(iii) If $M_X(\lambda) \le \exp(\nu\lambda + \frac{1}{2}\tau^2\lambda^2)$ for all real λ, show that $\mathbb{P}X = \nu$.

[2] Theorem <6> suggests that $\|X\|_2$ might be comparable to $\beta(X)$ if X has a subgaussian distribution. It is easy to deduce from inequality <3> that $\|X\|_2 \le 2\beta(X)$. Show that there is no companion inequality in the other direction by considering the bounded, symmetric random variable X for which $\mathbb{P}\{X = \pm M\} = \delta$ and $\mathbb{P}\{X = 0\} = 1 - 2\delta$. If $2\delta = M$ show that $\mathbb{P}X^2 = 1$ but

$$\log\left(1 + \lambda^2/2! + (\lambda^4 M^2)/4!\right) \le \log \mathbb{P}e^{X\lambda} \le \tau^2\lambda^2/2 \qquad \text{for all real } \lambda$$

would force $\tau^2 \ge 2\log\left(3/2 + M^2/24\right)$.

[3] Show that a random variable X that has either of the following two properties is subgaussian.

(i) For some positive constants c_1, c_2, and c_3,

$$\mathbb{P}\{|X| \ge x\} \le c_1 \exp(-c_2 x^2) \qquad \text{for all } |x| \ge c_3.$$

(ii) For some positive constants c_1, c_2, and c_3,

$$\mathbb{P}e^{\lambda X} \le c_1 \exp(c_2\lambda^2) \qquad \text{for all } |\lambda| \ge c_3$$

[4] Suppose $\mathbb{P}\{X = -1/2\} = 2/3$ and $\mathbb{P}\{X = 1\} = 1/3$. Show that $X \in \text{SUBG}(1)$. Also show that $\mathbb{P}X = 0$ and $\mathbb{P}X^2 = 1/2 < 1$. You might also find it informative to compare the higher order moments of X with the correponding moments of the $N(0,1)$.

[5] Suppose $\{X_n\}$ is a sequence of random variables for which

$$\mathbb{P}\exp(\lambda X_n) \le \exp(\tau^2\lambda^2/2) \qquad \text{for all } n \text{ and all real } \lambda,$$

where τ is a fixed positive constant. If $X_n \to X$ almost surely, show that $\mathbb{P}e^{\lambda X} \le \exp(\tau^2\lambda^2/2)$ for all real λ.

7.7 Notes

The essentials of the characterizations for subgaussians in Theorem <6> come from Kahane (1968, Exercise 6.10). He did not cite any source. Kahane (1960, pages 4-5) proved similar results. For that paper, it is not clear to me whether Kahane was merely reminding readers of well known facts or whether he was introducing subgaussianity as a new concept.

Apparently Azuma (1967) was unaware of Hoeffding's paper, although he did cite a 1966 paper of Chow, which in turn cited a 1963 paper of Kahane for the properties of "semi-gaussian" random variables.

The proof of Theorem <18> borrows ideas from Freedman (1975).

References

Alon, N. and J. H. Spencer (2000). *The Probabilistic Method* (second ed.). Wiley.

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tôhoku Mathematical Journal 19*(3), 357–367.

Barbour, A. D., L. Holst, and S. Janson (1992). *Poisson Approximation*. Oxford University Press.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Breiman, L. (1968). *Probability*. Reading, Massachusets: Addison-Wesley.

Erdös, P. and A. Rényi (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci 5*(1), 17–60.

Freedman, D. A. (1975). On tail probabilities for martingales. *Annals of Probability 3*(1), 100–118.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*, 13–30.

Janson, S., T. Łuczak, and A. Ruciński (2000). *Random Graphs*. Wiley.

Kahane, J.-P. (1960). Propriétés locales des fonctions à séries de Fourier aléatoires. *Studia Mathematica 19*(1), 1–25.

Kahane, J.-P. (1968). *Some Random Series of Functions*. Heath. Second edition: Cambridge University Press 1985.

Kim, J. H. and V. H. Vu (2000). Concentration of multivariate polynomials and its applications. *Combinatorica 20*(3), 417–434.

Ledoux, M. (2001). *The Concentration of Measure Phenomenon*, Volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society.

McDiarmid, C. (1989). On the method of bounded differences. In J. Siemons (Ed.), *Surveys in Combinatorics*, Volume 141 of *London Mathematical Society Lecture Notes*, pp. 148–188. Cambridge University Press.

McDiarmid, C. (1998). Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsen, and B. Reed (Eds.), *Probabilistic Methods for Algorithmic Discrete Mathematics*, pp. 195–248. Springer-Verlag.

Pollard, D. (2001). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.

Rademacher, H. (1922). Einige Sätze über Reihen von allgemeinen Orthogonalfunktionen. *Mathematische Annalen 87*(1), 112–138.

Steele, J. M. (1997). *Probability Theory and Combinatorial Optimization*. SIAM.

Talagrand, M. (1996). A new look at independence. *Annals of Probability 24*, 1–34.

Talagrand, M. (2021). *Upper and Lower Bounds for Stochastic Processes: Decomposition Theorems* (Second ed.), Volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag.

Vu, V. H. (2002). Concentration of non-Lipschitz functions and applications. *Random Structures and Algorithms 20*, 262–316.

Chapter 8
Between subgaussian and subexponential

SECTION 8.1 *discusses the* BENNETT *inequality for sums of independent random variables that are bounded above by a constant.*

SECTION 8.2 *illustrates the ideas from the previous Section by deriving a subgaussian/subexponential tail bound for quadratic forms in independent subgaussian random variables. The result is often called the* HANSON-WRIGHT *inequality.*

SECTION 8.3 *derives a* BENNETT *inequality for sums of martingale differences that are bounded above by a constant with high probability.*

*SECTION 8.4 *applies the martingale* BENNETT *inequality to derive a version of a* KIM-VU *inequality for polynomials in independent, bounded random variables.*

8.1 A BENNETT inequality for independent summands

As you saw in Chapter 7, the HOEFFDING inequalities provide exponential (SUBGAUSSIAN) tail bounds by means of a crude upper bound, $(b-a)^2/4$, on the variance (or conditional variance) for a random variable taking values in an interval $[a, b]$. This Chapter improves those inequalities by bringing second moment quantities into the bound. The resulting inequalities will mostly be of the form

$$\mathbb{P}\{X - \mathbb{P}X \geq r\} \leq \exp(-\Lambda(r)) \qquad \text{for } r \geq 0,$$

where Λ is an increasing function that behaves like $c_1 r^2$ for $r \approx 0$ and no worse than $c_2 r$ for large r, for positive constants c_1 and c_2.

You have already seen some instances of such behavior in Chapter 3: in §3.5, if $X \sim \text{POISSON}(\theta)$ then $\Lambda(r) = (r^2/2\theta)\psi_{\text{benn}}(r/\theta)$; and in §3.6, a BERNSTEIN inequality (a SUBGAMMA tail bound) for sums of independent random variables, $X = W_1 + \cdots + W_n$, with $\mathcal{V} = \sum_i \mathbb{P}W_i^2$ and satisfying a condition

<1> $$\sum_i \mathbb{P}(W_i^+/b)^k \leq \tfrac{1}{2}k!\mathcal{V}/b^2 \qquad \text{for } k \geq 3 \text{ and some constant } b.$$

In the BERNSTEIN case $\Lambda(r) \geq r^2/(2\mathcal{V} + 2br)$ for $r \geq 0$.

Remark. Note that condition <1> holds if $b \geq 0$ and $W_i \leq b$ for all i.

Bennett (1962) focussed mostly on improvements to the BERNSTEIN inequality for bounded random variables. He proved an inequality similar in form to the POISSON bound.

Bennett's argument can be simplified by exploiting properties of the FENCHEL-LEGENDRE (§2.4) dual pair of convex functions,

<2> $$\hbar(y) = \mathbb{f}^*(y) := \sup_{x\in\mathbb{R}} (yx - \mathbb{f}(x)) \qquad \text{for } y \in \mathbb{R},$$

where

<3> $$\mathbb{f}(x) := e^x - 1 - x = \tfrac{1}{2}x^2 \mathbb{A}(x),$$

<4> $$\hbar(y) := (1+y)\log(1+y) - y = \tfrac{1}{2}y^2 \psi_{\text{benn}}(y) \qquad \text{for } y \geq -1.$$

The value $\hbar(y)$ equals $+\infty$ for $y < -1$. The functions $\mathbb{A}$ and ψ_{benn} have integral representations (§ 2.2),

<5> $$\mathbb{A}(x) := \int_0^1 2(1-s)e^{xs}\, ds \qquad \text{for } x \in \mathbb{R},$$

<6> $$\psi_{\text{benn}}(y) := \int_0^1 \frac{2(1-s)}{(1+ys)}\, ds \qquad \text{for } y \geq -1,$$

which show that $\mathbb{A}$ is continuous, convex, and strictly increasing and ψ_{benn} is continuous, convex, and strictly decreasing on $[-1, \infty)$.

The function $\mathbb{f}$ enters the MGF bound for $\lambda \geq 0$ via the rearrangement

$$\begin{aligned} e^{\lambda W} &= 1 + \lambda W + \mathbb{f}(\lambda W) = 1 + \lambda W + \tfrac{1}{2}(\lambda W)^2 \mathbb{A}(\lambda W) \\ &\leq 1 + \lambda W + \tfrac{1}{2}(\lambda W)^2 \mathbb{A}(\lambda b) \qquad \text{when } W \leq b. \end{aligned}$$

If $\mathbb{P}W^2 = v < \infty$ and $\mathbb{P}\{W \leq b\} = 1$ for some constant b then

$$\begin{aligned} \mathbb{P}e^{\lambda W} &\leq 1 + \lambda \mathbb{P}W + \tfrac{1}{2}\lambda^2 v \mathbb{A}(\lambda b) \qquad \text{for } \lambda \geq 0 \\ &\leq \exp\left(\lambda \mathbb{P}W + \tfrac{1}{2}\lambda^2 v \mathbb{A}(\lambda b)\right) = \exp\left(\lambda \mathbb{P}W + v\mathbb{f}(\lambda b)/b^2\right), \end{aligned} \qquad \text{<7>}$$

the final equality assuming $b \neq 0$.

Remark. For the next Theorem it is important that summands X_i all share the same upper bound, $X_i \leq b$ for all i. If we were to subtract off the means the $\mathbb{P}X_i^2$'s would be replaced by true variances but the upper bounds might no longer be equal. We would then have to replace b by $\max_i(b - \mathbb{P}X_i)$. This trade-off has long intrigued me. Some authors require zero means to get true variances; others just put up with second moments. I wonder if there is some optimal compromise.

<8> **Theorem.** *(*BENNETT *inequality for independent summands) Let $X_1, \ldots, X_n$ be independent random variables with $X_i \leq b$, for a positive constant b, and $\mathbb{P}X_i^2 \leq v_i$ for each i. Then*

$$\mathbb{P}\left\{\sum\nolimits_{i\leq n}(X_i - \mathbb{P}X_i) \geq r\right\} \leq \exp\left(-\frac{r^2}{2\mathcal{V}}\psi_{\text{benn}}\left(\frac{br}{\mathcal{V}}\right)\right) \qquad \text{for } r \geq 0.$$

where $\mathcal{V}$ is any constant with $\mathcal{V} \geq \sum_{i\leq n} v_i$.

Remark. The case where $b = 0$, which gives a SUBGAUSSIAN tail, could be obtained as a limit as $b \downarrow 0$ if $X_i \le 0$ for all i.

Proof. Define $S = \sum_{i\le n}(X_i - \mathbb{P}X_i)$ so that multiple appeals to <7> give

$$\mathbb{P}e^{\lambda S} = \prod_{i\le n} \mathbb{P}e^{\lambda(X_i - \mathbb{P}X_i)} \le \prod_{i\le n} \exp\left(v_i \mathbb{f}(\lambda b)/b^2\right) \le \exp\left(\mathcal{V}\mathbb{f}(\lambda b)/b^2\right),$$

from which it follows that

$$\begin{aligned}\mathbb{P}\{S \ge r\} &\le \inf_{\lambda>0} e^{-\lambda r} \exp\left(\mathcal{V}\mathbb{f}(\lambda b)/b^2\right) = \exp\Big(-\frac{\mathcal{V}}{b^2}\sup_{\lambda>0}\Big(\frac{(\lambda b)br}{\mathcal{V}} - \mathbb{f}(\lambda b)\Big)\Big)\\ &= \exp\Big(-\frac{\mathcal{V}}{b^2}\mathbb{h}(br/\mathcal{V})\Big) = \exp\Big(-\frac{r^2}{2\mathcal{V}}\psi_{\text{benn}}(br/\mathcal{V})\Big),\end{aligned}$$

□ as asserted.

<9> **Example.** Consider the case where $X_i = \xi_i - p$ for $\xi_1, \ldots, \xi_n$ independent BER(p)'s and $S = \sum_i X_i$. Note that $\mathbb{P}X_i = 0$ and $\mathbb{P}X_i^2 = pq$ and $X_i \le q = 1 - p$. Theorem <8> gives

<10> $$\mathbb{P}\{S \ge r\} \le \exp\left(-\frac{r^2}{2npq}\psi_{\text{benn}}\left(\frac{r}{np}\right)\right) \qquad \text{for } r \ge 0.$$

Using the fact that $\sum_{i\le n}\xi$ has a BIN(n, p) distributions we also have (§3.7)

<11> $$\mathbb{P}\{S \ge r\} \le \exp\left(-\frac{r^2}{2npq}\left[q\psi_{\text{benn}}\left(\frac{r}{np}\right) + p\psi_{\text{benn}}\left(-\frac{r}{nq}\right)\right]\right) \qquad \text{for } r \ge 0.$$

As <11> comes from a minimization involving the actual MGF it is better than <10>, which starts from an upper bound for that MGF. The superiority can also be seen directly from the fact that a convex combination of two distinct real numbers is greater than the smaller of the two, together with the inequality $\psi_{\text{benn}}(-r/nq) > 1 > \psi_{\text{benn}}(r/np)$ for $r > 0$. The general BENNETT bound does not capture the skewness effect from □ $\mathbb{P}S^3 = npq(q - p)$.

8.2 Quadratics in independent subgaussians

The following result generalizes a theorem proved by Hanson and Wright (1971). See the Notes for some history.

<12> **Theorem.** *Suppose $W := X'AX$ with independent $X_1, \ldots, X_n$, for which $X_i \in$ SUBG(τ^2) for each i, and A an $n \times n$ matrix of real numbers with $A_{i,i} = 0$ for each i. Then*

$$W/\beta \in \text{SUBGAMMA}(\alpha) \qquad \text{for } \beta := 4\tau^2|A|_2 \text{ and } \alpha := \|A\|_{\text{F}}^2 / \|A\|_2^2,$$

where $\|A\|_{\text{F}}^2 := \sum_{i,j} A_{i,j}^2 = \text{trace}(A'A)$ and $\|A\|_2 := \sup\{|Au|_2 : |u|_2 \le 1\}$. Consequently,

□ $$\log \mathbb{P}\{W \ge \beta r\} \le -r^2/(2\alpha + 2r) \qquad \text{for } r \ge 0.$$

A similar result was proved in §3.6: if $Y := \sum_j b_j(Z_j^2 - 1)$ for constants $b_j > 0$ and independent Z_j's, each distributed $N(0,1)$, then

<13> $$\log \mathbb{P}e^{\lambda Y} \le \tfrac{1}{2}|b|_2^2\lambda^2/(1 - 2|b|_\infty\lambda) \qquad \text{for } 0 \le |b|_\infty\lambda < 1.$$

That is, $Y/(2|b|_\infty) \in \text{SUBGAMMA}(\alpha)$ for $\sqrt{2\alpha} = |b|_2/|b|_\infty$. As you will see from the Proof in this Section, $\|A\|_{\text{F}}$ corresponds to $|b|_2$ and $\|A\|_2$ corresponds to $|b|_\infty$.

The proof of <13> used the fact: if $Z_j \sim N(0,1)$ then $\mathbb{P}\exp(\lambda Z_j^2) = (1-2\lambda)^{-1/2}$ for $2\lambda < 1$, which implies (for $2\lambda < 1$)

$$\log \mathbb{P}\exp(\lambda Z_j^2) = -\tfrac{1}{2}\log(1-2\lambda) = \tfrac{1}{2}\left((2\lambda) + (2\lambda)^2/2 + (2\lambda)^3/2 \ldots\right)$$

For Chapter 3 the upper bound

<14> $$\log \mathbb{P}\exp(\lambda(Z_j^2 - 1)) \le -\lambda + \lambda + \lambda^2/(1-2\lambda) \qquad \text{for } 2\lambda < 1$$

was used. In this Section an apparently cruder bound,

<15> $$\log \mathbb{P}\exp(\lambda Z_j^2) \le \tfrac{1}{2}(2\lambda)/(1-2\lambda) = \lambda/(1-2\lambda) \qquad \text{for } 2\lambda < 1,$$

will suffice because it will be used with λ replaced by λ^2

The need for this subtle modification caught me by surprise.

The assumption $A_{ii} = 0$ removes a contribution $W_0 := \sum_i A_{ii}X_i^2$ from $X'AX$, leaving a sum $W := \sum_{i\ne j} A_{i,j}X_iX_j$ with $\mathbb{P}W = 0$. Problem [4] shows how to handle W_0 and Problem [5] shows how to recombine them to get $X'AX$. The zero expected value for W lies behind the switch from <14> and <15>.

Remark. The quantity $\|A\|_{\text{F}} := \left(\sum_{i,j} A_{i,j}^2\right)^{1/2} = (\text{trace}(A'A))^{1/2}$ is called the FROBENIUS (or HILBERT-SCHMIDT) norm of A and $\|A\|_2$ is the norm of A as a linear map from $\mathbb{R}^n$ back into itself: $\|A\|_2 := \sup\{|Aw|_2 : |w|_2 \le 1\}$.

If A has the SVD (singular value decomposition) $A = \sum_{j=1}^n \sigma_j u_j v_j'$, with $\sigma_1 \ge \cdots \ge \sigma_n \ge 0$ and both $\{u_i : i \in [\![n]\!]\}$ and $\{v_j : j \in [\![n]\!]\}$ being orthonormal bases for $\mathbb{R}^n$, then $\|A\|_2 = \sigma_1 = |Av_1|_2$. Also $A'A = \sum_j \sigma_j^2 v_j v_j'$, which implies $\|A\|_{\text{F}} = (\sum_j \sigma_j^2)^{1/2} = |\sigma|_2$.

See Axler (2026, §§7E,8D) for an excellent discussion of the SVD.

It simplifies notation if we regard X_i as the ith coordinate map on $\mathbb{R}^n$, thereby identifying $X := (X_1, \ldots, X_n)$ with an $x := (x_1, \ldots, x_n) \in \mathbb{R}^n$. That way FUBINI replaces conditioning arguments.

Initially $\mathbb{R}^n$ is equipped with a probability measure $\mathbb{P} = \otimes_{i\in[\![n]\!]} P_i$ under which $X_i \in \text{SUBG}(1)$ for each i. The first step in the Proof of <12> will replace $\mathbb{P}$ by $\gamma := N(0, I_n)$ by means of a very clever decoupling trick due to Bourgain and Tzafriri (1987, p149). The second step will then (effectively) use a rotation to transform $X'AX$ into a sum similar to the Y from <13>.

Remark. In what follows, $\mathbb{P}_J := \otimes_{i\in J} P_i$ and $\gamma_J := \otimes_{i\in J}\gamma_i$ for $\gamma_i := N(0,1)$, for each nonempty subset J of $[\![n]\!]$.

Proof (of Theorem <12>). The τ^2 part of the rescaling in <12> is equivalent to replacement of X_i by X_i/τ. More concisely, without loss of generality, we may assume $\tau = 1$.

Let $\mathbb{Q}$ denote the uniform distribution on $\mathbb{B}_n := \{0,1\}^n$. Under $\mathbb{Q}$, the coordinates of the generic point $\xi := (\xi_1,\ldots,\xi_n)$ in $\mathbb{B}_n$ are independent $\textsc{Ber}(1/2)$ random variables. For each such $\xi := (\xi_1,\ldots,\xi_n)$ define a new $n\times n$ matrix A_ξ by

$$A_\xi[i,j] := \xi_i(1-\xi_j)A[i,j].$$

Define $J := \{i : \xi_i = 1\}$ and $K = J^c$. If either J or K is empty the matrix A_ξ is filled with zeros. Otherwise define $H_\xi := A[J,K]$, the submatrix obtained by deleting from A rows not in J and columns not in K. If the rows and columns of A were permuted so that $j<k$ for all $j\in J$ and $k\in K$ then A_ξ and A would look like

<16> $$A_\xi = \begin{matrix} \\ J \\ K \end{matrix}\!\!\begin{matrix} \begin{matrix} J & K \end{matrix} \\ \begin{bmatrix} 0 & H_\xi \\ 0 & 0 \end{bmatrix} \end{matrix} \quad \text{AND} \quad A = \begin{matrix} \\ J \\ K \end{matrix}\!\!\begin{matrix} \begin{matrix} J & K \end{matrix} \\ \begin{bmatrix} B_1 & H_\xi \\ B_2 & B_3 \end{bmatrix} \end{matrix} \qquad \text{for some } B_i\text{'s.}$$

For simplicity of notation, let us assume from now on that A_ξ and A look like <16>.

For a fixed ξ, suppose H_ξ has SVD $\sum_{i\le m}\delta_i u_i v_i'$ with $\delta_1\ge\cdots\ge\delta_m>0$. Then $H_\xi' H_\xi = \sum_i \delta_i^2 v_i v_i'$, which gives

<17> $$\sum\nolimits_i \delta_i^2 = \operatorname{trace}(H_\xi' H_\xi) = \|H_\xi\|_{\mathrm{F}}^2 = \sum\nolimits_{j\in J,k\in K} A[j,k]^2 \le \|A\|_{\mathrm{F}}^2 .$$

The argument for $\|H_\xi\|_2$ is similar. Define $\widetilde{v}' := (0_J', v_1')$, a unit vector in $\mathbb{R}^n$. Then

<18> $$\delta_1 = \|H_\xi\|_2 = |H_\xi v_1|_2 \le \left|\begin{pmatrix} Hv_1 \\ B_3 v_1\end{pmatrix}\right|_2 = |A\widetilde{v}_1|_2 \le \|A\|_2 .$$

Most importantly,

<19> $$|H_\xi X_K|_2^2 = X_K' H_\xi' H_\xi X_K = \sum\nolimits_i \delta_i^2 Z_i^2 \qquad \text{where } Z_i := v_i' X_K.$$

Under γ the random variables $Z_1,\ldots,Z_m$ are independent and each has distribution $N(0,1)$ because $\{v_i : i\le m\}$ is a set of orthogonal unit vectors. If $2\theta^2\|A\|_2<1$,

$$\begin{aligned}
\log\gamma\exp\left(\theta^2|H_\xi X_K|_2^2\right) &= \log\gamma\exp\Big(\sum\nolimits_i \theta^2\delta_i^2 Z_i^2\Big) \qquad \text{by <19>} \\
&\le \theta^2\sum\nolimits_i \delta_i^2/(1-2\theta^2\delta_i^2) \qquad \text{by <15>} \\
&\le \theta^2\|A\|_{\mathrm{F}}^2/\left(1-2\theta^2\|A\|_2^2\right) \qquad \text{by <17> and <18>.}
\end{aligned} \tag{20}$$

Notice that the final upper bound does not depend on ξ.

So much for the preliminaries. For the clever part of the argument, note that $\mathbb{Q}\xi_i(1-\xi_j) = 1/4$ if $i \neq j$ but $\xi_i(1-\xi_i) = 0$. The assumption $A[i,i] = 0$ is needed so that $\mathbb{Q}A_\xi = \frac{1}{4}A$. Thus

$$\begin{aligned} \mathbb{P}e^{\lambda W} &= \mathbb{P}e^{\lambda X'AX} = \mathbb{P}e^{4\lambda X'(\mathbb{Q}A_\xi)X} = \mathbb{P}e^{4\lambda\mathbb{Q}X'_J H_\xi X_K} \\ &\le \mathbb{Q}\mathbb{P}e^{4\lambda X'_J H_\xi X_K} \qquad \text{by JENSEN and FUBINI.} \end{aligned} \tag{21}$$

Consider a fixed ξ, which also fixes J and K. Under both $\mathbb{P} = \mathbb{P}_J \otimes \mathbb{P}_K$ and $\gamma = \gamma_J \otimes \gamma_K$ the coordinate blocks X_J and X_K are independent—the whole point of the decoupling trick. For fixed $\eta \in \mathbb{R}^J$ the assumed SUBGAUSSIAN property gives

$$\mathbb{P}_J e^{4\lambda X'_J\eta} = \prod\nolimits_{i\in J} P_i e^{4\lambda X_i\eta_i} \le \exp\left(\tfrac{1}{2}(4\lambda)^2|\eta|_2^2\right) = \gamma_J e^{4\lambda X'_J\eta}.$$

As X_K is treated like a constant by $\mathbb{P}_J$ (and similarly for X_J and $\mathbb{P}_K$),

$$\begin{aligned} \mathbb{P}_K\mathbb{P}_J e^{4\lambda X'_J H_\xi X_K} &\le \mathbb{P}_K\gamma_J e^{4\lambda X'_J H X_K} = \gamma_J\mathbb{P}_K e^{4\lambda X'_J H_\xi X_K} \\ &\le \gamma_J\gamma_K e^{4\lambda X'_J H_\xi X_K} \\ &= \gamma_K e^{8\lambda^2|H_\xi X_K|_2^2} \\ &\le \exp\Bigl(8\lambda^2\|A\|_{\mathrm{F}}^2/(1-16\lambda^2\|A\|_2^2\Bigr) \qquad \text{by <20>.} \end{aligned}$$

As the final bound does not depend on ξ it would not be affected by integration with respect to $\mathbb{Q}$, which leaves us with a bound

$$\log\mathbb{P}e^{\lambda W} \le 8\lambda^2\|A\|_{\mathrm{F}}^2/\Bigl(1-16\lambda^2\|A\|_2^2\Bigr) \qquad \text{provided } 16\lambda^2\|A\|_2 < 1.$$

If we replace λ by $\lambda/(4\|A\|_2)$ we end up with

$$\log\mathbb{P}e^{\lambda W/(4\|A\|_2)} \le \frac{\lambda^2\|A\|_{\mathrm{F}}^2/\|A\|_2^2}{2(1-\lambda^2)} \le \frac{\lambda^2\|A\|_{\mathrm{F}}^2/\|A\|_2^2}{2(1-\lambda)} \qquad \text{for } 0 \le \lambda < 1.$$

That is, $W/(4\|A\|_2) \in \text{SUBGAMMA}(\|A\|_{\mathrm{F}}^2/\|A\|_2^2)$. The inequality displayed in the statement of the Theorem is one of the bounds derived in §3.6. □

8.3 A BENNETT inequality for martingales

Theorem <8> has a useful generalization where the assumption of independence of $X_1, \ldots, X_n$ is replaced by the assumption that they are martingale differences for a filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n$.

Remark. Note that Theorem <8> allowed the possibility that $\mathbb{P}X_i$ might not be zero, so that $\mathbb{P}X_i^2$ might be larger that $\operatorname{var}(X_i)$. This time the martingale difference property requires the conditional expectation $\mathbb{P}_{i-1}X_i := \mathbb{P}(X_i \mid \mathcal{F}_{i-1})$ to be zero so that $\mathbb{P}_{i-1}X_i^2$ is the conditional variance.

As before, I will omit the various 'almost sure' qualifications.

<22> **Theorem.** *(A BENNETT inequality for martingales) Suppose $X_1, \ldots, X_n$ are martingale differences for which there exist nonnegative, $\mathcal{F}_{j-1}$-measurable random variables M_j and v_j such that $X_j \le M_j$ and $\mathbb{P}_{j-1}X_j^2 \le v_j$ for each j. Define $S_n = X_1 + \cdots + X_n$ and $V_n = v_1 + \cdots + v_n$. For each choice of positive constants b and $\mathcal{V}$ define $G_{b,\mathcal{V}} := \{\max_{j\le n} M_j \le b \text{ and } V_n \le \mathcal{V}\}$. Then*

$$\mathbb{P}\left(\{S_n \ge t\} \cap G_{b,\mathcal{V}}\right) \le \exp\left(-\frac{t^2}{2\mathcal{V}}\psi_{\text{benn}}\left(\frac{bt}{\mathcal{V}}\right)\right) \qquad \text{for each } t \ge 0$$

so that

$$\mathbb{P}\{S_n \ge t\} \le \exp\left(-\frac{t^2}{2\mathcal{V}}\psi_{\text{benn}}\left(\frac{bt}{\mathcal{V}}\right)\right) + \mathbb{P}\{\max_{j\le n} M_j > b \text{ or } V_n > \mathcal{V}\}. \qquad \square$$

Remark. The constants b and $\mathcal{V}$ could be allowed to depend on t.

The sequences $\{v_j\}$ and $\{M_j\}$ are said to be ***predictable***, because both v_j and M_j are determined by what we learn up to step $j-1$. In the following proof they are used to define a ***predictable stopping time*** τ: the event $\{\tau \le i\}$ will belong to $\mathcal{F}_{i-1}$, a stronger property than the inclusion $\{\tau \le i\} \in \mathcal{F}_i$ required by the definition of a stopping time. The trick for using a predictable stopping time to create a better behaved martingale is well known in the martingale literature—see, for example, Pollard (1984, pages 172–179). The corresponding concept in continuous time is exceedingly subtle.

Proof. Define $S_i = \sum_{j\le i} X_j$ and $V_i = \sum_{j\le i} v_j$ for $1 \le i \le n$. To allow a clean recursive argument also define $S_0 = V_0 = 0$.

As with many martingale proofs, the main idea is to multiply the increments X_j by predictable weights, which sets up an appeal to the analog of inequality <7> for conditional expectations. The weights come from the stopping time $\tau := \inf\{j \ge 1 : M_j > b \text{ or } V_j > \mathcal{V}\}$. As usual, the infimum of an empty set is taken to be $+\infty$; that is, $\{\tau = +\infty\} = G_{b,\mathcal{V}}$. More generally,

<23> $$\{\tau > j\} = \{\max_{i\le j} M_i \le b \text{ and } V_j \le \mathcal{V}\} \in \mathcal{F}_{j-1} \qquad \text{for } 1 \le j \le n.$$

The $\mathcal{F}_{j-1}$-measurability of the event $\{\tau > j\}$ lets us stop the $\{S_i\}$ process before the event $G_{b,\mathcal{V}}^c$ occurs, without destroying the martingale

property. Define new increments $Y_j = X_j\{\tau > j\}$ for $1 \le j \le n$. Then, with probability one,

$$Y_j \le \{\tau > j\} M_j \le b, \tag{24}$$

$$\mathbb{P}_{j-1} Y_j = \mathbb{P}_{j-1}\{\tau > j\} X_j = \{\tau > j\}\mathbb{P}_{j-1} X_j = 0, \tag{25}$$

$$w_j := \mathbb{P}_{j-1} Y_j^2 = \mathbb{P}_{j-1}\{\tau > j\} X_j^2 = \{\tau > j\}\mathbb{P}_{j-1} X_j^2 \le \{\tau > j\} v_j. \tag{26}$$

These inequalities imply

$$\begin{aligned} \mathbb{P}_{j-1} e^{\lambda Y_j} &\le 1 + 0 + \tfrac{1}{2}\lambda^2 w_j \mathbb{\Delta}(\lambda b) \\ &\le \exp\left(\tfrac{1}{2}\lambda^2 w_j \mathbb{\Delta}(\lambda b)\right) = \exp\left(w_j \mathbb{f}(\lambda b)/b^2\right) \qquad \text{for } \lambda \ge 0, \end{aligned} \tag{27}$$

with

$$W_i := \sum\nolimits_{j \le i} w_j \le \sum\nolimits_{j \le n} \{\tau > j\} v_j \le \mathcal{V} \qquad \text{for all } i. \tag{28}$$

The partial sums $T_j = Y_1 + \cdots + Y_j$, with $T_0 = 0$, also define a martingale, for which $T_n = S_n$ on the set $\{\tau = +\infty\} = G_{b,\mathcal{V}}$. Thus

$$\begin{aligned} \mathbb{P}\left(\{S_n \ge t\} \cap G_{b,\mathcal{V}}\right) &= \mathbb{P}\left(\{T_n \ge t\} \cap G_{b,\mathcal{V}}\right) \\ &\le \mathbb{P}\{T_n \ge t\} \le \inf_{\lambda > 0} \mathbb{P} \exp\left(\lambda T_n - \lambda t\right). \end{aligned}$$

We are back to the MGF method for a martingale $\{T_j : 0 \le j \le n\}$ with well behaved increments.

Define $D_j = \exp\left(\lambda T_j - W_j \mathbb{f}(\lambda b)/b^2\right)$ with $D_0 = 1$. For $j \ge 1$ we have

$$\begin{aligned} \mathbb{P}_{j-1} D_j &= \mathbb{P}_{j-1} D_{j-1} \exp\left(\lambda Y_j - w_j \mathbb{f}(\lambda b)/b^2\right) \\ &= D_{j-1} \exp\left(-w_j \mathbb{f}(\lambda b)/b^2\right) \mathbb{P}_{j-1} \exp\left(\lambda Y_j\right) \le D_{j-1}. \end{aligned}$$

Take expected values then iterate to deduce that

$$\begin{aligned} \mathbb{P} \exp\left(\lambda T_n - \mathcal{V} \mathbb{f}(\lambda b)/b^2\right) &\le \mathbb{P} D_n \qquad \text{by inequality <28>} \\ &\le \mathbb{P} D_0 = 1. \end{aligned}$$

That is, $\mathbb{P}\exp(\lambda T_n) \le \exp\left(\mathcal{V}\mathbb{f}(\lambda b)/b^2\right)$, the same as the bound for $\mathbb{P}\exp(\lambda S)$ in the proof of Theorem <8>. The rest of the proof is the same as the proof for that Theorem. □

If, in addition to the assumptions of the Theorem we actually have two-sided control over the increments, $|X_j| \le M_j$, then we get a two-sided bound covering both tails,

$$\mathbb{P}\{|S_n| \ge t\} \le 2 \exp\left(-\frac{t^2}{2\mathcal{V}} \psi_{\text{benn}}\left(\frac{bt}{\mathcal{V}}\right)\right) + \mathbb{P}\{\max\nolimits_{j \le n} M_j > b \text{ or } V_n > \mathcal{V}\}. \tag{29}$$

See the next Section for a surprising application of this two-sided bound.

*8.4 Concentration of random polynomials

Although slightly disappointed with the current form of the Section, even after so much effort, I decided to include it in the book because I regard the topic as important. Maybe some reader can come up with a better treatment.

The main ideas for the following material come from the papers by Kim and Vu (2000) and Vu (2002). My Theorem <34> is a minor variation on Theorem 4.1 of Vu (2002). See the Notes for an explanation of the liberties I am taking in referring to the KIM-VU method.

Consider a random variable expressible as a polynomials$f(\xi)$ for a vector $\xi = (\xi_1, \ldots, \xi_n)$ of independent random variables, each taking values in $[0,1]$. For example, we might have $f(x) = 3x_1^2x_2^5 + 2x_1^3x_2^3x_3x_4 + 7x_1^3x_4^5x_5x_9 + 6x_8^8x_9^9$ for $n = 9$. The degree of f is defined to be the largest of the degrees of its terms, $17 = \max(2{+}5, 3{+}3{+}1{+}1, 3{+}5{+}1{+}1, 8{+}9)$ for the example just given. At the cost of an extra factor of 2 we may assume that all the coefficients in the polynomial are nonnegative.

To simplify notation, I will yet again identify the ξ_i's with the coordinate maps on $[0,1]^n$ equipped with a product measure $\mathbb{P} = \otimes_{i\le n} P_i$. For $\mathcal{F}_j := \sigma\{\xi_1, \ldots, \xi_j\}$ the conditional expectation $\mathbb{P}_j := \mathbb{P}_{\mathcal{F}_j}$ corresponds to averaging over the variables $\xi_{j+1}, \ldots, \xi_n$; it is achieved by integration with respect to $P_{j+1} \otimes \cdots \otimes P_n$.

In general, the polynomials take the form

<30> $$f(x) = \sum\nolimits_{a\in\mathbb{A}} w_a x^a \qquad \text{for } x \in [0,1]^n,$$

where $\mathbb{A}$ is a finite subset of $\mathbb{Z}_+^n$, the set of all n-tuples $a = (a_1, \ldots, a_n)$ of nonnegative integers, and

$$x^a = \prod\nolimits_{i\le n} x_i^{a_i} \qquad \text{with the convention that } x_i^0 = 1.$$

Note that the degree of f equals $\max_{a\in\mathbb{A}}\{w_a > 0\} \sum_{j\le n} a_j$.

To avoid an excess of subscripted superscripts I will also use a notation borrowed from ***R***,

$$x^{a[I]} := \prod\nolimits_{i\in I} x_i^{a_i} \qquad \text{for } I \subset [[n]],$$

with abbreviations $a[< j]$ if $I = \{i \in [[n]] : i < j\}$, and so on. If $I = \emptyset$ interpret $x^{a[I]}$ to be identically equal to 1. For example,

<31> $$\mathbb{P}_j f(\xi) = \sum\nolimits_{a\in\mathbb{A}} w_a \mathbb{P}_j \xi^a = \sum\nolimits_{a\in\mathbb{A}} w_a \xi^{a[\le j]} \mathbb{P}\xi^{a[>j]}.$$

The main result (Theorem <34>) provides a useful exponential tail bounds for $|f(\xi) - \mathbb{P}f(\xi)|$. The statement of that result involves two carefully chosen scaling quantities, $\mathcal{E}_0(f)$ and $\mathcal{E}_1(f)$, whose definition greatly puzzled me when I first looked at the Kim & Vu and Vu papers. Only after extracting the main ideas from their proofs did I understand the motivation for the following definition.

<32> **Definition.** *For each nonempty subset H of the index set $[\![n]\!] := \{1, 2, \ldots, n\}$ and $x \in [0,1]^n$ define*

$$D_H x^a = \begin{cases} 0 & \text{if there exists an } i \in H \text{ for which } a_i = 0 \\ x^{a'} & \text{where } a'_i = 0 \text{ for } i \in H \text{ and } a'_i = a_i \text{ otherwise} \end{cases}.$$

If $H = \{j\}$ is a singleton write D_j instead of $D_{\{j\}}$. For a polynomial f as in <30> define

$$D_H f = \sum_{a \in \mathbb{A}} w_a D_H x^a = \sum_{a \in \mathbb{A}} w_a \{a_i > 0 \text{ for each } i \text{ in } H\} \prod_{j \in [\![n]\!] \setminus H} x_j^{a_j}.$$

□ *Define $\mathcal{E}_1(f) = \max_{H \neq \emptyset} \mathbb{P} D_H f$ and $\mathcal{E}_0(f) = \max(\mathbb{P} f(\xi), \mathcal{E}_1(f))$.*

Remark. Notice that $D_H f$ has a strictly smaller degree than f, because only terms with $\min_{i \in H} a_i > 0$ survive and those terms all lose a factor $\prod_{i \in H} x_i^{a_i}$.

The effect of D_j is similar to the effect of the partial derivatives $\partial^r / \partial x_j^r$, except for the introduction of some extra constants. If $a_j = 0$ then both D_j and $\partial / \partial x_j$ would kill the whole x^a term. If $a_j = r > 0$ then both D_j and $\partial^r / \partial x_j^r$ would neatly remove the $x_j^{a_j}$ factor, at least from the ath term. I find it much cleaner to work with D_H rather than with partial derivatives, as Vu (2002) did.

<33> **Example.** For $f(x) = 3x_1^2 x_2^5 + 2x_1^3 x_2^3 x_3 x_4 + 7x_1^3 x_4^5 x_5 x_9 + 6x_8^8 x_9^9$,

$$D_H f(x) = 3 + 2x_3 x_4 + 0 + 0 \qquad \text{if } H = \{1,2\}$$
$$D_H f(x) = 0 + 2x_1^3 x_4 + 0 + 0 \qquad \text{if } H = \{2,3\}$$

whereas $\partial^2 \partial^5 f(x) / \partial x_1^2 \partial x_2^5 = 720 + 0 + 0 + 0$ and $\partial^2 \partial^3 f(x) / \partial x_1^2 \partial x_2^3 = 360 x_2^2 + 72 x_1 x_2 x_3$. Even ignoring the coefficients, it is not possible to achieve the
□ effect of $D_{\{1,2\}}$ (or even D_1) with partial derivatives.

<34> **Theorem.** *For every kth degree polynomial f (with nonnegative coefficients),*

<35> $$\mathbb{P}\{|f(\xi) - \mathbb{P} f(\xi)| \geq C_k t^k \sqrt{\mathcal{E}_0(f) \mathcal{E}_1(f)}\,\} \leq C_{n,k} e^{-\mathbb{h}(t)} \qquad \text{for all } t > 0,$$

where the constants are defined recursively by $C_{n,k} = 2 + (n+1) C_{n,k-1}$
□ *with $C_{n,1} = 2$ and $C_k = 2k(1 + C_{k-1})$ with $C_1 = 1$.*

Remark. The original Kim-Vu and Vu proofs gave tail bounds involving factors $\exp(-t/4)$ for $t > 0$. For my version of their results the $t/4$ is replaced by

$$\hbar(t) := (1+t)\log(1+t) - t = \tfrac{1}{2}t^2\psi_{\text{benn}}(t),$$

which fits better with appeals to BENNETT inequalities. I do not regard the extra log factor as a particulary significant improvement.

The recursion for $C_{n,k}$ implies $C_{n,k}(\lambda) \le 4(n+1)^{k-1}$. The appearance of such powers of n in the final bound will be an inevitable consequence of the appeal to a union bound for $\max_j M_j$ in <40>.

While trying to understand the Vu paper, I found it helpful to work through the following Example before tackling the proofs. Both Kim & Vu and Vu used the same example to illustrate the use of their new theorems.

<36> **Example.** Recall the Example from §7.5, which pointed out the deficiencies of the bounded-differences inequality in establishing a concentration bound for the number of triangles in an ERDŐS-RÉNYI random graph. For your convenience, I repeat the introduction to that Example:

> Consider a graph with vertex set $[\![n]\!] := \{i \in \mathbb{N} : i \le n\}$ whose edges are a subset of the set $\mathcal{E}$ of all $|\mathcal{E}| := \binom{n}{2}$ pairs of distinct vertices. Two edges are said to be adjacent if they share an endpoint. Three distinct edges form a triangle if together they contain only three vertices: that is, the edges are $\{i,j\}$, $\{j,k\}$, and $\{k,i\}$ for distinct vertices i, j, k. Write $\mathcal{T}$ for the set of all triangles, subsets of $\mathcal{E}$ of size 3 that involve exactly 3 vertices. The set $\mathcal{T}$ has size $\binom{n}{3}$.
>
> The ERDŐS-RÉNYI random graph $G_n(p)$ chooses its edges by means of a set of independent random variables $\{\xi_e : e \in \mathcal{E}\}$, with each ξ_e distributed $\text{BER}(p)$. That is, $G_n(p)$ includes e when $\xi_e = 1$, an event with probability p.
>
> The number of triangles in $G_n(p)$ equals $f(\xi)$ where
>
> $$f(y) := \sum\nolimits_{\{e_1,e_2,e_3\}\in\mathcal{T}} y_{e_1}y_{e_2}y_{e_3} \qquad \text{for } y \in \{0,1\}^{\mathcal{E}}.$$
>
> The expected number of triangles is $\theta := \mathbb{P}f(\xi) = \binom{n}{3}p^3$.

Remember that the set $\mathcal{E}$ was enumerated in an arbitrary fashion as a sequence $(e_j : 1 \le j \le |\mathcal{E}|)$.

The CHEN-STEIN method had suggested that, under the extraneous assumption that np^2 is small, for each $\epsilon > 0$ there should be some $s > 0$ for which

<37> $$\mathbb{P}\{|f(\xi) - \theta| > s\sqrt{\theta}\,\} < \epsilon.$$

Let us see how close we can get to such an inequality using Theorem <34>. Calculate.

(i) If $H = \{e_i\}$ for some edge e_i then D_H kills all triangles except for those that have e_i as one of their edges. As explained back in §7.5, there are $n-2$ such triangles, corresponding to the choices of the third vertex for the triangle. Thus $D_H f(x)$ is a sum of the form $\sum x_{e_j} x_{e_k}$ over all $n-2$ pairs of edges for which $\{e_i, e_j, e_k\} \in \mathcal{T}$. It follows that $\mathbb{P} D_H f = (n-2)p^2$, the magic quantity identified in the post mortem for the previous attempt at Erdös-Rényi.

(ii) If $H = \{e_i, e_j\}$ then only triangles containing both e_i and e_j can survive D_H. If these two edges do not share a common vertex there are no such triangles; otherwise, there is a single triangle containing both edges. It follows that $\mathbb{P} D_H f(\xi)$ is either 0 or p.

(iii) If $H = \{e_i, e_j, e_k\}$ then $D_H f$ is 1 if $H \in \mathcal{T}$ and 0 otherwise.

(iv) If $|H| \geq 4$ then $D_H f = 0$.

It follows that

$$\mathcal{E}_1(f) = \max\left((n-2)p^2, p, 1\right) \quad \text{AND} \quad \mathcal{E}_0(f) = \max\left(\theta, \mathcal{E}_1(f)\right).$$

To simplify the discussion, suppose $np = R$ for an R large enough that θ is the largest of the quantities

$$\theta \approx R^3/6, \quad (n-2)p^2 \approx R^2/n, \quad p = R/n, \quad 1.$$

In that case, $\mathcal{E}_0(f) = \theta$ and $\mathcal{E}_1(f) = (n-2)p^2$ and

$$\gamma := \sqrt{\mathcal{E}_0(f)\mathcal{E}_1(f)} \approx \text{const} \times \sqrt{\theta} \times \max(1, \sqrt{np^2})$$

and the bound from Theorem <34> simplifies to

$$\mathbb{P}\{|f(\xi) - \theta| > c_1 t^3 \sqrt{\theta} \max(1, \sqrt{np^2})\} \leq c_2 n^2 e^{-\mathbb{h}(t)}$$

for constants c_1 and c_2. This inequality is remarkably close to the desired <37> with $c_1 t^3$ playing the role of s. The extra n^2 on the right-hand side means that we need t large enough that $\mathbb{h}(t) \approx t \log(t)$ kills off a $2 \log n$ term. The presence of the $\sqrt{np^2}$ factor suggests that the requirement that np^2 is small, which came from the CHEN-STEIN argument in §7.5, might not be as extraneous as I suspected.

Maybe one of the other Theorems proved by Vu (2002) could get us even closer to the desired <37>. I have not checked. □

The proof of the Theorem will involve an induction on k, or a recursion if one thinks of starting with polynomials of degree k and working down to polynomials of degree 1. At the first step (with the degree of f equal to k) we start from the usual representation via a sum of martingale differences,

<38> $$S_n = f(\xi) - \mathbb{P} f(\xi) = X_1 + \cdots + X_n$$

where

$$\begin{aligned} X_j &= \mathbb{P}_j f(\xi) - \mathbb{P}_{j-1} f(\xi) \\ &= \sum\nolimits_{a\in\mathbb{A}} w_a \xi^{a[<j]} \left(\xi_j^{a_j} - \mathbb{P}\xi_j^{a_j}\right) \mathbb{P}\xi^{a[>j]} \qquad \text{via } <31>. \end{aligned}$$

The tails of S_n can be bounded by an appeal to the martingale BENNETT inequality <29>:

$$<39> \qquad \mathbb{P}\{|S_n| \ge t\} \le 2\exp\left(-\frac{t^2}{2\mathcal{V}}\psi_{\text{benn}}\left(\frac{bt}{\mathcal{V}}\right)\right) + \mathbb{P}\{\max_{j\le n} M_j > b \text{ or } V_n > \mathcal{V}\},$$

where

$$\begin{aligned} M_j &= M_j(\xi_1,\ldots,\xi_{j-1}) \ge |X_j| \\ V_n &= \sum\nolimits_{j\le n} v_j \qquad \text{with } v_j = v_j(\xi_1,\ldots,\xi_{j-1}) \ge \mathbb{P}_{j-1}X_j^2. \end{aligned}$$

If we could choose all M_j's and v_j's as polynomials of degree at most $k-1$ then we could hope to bound the second term on the right-hand side of <39> by an inductive appeal to the Theorem with k replaced by $k-1$. To this end it helps—for a reason soon to be made clear—to bound the second term by

$$<40> \qquad \sum\nolimits_{j\le n} \mathbb{P}\{M_j > b\} + \mathbb{P}\{\max_j M_j \le b \text{ and } V_n > \mathcal{V}\}.$$

It is easy to find a suitable M_j by noting that $\xi_j^{a_j} - \mathbb{P}\xi_j^{a_j}$ is zero if $a_j = 0$ and is bounded in absolute value by 1 otherwise. Thus

$$<41> \qquad |X_j| \le M_j := \sum\nolimits_{a\in\mathbb{A}} w_a \{a_j > 0\} \xi^{a[<j]} \mathbb{P}\xi^{a[>j]}.$$

As the sum includes only terms with $a_j \ge 1$ and the $\xi_j^{a_j}$ is omitted, this M_j has degree at most $k-1$. We then need b large enough that

$$\mathbb{P}\{M_j > b\} \le \mathbb{P}\{M_j - \mathbb{P}M_j \ge C_{k-1} t^{k-1} \sqrt{\mathcal{E}_0(M_j)\mathcal{E}_1(M_j)}\,\},$$

which, by the inductive hypthesis, is less than $C_{n,k-1}e^{-\mathbb{h}(t)}$.

Remark. With the recursive interpretation of the proof we would need a martingale representation of $M_j - \mathbb{P}M_j$ followed by another appeal to <29> with the polynomial f replaced by the polynomial M_j. And so on. As you will see from Lemma <45>, the cunning choices made in Definition <32> allow us to bound $\mathcal{E}_0(M_j)\mathcal{E}_1(M_j)$ by $\mathcal{E}_1(f)^2$. In fact the definition for $\mathcal{E}_0(f)$ and $\mathcal{E}_1(f)$ would allow us to control all $\mathcal{E}_0$'s and $\mathcal{E}_1$'s down to the base case where f is replaced by a polynomial of degree 1.

The analysis for the contribution from V_n is a little more complicated because of the squaring involved in the calculation of $\mathbb{P}_{j-1}X_j^2$, which would seem to make the degree larger than k. However, there is a neat trick to avoid this difficulty. First bound $\mathbb{P}_{j-1}X_j^2$ by $M_j\mathbb{P}_{j-1}|X_j|$ then argue that

$$\begin{aligned}\mathbb{P}_{j-1}|X_j| &\le \sum\nolimits_{a\in\mathbb{A}} w_a\{a_j>0\}\xi^{a[<j]}\mathbb{P}_{j-1}\left(\xi_j^{a_j}+\mathbb{P}\xi_j^{a_j}\right)\mathbb{P}\xi^{a[>j]}\\ &= 2u_j := \sum\nolimits_{a\in\mathbb{A}} w_a\{a_j>0\}\xi^{a[<j]}2\mathbb{P}\xi^{a[\ge j]}\end{aligned}$$

As with M_j, the exclusion of the $\xi_j^{a_j}$ factor ensures that u_j is a polynomial of degree at most $k-1$ in $\xi_1,\ldots,\xi_{j-1}$ for which $\mathbb{P}_{j-1}X_j^2 \le v_j := 2M_ju_j$. On any event where $\max_j M_j \le b$ we have

<42> $$V_n \le 2bU_n \qquad \text{where } U_n = u_1+\cdots+u_n = \sum_{j\le n}\sum_{a\in\mathbb{A}} w_a\{a_j>0\}\xi^{a[<j]}\mathbb{P}\xi^{a[\ge j]}.$$

In particular, $\mathbb{P}\{\max_j M_j \le b \text{ and } V_n > \mathcal{V}\} \le \mathbb{P}\{2bU_n > \mathcal{V}\}$ so that the upper bound <39> can be replaced by

<43> $$\begin{aligned}\mathbb{P}\{|S_n|\ge t\} &\le 2\exp\left(-\tfrac12(t^2/\mathcal{V})\psi_{\text{benn}}\left(bt/\mathcal{V}\right)\right)\\ &\quad+\sum\nolimits_{j\le n}\mathbb{P}\{M_j>b\}+\mathbb{P}\{2bU_n>\mathcal{V}\},\end{aligned}$$

a form well suited to an induction/recursion on k.

Remark. You might be wondering why I bounded $\mathbb{P}_{j-1}X_j^2$ by $M_j\mathbb{P}_{j-1}|X_j|$ instead of by M_j^2. The answer is quite subtle. I leave to you the pleasure of discovering why the argument for Lemma <45>(ii) runs into trouble if we try to use $v_j = M_j^2$ instead of $v_j = 2M_ju_j$. Hint: Look carefully at the difference between M_j and u_j.

The proof of the Theorem is now reduced to the following two Lemmas plus an exercise in juggling the constants b and $\mathcal{V}$.

<44> **Lemma.** *The Theorem is true for $k = 1$.*

Proof. The most general f with degree 1 has the form $f(\xi) = w_0 + \sum_{i\le n} w_i\xi_i$ with $0 \le w_i$ for all i. The constant w_0 has no effect on the centered variable $f(\xi) - \mathbb{P}f(\xi)$, so we may as well assume $w_0 = 0$, in which case $\mathbb{P}f(\xi) = \sum_{j\le n} w_j\mathbb{P}\xi_j$.

If $H \subset [\![n]\!]$ has size 2 or more then D_H kills all the terms in f. If $H = \{j\}$ then $D_Hf = w_j$. It follows that $\mathcal{E}_1(f) = b := \max_{1\le j\le n} w_j$ and $\mathcal{E}_0(f) = \max(b, \mathbb{P}f(\xi))$, which leads to the scale factor

$$\gamma := \sqrt{\mathcal{E}_0(f)\mathcal{E}_1(f)} = \max\left(b, \sqrt{b\mathbb{P}f(\xi)}\right).$$

We need to show that $\mathbb{P}\{|f(\xi) - \mathbb{P}f(\xi)| \ge t\gamma\} \le 2e^{-\mathbb{h}(t)}$ for $t > 0$. For independent summands $X_1, \ldots, X_n$, Corollary <8> (the two-sided version of the BENNETT inequality) implies

$$\mathbb{P}\left\{\left|\sum_{i\le n}(X_i - \mathbb{P}X_i)\right| \ge \gamma t\right\} \le 2\exp\left(-\frac{\gamma^2 t^2}{2\mathcal{V}}\psi_{\text{benn}}\left(\frac{b\gamma t}{\mathcal{V}}\right)\right) \qquad \text{for } t \ge 0$$

if $\max_i |X_i| \le b$ and $\mathcal{V} \ge \sum_i v_i$ where $v_i = \mathbb{P}X_i^2$. If we choose $X_i = w_i\xi_i$ and $\mathcal{V} = \gamma^2$ then $\max_i |X_i| \le b$ because $0 \le \xi_i \le 1$ and

$$\sum_{i\le n}\mathbb{P}(w_i\xi_i)^2 \le \sum_{i\le n} bw_i\mathbb{P}\xi_j = b\mathbb{P}f(\xi) \le \gamma^2 = \mathcal{V}$$

because $w_i^2 \le bw_i$ and $\xi_i^2 \le \xi_i$. The upper bound then becomes

$$2\exp\left(-\tfrac{1}{2}t^2\psi_{\text{benn}}(tb/\gamma)\right) \le 2\exp\left(-\tfrac{1}{2}t^2\psi_{\text{benn}}(t)\right) = 2\exp(-\mathbb{h}(t)),$$

because $b/\gamma \le 1$ and ψ_{benn} is monotone decreasing. That is, Theorem <34> for $k = 1$ follows directly from the BENNETT inequality for independent summands. □

The next Lemma shows how $\mathcal{E}_0(f)$ and $\mathcal{E}_1(f)$ control all the $\mathcal{E}_0$ and $\mathcal{E}_1$ terms that would arise from a detailed recursive argument. You can thank Definition <32> for this property.

<45> **Lemma.** *For M_j as in <41> and U_n as in <42>,*

(i) $\mathcal{E}_0(M_j) \le \mathcal{E}_1(f)$ for each j,

(ii) $\mathcal{E}_0(U_n) \le k\mathcal{E}_0(f)$.

Proof. Start with the inequality for the expected value of $M_j(\xi_1, \ldots, \xi_{j-1}) = \sum_{a\in\mathbb{A}} w_a\{a_j > 0\}\xi^{a[<j]}\mathbb{P}\xi^{a[>j]}$, which illustrates the general pattern of the arguments. By independence of the ξ_i's,

$$\mathbb{P}M_j = \sum_{a\in\mathbb{A}} w_a\{a_j > 0\}\mathbb{P}\xi^{a[<j]}\xi^{a[>j]}.$$

When $a_j > 0$ the product $\xi^{a[<j]}\xi^{a[>j]}$ equals $D_j\xi^a$. When $a_j = 0$ the ath term makes no contribution to M_j. Thus $\mathbb{P}M_j \le \mathbb{P}D_jf(\xi) \le \mathcal{E}_1(f)$. The final inequality comes from the fact that $\{j\}$ is one of H's over which the max for $\mathcal{E}_0(f)$ is taken.

Similarly, for a nonempty subset H of $[[n]] = \{i \in \mathbb{N} : i \le n\}$,

$$D_HM_j = \sum_{a\in\mathbb{A}} w_a\{a_j > 0\}\left(D_H\xi^{a[<j]}\right)\mathbb{P}\xi^{a[>j]}.$$

If H is not a subset of $[[j-1]] := \{i \in \mathbb{N} : i < j\}$ then $D_H\xi^{a[<j]} = 0$ by definition. If $a_j > 0$ and $H \subset [[j-1]]$ then, by independence,

$$\mathbb{P}\left(D_H\xi^{a[<j]}\right)\mathbb{P}\xi^{a[>j]} = \mathbb{P}\left(D_H\xi^{a[<j]}\right)\xi^{a[>j]} = \mathbb{P}D_{H\cup\{j\}}\xi^a.$$

Thus

$$\mathbb{P}D_H M_j \le \sum\nolimits_{a\in\mathbb{A}} w_a\{a_j>0\}\mathbb{P}D_{H\cup\{j\}}\xi^a \le \mathbb{P}D_{H\cup\{j\}}f(\xi) \le \mathcal{E}_1(f).$$

The argument for $U_n := \sum_{j\le n}\sum_{a\in\mathbb{A}} w_a\{a_j>0\}\xi^{a[<j]}\mathbb{P}\xi^{a[\ge j]}$ is similar. Again start with the expected value. Again by independence,

$$\mathbb{P}U_n = \sum_{j\le n}\sum_{a\in\mathbb{A}} w_a\{a_j>0\}\mathbb{P}\xi^{a[<j]}\xi^{a[\ge j]} = \sum_{j\le n}\sum_{a\in\mathbb{A}} w_a\{a_j>0\}\mathbb{P}\xi^a.$$

Notice how this time we have no D_j inside the sum. Interchange the order of summation and then note that

$$\sum\nolimits_{j\le n}\{a_j>0\} \le \sum\nolimits_{j\le n} a_j \le k$$

to deduce that $\mathbb{P}U_n \le k\mathbb{P}f(\xi)$.

Again suppose H is a nonempty subset of $[\![n]\!]$. Again $D_H\xi^{a[<j]} = 0$ if $H \not\subseteq [\![j-1]\!]$. And if $H \subseteq [\![j-1]\!]$ then

$$\begin{aligned}
\mathbb{P}D_H U_n &= \sum\nolimits_{j\le n}\sum\nolimits_{a\in\mathbb{A}} w_a\{a_j>0\}\mathbb{P}\left(D_H\xi^{a[<j]}\right)\xi^{a[\ge j]} \\
&= \sum\nolimits_{j\le n}\sum\nolimits_{a\in\mathbb{A}} w_a\{a_j>0\}\mathbb{P}\left(D_H\xi^{a[<j]}\right)\xi^{a[\ge j]} \\
&\le \mathbb{P}\sum\nolimits_{j\le n}\sum\nolimits_{a\in\mathbb{A}} w_a\{a_j>0\}D_H\xi^a \\
&\le k\mathbb{P}\sum\nolimits_{a\in\mathbb{A}} w_a D_H\xi^a = k\mathbb{P}D_H f(\xi) \le \mathcal{E}_1(f).
\end{aligned}$$

□ Assertion (ii) follows.

Remark. As an exercise, try to write $M_j - \mathbb{P}M_j$ as a sum of martingale differences $\sum_{i=1}^n X_{ij}$ with $|X_{i,j}| \le M_{i,j}$ for an $\mathcal{F}_{j-1}$-measurable $M_{i,j}$. If you choose the $M_{i,j}$'s by mimicking the construction for the M_j's then the previous Lemma should tell you that $\mathcal{E}_0(M_{i,j}) \le \mathcal{E}_1(M_j) \le \mathcal{E}_1(f)$.

At last we have all our mathematical ducks in a row; we now know enough to make good choices for the 'constants' b and $\mathcal{V}$ derived from inequality <43> with $s = C_k t^k\sqrt{\mathcal{E}_0(f)\mathcal{E}_1(f)}$:

$$\begin{aligned}
\mathbb{P}\{|S_n| \ge C_k t^k\sqrt{\mathcal{E}_0(f)\mathcal{E}_1(f)}\,\} &\le 2\exp\left(-\tfrac{1}{2}(s^2/\mathcal{V})\psi_{\text{benn}}(bs/\mathcal{V})\right) \\
&\quad + \sum\nolimits_{j\le n}\mathbb{P}\{M_j>b\} + \mathbb{P}\{2bU_n > \mathcal{V}\},
\end{aligned} \tag{<46>}$$

I wrote 'constant' because both choices will depend on t.

Proof (of Theorem <34>). In what follows I always assume that $t \ge 1$ because the upper bounds for $\mathbb{P}\{|S_n| \ge t\}$ will always involve functions larger than $2e^{-\mathbb{h}(t)}$ and $2e^{-\mathbb{h}(t)} \ge 2e^{-\mathbb{h}(1)} \approx 1.4$ for $0 \le t \le 1$.

For the inductive step, assume the Theorem holds for degree up to $k-1$. To receive the inductive benefits

$$\mathbb{P}\{M_j > b\} \le C_{n,k-1} e^{-\mathbb{h}(t)} \qquad \text{for each } j$$
$$\mathbb{P}\{U_n > \mathcal{V}/(2b)\} \le C_{n,k-1} e^{-\mathbb{h}(t)}$$

we need

<47> $$b \ge \mathbb{P}M_j + C_{k-1} t^{k-1} \sqrt{\mathcal{E}_0(M_j)\mathcal{E}_1(M_j)} \qquad \text{for each } j,$$

<48> $$\mathcal{V}/(2b) \ge \mathbb{P}U_n + C_{k-1} t^{k-1} \sqrt{\mathcal{E}_0(U_n)\mathcal{E}_1(U_n)}.$$

By Lemma <45>, the right-hand side of <47> is bounded above by

$$\mathcal{E}_1(f) + C_{k-1} t^{k-1} \mathcal{E}_1(f) \le (1 + C_{k-1}) t^{k-1} \mathcal{E}_1(f) \qquad \text{for } t \ge 1$$

and the right-hand side of <47> is bounded above by

$$k\mathcal{E}_0(f) + C_{k-1} t^{k-1} k \mathcal{E}_0(f) \le k(1 + C_{k-1}) t^{k-1} \mathcal{E}_0(f) \qquad \text{for } t \ge 1.$$

If we choose $C_k = 2k(1 + C_{k-1})$ and

$$b = C_k t^{k-1} \mathcal{E}_1(f),$$
$$\mathcal{V} = 2bk(1 + C_{k-1}) t^{k-1} \mathcal{E}_0(f) = C_k^2 t^{2k-2} \mathcal{E}_0(f)\mathcal{E}_1(f)$$

then

$$s^2/\mathcal{V} = \frac{C_k^2 t^{2k} \mathcal{E}_0(f)\mathcal{E}_1(f)}{C_k^2 t^{2k-2} \mathcal{E}_0(f)\mathcal{E}_1(f)} = t^2,$$
$$bs/\mathcal{V} = \frac{C_k t^{k-1} \mathcal{E}_1(f) C_k t^k \sqrt{\mathcal{E}_0(f)\mathcal{E}_1(f)}}{C_k^2 t^{2k-2} \mathcal{E}_0(f)\mathcal{E}_1(f)} = t\sqrt{\mathcal{E}_1(f)/\mathcal{E}_0(f)} \le t.$$

By virtue of the monotonicity of ψ_{benn}, the right-hand side of <46> is then bounded above by

$$2\exp\left(-\tfrac{1}{2}t^2 \psi_{\text{benn}}(t)\right) + C_{n,k-1}(n+1) e^{-\mathbb{h}(t)} = \left(2 + C_{n,k-1}(n+1)\right) e^{-\mathbb{h}(t)}.$$

The choice $C_{n,k} = 2 + C_{n,k-1}(n+1)$ then completes the proof of the Theorem. □

8.5 Problems

Remember that $\mathbb{f}(t) := e^t - 1 - t = \Psi_1(t) - t$ *and* $\Psi_\alpha(x) = e^{x^\alpha} - 1$ *for* $\alpha \ge 1$ *are* ORLICZ *functions, with corresponding* ORLICZ *norms* $\|\cdot\|_{\mathbb{f}}$ *and* $\|\cdot\|_{\Psi_\alpha}$. *Also remember the* STIRLING *formula,*

$$k! = \sqrt{2\pi k}\,(k/e)^k e^{r_k} \qquad \text{where } (1+12k)^{-1} < r_k < (12k)^{-1}$$

and its weak analog: $(k/e)^k \le k! \le k^k$.

[1] Facts about $\mathcal{L}^{\Psi_1}(\mathbb{P})$:

(i) If $X \in \mathcal{L}^{\Psi_1}(\mathbb{P})$ with $\gamma_1 := \|X\|_{\Psi_1} > 0$, show that

$$\mathbb{P}\{|X| \ge \gamma_1 t\} \le \mathbb{P}\exp\left((|X| - \gamma_1 t)/\gamma_1\right) \le 2\exp(-t) \qquad \text{for } t \ge 0.$$

Conversely, if $\mathbb{P}\{|X| \ge \beta_1 t\} \le 2\exp(-t)$ for all $t \ge 0$ and some $\beta_1 > 0$, show that

$$\mathbb{P}\Psi_1(|X|/3\beta_1) = \mathbb{P}_0^\infty\{|X| \ge 3\beta_1 t\}\, dt \le \int_0^\infty 2e^{-2t}dt = 1.$$

Deduce that $\gamma_1 \le 3\beta_1$.

(ii) Again, if $\infty > \gamma_1 := \|X\|_{\Psi_1} > 0$, deduce from

$$1 \ge \mathbb{P}\Psi_1(|X|/\gamma_1) = \sum\nolimits_{k\in\mathbb{N}} \mathbb{P}|X/\gamma_1|^k/k!$$

that $\|X\|_k \le \gamma_1(k!)^{1/k} \le \gamma_1 k$ for each $k \in \mathbb{N}$. That is,

$$L_1(X) := \sup\nolimits_{k\in\mathbb{N}} \|X\|_k / k \le \|X\|_{\Psi_1}.$$

Conversely, suppose $L_1 = L_1(X) < \infty$. Show that

$$\mathbb{P}\Phi_1(|X|/c) \le \sum\nolimits_{k\in\mathbb{N}} \mathbb{P}(L_1 k/c)^k/k! \le \sum\nolimits_{k\in\mathbb{N}} (L_1 e/c)^k = 1 \qquad \text{if } c = 2L_1/e.$$

Deduce that $\|X\|_{\Psi_1}/(2e) \le L_1(X) \le \|X\|_{\Psi_1}$.

(iii) Suppose $\|X\|_k \le L_1 k$ for all $k \in \mathbb{N}$, with L_1 a finite constant. For all real λ with $|\lambda| \le 1/(L_1 e)$ show that

$$\mathbb{P}e^{\lambda X} \le 1 + \lambda\mathbb{P}X + \sum\nolimits_{k\ge 2} \frac{(|\lambda|L_1 k)^k}{k!} \le \exp\left(\lambda\mathbb{P}X + \frac{(|\lambda|L_1 e)^2}{1 - |\lambda|L_1 e}\right).$$

Deduce that $\pm(X - \mathbb{P}X)/(L_1 e) \in \text{subGamma}(1)$. Conversely, suppose Y is a random variable for which $\pm Y \in \text{subGamma}(1)$. Show that

$$1 + \sum_{k\in\mathbb{N}} \lambda^{2k}\mathbb{P}Y^{2k} = \tfrac{1}{2}\left(\mathbb{P}e^{\lambda Y} + \mathbb{P}e^{-\lambda Y}\right) \le \exp\left(\frac{\lambda^2/2}{1-\lambda}\right) \qquad \text{for } 0 \le \lambda < 1.$$

Deduce that $\|Y\|_{2k}^{2k} \le 2^k e^{1/4}(2k)!$ for $k \in \mathbb{N}$, implying $\sup_{p\ge 1} \|Y\|_p / p \le 5.2$.

[2] Deducing $\mathcal{L}^{\Psi_\alpha}$ facts from $\mathcal{L}^{\Psi_1}$ facts:

(i) For each $\alpha > 1$ show that $X \in \mathcal{L}^{\Psi_\alpha}(\mathbb{P})$ if and only if $|X|^\alpha \in \mathcal{L}^{\Psi_1}(\mathbb{P})$. Define $\gamma_\alpha := \|X\|_{\Psi_\alpha}$. For such an X show that $\gamma_\alpha^\alpha = \| |X|^\alpha \|_{\Psi_1}$.

(ii) Deduce from Problem [1] that

$$\mathbb{P}\{|X| \ge \gamma_\alpha t\} = \mathbb{P}\{|X|^\alpha \ge (\gamma_\alpha t)^\alpha\} \le 2\exp(-t^\alpha) \qquad \text{for } t \ge 0.$$

Conversely, if there is a constant $\beta > 0$ for which $\mathbb{P}\{|X| \geq \beta t\} \leq 2\exp(-t^\alpha)$ for all $t \geq 0$ deduce that $\gamma_\alpha \leq 3^{1/\alpha}\beta$.

[3] Suppose $X_1, \ldots, X_n$ are independent random variables in $\mathcal{L}^{\mathbb{f}}$, with $\gamma_i := \|X_i\|_{\mathbb{f}} > 0$. Define $B = \max_i \gamma_i$ and $\mathcal{V} = 2\sum_i \gamma_i^2$ and $W = \sum_i (X_i - \mathbb{P}X_i)$.

(i) From the inequality $\mathbb{P}\mathbb{f}(|X_i|/\gamma_i) \leq 1$ deduce that $\mathbb{P}|X|^k \leq \gamma_i^k k!$ for $k \geq 2$ so that

$$\mathbb{P}e^{\lambda(X_i - \mathbb{P}X_i)} \leq \exp\left(\sum\nolimits_{k\geq 2} \lambda^k \gamma_i^k\right) \leq \exp\left(\frac{\lambda^2\gamma_i^2}{1 - B\lambda}\right) \qquad \text{for } 0 \leq \lambda B < 1.$$

(ii) Deduce that $\pm W/B \in \text{subGamma}(\alpha)$ where $\alpha = \mathcal{V}/B^2$.

(iii) Deduce that

$$\mathbb{P}\{|W| \geq t\} \leq 2\exp\left(-\frac{t^2}{4\sum_i \gamma_i^2 + 2t\max_i \gamma_i)}\right) \qquad \text{for } t \geq 0.$$

[4] This problem handles the diagonal term that was omitted from Theorem <12>. Suppose $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and $X_1, \ldots, X_n$ are independent subgaussians for which $\max \tau(X_i) \leq 1$. Define $Y_i = X_i^2 - \mathbb{P}X_i^2$ and $W = \sum_{i\leq n} a_i Y_i$. As usual, $|a|_2 := \sqrt{\sum_{i\leq n} a_i^2}$ and $|a|_\infty = \max_{i\leq n} |a_i|$.

(i) Use the inequality $\sup_{k\in\mathbb{N}} \|X_i\|_{2k}/\sqrt{k} := L(X_i) \leq \sqrt{6}\,\tau(X_i)$ from §7.2 to show that

$$\mathbb{P}|Y_i|^k = \|Y_i\|_k^k \leq \left(2\left\|X_i^2\right\|_k\right)^k = 2^k \mathbb{P}X_i^{2k} \leq 2^k\left(\sqrt{6k}\right)^{2k} = (12k)^k.$$

(ii) For $0 \leq 12e|a|_\infty\lambda < 1$ show that

$$\begin{aligned}\mathbb{P}e^{\lambda a_i Y_i} &= 1 + \lambda a_i \mathbb{P}Y_i + \sum\nolimits_{k\geq 2} \frac{|a_i|^k\lambda^k(12k)^k}{k!} \\ &\leq 1 + \sum\nolimits_{k\geq 2}(12e|a_i|\lambda)^k \leq \exp\left(\frac{(12e\lambda)^2 a_i^2}{1 - 12e|a|_\infty\lambda}\right),\end{aligned}$$

so that $\log \mathbb{P}e^{\lambda W/d} \leq \alpha\lambda^2/2(1-\lambda)$ for $0 \leq \lambda < 1$, where $B := 12e|a|_\infty$ and $\alpha := 2|a|_2^2/|a|_\infty^2$. That is, $W/B \in \text{subGamma}(\alpha)$, in the sense of §3.6.

(iii) Deduce that

$$\mathbb{P}\{S \geq t\} \leq \exp\left(\frac{-t^2/d^2}{2(\alpha + t/d)}\right) = \exp\left(\frac{-t^2}{576|a|_2^2 + 12e|a|_\infty t}\right) \qquad \text{for } t \geq 0.$$

[5] Suppose W_1 and W_2 are random variables, not necessarily independent, with $W_i \in \text{subGamma}(\alpha_i)$.

(i) For $0 \le \lambda < 1$, show that

$$\log \mathbb{P} e^{\lambda(W_1+W_2)/2} \le \tfrac{1}{2} \log \mathbb{P} e^{\lambda W_1} + \tfrac{1}{2} \log \mathbb{P} e^{\lambda W_2} \le \frac{\lambda^2(\alpha_1+\alpha_2)/2}{1-\lambda}$$

(ii) Deduce that $(W_1 + W_2)/2 \in \text{subGamma}(\alpha_1 + \alpha_2)$.

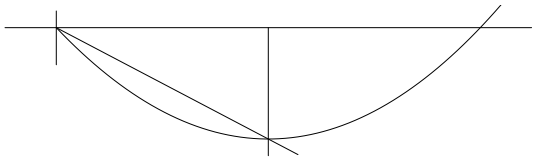
[6] If $\mathcal{Q}(\lambda) = D_1\lambda^2 - \lambda t$ for positive constants D_1, D_2, and t, show that

$$\min_{0\le\lambda\le D_2} \mathcal{Q}(\lambda) \le \max\left(-t^2/(4D_1), -tD_2/2\right).$$

Remark. Vershynin (2018) consistently wrote his BERNSTEIN-like tail bounds as maxima of two terms, because he reduced the MGF calculations to this constrained minimization problem rather than following the subGAMMA route.

(i) Show that the global minimum of $\mathcal{Q}(\lambda)$ is $-t^2/(4D_1)$, which is achieved at $\lambda_1 = t/(2D_1)$. If $D_2 \ge \lambda_1$ then the constrained minimum equals the global minimum, otherwise it equals $\mathcal{Q}(D_2)$.

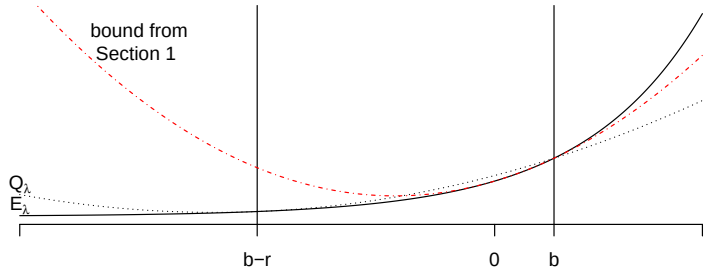
(ii) Show that $\mathcal{Q}(\lambda) \le -t\lambda/2$ for $0 \le \lambda \le \lambda_1$. Deduce that $\mathcal{Q}(D_2) \le -tD_2/2$ if $D_2 < \lambda_1$.



[7] Here is an outline of the key step in the original proof for Theorem <8> given by Bennett (1962, page 42). The main task was to bound the MGF for a random variable W with $\mathbb{P}W = 0$, $\text{var}(W) = \sigma^2$, and $W \le b$ with probability one. Bennett showed that

<49> $$\mathbb{P} e^{\lambda W} \le \frac{b^2}{b^2+\sigma^2} e^{-\lambda\sigma^2/b} + \frac{\sigma^2}{b^2+\sigma^2} e^{\lambda b} \qquad \text{for } \lambda > 0.$$

With some brute-force calculus this inequality can be used to derive an upper bound similar to my inequality <7>. See Hoeffding (1963, page 23) or Pollard (1984, page 194) for condensed accounts of the gruesome details.



Bennett's ingenious idea (for which he acknowledged the help of a referee) was to construct, for a fixed $\lambda > 0$, a quadratic $\mathcal{Q}_\lambda(x) = \alpha x^2 + \beta x + \gamma$ such that

<50> $\mathcal{Q}_\lambda(x) \geq E_\lambda(x)$ for $x \leq b$, with equality only at $x = -\sigma^2/b$ and $x = b$.

Here and subsequently, $E_\lambda(x) = e^{\lambda x}$, considered as a function of x for a fixed λ. The following steps show you how to find the $\mathcal{Q}_\lambda$ that satisfies <50> and then how that inequality implies <49>.

(i) Let P denote the distribution that puts probability θ at $b - r$ and probability $1 - \theta$ at b. Show that P has expected value 0 and variance σ^2 if

$$r = (b^2 + \sigma^2)/b \quad \text{AND} \quad b - r = -\sigma^2/b \quad \text{AND} \quad \theta = b^2/(b^2 + \sigma^2).$$

Remark. Note that the right-hand side of <49> equals $Pe^{\lambda x}$. This P is extremal amongst those distributions on $(-\infty, b]$ with expected value 0 and variance σ^2.

(ii) Write $E_\lambda(x)$ for $e^{\lambda x}$ and x_0 for $b - r$, with r as in the definion of P. Find α, β, γ as functions of λ for which

$$\begin{aligned} E_\lambda(x_0) = \mathcal{Q}_\lambda(x_0) &= \alpha\left(b^2 - 2br + r^2\right) + \beta(b - r) + \gamma, \\ \dot{E}_\lambda(x_0) = \mathcal{Q}'_\lambda(x_0) &= 2\alpha(b - r) + \beta, \\ E_\lambda(b) = \mathcal{Q}_\lambda(b) &= \alpha b^2 + \beta b + \gamma. \end{aligned}$$

Hint: Define

$$\begin{aligned} e_2(y) &:= E_\lambda(x_0 + y) - E_\lambda(x_0) - y\dot{E}_\lambda(x_0) \\ q_2(y) &:= \mathcal{Q}_\lambda(x_0 + y) - \mathcal{Q}_\lambda(x_0) - y\dot{\mathcal{Q}}_\lambda(x_0). \end{aligned}$$

Use the fact that $e_2(r) = q_2(r) = \alpha r^2$, to show that

$$\alpha = e_2(r)/r^2, \qquad \beta = \dot{E}_\lambda(x_0) - 2\alpha x_0, \qquad \gamma = e^{\lambda b} - \alpha b^2 - \beta b.$$

(iii) Use the two integral representations (cf. §2.2)

$$\begin{aligned} e_2(y) &= y^2 \int_0^1 (1 - s)\ddot{E}_\lambda(x_0 + sy)\, ds, \\ q_2(y) &= y^2 \int_0^1 (1 - s)2\alpha\, ds, \end{aligned}$$

to show that

$$\begin{aligned} \mathcal{Q}_\lambda(x_0 + y) - E_\lambda(x_0 + y) &= q_2(y) - e_2(y) \\ &= y^2\mathcal{R}(y) \qquad \text{where } \mathcal{R}(y) = \int_0^1 (1 - s)\left(2\alpha - \ddot{E}_\lambda(x_0 + sy)\right) ds. \end{aligned}$$

Use the fact that $\ddot{E}_\lambda$ is strictly increasing to deduce that the function $\mathcal{R}(y)$ is strictly decreasing.

(iv) From the equality $E_\lambda(b) = \mathcal{Q}_\lambda(b)$ show that $0 = r^2\mathcal{R}(r)$. Then deduce that $\mathcal{R}(y) > 0$ for $y < r$, implying $\mathcal{Q}_\lambda(x) \geq E_\lambda(x)$ for $x \leq b$, with equality only at $x = x_0$ and $x = b$.

(v) Finally, argue that

$$\begin{aligned} \mathbb{P}e^{\lambda W} &\leq \mathbb{P}\mathcal{Q}_\lambda(W) = \alpha\mathbb{P}W^2 + \beta\mathbb{P}W + \gamma \\ &= \alpha\sigma^2 + \beta.0 + \gamma \\ &= P\mathcal{Q}_\lambda(x) = Pe^{\lambda x}. \end{aligned}$$

8.6 Notes

The BENNETT inequality for independent summands comes from Bennett (1962), who gave a different proof, which I outlined in Problem [7] because I suspect his ingenious argument might be useful for other constrained optimization problems.

Section 8.2 is based on an elegant little paper by Rudelson and Vershynin (2013) supplemented by Vershynin (2018, Chapter 6), both of which described applications and discussed some of the history. The original form of Theorem <12> is due to Hanson and Wright (1971), who assumed symmetry of the matrix A and symmetry around zero for the distributions of the X_i's. They commented: "We would very much like to remove the restriction that the distributions of the X's be symmetric. Unfortunately, our proof depends heavily on this symmetry." Interestingly, that method involved a bounding of moment quantities for the X_i's by analogous moment quantities for the $N(0,1)$, leading to a bound on the MGF for the centered quadratic form by an analogous MGF for a quadratic form in standard normals. From that point on, their proof was similar to the proof of Theorem <12>. Wright (1973) removed the symmetry assumption for the random variables. For the clever decoupling idea with the δ_i's in Section 8.2, Vershynin (2018, Chapter 6) (which I highly recommend) cited Bourgain and Tzafriri (1987, page 149), who commented:

> The proof of Theorem 1.6 requires the use of a variant of the so-called decoupling principle.
>
> This principle can be found in literature, mostly for symmetric matrices. For sake of completeness, we give here a proof of the version needed below.

As B&T noted, the decoupling idea is much more general and has a longer history. See the Notes to de la Peña and Giné (1999, Chap 3).

The extension of the BENNETT inequality to sums of martingale differences with increments that are bounded in absolute value by 1 is largely

due to Freedman (1975). The relaxation to predictable upper bounds is implicit in Sections 3.3 and 3.4 of the paper of Vu (2002), although he did not express it via stopping times. He did not appeal to the BENNETT inequality, but instead derived the necessary exponential bounds by direct arguments. Something similar to inequality <35> appeared in the Kim and Vu (2000) paper, but only for X_i's taking values in $\{0, 1\}$. Vu (2002) developed more elaborate versions of the inequality. Most of my discussion in Section 8.4 is a translation of Vu's paper into the framework of stopping times and the BENNETT inequality. Also, for my D_H I used a definition closer to that in the Kim-Vu paper rather than the $\partial_A Y$ of Vu (2002, page 275).

References

Axler, S. (2026). *Linear Algebra Done Right* (Fourth ed.). Undergraduate Texts in Mathematics. Springer. Free online at https://linear.axler.net/LADR4e.pdf.

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association 57*, 33–45.

Bourgain, J. and L. Tzafriri (1987). Invertibility of "large" submatrices with applications to the geometry of banach spaces and harmonic analysis. *Israel journal of mathematics 57*(2), 137–224.

de la Peña, V. H. and E. Giné (1999). *Decoupling: From dependence to independence*. Springer-Verlag.

Freedman, D. A. (1975). On tail probabilities for martingales. *Annals of Probability 3*(1), 100–118.

Hanson, D. L. and F. T. Wright (1971). A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics 42*(3), pp. 1079–1083.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*, 13–30.

Kim, J. H. and V. H. Vu (2000). Concentration of multivariate polynomials and its applications. *Combinatorica 20*(3), 417–434.

Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer.

Rudelson, M. and R. Vershynin (2013). Hanson-Wright inequality and sub-gaussian concentration. *Electron. Commun. Probab 18*(82), 1–9.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.

Vu, V. H. (2002). Concentration of non-Lipschitz functions and applications. *Random Structures and Algorithms 20*, 262–316.

Wright, F. T. (1973). A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric. *The Annals of Probability 1*(6), 1068 – 1070.

Chapter 9
Measurability woes

SECTION *9.1 presents an example to illustrate some of the measurabilty difficulties that beset stochastic processes with uncountable index sets.*
SECTION *9.2 describes the the classical theory of convergence in distribution for random,* BOREL*-measurable elements of a metric space.*
SECTION *9.3 introduces the idea that a stochastic process defines a random element of a space of functions.*
SECTION *9.4 explains how to replace supremum and infimum of uncountable families of random variables by weaker concepts—the essential supremum and infimum—that restore measurabilty at the cost of a few almost sure qualifiers.*
SECTION *9.5 explains how outer measure and outer integrals are related to essential suprema.*
SECTION *9.6 describes how to make negligible modifications to each member of an uncountable set of random variables, with each random variable being changed on its own negligible set, to produce a version with better sample path properties.*
SECTION *9.7 describes the currently popular method for handing the measurability difficulties associated with the concept of convergence in distribution for stochastic processes as maps into non-separable metric spaces of functions.*
SECTION *9.8 derives a necessary and sufficient condition for convergence in distribution of a sequence of processes: oscillation control plus convergence of finite dimensional distributions.*

9.1 The difficulty

Measurability issues often complicate the handling of stochastic processes with uncountable index sets, where typical operations (such as taking the supremum of a collection of random variables) are not guaranteed to preserve measurability. For specific problems such difficulties can sometimes be handled by ad hoc approximation arguments using countable subsets of the index set. More general strategies do exist but they involve ideas that might seem unnecessarily complicated at first sight. This Chapter will discuss some of those strategies after describing the obstacles they were designed to sidestep.

The following Example illustrates some of the important issues that can arise with an uncountable index set T.

<1> **Example.** Suppose $U_1, U_2, \ldots$ are independent random variables, each distributed UNIF$[0,1]$, all living on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Together they define a sequence of ***empirical distribution functions***,

$$F_n(t,\omega) := n^{-1} \sum\nolimits_{i \le n} \{U_i(\omega) \le t\} \qquad \text{for } 0 \le t \le 1,$$

and, after appropriate standardization, ***empirical processes***,

<2> $$\nu_n(t,\omega) := n^{1/2} \left(F_n(t,\omega) - t\right) = n^{-1/2} \sum\nolimits_{i \le n} \Big(\{U_i(\omega) \le t\} - t \Big).$$

Using the fact that $nF_n(t)$ has a BIN(n,t) distribution and each pair $(nF_n(s), nF_n(t))$ has a multinomial distribution we get

$$\mathbb{P}\nu_n(t) = 0 \qquad \text{for } t \in [0,1],$$

<3> $$\operatorname{cov}(\nu_n(s), \nu_n(t)) = \min(s,t) - st \qquad \text{for } s,t \in [0,1].$$

Also, by the multivariate CLT (UGMTP §7.3), for each finite subset J of T the random vector $(\nu_n(t,\omega) : t \in J)$ converges in distribution to a centered multivariate normal distribution on $\mathbb{R}^J$ with the same covariances.

As Doob (1949) pointed out, the limiting mutivariate normals correspond to the finite-dimensional distributions of what is now called a BROWNIAN bridge, the gaussian process $\{B_0(t,\omega) : 0 \le t \le 1\}$ with continuous sample paths, zero expected values, and the same covariances as ν_n. On that basis he made a truly inspired suggestion (p395):

> "We shall assume, until a contradiction frustrates our devotion to heuristic reasoning, that *in calculating asymptotic [$\nu_n(t)$] process distributions when $n \to \infty$ we may simply replace the [$\nu_n(t)$] processes by the [$B_0(t)$] process.* It is clear that this cannot be done in all possible situations, but let the reader who has never used this sort of reasoning exhibit the first counter example."

He then proceeded to show how the heuristic leads to a limit result proved in 1933 by Kolmogorov, who used a more laborious direct argument.

Remark. See item 15 (pp. 139–146) of volume 2 of Kolmogorov's Selected Works (Kolmogorov, 1993) for an English translation of his 1933 paper. For a discussion (with pictures) of Doob's argument see Pollard (1984, §V.6). I still find Doob's approach instructive, even though there now exist simpler martingale methods for deriving the same result.

Subsequently, Donsker (1952) turned Doob's suggestion into a rigorous argument, rederiving the limiting distribution of $\max_{0 \le t \le 1} \nu_n(t)$ by a direct asymptotic analysis of the empirical process. He also asserted that convergence of more general functionals of ν_n could be proved by means of the method he had established in his doctoral dissertation.

Remark. Donsker sandwiched the event $\{\sup_t \nu_n(t,\omega) \le \alpha\}$ between two events depending only on the values $\nu_n(i/n,\omega)$ for $i \in [\![n]\!]$, an approximation similar to one used by Kolmogorov in the 1933 paper. This approach avoids the measurability difficulties described in the rest of this Example. However, the general Theorem that he stated on the final page of the 1952 paper made no mention of measurabilty issues.

The sample path $\nu_n(\cdot,\omega)$ has jumps at the $U_i(\omega)$'s; it belongs to the set $D[0,1]$ of all real-valued functions on $[0,1]$ that are right-continuous at each t in $[0,1)$ and have limits from the left at each t in $(0,1]$. Under the uniform metric

$$d_u(x,y) := \sup_t |x(t) - y(t)| \qquad \text{for } x, y \in D[0,1], \tag{<4>}$$

the set $D[0,1]$ becomes a complete metric space. For future reference, note that it is not separable: the indicator functions of the uncountable collection of intervals $[0,a]$, for $0 \le a \le 1$, are at d_u-distance 1 from each other.

We might hope that ν_n defines a random element of that metric space, meaning that $\omega \mapsto \nu_n(\cdot,\omega)$ defines an $\mathcal{F}\backslash\mathcal{B}$-measurable map for $\mathcal{B}$ equal to the BOREL sigma-field on $D[0,1]$ under d_u. Unfortunately, it does not. To understand this failure it helps (for a fixed n) to take the U_i's as the coordinate maps on $\Omega = [0,1]^n$ equipped with $\mathbb{P}$ equal to LEBESGUE measure, $\mathfrak{m}_n$, on the BOREL sigma-field (or its completion with respect to $\mathfrak{m}_n$). Let A be a subset of $[0,1]$ that is not LEBESGUE-measurable. (Existence of such a set involves questions related to the axioms for set theory: see Dudley, 2003, §3.4 for a most insightful discussion.) The set

$$D_A := \{x \in D[0,1] : \text{ all points of discontinuity of } x \text{ (if any) lie in } A\}$$

is closed under the d_u metric; it belongs to $\mathcal{B}$. If ν_n were $\mathcal{F}\backslash\mathcal{B}$-measurable the set

$$\Omega_A := \{\omega \in \Omega : \nu_n(\cdot,\omega) \in D_A\} = \{\omega : U_i(\omega) \in A \text{ for all } i\} = A^n$$

would have to belong to $\mathcal{F}$. The FUBINI theorem (Dudley, 2003, p137) would then require almost all one-dimensional cross-sections of A^n (which are either A or $\emptyset$) to be measurable. The nasty choice for A rules out that possibility. □

9.2 Convergence in distribution, traditional style

You might be wondering why it matters whether the ν_n in Example <1> is measurable with respect to the BOREL sigma-field on $D[0,1]$, or not. If it were measurable then, with slight rearrangements, the result established by Donsker would fit into a well-established theory for convergence in distribution of random elements of metric spaces.

In general, for BOREL-measurable random elements $\{X_n : n \in \mathbb{N}\}$ of a general metric space $(\mathcal{X}, d)$, convergence in distribution to a probability measure Q on the BOREL-sigma field $\mathcal{B}(\mathcal{X}, d)$ can be defined by

<5> $$X_n \rightsquigarrow Q \text{ means } \mathbb{P}f(X_n) \to Qf \qquad \text{for each } f \text{ in } \mathrm{BL}(\mathcal{X}, d),$$

where $\mathrm{BL}(\mathcal{X}, d)$ denotes the set of all bounded, LIPSCHITZ-continuous, real-valued functions on $\mathcal{X}$. That is, $f \in \mathrm{BL}(\mathcal{X}, d)$ iff

$$\|f\|_{\mathrm{BL}} := \max\left(2\|f\|_\infty, \|f\|_{\mathrm{Lip}}\right) \text{ is finite,}$$
$$\text{where } \|f\|_\infty := \sup_{x\in\mathcal{X}} |f(x)| \quad \text{AND} \quad \|f\|_{\mathrm{Lip}} := \sup_{x\neq y} |f(x) - f(y)|/d(x, y).$$

More concisely, if $f \in \mathrm{BL}(\mathcal{X}, d)$ then

<6> $$|f(x) - f(y)| \le \|f\|_{\mathrm{BL}} \Big(1 \wedge d(x, y)\Big) \qquad \text{for all } x, y \text{ in } \mathcal{X}.$$

Remark. This definition of $\|\cdot\|_{\mathrm{BL}}$ comes from UGMTP §7.1. Dudley (2014, §3.6) defined $\|f\|_{\mathrm{BL}} := \|f\|_\infty + \|f\|_{\mathrm{Lip}}$. Both definitions lead to the same $\mathrm{BL}(\mathcal{X}, d)$. The definitions could also have started with a sequence of probability measures $\{Q_n\}$ on $\mathcal{B}(\mathcal{X}, d)$ and a concept often called ***weak convergence***: $Q_n \rightsquigarrow Q$ means $Q_n f \to Qf$ for all f in $\mathrm{BL}(\mathcal{X}, d)$. Similarly, we could eliminate probability measures from the definition: if X_∞ is a random element of $\mathcal{X}$ with distribution Q then $X_n \rightsquigarrow X_\infty$ means the same thing as $X_n \rightsquigarrow Q$.

It is not difficult to show (UGMTP §7.1) that property <5> is equivalent to each of the following conditions:

(CID.i) $\liminf_{n\to\infty} \mathbb{P}f(X_n) \ge Qf$ for all lower-semicontinuous functions f that are bounded below by a (finite) constant.

(CID.ii) $\limsup_{n\to\infty} \mathbb{P}f(X_n) \le Qf$ for all upper-semicontinuous functions f that are bounded above by a (finite) constant.

(CID.iii) $\mathbb{P}f(X_n) \to Qf$ for all (two-sided) bounded real-valued functions on $\mathcal{X}$ with $Q\{x \in \mathcal{X} : \ f \text{ is continuous at } x\} = 1$.

As an easy consequence: if $X_n \rightsquigarrow Q$ and if a BOREL-measurable map ψ from $(\mathcal{X}, d)$ into another metric space $(\mathcal{X}', d')$ is continuous at Q almost all x in $\mathcal{X}$ then $\psi(X_n) \rightsquigarrow \psi(Q)$, where $\psi(Q)$ denotes the image of Q under ψ (see UGMTP §2.9). In particular, convergence in distribution of $\{\nu_n\}$ would have generated a host of limit theorems. For example, the map $\psi(x) = \sup_t x(t)$ would have given convergence in distribution of $\sup_t \nu_n(t)$ and the map $\psi(x) = \int_0^1 x(t)^2 dt$ would have given convergence in distribution of $\int_0^1 \nu_n(t, \omega)^2\, dt$.

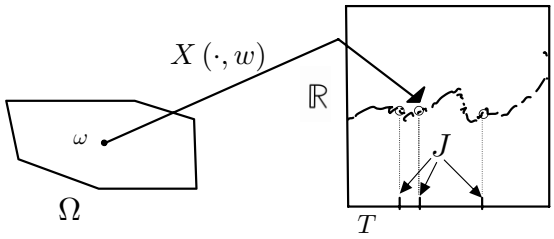
Despite the discouraging experience with ν_n, the general idea of treating a stochastic process as a random element of a set of functions has become one of the most fruitful ways to think about such processes. The remaining Sections of this Chapter will explain methods that can work around the measurabilty obstacles. In particular, Section 9.7 will show how <5> can be generalized to provide a useful theory for X_n's that are not BOREL-measurable.

Remark. We could avoid the measurability difficulties associated with treating ν_n from Example <1> as a map into $D[0,1]$ if we replaced d_u by the J_1 metric, d_J, defined by Skorohod (1956). The process ν_n is $\mathcal{F}\backslash\mathcal{B}_J$-measurable, for $\mathcal{B}_J$ equal to the J_1-BOREL sigma-field on $D[0,1]$. By a slight tweaking of the J_1 definition, Billingsley (1968, §14) defined another metric, which generates the same open sets as the d_J and under which $D[0,1]$ becomes a complete, separable metric space. The analysis of the empirical process from Example <1> would then fit neatly into the theory built around <5>. Unfortunately, such alternatives depend heavily on the ordering properties of the index set $[0,1]$; they have no obvious extension to more general index sets.

9.3 Stochastic processes as random functions

For most of this Chapter a stochastic process will be just a collection of (real-valued) random variables $X := \{X_t : t \in T\}$, all defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. That is, each X_t will be an $\mathcal{F}\backslash\mathcal{B}(\mathbb{R})$-measurable map from Ω into $\mathbb{R}$, where $\mathcal{B}(\mathbb{R})$ denotes the BOREL sigma-field on $\mathbb{R}$. It could also be thought of as a function of two arguments, $\{X(t,\omega) : t \in T, \omega \in \Omega\}$, in which case $X_t(\cdot) := X(t,\cdot)$, or as a map $\omega \mapsto X_\omega(\cdot) := X(\cdot,\omega)$ from Ω into the set $\mathbb{R}^T$, of all real-valued functions on T. The function X_ω is usually called a ***sample path*** of the process.

Remark. The arguments would change little if $\mathbb{R}$ were replaced by its compactification $\overline{\mathbb{R}} := [-\infty, \infty]$ under its BOREL sigma-field (c.f. UGMTP page 23). As you will see in Section 9.6, the quest for nice sample paths leads to such an extension. For many purposes we could also allow X_t to be a measurable function with values in $\mathbb{R}^k$ (or $\overline{\mathbb{R}}^k$) or even in some abstract measurable space $(\mathcal{Y}, \mathcal{D})$.



Traditionally the index set T was always a subset of $\mathbb{R}$ or $\overline{\mathbb{R}}$. In that case, if $J = \{t_1, t_2, \ldots, t_k\}$ were a finite subset of $\mathbb{R}$ then it was natural to regard $(X_t : t \in J)$ as a random element of $\mathbb{R}^k$ given by $\omega \mapsto (X_{t_1}(\omega), \ldots, X_{t_k}(\omega))$ instead of as a random element of $\mathbb{R}^J$, the set of all real-valued functions on J. However, a small notational problem then arises: What happens if we relabel the points of J as $s_1, \ldots, s_k$ with $s_i \neq t_i$? In that case the vector $(X_{s_1}(\omega), \ldots, X_{s_k}(\omega))$ would usually be different from the vector $(X_{t_1}(\omega), \ldots, X_{t_k}(\omega))$. Such an ambiguity affects some expositions of the basic existence theorem due to Kolmogorov (1933, §III.4).

Consider first the version of that theorem stated by Dudley (2003, Thm 12.1.2). (Actually Dudley allowed coordinate spaces more general than the real line, so I should say a 'special case of a generalization of the original theorem'.) As starting point we have a consistent family of ***finite-dimensional distributions*** (or 'fidis', for short). That is, for each finite subset J of an (infinite) index set T we have a probability measure P_J on $\mathcal{B}(\mathbb{R}^J)$ satisfying a consistency condition: for $K \subset J$ the measure P_K is the image of P_J under the projection map $\pi_{J\to K}$ that takes each function x in $\mathbb{R}^J$ onto its restriction $x|_K$ in $\mathbb{R}^K$. That is,

<7> $$P_K B = P_J\left(\pi_{J\to K}^{-1} B\right) = P_J\{x \in \mathbb{R}^J : x|_K \in B\} \qquad \text{for each } B \text{ in } \mathcal{B}(\mathbb{R}^K).$$

The theorem asserts existence of a probability measure $\mathbb{P}$ on $\mathcal{C}(\mathbb{R}^T)$, the product (or cylinder) sigma-field on $\mathbb{R}^T$ generated by the coordinate projections $\{\pi_t : t \in T\}$, where $\pi_t(x) := x(t)$ for $x \in \mathbb{R}^T$. Equivalently, $\mathcal{C}(\mathbb{R}^T)$ is generated by the sets $\pi_J^{-1} B := \{x \in \mathbb{R}^T : x|_J \in B\}$ as J ranges over all finite subsets J of T and B ranges over all BOREL subsets of $\mathbb{R}^J$.

<8> **Theorem.** *For any (infinite) set T let $\{P_J :$ finite $J \subset T\}$ be a consistent family of fidis. Then there exists a probability measure $\mathbb{P}$ defined on the product sigma-field $\mathcal{C}(\mathbb{R}^T)$ such that $\mathbb{P}\{x \in \mathbb{R}^T : x|_J \in B\} = P_J B$ for each J and $B \in \mathcal{B}(\mathbb{R}^J)$.* □

Billingsley (1986, §36) followed Kolmogorov's presentation more closely, indexing the fidis by k-tuples for each finite k rather than by subsets of size k. He assumed that for each $(t_1, \ldots, t_k) \in \mathbb{R}^{[\![k]\!]}$ there is a probability measure $\mu_{t_1,\ldots,t_k}$ on $\mathcal{B}(\mathbb{R}^{[\![k]\!]})$ for which

$$\mu_{t_1,\ldots,t_{k-1}}(H_1 \times H_2 \times \cdots \times H_{k-1}) = \mu_{t_1,\ldots,t_k}(H_1 \times H_2 \times \cdots \times H_{k-1} \times \mathbb{R}),$$

for each $(t_1, \ldots, t_k) \in \mathbb{R}^{[\![k]\!]}$ and $H_j \in \mathcal{B}(\mathbb{R})$ for each j. (Essentially this requirement is just <7> when restricted to a generating class for $\mathcal{B}(\mathbb{R}^k)$.) He also needed to assume

$$\mu_{t_1,\ldots,t_k}(H_1 \times H_2 \times \cdots \times H_k) = \mu_{t_{\pi(1)},\ldots,t_{\pi(k)}}\left(H_{\pi(1)} \times H_{\pi(2)} \times \cdots \times H_{\pi(k)}\right)$$

for every permutation π (that is, a bijection $\pi : [\![k]\!] \to [\![k]\!]$). The next Example shows why the second condition is needed.

<9> **Example.** BROWNIAN motion$\{B_t : t \in T\}$ for $T := [0,1]$ is a gaussian process with $\mathbb{P}B_t = 0$ for all t and $\operatorname{cov}(B_s, B_t) = \min(s,t)$ for all s,t. Equivalently, if $J = \{t_0, t_1, \ldots, t_k\}$ with $0 = t_0 < t_1 < \cdots < t_k \le 1$ then we can just require that the increments $B_{t_i} - B_{t_{i-1}}$ for $i \in [\![k]\!]$ be independent with $B_{t_i} - B_{t_{i-1}}$ distributed $N(0, t_i - t_{i-1})$. Either way, these assumptions completely determine a consistent family of fidis.

To be more specific, suppose $J = \{0, 1/3, 1/2, 3/4\}$. If we list the members of J as $t_0 = 0$, $t_1 = 1/3$, $t_2 = 1/2$, $t_3 = 3/4$ then put $X_i = B_{t_i}$ we get a joint distribution with (X_0, X_1, X_2, X_3) having a centered multivariate normal distribution with X_0 degenerate at 0. However, if we define $Y_0 = B_{3/4}$, $Y_1 = B_{1/2}$, $Y_2 = B_{1/3}$, $Y_3 = B_0$ then (Y_0, Y_1, Y_2, Y_3) has a different multivariate normal distribution: same J but different ways of referring to its members. I strongly feel that is better to have a single P_J rather than 24 different measures $\mu_{0,1/3,1/2,3/4}$ and $\mu_{0,1/3,1/2,3/4}$ and ... and $\mu_{3/4,1/2,1/3,0}$ on $\mathcal{B}(\mathbb{R}^4)$ obtained by permuting coordinates. The moral: It pays to have a notation that automatically keeps track of the coordinates.

It is usually also assumed that the sample paths B_ω are continuous. Contrary to what some authors might have you believe, this property does not follow automatically from specification of the fidis (UGMTP p214). Theorem <8> does give us a probability measure $\mathbb{P}$ on $\mathcal{C}(\mathbb{R}^T)$ with the given fidis. It might seem that we have only to check that $\mathbb{P}C[0,1] = 1$, where $C[0,1]$ denotes the set of all continuous functions on $[0,1]$. That approach dies immediately because $C[0,1] \notin \mathcal{C}\left(\mathbb{R}^{[0,1]}\right)$, a failure clearly pointed out by Doob (1947). Kolmogorov was also aware of this measurability problem—see his discussion near the start of §III.4.

In general, for an uncountably infinite T a simple argument (using the fact that a countable union of countable sets is still countable; see Billingsley, 1986, Theorem 36.3) shows that

$$\mathcal{C}(\mathbb{R}^T) = \cup\{\mathcal{C}_S : S \text{ a countable subset of } T\},$$

where $\mathcal{C}_S$ denotes the sigma-field on $\mathbb{R}^T$ generated by the countable set of coordinate projections $\{\pi_t : t \in S\}$. Consequently, every $\mathcal{C}(\mathbb{R}^T)$-measurable, real-valued function g on $\mathbb{R}^T$ must be of the form $G_S(x|_S)$ for some countable subset S and some $\mathcal{C}(\mathbb{R}^S)$-measurable function G_S on $\mathbb{R}^S$. Unfortunately the continuous members of $\mathbb{R}^T$ cannot be determined just by examining $x|_S$ for a single countable subset of T.

Most methods for getting a BROWNIAN motion with continuous sample paths need more than Theorem <8>. One approach uses a "DOOB-separable version" of the process constructed by Kolmogorov's theorem. (As you might guess from the name, this method was invented by Doob. See Section 9.6 for details.) Another approach builds the gaussian process by means of an isometry and an orthogonal series expansion. See UGMTP §9.3 for details.

Under the uniform metric, $d_u(x,y) := \sup_t |x(t) - y(t)|$, the set $C[0,1]$

becomes a complete, separable metric space with the useful property (Problem [6]) that its d_u-BOREL sigma-field is the same as the cylinder sigma-field generated by the coordinate projections $\{\pi_t : 0 \le t \le 1\}$. Thus a BROWNIAN motion with continuous sample paths defines a BOREL-measurable, random element of $C[0,1]$. The distribution of B_ω, as an $\mathcal{F}\backslash\mathcal{B}$-measurable map into $C[0,1]$, is called WIENER measure. It can serve as the probability measure Q in <5>. □

9.4 Essential supremum and infimum

The key ideas for handling stochastic processes with uncountable index sets came from Doob (1937) and Doob (1953, Chap II), who used a very neat measure-theoretic trick for reducing an uncountable collection of random variables down to a countable subcollection, at the cost of a host of negligible sets. The trick will be described in the current Section and the closely related constructions will be described in Sections 9.5 and 9.6.

For a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ write $\overline{\mathcal{M}}$ for the set of all $\mathcal{F}\backslash\mathcal{B}(\overline{\mathbb{R}})$-measurable maps from Ω into $\overline{\mathbb{R}} := [-\infty, +\infty]$.

The pointwise supremum H of a nonempty subset $\mathcal{H}$ of $\overline{\mathcal{M}}$ is the smallest function for which $H(\omega) \ge h(\omega)$ for all $\omega \in \Omega$ and all $h \in \mathcal{H}$. If $\mathcal{H}$ is uncountable then H need not be measurable. Indeed, there need not be a smallest member of $\overline{\mathcal{M}}$ satisfying the same pointwise inequalities. However, if we relax the pointwise requirement by allowing just an almost sure inequality, with the negligible set allowed to depend on h, there will be such an H in $\overline{\mathcal{M}}$. That is, there exists an H in $\overline{\mathcal{M}}$ for which:

(ESi) $\mathbb{P}\{H \ge h\} = 1$ for all h in $\mathcal{H}$.

(ESii) If $\mathbb{P}\{G \ge h\} = 1$ for all h in $\mathcal{H}$ then $\mathbb{P}\{G \ge H\} = 1$,

Property (ESii) ensures that $\mathbb{P}\{H = H_1\} = 1$ if H_1 is another member of $\overline{\mathcal{M}}$ satsifying the analogs of (ESi) and (ESii). Such an H is called an ***essential supremum of*** $\mathcal{H}$ and the equivalence class of such functions is denoted by $\operatorname{ess\,sup}\mathcal{H}$. The qualifier "essential" serves as a reminder that the negligible sets $\{\omega : H(\omega) < h(\omega)\}$ need not be empty. The next Theorem shows how to construct a member of $\operatorname{ess\,sup}\mathcal{H}$.

<10> **Theorem.** *Each nonempty subset $\mathcal{H}$ of $\overline{\mathcal{M}}$ (possibly uncountably infinite) has an essential supremum of the form $H_{\mathcal{G}}(\omega) := \sup_{h\in\mathcal{G}} h(\omega)$ for some countable subset $\mathcal{G}$ of $\mathcal{H}$.*

Proof. The defining properties (ESi) and (ESii) are unaffected by a one-to-one, increasing transformation (such as the arctan). Thus there is no loss of generality in assuming that $\sup_{\omega\in\Omega,\, h\in\mathcal{H}} |h(\omega)| \le c$ for some finite constant c.

Write $\mathbb{G}$ for the collection of all countable subsets of $\mathcal{H}$. The boundedness assumption ensures that $M_{\mathcal{H}} := \sup\{\mathbb{P}H_{\mathcal{G}} : \mathcal{G} \in \mathbb{G}\} < \infty$. Thus

there is a $\mathcal{G}_n$ in $\mathbb{G}$ for each n in $\mathbb{N}$ such that $\mathbb{P}H_{\mathcal{G}_n} > M_{\mathcal{H}} - n^{-1}$. The set $\mathcal{G}_\infty := \cup_n \mathcal{G}_n$ is countable; it belongs to $\mathbb{G}$. Define $H := H_{\mathcal{G}_\infty}$. Then $M_{\mathcal{H}} \geq \mathbb{P}H > M_{\mathcal{H}} - n^{-1}$ for every n, implying $\mathbb{P}H = M_{\mathcal{H}}$.

For property (ESi): if $h \in \mathcal{H}$ then $\mathcal{G} := \mathcal{G}_\infty \cup \{h\} \in \mathbb{G}$ so that

$$\mathbb{P}H = M_{\mathcal{H}} \geq \mathbb{P}H_{\mathcal{G}} = \mathbb{P}\max(h, H) = \mathbb{P}h\{h > H\} + \mathbb{P}H\{h \leq H\},$$

which rearranges to $0 \geq \mathbb{P}(h - H)\{h > H\}$, an inequality that can hold only if $\mathbb{P}\{h > H\} = 0$.

For property (ESii): $\{\omega : G(\omega) < H(\omega)\} = \cup_{h \in \mathcal{G}_\infty} \{\omega : G(\omega) < h(\omega)\}$, a countable union of $\mathbb{P}$-negligble sets. □

The ***essential infimum*** could be defined in a similar fashion, by changing all the $\geq$'s in (ESi) and (ESii) to $\leq$'s. Equivalently, just note that $K \in \operatorname{ess\,inf} \mathcal{H}$ iff $-K \in \operatorname{ess\,sup}\{-h : h \in \mathcal{H}\}$.

9.5 Upper/lower integrals and measurable cover functions

The ***outer measure*** of a subset A of Ω is defined as

$$\mathbb{P}^*A := \inf\{\mathbb{P}F : F \in \mathcal{F}_A\} \qquad \text{where } \mathcal{F}_A := \{F \in \mathcal{F} : F \supseteq A\}.$$

As the intersection of countably many sets in $\mathcal{F}_A$ also belongs to $\mathcal{F}_A$, Theorem <10> ensures that there exists an A^* in $\mathcal{F}_A \cap \operatorname{ess\,inf} \mathcal{F}_A$. That is, there exists an $\mathcal{F}$-measurable set A^* for which $A^* \supseteq A$, with *no almost sure qualification*, and $\mathbb{P}A^* = \mathbb{P}^*A$. Such an A^* is called a ***measurable cover*** for A. It is unique up to almost sure equivalence. Of course if $A \in \mathcal{F}$ then we can choose A^* equal to A.

Remark. If $N \in \mathcal{F}$ with $\mathbb{P}N = 0$ and $\emptyset \neq N \subset A$ then we have $H := A^* \backslash N \in \operatorname{ess\,inf} \mathcal{F}_A$ and $\mathbb{P}H = \mathbb{P}^*A$ but H does not include A as a subset.

Similarly, for each (possibly non-measurable) function f from Ω into $\overline{\mathbb{R}}$ the pointwise infimum of countable many functions in the set of measurable functions

$$\overline{\mathcal{M}}_{\geq f} := \{h \in \overline{\mathcal{M}} : h(\omega) \geq f(\omega) \text{ for all } \omega \in \Omega\}$$

also belongs to $\overline{\mathcal{M}}_{\geq f}$. Again, by Theorem <10>, there will exist an f^* in $\overline{\mathcal{M}}_f \cap \operatorname{ess\,inf} \overline{\mathcal{M}}_f$. Such an f^* is called a ***measurable cover*** for f. (I think the name "upper measurable envelope" would be better—cover suggests only set inclusion, as in $A^* \supseteq A$.) It too is unique only up to almost sure equivalence, being characterized by:

(MCi) $f^*(\omega) \geq f(\omega)$ for all $\omega \in \Omega$; that is, $f^* \in \overline{\mathcal{M}}_{\geq f}$.

(MCii) $\mathbb{P}\{h \geq f^*\} = 1$ for all $h \in \overline{\mathcal{M}}_{\geq f}$.

Remark. The difference between an inequality that holds $\mathbb{P}$-almost everywhere and an inequality that holds everywhere becomes quite important if we wish to take an expected value with respect to another probability measure whose negligible sets do not coincide with the $\mathbb{P}$-negligible sets.

The analog of the outer measure of a set is the ***upper integral*** of a function, but here the story gets a bit more complicated. For most functions f that map Ω into $\overline{\mathbb{R}}$, Dudley (2014, page 136) defined

<11> $$\mathbb{P}^* f := \inf\{\mathbb{P}h : h \in \overline{\mathcal{M}}_{\geq f} \text{ and } \mathbb{P}h \text{ is well defined}\},$$

allowing some $\mathbb{P}^* f$'s to remain undefined. You might be expecting $\mathbb{P}^* f = \mathbb{P} f^*$ but the dreaded $\infty - \infty$ can interfere. Consider, for example, the case where f is actually $\mathcal{F}$-measurable and $\mathbb{P}f^+ = \infty = \mathbb{P}f^-$, which leaves $f^* = f$ and $\mathbb{P}f^*$ undefined. Also in that case, if $h \in \overline{\mathcal{M}}_{\geq f}$ then $h^+ \geq f^+$ and $h^- \leq f^-$. If $\mathbb{P}h^- = \infty$ then $\mathbb{P}h$ is not defined; if $\mathbb{P}h^- < \infty$ then $\mathbb{P}h = \infty$. Thus $\mathbb{P}^* f = \infty$. Reasonable or not?

Fortunately this sort of thing cannot happen for an f bounded below (or above) by a finite constant, in which case we do have $\mathbb{P}^* f = \mathbb{P} f^*$.

Similar arguments and caveats apply to lower bounds. The lower measurable envelope can be defined similarly as an f_* in $\overline{\mathcal{M}}_{\leq f} \cap \operatorname{ess\,sup} \overline{\mathcal{M}}_{\leq f}$ where

$$\overline{\mathcal{M}}_{\leq f} := \{h \in \overline{\mathcal{M}} : h(\omega) \leq f(\omega) \text{ for all } \omega \in \Omega\},$$
$$\mathbb{P}_* f := \sup\{\mathbb{P}h : h \in \overline{\mathcal{M}}_{\leq f} \text{ and } \mathbb{P}h \text{ is well defined}\}.$$

Remark. Some authors refer to $\mathbb{P}^* h$ and $\mathbb{P}_* h$ as outer and inner integrals, by analogy with outer and inner measures for sets.

Apart from the $\infty - \infty$ difficulty, arguments involving upper integrals and measurable covers tend to be straightforward (but perhaps a little tedious) once the desired properties are reduced to pointwise assertions about measurable sets and measurable functions.

<12> **Example.** Here is an argument to show that $\mathbb{P}\{f^* > r\} = \mathbb{P}^*\{f > r\}$ for each real-valued function f on Ω and $r \in \mathbb{R}$. (Compare with Dudley, 2014, Lemma 3.8.)

Define $A := \{\omega : f(\omega) > r\}$, which need not be $\mathcal{F}$-measurable, and $D := \{\omega : f^*(\omega) > r\}$. Clearly $D \supseteq A$ because $f^* \in \overline{\mathcal{M}}_{\geq f}$. Choose any B with $A \subseteq B \in \mathcal{F}$ and $\mathbb{P}B = \mathbb{P}^* A$. Because $A \subseteq B \cap D \in \mathcal{F}$ we also have $\mathbb{P}(B \cap D) = \mathbb{P}^* A$, so we may as well assume $D \supseteq B$. It suffices to show $\mathbb{P}(D \backslash B) = 0$. To that end consider the measurable function

$$\begin{aligned} h(\omega) &:= f^*(\omega)\{\omega \in B\} + (r \wedge f^*(\omega))\{\omega \in B^c\} \\ &\geq f(\omega)\{\omega \in B\} + (r \wedge f(\omega))\{\omega \in B^c\} \qquad \text{because } f^* \geq f \\ &= f(\omega) \qquad \text{because } B^c \subseteq A^c \text{ and } f(\omega) \leq r \text{ on } A^c. \end{aligned}$$

It follows by property (MCii) that $h \geq f^*$ almost surely. However, for $\omega \in D\backslash B$ we must have $f^*(\omega) > r \geq h(\omega)$. Consequently, $\mathbb{P}(D\backslash B) = 0$. □

9.6 DOOB-separability: reduction to countable collections

For many probabilistic purposes, two random variables that differ only on a $\mathbb{P}$-negligble set are essentially the same. Indeed, many random variables can only be defined up to some sort of almost sure equivalence and, in isolation, there is usually no good reason to prefer one choice from the almost sure equivalence class over another. However, for a stochastic process $\{X_t : t \in T\}$ indexed by an uncountable T, if we need good behavior for the sample paths $X(\cdot, \omega)$ for each fixed ω then the selection from an equivalence class becomes much more important. In such situations the measurability difficulties can often be handled by taking limits along a countable dense subset of T if the process is "separable", in the following sense.

<13> **Definition.** *Let S be a countable, dense subset of a metric space (T, ρ). Say that a process $X = \{X_t : t \in T\}$ is* **doob-*separable, with approximating set*** *S, if every sample path $X(\cdot, \omega)$ has the property: for each t in $T\backslash S$ there exists a sequence $\{s_n : n \in \mathbb{N}\}$ (depending on ω and t) in S for which $\rho(s_n, t) \to 0$ and $X(s_n, \omega) \to X(t, \omega)$.* □

Remark. Of course the concept only makes sense for uncountable T.

The word "separable" is heavily overworked, particularly when the conversation also involves topological spaces that are separable in the sense of possessing a countable, dense subset. To avoid ambiguity, I have attached the name of its inventor.

Some authors refer to S as a "separant" or a "separating set". I am mostly following the exemplary exposition by Billingsley (1986, §38), who referred to processes "separable with respect to S" .

Many authors use a slightly more general notion of separability, by requiring that the approximation property need hold only for ω in N^c, for some $\mathbb{P}$-negligible set N. In that case, we have only to replace $X(\cdot, \omega)$ by a constant function for all ω in N to get a DOOB-separable process in the sense of the Definition <13>.

Not every process is DOOB-separable: the classic example is the stochastic process $X(t, \omega) = 1$ if $t = \omega$, and zero otherwise, when $T = \Omega = [0, 1]$. However, minor surgery can usually correct such difficulties.

<14> **Definition.** *Say that a stochastic process $\widetilde{X} = \{\widetilde{X}_t : t \in T\}$ is a version of $X = \{X_t : t \in T\}$ if they are both defined on the same probability space and $\mathbb{P}\{\omega : X_t(\omega) \neq \widetilde{X}_t(\omega)\} = 0$ for each t in T.* □

Under mild toplogical assumptions, a stochastic process X will have a version $\widetilde{X}$ that is DOOB-separable. For the traditional case where each X_t

is an $\mathcal{F}$-measurable map into the real line we must allow $\widetilde{X}_t$ to take values in $\overline{\mathbb{R}} := [-\infty, +\infty]$. See Problem [2]. Similarly, if X takes values in $\mathbb{R}^k$ we must allow $\widetilde{X}_t$ to take values in $\overline{\mathbb{R}}^k$. We can cover both cases just by assuming that X_t takes values in some compact metric space (E, d). Definitions <13> and <14> together require:

(DSVi) There exist $\mathbb{P}$-negligible sets $\{N_t : t \in T\}$ such that $\widetilde{X}(t,\omega) = X(t,\omega)$ for $\omega \in N_t^c$.

(DSVii) For each t in $T\backslash S$ there exists a sequence $\{s_j(\omega) : j \in \mathbb{N}\}$ in S with $\rho(s_j(\omega), t) \to 0$ and $d\left(\widetilde{X}(s_j(\omega),\omega), \widetilde{X}(t,\omega)\right) \to 0$ as $j \to \infty$.

<15> **Theorem.** *Suppose:*
(i) $(\Omega, \mathcal{F}, \mathbb{P})$ *is a complete probability space.*
(ii) (T, ρ) *is a separable metric space.*
(iii) (E, d) *is a compact metric space.*
(iv) $\{X_t : t \in T\}$ *is an* E*-valued stochastic process, that is, each* X_t *is an* $\mathcal{F}\backslash\mathcal{B}(E)$*-measurable map from* Ω *into* E.
□ *Then* X *has a* DOOB*-separable version.*

Remark. Remember that completeness of $\mathbb{P}$ means: if $N \subseteq F \in \mathcal{F}$ and $\mathbb{P}F = 0$ then $N \in \mathcal{F}$ with $\mathbb{P}N = 0$.

Proof. Let me start with some heuristics before adding the refinements needed for the asserted DOOB-separability. First we need to make use of the topological separability of both E and T. Let A be any countable, d-dense subset of E and T_0 be any countable, ρ-dense subset of T. The countable set $\mathcal{G}$ of all open balls with centers in T_0 and rational radii forms a base for the ρ-topology, meaning that for each open subset H of T and each t in H there exists some G in $\mathcal{G}$ for which $t \in G \subset H$. In particular, for each t in T there exists a sequence $\{G_{j,t} : j \in \mathbb{N}\}$ in $\mathcal{G}$ with $\operatorname{diam}(G_{j,t}) < 1/j$ and $G_{j,t} \searrow \{t\}$ as $j \nearrow \infty$.

Suppose for the moment that for each a in A and G in $\mathcal{G}$ there exists a countable dense subset $S(a, G)$ of G for which

<16> $$\inf_{s\in S(a,G)} X(s,\omega) = \inf_{t\in G} X(t,\omega) \qquad \text{for every } \omega \text{ in } \Omega.$$

Then $S := \cup\{S(a, G) : a \in A, G \in \mathcal{G}\}$ is a countable, ρ-dense subset of T.

Now focus on a particular (t,ω) pair. By the density of A there exists a sequence $a_j = a_j(t,\omega)$ in A such that $d(a_j, X(t,\omega)) < 1/j$. By <16> there exists an $s_j = s_j(t,\omega)$ in $S(a_j, G_{j,t})$ for which $d(a_j, X(s_j,\omega)) < 1/j$. Consequently $\rho(s_j, t) \to 0$ and $d\left(X(s_j,\omega), X(t,\omega)\right) < 2/j$, which gives the desired (DSVii) property to make S an approximating set.

To tighten up the argument we need to cast out a bunch of $\mathbb{P}$-negligible sets. To replace <16>, we should invoke the ess inf analog of Theorem <10> to choose $S(a,G)$ as a countable, ρ-dense subset of G for which

$$\inf_{s\in S(a,G)} d(a,X_s(\omega)) \in \operatorname{ess\,inf}\{d(a,X_t(\omega)) : t\in G\},$$

then define $S := \cup\{S(a,G) : a\in A, G\in \mathcal{G}\}$, as before. For any (t,a,G) triple with $t\in G$ we then have a $\mathbb{P}$-negligible set $N_{t,a,G}$ with

$$\inf_{s\in S(a,G)} d(a,X_s(\omega)) \le d(a,X_t(\omega)) \qquad \text{for } \omega\in N^c_{t,a,G}.$$

The set $N_t := \cup_{a,G}\{N_{t,a,G} : a\in A, t\in G\}$ is also $\mathbb{P}$-negligible. For ω in N_t^c the same argument as before gives a sequence $s_j = s_j(t,\omega)$ in S for which $\rho(s_j,t)\to 0$ and $d(X(s_j,\omega),X(t,\omega)) < 2/j$, which suggests we should define $\widetilde{X}(t,\omega) = X(t,\omega)$ if $\omega\in N_t^c$. For ω in N_t define $\widetilde{X}(t,\omega)$ to be a cluster point of the sequence $\{X(\sigma_j,\omega) : j\in\mathbb{N}\}$ for any sequence $\{\sigma_j : j\in\mathbb{N}\}$ in S that converges to t. □

In one special case, which applies to most of the processes considered in the next six Chapters, we can simplify the ess inf argument. Suppose the process $X = \{X_t : t\in T\}$ is continuous in probability, that is, for each t in T and each $\epsilon>0$, we have $\mathbb{P}\{d(X_{t_n},X_t)>\epsilon\}\to 0$ if $\rho(t_n,t)\to 0$. Then for every countable, ρ-dense subset S of T, there exists a version of the process that is DOOB-separable *with S as the approximating set*. See Problem [3] for the simple proof.

9.7 Convergence in distribution without BOREL measurability

Suppose $X^{(n)} := \{X^{(n)}(t,\cdot) : t\in T\}$ for $n\in\mathbb{N}$ is a sequence of stochastic processes, all defined on some $(\Omega,\mathcal{F},\mathbb{P})$, whose sample paths all belong to the set $\mathbb{B}(T)$of all bounded, real-valued functions on an index set T. Under its uniform metric,

<17> $$d_u(x,y) := \sup_{t\in T}|x(t)-y(t)| \qquad \text{for } x,y\in\mathbb{B},$$

the set $\mathbb{B}(T)$ becomes a complete metric space with BOREL-sigma-field $\mathcal{B} := \mathcal{B}(\mathbb{B}(T),d_u)$ generated by the d_u-open sets. If the sample path map $\omega\mapsto X^{(n)}_\omega := X^{(n)}(\cdot,\omega)$ is not $\mathcal{F}\backslash\mathcal{B}$-measurable then $\omega\mapsto f(X^{(n)}_\omega)$ need not be $\mathcal{F}\backslash\mathcal{B}(\mathbb{R})$-measurable for all f in $\mathrm{BL}(\mathbb{B}(T),d_u)$, so that the defining property <5> for $X^{(n)}\rightsquigarrow Q$ makes no sense.

Fortunately we can fix this problem by adding a few stars (upper/lower integrals) to the definitions in Section 9.2 and by explicity requiring some extra regularity for the limit distribution.

Remark. Dudley (1966, 1967) recognized that non-separabilty (in the metric space sense) lay behind the failure of BOREL-measurability, as in the case of the empirical process ν_n from Example <1>, whose sample paths belong to the space $D[0,1]$ but are not measurable with respect to its d_u-BOREL sigma-field. He suggested that measurability with respect to the sigma-field $\mathcal{U}$ generated by the d_u-open balls would suffice for processes, so that their distributions live on $\mathcal{U}$. For probability measures μ_n on $\mathcal{U}$ and a BOREL measure μ, he defined $\mu_n \rightsquigarrow \mu$ to mean $\mu_n^* f \to \mu f$ for all bounded, continuous real-valued functions. Clearly his proposed fix inspired the approach taken in Definition <18>, although Dudley (2014, pp173–174) seemed to be modestly giving all the credit to Hoffmann-Jørgensen.

<18> **Definition.** *For a probability measure Q on the* BOREL *sigma-field $\mathcal{B} := \mathcal{B}(\mathbb{B}(T), d_u)$, define $X^{(n)} \overset{*}{\rightsquigarrow} Q$ to mean that $\mathbb{P}^* f(X_\omega^{(n)}) \to Qf$ for each f in the set $\mathrm{BL}(\mathbb{B}(T), d_u)$ of all bounded, real-valued functions on $\mathbb{B}(T)$ that are* LIPSCHITZ*-continuous with respect to d_u.* □

Remark. Sometimes it simplifies notation to replace Q by a process $\{X^{(\infty)}(t,\cdot) : t \in T\}$ with distribution Q. We could require that all the sample paths $X_\omega^{(\infty)}$ belong to the separable set $\cup_m K_m$, thereby ensuring that $\omega \mapsto X_\omega^{(\infty)}$ is $\mathcal{F}\backslash\mathcal{B}(B(T), d_u)$-measurable. In that case $X^{(n)} \overset{*}{\rightsquigarrow} X^{(\infty)}$ means the same thing as $X^{(n)} \overset{*}{\rightsquigarrow} Q$.

Definition <18> could also be written using lower integrals because, if $X^{(n)} \overset{*}{\rightsquigarrow} Q$, then $-\mathbb{P}_* f(X_\omega^{(n)}) = \mathbb{P}^*[-f(X_\omega^{(n)})] \to Q(-f)$ for $f \in \mathrm{BL}(\mathbb{B}, d_u)$. That is, $\mathbb{P}_* f(X_\omega^{(n)}) \to Qf$.

If $X^{(n)} \overset{*}{\rightsquigarrow} Q$ then the consequences (CID.j) for $j \in \{i, ii, iii\}$ from Section 9.2 also gain a few stars:

(CID*.i) $\liminf_{n\to\infty} \mathbb{P}_* f(X_\omega^{(n)}) \geq Qf$ for all lower-semicontinuous functions f that are bounded below by a (finite) constant.

(CID*.ii) $\limsup_{n\to\infty} \mathbb{P}^* f(X_\omega^{(n])}) \leq Qf$ for all upper-semicontinuous functions f that are bounded above by a (finite) constant.

(CID*.iii) $\mathbb{P}^* f(X_\omega^{(n)})) \to Qf$ and $\mathbb{P}_* f(X_\omega^{(n)})) \to Qf$ for all (two-sided) bounded real-valued functions on $\mathcal{X}$ with $Q\{x \in \mathcal{X} : f \text{ is continuous at } x\} = 1$.

The proofs from UGMTP §7.1 carry over with only minor notational changes. Look at van der Vaart and Wellner (1996, p18) if you want to see the $*$ details explicitly.

Amazingly (Dudley, 2014, Appendix C): "It is consistent with the usual axioms of set theory (including the axiom of choice) that [statement about measurable cardinals]. So, for practical purposes, a probability measure defined on the Borel sets of a metric space is always concentrated in some separable subspace", a result due to Marczewski and Sikorski (1948).

Once we have separable support an easy argument using the completeness of $\mathbb{B}$ under d_u shows (Dudley, 2003, p225) that Q must be ***tight***: *for each $\epsilon > 0$ there exists a compact subset K of $\mathbb{B}(T)$ such that $QK > 1 - \epsilon$.* In particular, there exists a sequence of compact subsets $\{K_m : m \in \mathbb{N}\}$ of $\mathbb{B}(T)$ such that $QK_m > 1 - m^{-1}$, implying $Q(\cup_m K_m) = 1$. Tightness also gives us (Problem [4]) a rather surprising fact:

(i) There exists a semimetric ρ on T under which T is totally bounded.

(ii) Every x in $\cup_m K_m$ is ρ-uniformly continuous, that is: for each $\eta > 0$ there exists a $\delta > 0$ for such that

<19> $$\text{OSC}(\delta, x, T) := \sup\{|x(s) - x(t)| : s, t \in T \text{ and } \rho(s,t) < \delta\} < \eta.$$

In fact $\delta := \eta/2^m$ works for all x in K_m.

In summary: tightness of Q implies, for each $\eta > 0$,

<20> $$Q\{x \in \mathbb{B}(T) : \text{OSC}(\delta, x, T) \geq \eta\} \to 0 \qquad \text{as } \delta \to 0.$$

The OSC in <19> stands for ***oscillation***, a uniform way to describe how much x might jump about in (small) neighborhoods of each point in T. A two line argument shows that

$$|\text{OSC}(\delta, x, T) - \text{OSC}(\delta, y, T)| \leq 2d_u(x, y) \qquad \text{for } x, y \in \mathbb{B}(T).$$

Consequently, $\text{OSC}(\delta, \cdot, T)$ is d_u-uniformly continuous on $\mathbb{B}(T)$.

In short, the first line of Definition <18> could be replaced by the assumption that Q is a tight probability measure on $\mathcal{B}(\mathbb{B}(T), d_u)$ that concentrates on the set

<21> $$\text{UC}(T, \rho) = \{x \in \mathbb{B}(T) : \text{OSC}(\delta, x, T) \searrow 0 \text{ as } \delta \searrow 0\}$$

of all ρ-uniformly continuous functions in $\mathbb{B}(T)$, for some semimetric ρ on T under which T is totally bounded.

The semimetric ρ will not usually be constructed by the method described in Problem [4]. The next Example illustrates the more typical situation.

<22> **Example.** Suppose $\xi_1, \ldots, \xi_n$ are independent observations on a BOREL probability measure P on the real line. By analogy with the empirical process from Example <1>, define

$$\nu_n(t, \omega) := n^{-1/2} \sum\nolimits_{i \leq n} \Big(\{\xi_i(\omega) \leq t\} - P(-\infty, t] \Big) \qquad \text{for } t \in T = \mathbb{R}.$$

Then $\mathbb{P}\nu_n(t) = 0$ for all t and

$$\text{cov}\,(\nu_n(s), \nu_n(t)) = \sigma(s,t) := P(-\infty, s \wedge t] - P(-\infty, s]P(-\infty, t].$$

The semimetric

$$\rho(s,t) := \sqrt{P(s,t] - (P(s,t])^2} \qquad \text{for } s \le t$$

controls the increments of this process because $\operatorname{var}(\nu_n(s) - \nu_n(t)) = \rho(s,t)^2$.

The limit process will be a centered gaussian process $\{G_t : t \in \mathbb{R}\}$ with $\operatorname{cov}(G_s, G_t) = \rho(s,t)$, with bounded sample paths that are uniformly continuous with respect to ρ. The existence of such a gaussian process will be proved in Chapter 10.

Remark. For this particular case we could also exploit the ordering of the real line by defining $G_t = B_0(F(t))$ with B_0 a BROWNIAN bridge, as in Example <1>, and $F(t) := P(-\infty, t]$. However this quick fix has no obvious extension if $\mathbb{R}$ is replaced by another space such as $\mathbb{R}^k$. I think it is better to invoke the methods from Chapter 10 through 12 than to repeat the history that for so long kept empirical processes theory tied to the real line.

Continuity with respect to ρ need not coincide with continuity with respect to the usual euclidean distance on $\mathbb{R}$. For Example, if there were some point a with $P\{a\} = \theta > 0$ then $\rho(a-\delta, a) \to \sqrt{\theta - \theta^2} > 0$ as $\delta \searrow 0$. All the sample paths of G would have a jump at a in the euclidean sense but not in the ρ sense because $a - \delta$ does not converge to a in the ρ sense. In addition, if P happened to give zero measure to some interval $[a,b]$ then $\rho(s,t) = 0$ for $a \le s \le t \le b$ and

$$\mathbb{P}\{\nu_n(t) = \nu_n(a) \text{ for } a \le t \le b\} = 1 = \mathbb{P}\{G_t = G_a \text{ for } a \le t \le b\}.$$

□ Consequently ρ would be only a semimetric, not a metric, for such a P.

Remark. In the final paragraph of the previous Example we could eliminate the flat spot by removing the interval (a,b) from the index set. This surgery would turn ρ into a metric, provided there were no more open intervals with zero P measure.

A more systematic approach would be to partition the original index set into equivalence classes of points at zero ρ distance then construct a new index set consisting of a single point from each equivalence class. This approach fits nicely with results from earlier in this Chapter, such as Theorem <15>, provided we started from a stochastic process $\{X(t,\cdot) : t \in T\}$ for which $X_s = X_t$ almost surely whenever $\rho(s,t) = 0$. Indeed, it would be a good exercise to determine which Definitions and Theorems should be modified if it were possible to have $\rho(s,t) = 0$ with $s \ne t$. Hint: Look at the definition of $\|\cdot\|_{\text{Lip}}$.

By d_u-continuity of the oscillation function and <21>, for each $\delta > 0$ and $\eta > 0$ the set $F_{\delta,\eta} := \{x \in \mathbb{B}(T) : \text{OSC}(\delta, x, T) \ge \eta\}$ is d_u-closed and $F_{\delta,\eta} \cap \text{UC} \searrow \emptyset$ as $\delta \searrow 0$. By inequality (CID*.ii) and the fact that

indicator functions of closed set are upper-semicontinuous, if $X^{(n)} \overset{*}{\rightsquigarrow} Q$ and $\delta = \delta(\epsilon, \eta)$ is small enough then

<23> $$\limsup_n \mathbb{P}^*\{X^{(n)}_\omega \in F_{\delta,\eta}\} \le QF_{\delta,\eta} < \epsilon.$$

Thus a necessary condition for $X^{(n)} \overset{*}{\rightsquigarrow} Q$ is: for every $\eta > 0$ and $\epsilon > 0$ there exists a $\delta > 0$ such that

<24> $$\limsup_n \mathbb{P}^*\{\text{OSC}(\delta, X^{(n)}_\omega, T) \ge \eta\} < \epsilon$$

In the literature condition <24> also goes by the name ***asymptotic equicontinuity*** or ***asymptotic tightness***. See van der Vaart and Wellner (1996, §1.5), for example. As you will soon learn from Theorem <26>, this condition plus convergence of the finite-dimensional distributions together imply convergence in the sense of Definition <18>.

Of course for finite subsets J of T no stars are needed when describing the convergence in distribution of the projection $\pi_J X^{(n)}$, a random element of $\mathbb{R}^J$, because the map $\omega \mapsto \pi_J X^{(n)}_\omega \in \mathbb{R}^J$ inherits $\mathcal{F}$-measurability from the assumed measurability of each X_t. We'll be needing these random vectors to converge in distribution, in the sense of <5>, to $\pi_J Q$, the image of the measure Q under π_J defined by

$$(\pi_J Q)B := Q\{x \in \mathbb{B}(T) : x|_J \in B\} \qquad \text{for each } B \in \mathcal{B}(\mathbb{R}^J).$$

To emphasize these distinctions I'll write

<25> $$X^{(n)} \rightsquigarrow_{\text{fidi}} Q \qquad \text{to mean } \pi_J X^{(n)} \rightsquigarrow Q_J := \pi_J Q \text{ in the sense } <5>$$

for each finite subset J of T, while $X^{(n)} \overset{*}{\rightsquigarrow} Q$ refers to the stronger property from Definition <18>.

The remaining six Chapters of this book will largely focus on methods for proving the maximal inequalities needed to check things like <24> for many useful processes. As far as "uniform central limit theorems" (aka DONSKER theorems) are concerned, with only slight exaggeration I claim that much of the field of empirical process theory inspired by Dudley (1978) rests on methods developed to establish special cases of <24>.

9.8 An oscillation condition for convergence in distribution

The oscillation function from <19> plays a central role in establishing that $X^{(n)} \overset{*}{\rightsquigarrow} Q$. It is the key to transforming fidi convergence into convergence in distribution of processes, as shown by the following Theorem.

<26> **Theorem.** *Let $\{X^{(n)}(t, \cdot) : t \in T, n \in \mathbb{N}\}$ be a sequence of stochastic processes with sample paths in $\mathbb{B}(T)$ for some totally bounded semimetric space (T, ρ). Suppose:*

(i) For each finite subset J of T there is a probability measure Q_J on $\mathcal{B}(\mathbb{R}^J)$ such that $\pi_J X^{(n)} \rightsquigarrow Q_J$.

(ii) For each $\eta > 0$ and $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\limsup_n \mathbb{P}^*\{\text{OSC}(\delta, X^{(n)}_\omega, T) \geq \eta\} < \epsilon.$$

Then there exists a probability measure Q concentrated on the set $\text{UC}(T,\rho)$ with $\pi_J Q = Q_J$ for each finite subset J of T such that $X^{(n)} \overset{*}{\rightsquigarrow} Q$. □

Remark. The statement of the Theorem makes no mention of where ρ comes from. As Example <22> suggests, the fidi convergence of the $X^{(n)}$ typically tells us the covariances we should expect for Q, particularly so if we rely on a multivariate CLT to establish the fidi convergence. In that case Q will be the distribution of a centered gaussian process. Compare with Dudley (2014, Thm 3.34), where gaussianity appears as a part of the definition of a DONSKER class.

The Proof of the Theorem will involve three steps:

1) Prove the existence of Q by means of Theorem <8> (Kolmogorov's existence theorem) plus an appeal to DOOB-separability.

2) Show how to approximate uniformly continuous functions by combining total boundedness, approximation of projections into $\mathbb{R}^J$, and oscillation control (Lemma <27>).

3) Use the results from 1) and 2) to approximate $\mathbb{P}^* f(X^{(n)}_\omega)$ for a bounded LIPSCHITZ function f.

Step 1: existence of Q. For finite subsets $K \subset J$ of T and $g \in \text{BL}(\mathbb{R}^K, \ell_2)$ the composition $f := g \circ \pi_{J\to K}$ belongs to $\text{BL}(\mathbb{R}^J, \ell_2)$: for $x, y \in \mathbb{R}^J$,

$$\begin{aligned}|f(x) - f(y)| &= |g(\pi_{J\to K}x) - g(\pi_{J\to K}y)| = |g(x|_K) - g(y|_K)| \\ &\leq \|g\|_{\text{BL}} \min\Big(1, \big|\, x|_K - y|_K \,\big|_2\Big) \leq \|g\|_{\text{BL}} \left(1 \wedge |x-y|_2\right).\end{aligned}$$

Thus $Q_J f = \lim_n \mathbb{P} g(\pi_{J\to K} X^{(n)}_\omega) = Q_K g$, the consistency condition $Q_K = \pi_{J\to K} Q_J$ for Theorem <8>. There exists a probability measure $\mathbb{Q}$ on the cylinder sigma-field of $\mathbb{R}^T$ with $\pi_J \mathbb{Q} = Q_J$ for each finite J.

As explained in Section 9.3, we cannot assert that $\mathbb{Q}\text{UC}(T,\rho)$ equals 1. However, we can think of $\{x(t) : t \in T\}$ as a stochastic process defined on $(\mathbb{R}^T, \mathcal{C}(\mathbb{R}^T), \mathbb{Q})$, so that by Theorem <15> it has a DOOB-separable version $\{X^{(\infty)}(t,\cdot) : t \in T\}$ with fidis Q_J. Let S be the approximating set for $X^{(\infty)}$. For each finite subset J of S the fidi convergence tells us that

$$\text{OSC}(\delta, X^{(n)}_\omega, J) := \max\{|X^{(n)}(s_1,\omega) - X^{(n)}(s_2,\omega)| : s_i \in J,\, \rho(s_1,s_2) < \delta\}$$

converges in distribution to $\text{OSC}(\delta, X^{(\infty)}_\omega, J)$, implying

$$\epsilon > \liminf_n \mathbb{P}\{\text{OSC}(\delta, X^{(n)}_\omega, J) > \eta\} \geq \mathbb{Q}\{\text{OSC}(\delta, X^{(\infty)}, J) > \eta\},$$

by virtue of (CID.i) from Section 9.2.

Remark. You should convince yourself, using the continuity of OSC, that $G_{\delta,\eta} := \{y \in \mathbb{R}^J : \text{OSC}(\delta, y, J) > \eta\}$ is an open subset of $\mathbb{R}^J$. I needed to replace the closed set $F_{\delta,\eta}$ from <23> by the open $G_{\delta,\eta}$ to reverse the direction of the inequality—a common trick.

If we let J expand up to S we get $\mathbb{Q}\{\text{OSC}(\delta, X^{(\infty)}, S) > \eta\} \le \epsilon$, which DOOB-separability translates into $\mathbb{Q}\{\text{OSC}(\delta, X^{(\infty)}, T) > \eta\} \le \epsilon$. It follows that Q, the distribution of $X_\omega^{(\infty)}$, concentrates on the set $\text{UC}(T, \rho)$.

Step 2: approximation of uniformly continuous functions. The combination of a totally bounded (T, ρ) and processes whose sample paths lie close to $\text{UC}(T, \rho)$ makes it easy to approximate sample paths from their values at a fixed finite subset of T, thereby obtaining good control of the processes by means of their finite-dimensional distributions. The following elementary Lemma provides the main technical tool.

<27> **Lemma.** *For a given $\delta > 0$ let J be a finite subset of T with the property that $\min_{\tau\in J} \rho(t,\tau) \le \delta/2$ for all t in T. Then:*

(a) There exist nonnegative functions $\{\psi_\tau : \tau \in J\}$ in $\text{UC}(T,\rho)$ with $\sum_{\tau\in J} \psi_\tau(t) = 1$ for each t.

(b) The function $A_J(y,t) := \sum_{\tau\in J} y_\tau \psi_\tau(t)$ maps $\mathbb{R}^J$ into $\text{UC}(T,\rho)$ and satisfies $d_u(A_J(y,\cdot), A_J(z,\cdot)) \le \max_{\tau\in J} |y_\tau - z_\tau|$.

(c) $d_u(x, A_J(x|_J)) \le \text{OSC}(\delta, x, T)$ for each x in $\mathbb{B}(T)$. Consequently,

$$|f(x) - f(A_J(x|_J))| \le \|f\|_{\text{BL}} \left[1 \wedge \text{OSC}(\delta, x, T)\right] \qquad \text{for each } f \in \text{BL}(\mathbb{B}, d_u).$$

□

Proof. For each τ in J define $f_\tau(\cdot) := (1 - \rho(\cdot,\tau)/\delta)^+$. Note that

$$|f_\tau(s) - f_\tau(t)| \le |\rho(s,\tau) - \rho(t,\tau)|/\delta \le \rho(s,t)/\delta,$$

which ensures $f_\tau \in \text{UC}(T,\rho)$. The sum $F(t) := \sum_{\tau\in J} f_\tau(t)$ inherits uniformly continuity: $|F(s) - F(t)| \le |J|\rho(s,t)/\delta$. The approximating property of J ensures that for each t there is a τ for which $\rho(t,\tau) \le \delta/2$ so that $f_\tau(t) \ge 1/2$ and $F(t) \ge 1/2$.

For (a): define $\psi_\tau(t) := f_\tau(t)/F(t)$ to get $\sum_\tau \psi_\tau(t) = 1$. The uniform coninuity comes from:

$$\begin{aligned} |\psi_\tau(s) - \psi_\tau(t)| \times F(s)F(t) &= |f_\tau(s)F(t) - f_\tau(t)F(s)| \\ &\le |f_\tau(s) - f_\tau(t)|F(t) + f_\tau(t)|F(t) - F(s)|, \end{aligned}$$

implying $|\psi_\tau(s) - \psi_\tau(t)| \le (2 + |J|)\rho(s,t)/\delta$.

For (b): $A_J(y,\cdot)$ inherits uniform continuity from the $\psi_\tau(\cdot)$'s and

$$|A_J(y,t) - A_J(z,t)| \le \sum\nolimits_\tau \psi_\tau(t) \max_\tau |y_\tau - z_\tau| \le \max_\tau |y_\tau - z_\tau|.$$

Take a supremum over t.

For (c): consider any x in $\mathbb{B}(T)$ and any t in T. From the fact that $\sum_\tau \psi_\tau = 1$ we get

$$|x(t) - \sum\nolimits_\tau x(\tau)\psi_\tau(t)| \le \sum\nolimits_\tau |x(t) - x(\tau)|\psi_\tau(t).$$

As $\psi_\tau(t) = 0$ if $\rho(t,\tau) \ge \delta$, only those τ with $\rho(t,\tau) < \delta$ contribute to the sum. For those τ's we have $|x(t) - x(\tau)| \le \text{OSC}(\delta, x, T)$. The consequence for f comes from property <6> for bounded LIPSCHITZ functions. □

Step 3: putting it all together. We need to show that $\mathbb{P}^* f(X_\omega^{(n)}) \to Qf$ for each fixed f in $\text{BL}(\mathbb{B}, d_u)$. To simplify notation assume $\|f\|_{\text{BL}} \le 1$.

Step 2) will turn an approximation on a finite subset J into a uniform approximation. For any given $\eta > 0$ and $\epsilon > 0$ choose $\delta > 0$ as in assumption (ii) of Theorem <26> then choose a finite subset J as in Lemma <27>. By <20> we may assume δ is small enough that

<28> $$Q\{\text{OSC}(\delta, x, T) \ge \eta\} < \epsilon.$$

By Lemma <27>(b), $g_J := f \circ A_J$ is a bounded LIPSCHITZ function for any of the usual norms on $\mathbb{R}^J$. Thus fidi convergence gives

<29> $$\mathbb{P} g_J(\pi_J X_\omega^{(n)}) < \epsilon + Q^x g_J(x|_J) \qquad \text{for all large enough } n.$$

Part (c) of the same Lemma with $\|f\|_{\text{BL}} \le 1$ gives

$$\begin{aligned} |f(x) - g_J(x|_J)| &\le 1 \wedge \text{OSC}(\delta, x, T) \qquad \text{for each } x \text{ in } \mathbb{B}(T) \\ &\le \eta + \{\text{OSC}(\delta, x, T) \ge \eta\}. \end{aligned} \tag{<30>}$$

Equations <28>, <29>, and <30> provide everything needed to complete the Proof of the Theorem:

$$\begin{aligned} \mathbb{P}^* f(X_\omega^{(n)}) &\le \mathbb{P} g_J(X_\omega^{(n)}|_J) + \eta + \mathbb{P}^*\{\text{OSC}(\delta, X_\omega^{(n)}, T) \ge \eta\} \\ &\le \epsilon + Q g_J(x|_J) + \eta + \epsilon \qquad \text{for all large enough } n \\ &\le Qf + 2\eta + Q\{\text{OSC}(\delta, x, T) \ge \eta\} + 2\epsilon \\ &\le Qf + 2\eta + 3\epsilon. \end{aligned}$$

It follows that $\limsup_n \mathbb{P}^* f(X_\omega^{(n)}) \le Qf$ because the argument works for all $\eta > 0, \epsilon > 0$. The same argument applied to $-f$ delivers the companion lower bound:

$$\limsup_n \mathbb{P}^*(-f(X_\omega^{(n)})) = -\liminf_n \mathbb{P}_* f(X_\omega^{(n)}) \le -Qf,$$

or $\liminf_n \mathbb{P}_* f(X_\omega^{(n)}) \ge Qf$. Thus $\mathbb{P}^* f(X_\omega^{(n)}) \to Qf$.

9.9 Problems

[1] Suppose h is a bounded real function on Ω and $(\Omega, \mathcal{F}, \mathbb{P})$ is complete. Show that h is $\mathcal{F}$-measurable if and only if $\mathbb{P}^* h = \mathbb{P}_* h$.

[2] (from Billingsley, 1986, Ex 38.8) Suppose U is UNIF$[0,1]$ distributed and a process $\{X_t : t \in [0,1]\}$ is defined by

$$X(t,\omega) := 0\{t = U(\omega)\} + |U(\omega) - t|^{-1}\{t \neq U(\omega)\}.$$

(i) Show that X is not DOOB-separable.

(ii) Let $\{\widetilde{X}_t : t \in [0,1]\}$ be a DOOB-separable version of X, with approximating set S and exceptional $\mathbb{P}$-negligible set N. Show that $\widetilde{X}(t,\omega)$ must equal ∞ for $t = U(\omega)$ when $\omega \notin N$.

[3] Suppose a DOOB-separable process $X = \{X_t : t \in T\}$, with separating set S, is actually continuous in probability. That is, for each t in T and each sequence $\{t_n : n \in \mathbb{N}\}$ converging to t the sequence $\{X_{t_n}\}$ converges in probability to X_t. Let R be any countable ρ-dense subset of T. Justify the following assertions to show that the process is also DOOB-separable with respect to R.

(i) Consider an s in S. There exists a sequence $\{r_j(s) : j \in \mathbb{N}\}$ in R for which $\rho(r_j(s), s) \leq j^{-1}$ and $\mathbb{P}\{d(X(r_j(s),\omega), X(s,\omega)) > j^{-1}\} < 2^{-j}$. Thus (BOREL-CANTELLI?) there exists a $\mathbb{P}$-negligible set N_s such that $d(X(r_j(s),\omega), X(s,\omega)) \leq j^{-1}$ for all large enough j if $\omega \in N_s^c$.

(ii) Consider a t in T and $\omega \in (\cup_{s\in S} N_s)^c$. Then there exists $s_n = s_n(\omega, t)$ such that $\rho(s_n, t) \to 0$ and $d(X(s_n,\omega), X(t,\omega)) \to 0$. By (i) there exists $r_n = r_n(s_n,\omega) \in R$ such that $\rho(r_n, s_n) < n^{-1}$ and $d(X(r_n,\omega), X(s_n,\omega)) < n^{-1}$. It follows that $\rho(r_n, t) \to 0$ and $d(X(r_n,\omega), X(t,\omega)) \to 0$.

(iii) For ω in the negligible set $\cup_{s\in S} N_s$, redefine $X(\cdot,\omega)$ appropriately.

> For the following Problems, $\mathbb{B}(T)$ denotes the set of all bounded, real-valued functions on a set T equipped with its uniform norm, $\|x\|_u := \sup_{t\in T} |x(t)|$ and uniform metric $d_u(x,y) := \|x - y\|_u$. If ρ is a semimetric on T then $\mathrm{UC}(T,\rho)$ denotes the set of all ρ-uniformly continuous members of $\mathbb{B}(T)$.

[4] Suppose Q is a tight probability measure on $\mathcal{B} := \mathcal{B}(\mathbb{B}(T), d_u)$, with an increasing sequence of compact sets $\{K_m : m \in \mathbb{N}\}$ for which $QK_m > 1 - 1/m$. Then there exists a semimetric ρ on T for which:

(a) $x \in \mathrm{UC}(T,\rho)$ for each x in $\cup_{m\in\mathbb{N}} K_m$, a set with Q probability 1.

(a) T is totally bounded under ρ.

Prove (a) and (b) by justifying the following assertions. (You might find it helpful to glance at §10.4 before handling the total boundedness.)

(i) For each m define $\rho_m(s,t) := \sup_{x\in K_m} |x(s)-x(t)|$. By compactness, there exists a finite subset F_m of K_m for which $\min_{z\in F_m} \|x-z\|_u \le 1$ for each x in K_m. Thus

$$\|x\|_u \le C_m := 1 + \max_{z\in F_m} \|z\|_u \qquad \text{for every } x \text{ in } K_m.$$

It follows that $\rho_m(s,t)$ is finite for all $s,t \in T$ and ρ_m is a semimetric on T. Note that $\rho_m(s,t) \ge \rho_{m-1}(s,t)$ because $K_{m-1} \subseteq K_m$.

(ii) Combine the ρ_m into single semimetric:

$$\begin{aligned}\rho(s,t) &:= \sum\nolimits_{m\in\mathbb{N}} 2^{-m} \min(1,\rho_m(s,t)) \\ &\le \sum\nolimits_{j\le m} 2^{-j}\rho_m(s,t) + \sum\nolimits_{j>m} 2^{-j} \le \rho_m(s,t) + 2^{-m}\end{aligned}$$

for each m.

(iii) For any fixed m, if $x \in K_m$ and $\rho(s,t) < \epsilon/2^m$ for some ϵ in $(0,1)$ then $1 \wedge \rho_m(s,t) < \epsilon$, implying $|x(s)-x(t)| \le \rho_m(s,t) < \epsilon$. Thus $x \in \text{UC}(T,\rho)$.

(iv) For any given $\delta > 0$ let S be a finite subset of T that is 4δ-separated, meaning that $\rho(s,t) > 4\delta$ for distinct s,t in S. To show that T is totally bounded it suffices (cf. §10.4) to find finite constant C_δ such that $|S| \le C_\delta$ for all such S. Choose m so that $2^{-m} < \delta$ then choose a finite subset $\{z_i : i \in [\![k]\!]\}$ of K_m for which $\min_i d_u(x,z_i) \le \delta$ for each x in K_m. Define $Z(t) := (z_1(t),\ldots,z_k(t))$, a point in the compact subset $[-C_m,C_m]^k$ of $\mathbb{R}^k$. Then for all $s,t \in T$,

$$\rho_m(s,t) \le 2\delta + \max_i |z_i(s)-z_i(t)| = 2\delta + |Z(s)-Z(t)|_\infty,$$

where $|\cdot|_\infty$ denotes the uniform norm on $\mathbb{R}^k$. For distinct s,t in S we have

$$4\delta < \rho(s,t) \le \delta + \rho_m(s,t) \le 3\delta + |Z(s)-Z(t)|_\infty.$$

That is, $\{Z(s) : s \in S\}$ is a δ-separated subset of $[-C_m,C_m]^k$. It follows that $|S| \le ((2C_m+\delta)/\delta)^k$. Hint: The sets $\{w \in \mathbb{R}^k : |w-Z(s)|_\infty \le \delta/2\}$ for $s \in S$ are disjoint and each has LEBESGUE measure δ^k.

[5] Suppose (T,ρ) is a semimetric space and $\overline{T}$ is its completion.

(i) Show that every function x in $\text{UC}(T,\rho)$ has a unique extension to a function in $\text{UC}(\overline{T},\rho)$. (Here ρ denotes both the semimetric on T and its unique extension to a semimetric on $\overline{T}$.)

(ii) Show that $\overline{T}$ is compact if T is totally bounded.

[6] Justify the following assertions to understand why $\mathbb{U} := \text{UC}(T,\rho)$ is a closed, separable subset of $\mathbb{B}(T)$ whose BOREL sigma-field is generated by the coordinate projections.

(i) $\mathbb{B}(T)$ is complete under the d_u metric and each coordinate projection π_t is d_u-continuous. Thus $\mathcal{C}(\mathbb{R}^T) \subseteq \mathcal{B}(\mathbb{B}(T), d_u))$ and $\mathcal{C}(\mathbb{U}) \subseteq \mathcal{B}(\mathbb{U}, d_u)$.

(ii) If $\{x_n : n \in \mathbb{N}\} \subset \mathbb{U}$ and $d_u(x_n, y) \to 0$ then $y \in \mathbb{U}$. Thus $\mathbb{U}$ is a closed subset of $\mathbb{B}(T)$.

(iii) Suppose (T, ρ) is a totally, bounded. By Lemma <27>, if $x \in \mathbb{U}$ and $\text{OSC}(\delta, x, T) < \epsilon < 1$ then $d_u(x, A_J y) < \epsilon$ for some y in $\mathbb{R}^J$ with rational coordinates. It follows that $\mathbb{U}$ is separable under d_u.

(iv) By separability every d_u-open subset of $\mathbb{U}$ can be written as a countable union of closed balls $B[x, r] := \{y \in \mathbb{U} : d_u(x, y) \le r\}$ with centers x in a countable, dense subset of $\mathbb{U}$. Moreover, if S is a countable ρ-dense subset of T then $B[x, r] = \{y \in \mathbb{U} : \sup_{s \in S} |x(s) - y(s)| \le r\} \in \mathcal{C}(\mathbb{U})$. It follows that the d_u-BOREL sigma-field on $\mathbb{U}$ coincides with its cylinder sigma-field.

9.10 Notes

Back in the days when probability theory was first developing serious measure theoretic credentials, the idea of viewing a stochstic process as a random function as in Section 9.3) seemed to require some vigorous promotion before it became widely accepted. Take a look at Doob (1942), Doob (1947), and Chung (1998) to get an inkling of why this idea was originally less obvious than it now might seem.

For measurable covers (Section 9.5) I mostly followed Dudley (2014, Section 3.2). Section 5.3 of that wonderful book also shows how to apply the heavy machinery related to the theory of analytic sets to attack the measurability problems caused by uncountable index sets.

Doob (1953, Section II.2) described the virtues of working with separable versions. See the Notes on page 625 of his book for the history of the concept. I learned about the extension to processes taking values in compact metric spaces from Meyer (1966, Chapter IV.2). My account in Section 9.6 also borrows from Neveu (1965, Section III.4), Gihman and Skorohod (1974, Section III.2) and Billingsley (1986, §38).

Billingsley (1968, Chap 1) and Dudley (2003, Chap 11)) both provided very clear expositions of traditional "$\rightsquigarrow$" weak convergence theory, of the type described in Section 9.1. It helps to have some familiarity with that approach before engaging with the "$\overset{*}{\rightsquigarrow}$" theory described in Section 9.7, for which Andersen (1985), Andersen and Dobrić (1987, 1988), Dudley (2014, Chap 3), and van der Vaart and Wellner (1996, Parts 1 & 2) are good sources. My Problem [4] combines ideas from those five sources.

The approach in Theorem <26> to convergence in distribution in function spaces is almost part of the probability folklore: the conditions for convergence in distribution are just the conditions for existence (of versions with sample paths in that function space) somehow made uniform over the sequence of processes. I learned to appreciate this point of view from discussions with Peter Gänssler and Winfried Stute.

References

Andersen, N. T. (1985). The central limit theorem for non-separable valued functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 70*, 445–455.

Andersen, N. T. and V. Dobrić (1987). The central limit theorem for stochastic processes. *Annals of Probability 15*, 164–177.

Andersen, N. T. and V. Dobrić (1988). The central limit theorem for stochastic processes. ii. *J. Theoret. Probab. 1*, 287–303.

Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley.

Billingsley, P. (1986). *Probability and Measure* (Second ed.). New York: Wiley. (First edition 1979; third edition, 1995).

Chung, K. L. (1998). Probability and Doob. *The American Mathematical Monthly 105*(1), 28–35.

Donsker, M. D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Annals of Mathematical Statistics 23*, 277–281.

Doob, J. L. (1937). Stochastic processes depending on a continuously varying parameter. *Transactions of the American Mathematical Society 42*(1), 107–140.

Doob, J. L. (1942). What is a stochastic process? *American Mathematical Monthly 49*(10), 648–653.

Doob, J. L. (1947). Probability in function space. *Bulletin of the American Mathematical Society 53*(1), 15–30.

Doob, J. L. (1949). Heuristic approach to the Kolmogorov-Smirirnov theorems. *Annals of Mathematical Statistics 20*, 393–403.

Doob, J. L. (1953). *Stochastic Processes*. New York: Wiley.

Dudley, R. M. (1966). Weak convergence of probabilities on nonseparable metric spaces and empirical measures on euclidean spaces. *Illinois J. Math. 10*, 109–126.

Dudley, R. M. (1967). Measures on non-separable metric spaces. *Illinois J. Math. 11*, 449–453.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability 6*, 899–929.

Dudley, R. M. (2003). *Real Analysis and Probability* (2nd ed.), Volume 74 of *Cambridge studies in advanced mathematics*. Cambridge University Press.

Dudley, R. M. (2014). *Uniform Central Limit Theorems* (2nd ed.), Volume 142 of *Cambridge studies in advanced mathematics*. Cambridge University Press. (First edition, 1999).

Gihman, I. I. and A. V. Skorohod (1974). *The Theory of Stochastic Processes*, Volume I. Springer.

Kolmogorov, A. N. (1933). *Foundations of the Theory of Probability*. Springer/Chelsea. (The Chelsea 1950 translation of 'Grundbegriffe der Wahrscheinlichkeitsrechnung', published in 1933 by Springer.).

Kolmogorov, A. N. (1993). *Selected Works of A. N. Kolmogorov: Probability Theory and Mathematical Statistics*, Volume 2. Springer. Edited by A. N. Shiryaev. Translated from the 1987 Russian edition.

Marczewski, E. and R. Sikorski (1948). Measures in non-separable metric spaces. *Colloquium Mathematicum 1*(2), 133–139. (Available online at at https://eudml.org/doc/209895).

Meyer, P. A. (1966). *Probability and Potentials*. Blaisdell.

Neveu, J. (1965). *Mathematical Foundations of the Calculus of Probability*. San Francisco: Holden-Day.

Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer.

Skorohod, A. V. (1956). Limit theorems for stochastic processes. *Theory of Probability and Its Applications 1*, 261–290.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Process: With Applications to Statistics*. Springer-Verlag.

Chapter 10
Chaining

SECTION *10.1 introduces the idea of a maximal inequality.*
SECTION *10.2 introduces chaining by means of a classical example.*
SECTION *10.3 describes a chaining strategy for approximation by means of a sequence of finite sets linked together into a tree, which serves to decompose a process into a sum of increments.*
SECTION *10.4 defines covering and packing numbers, which are often used in the construction of chaining approximations.*
SECTION *10.5 summarizes some facts about* ORLICZ *norms that are particularly useful for chaining arguments.*
SECTION *10.6 gives four examples of chaining with packing numbers, each based on control of increments by an* ORLICZ *norm.*
SECTION *10.7 shows how chaining can be used to prove uniform continuity of sample paths.*
*SECTION *10.8 presents an example to point out some some subtle differences between the approach described in Section 10.2 and the way chaining was employed in the earlier literature.*

10.1 Maximal inequalities

Solutions for many probabilistic and statistical problems involve construction of good upper bounds for quantities involving a stochastic process $X := \{X_t : t \in T\}$, quantities such as $\sup_{t\in T} |X_t|$ or—if T is equipped with a metric d—the ***oscillation***

<1> $$\text{OSC}(\delta, X, T) := \sup\{|X_s - X_t| : s, t \in T,\, d(s,t) < \delta\}.$$

For example, we might seek bounds on $\|\sup_{t\in T} |X_t|\,\|$ or $\|\text{OSC}(\delta, X, T)\,\|$ for some norm $\|\cdot\|$, or bounds on their tail probability analogs

$$\mathbb{P}\{\sup_{t\in T} |X_t| > \eta\} \qquad \text{or} \qquad \mathbb{P}\{\text{OSC}(\delta, X, T) > \eta\}.$$

Such results are often referred to as ***maximal inequalities*** because they involve maxima (or suprema) over potentially very large sets of random variables.

Maximal inequalities play an important role in the theory of stochastic processes, empirical process theory, and statistical asymptotic theory. Bounds for the oscillation are essential for the construction of processes with continuous sample paths (§10.7) and for the study of convergence in

distribution of sequences of stochastic processes (§9.7). In particular, oscillation control is the key ingredient in the proofs of DONSKER theorems for empirical processes. For the asymptotic theory of estimators defined by optimization over random functions, oscillation bounds (or something similar) have played a major role under various names, such as "stochastic asymptotic equicontinuity" or "uniform tightness".

In this Chapter you will learn about a powerful method for deriving such probabilistic inequalities. The method assumes that the index set T comes equipped with a metric d that gives some control over the increments of a process. For example, we might have

<2> $$\|X_s - X_t\| \le K_0 d(s,t)$$

for some norm $\|\cdot\|$ and some constant K_0, or we might have a tail bound

<3> $$\mathbb{P}\{|X_s - X_t| > K_1 d(s,t) r\} \le \beta(r) \qquad \text{for all } s,t \in T, \text{ and } r \ge 0.$$

for some decreasing function β with $\lim_{r\to\infty} \beta(r) = 0$ and some other constant K_1.

To simplify the exposition, in this Chapter I'll always take the norm in <2> to be $\|\cdot\|_\Psi$ for some ORLICZ function Ψ, perhaps with β derived from Ψ. To avoid having to fuss about some special cases I'll also assume that Ψ is strictly increasing on $\mathbb{R}^+$, with no flat spot. That is, $\Psi(x) = 0$ only at $x = 0$. For your convenience some relevant properties from Chapter 5 are summarized in Section 10.5.

If the index set T were uncountably infinite we could run into measurability difficulties of the type described in Chapter 9. In such cases we could resort to one of the remedies described in that Chapter, which would reduce the problem to one involving only countably infinite sets. Accordingly, we will lose little generality by assuming that T itself is at worst countably infinite. Even better, we can focus on quantities involving only finite subsets of T with the hope that the resulting bounds do not blow up as the size of that set increases.

Remark. It is not essential that $d(s,t) = 0$ should imply $s = t$. That is, d might be a semi-metric, in which case, it suffices to have $X_s = X_t$ almost surely if $d(s,t) = 0$. Inequalities <2> or <3> would ensure that this property holds. Similarly, $\|\cdot\|$ could be a semi-norm provided $\|W\| = 0$ if and only if $W = 0$ almost surely, as happens with the usual $\mathcal{L}^p$ and ORLICZ "norms". I leave it to the fastidious reader to modify the arguments to accommodate semi-metrics and semi-norms.

10.2 What is chaining?

The workhorse of one modern approach to approximating stochastic processes is called ***chaining***, which seems to have an undeserved reputation as a mysterious and complicated idea. In this Chapter I hope to dispel

some of the mysteries by departing from the historical order in which chaining ideas first emerged, proceeding instead from simpler to more subtle constructions.

Remark. A wise probabilist once told me that he couldn't understand all the mystery surrounding chaining, for 'it is really not much more than the triangle inequality'. It is indeed (mostly) just the triangle inequality, applied many times. The cleverness comes from the choices of when to invoke the inequality and how to collect terms into manageable chunks.

The following Example introduces, in a quite classical setting, the key to the chaining idea's success.

<4> **Example.** Consider the problem of bounding $\mathbb{P}\sup_t |B_t|$ for a BROWNIAN motion process $\{B_t : 0 \le t \le 1\}$. Very sharp inequalities can be derived using special properties of the process, such as linearity of the index set and independence of the increments across disjoint intervals. Less precise bounds, to show that the expected supremum is finite, can be derived using only the fact that $B_t - B_s$ has a SUBGAUSSIAN distribution with scale parameter equal to $\sqrt{|t-s|}$.

The main tool is a simple maximal inequality: if $X_1, \ldots, X_N$ are random variables with $X_i \in \text{SUBG}(\tau_i^2)$ for each i then

<5> $$\mathbb{P}\max_i X_i \le \tau\sqrt{2\log N} \qquad \text{where } \tau := \max_i \tau_i.$$

To derive this bound, argue for $\lambda > 0$ that

$$\begin{aligned}\exp\left(\lambda \mathbb{P}\max_i X_i\right) &\le \mathbb{P}\exp\left(\lambda \max_i X_i\right) \qquad \text{by the JENSEN inequality} \\ &= \mathbb{P}\max_i e^{\lambda X_i} \qquad \text{monotonicity of } \exp(\cdot) \\ &\le \sum\nolimits_i \mathbb{P}e^{\lambda X_i} \le \sum\nolimits_i e^{\lambda^2\tau_i^2/2} \le N e^{\lambda^2\tau^2/2}.\end{aligned}$$

Take logs, divide by λ, then choose $\lambda = \tau^{-1}\sqrt{2\log N}$ to minimize.

Remark. Of course we should not expect anything close to a sharp bound if the inequality $\tau_i \le \tau$ were too crude. For example, if $\tau_i = 1$ for $i < N$ and $\tau_N = 10^3$ inequality <5> might not be of much use.

It is natural to approximate the supremum by a maximum taken over a large, finite subset such as

$$T_i := \{j\delta_i : j = 0, 1, \ldots, 2^i\} \qquad \text{where } \delta_i := 2^{-i}.$$

Then we can let i tend to infinity, hoping that the limit will be finite.

To emphasize the fact that the inequality will only depend on information about the increments of the process I'll write $B_t - B_0$ instead of B_t, using the fact that $B_0 = 0$. Then <5> with $\tau^2 = 1$ gives

$$\mathbb{P}\max_{t\in T_i}(B_t - B_0) \le \sqrt{2\log(1+2^i)},$$

which blows up as $i \to \infty$. You might protest that $\tau^2 = 1$ is a gross overestimate when $j\delta_i$ is close to zero. However, half of the members of T_i need a τ^2 of at least $1/2$, which suffices for failure.

For the next part of the discussion it will be cleaner to work with two-sided bounds. This changes the problem very little. If one of the X_i's in <5> is identically zero, as will always be the case in what follows, then

<6> $$\mathbb{P} \max_i |X_i| \le \mathbb{P} \left(\max_i X_i + \max_i(-X_i)\right) \le 2\tau\sqrt{2 \log N}.$$

Remark. Another way to derive a two-sided bound like <6> is: for the ORLICZ function $\Psi(x) = \exp(x^2) - 1$ we have $\|X_i\|_\Psi \le C_0\tau$ for some constant C_0. A simple convexity argument then gives

$$\mathbb{P} \max_i |X_i| \le C_0\tau\Psi^{-1}(N) = C_0\tau\sqrt{\log(1+N)}.$$

See Section 10.5 for details.

It does indeed waste a lot to apply the SUBGAUSSIAN bound directly to every $B_t - B_0$ increment. It would be better to relate each $B_{j\delta_i}$ to a value $B_{\ell_{i-1}(j\delta_i)}$, where ℓ_{i-1} rounds down to the nearest multiple of δ_{i-1}. The function ℓ_{i-1} maps T_i into the set $T_{i-1} := \{j\delta_{i-1} : 0 \le j \le 2^{i-1}\}$ with $|t - \ell_{i-1}(t)| \le \delta_i$. The triangle inequality,

$$|B_t - B_0| \le |B_t - B_{\ell_{i-1}(t)}| + |B_{\ell_{i-1}(t)} - B_0| \le \max_{t\in T_i} |B_t - B_{\ell(t)}| + \max_{s\in T_{i-1}} |B_s - B_0|,$$

then leads us to the bound

$$\begin{aligned}\mathbb{P} \max_{t\in T_i} |B_t - B_0| &\le \mathbb{P} \max_{t\in T_i} |B_t - B_{\ell_{i-1}(t)}| + \mathbb{P} \max_{s\in T_{i-1}} |B_s - B_0| \\ &\le C_0\sqrt{\delta_i}\sqrt{2\log(1+2^i)} + \mathbb{P} \max_{s\in T_{i-1}} |B_s - B_0|.\end{aligned}$$

By matching each t in T_i with an s in T_{i-1} with $|t - s|$ significantly smaller than $|t - 0|$ we have have solved one explosive problem (for T_i) at the cost of creating an analogous problem with a slightly smaller index set T_{i-1}.

So how do we handle the T_{i-1} problem? Of course the answer is: use the same idea to approximate T_{i-1} by an smaller T_{i-2}. And so on. Eventually we get a bound

$$\mathbb{P} \max_{t\in T_i} |B_t - B_0| \le C_0 \sum_{j=1}^{i} \sqrt{2^{-j}}\sqrt{2\log(1+2^j)},$$

which does not blow up as $i \to \infty$.

An appeal to monotone convergence then bounds $\mathbb{P} \sup_{t\in T_\infty} |B_t - B_0|$ for a countable, dense subset T_∞ of $[0, 1]$. Continuity of sample paths, or even just DOOB-separabilty, then takes care of the whole interval $[0, 1]$. □

In the previous Example it was natural to use the points of equally spaced grids as the approximating sets because of the stationarity property of BROWNIAN motion: the distribution of $B_{t+h} - B_t$ is the same for all t in $[0, 1-h]$. In the early history of chaining, where the index set was usually a bounded interval of the real line and the process was controlled by an assumption such as $\mathbb{P}|X_s - X_t|^p \le C|s-t|^{1+r}$ (see Chentsov, 1956, for example), the approximating sets usually consisted of equi-spaced points. Not surprisingly, uniform approximating grids later showed up when index sets expanded to things like $[0,1]^d$.

However, once we try to generalize to processes indexed by a more complicated metric space (T, d), the choice of the T_i's becomes less straightforward. One obvious analog of a uniform grid is a T_i for which every point of T lies within a specified distance δ_i of a some member of T_i. Section 10.4 will describe that approach. In the probability literature such a choice was long considered the "natural" way to proceed until Xavier Fernique pointed out some difficulties arising in the study of sample paths of general gaussian processes. (See Chapter 11.)

It currently seems that selection of the approximating T_i's can be the major challenge for some of the more delicate problems involving control of the sample paths of a general stochastic process.

10.3 Chaining frameworks

Keeping the message from Example <4> in mind, let us now return to the problem of obtaining maximal inequalities for a process $\{X_t : t \in T\}$ indexed by a general metric space (T, d) with probabilistic control over the increments of the process, as in <2> or <3>.

As a typical example, suppose we wish to bound $\mathbb{P}\{\sup_{t\in T} |X_t| > \eta\}$, for a countably infinite or finite T. We might try the union bound,

$$\mathbb{P}\{\sup_{t\in T} |X_t| > \eta\} \le \sum\nolimits_{t\in T} \mathbb{P}\{|X_t| > \eta\},$$

Unfortunately, the sum might be larger than 1, especially if T is big. Instead we could creep up on T through a sequence of finite subsets, $T_0, T_1, \ldots$ (chosen in some as yet unspecified manner) such that, for each s in T,

<7> $$d(s, T_m) := \min\{d(s,t) : t \in T_m\} \to 0 \qquad \text{as } m \to \infty.$$

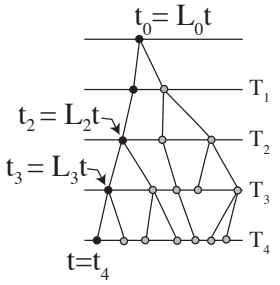
It simplifies some notation if we insist that $T_0 = \{t_0\}$, a singleton set.

The chaining method works by breaking the process on T_m into a contribution from t_0 plus a sum of increments across each T_i to T_{i-1} pair. To achieve such a decomposition we need a map $\ell_{i-1} : T_i \to T_{i-1}$ for each i, thereby arranging the T_i's into a tree. That tree then defines maps $L_i : T_m \to T_i$ for $i \le m$ by following paths towards t_0: for each t in T_m,

$$L_i t := \text{the point of } T_i \text{ on the tree path from } t \text{ to } t_0.$$

More formally, for $t \in T_m$ and $i < m$ the map L_i is given by $L_i t := \ell_i \circ \ell_{i+1} \circ \cdots \circ \ell_{m-1}(t)$.

Remark. To cut down on parenthetic clutter I'll abbreviate $\ell_i(t)$ to $\ell_i t$ and $L_i(t)$ to $L_i t$ whenever there is no ambiguity. Parentheses seem advisable for things like $\ell_{i-1}(j\delta_i)$.



The edges of the tree correspond to the increments of the process,

$$X_t - X_{t_0} = \sum_{i=1}^{m} X(L_i t) - X(L_{i-1} t) \qquad \text{if } t \in T_m,$$

with a corresponding bound derived from the triangle inequality,

<8> $$|X_t - X_{t_0}| \le \sum_{i=1}^{m} |X(L_i t) - X(L_{i-1} t)| \qquad \text{if } t \in T_m.$$

Remark. Here the $|\cdot|$ stands for the absolute value of a real-valued random variable. The same argument works if X takes values in some more exotic normed vector spaces (such as $\mathbb{R}^N$) with $|\cdot|$ denoting the corresponding norm.

For future reference, I'll give a name to the approximation scheme just described.

<9> **Definition.** *A **chaining framework** $\{(T_i, \ell_i) : i = 0, 1, \dots\}$ for a subset S of a metric space (T, d) consists of:*

- *a collection $T_0, T_1, \dots$ of finite subsets of T with $T_0 = \{t_0\}$, a singleton set*
- *maps $\ell_{i-1} : T_i \to T_{i-1}$ for each i*

such that

(i) If the sequence $\{T_i\}$ is infinite then, for each fixed s in S,

$$d(s, T_m) := \min\{d(s,t) : t \in T_m\} \to 0 \qquad \text{as } m \to \infty.$$

(ii) The $\{\ell_i\}$ define a directed tree rooted at t_0, with directed edges $(s, \ell_{i-1}s)$ for $s \in T_i$. The paths along the tree toward the root define maps $L_i : \cup\{T_j : j > i\} \to T_i$ such that $\ell_{i-1} \circ L_i = L_{i-1}$ for each i. □

The framework is said to be ***nested*** if $T_{i-1} \subset T_i$ for each i. In that case it makes sense to require $\ell_{i-1} t = t$ when $t \in T_{i-1}$.

Remark. It turns out that for some purposes (such as the construction in §11.6) it is useful to allow frameworks $\{(T_i, \ell_i) : i = 0, \ldots, m\}$ for a fixed finite m. Obviously the assumption that $d(s, T_m) \to 0$ should be ignored in such a case.

Clearly we must lose something by using inequality <8>. (See Problem [2] for an example.) Against that cost we receive a payoff from the simple fact that many different t's in T_m can share the same increment:

$$X(L_i t) - X(L_{i-1} t) = X(s) - X(\ell_{i-1} s) \qquad \text{if } L_i t = s.$$

We need to control that increment only once, not separately for each path. Inequalities <2> and <3> are tailor-made for that purpose. This sharing of increments is the main reason for the effectiveness of chaining arguments.

Sometimes—particularly when the ultimate aim is control of the oscillation of a process—it pays to stop following the links of the chain (= the edges in the tree) before they reach the T_0 level. For $0 \le k < m$ we then have

$$<10> \qquad |X(t) - X(L_k t)| \le \sum\nolimits_{i=k+1}^{m} |X(L_i t) - X(L_{i-1} t)| \qquad \text{for } t \in T_m,$$

with a corresponding inequality that provides a uniform approximation for $\{X_t : t \in T_m\}$ via the simpler process $\{X_t : t \in T_k\}$:

$$<11> \qquad \mathcal{M}_{k,m} := \max_{t \in T_m} |X(t) - X(L_k t)| \le \max_{t \in T_m} \sum_{i=k+1}^{m} |X(L_i t) - X(L_{i-1} t)|.$$

A variety of maximal inequalities have <11> as their starting point.

The early literature typically bounded the sum in <11> by taking the max inside the sum,

$$<12> \qquad \mathcal{M}_{k,m} \le \sum\nolimits_{i=k+1}^{m} \max\nolimits_{s \in T_i} |X(s) - X(\ell_{i-1} s)|,$$

and choosing the T_i's so that $\max_{s \in T_i} d(s, \ell_{i-1} s) \le \delta_i$, usually with the δ_i's decreasing geometrically as i increased. It was then mostly a matter of controlling the maximum over each T_i. For example, if for each s in T_i we have $X(s) - X(\ell_{i-1} s) \in \text{SUBG}(K\delta_i)$ then, as you saw for Example <4>,

$$\mathbb{P}\mathcal{M}_{k,m} \le \sum\nolimits_{i=k+1}^{m} \sqrt{K\delta_i}\sqrt{2\log(1 + |T_i|)}.$$

Section 10.6 will explain in more detail how this idea, and similar approaches, can lead to useful maximal inequalities.

More recent work by Michel Talagrand (see Chapters 11 and 12) has shown that in some cases we get much better maximal inequalities if we

do not require a uniform bound on the $d(s, \ell_{i-1}s)$ distances. Such a relaxation suggests a preliminary standardization: for s in T_i,

$$\Delta_i(s) := \begin{cases} \dfrac{|X(s) - X(\ell_{i-1}s)|}{\rho_i(s)} & \text{if } \rho_i(s) := \|X(s) - X(\ell_{i-1}s)\|_\Psi \neq 0 \\ 0 & \text{otherwise} \end{cases} . \tag{13}$$

With $\mathbb{M}_i := \max_{s\in T_i} \Delta_i(s)$, inequality <11> then gives

$$\begin{aligned} \mathcal{M}_{k,m} &\le \max_{t\in T_m} \sum_{i=k+1}^{m} \|X(L_i t) - X(L_{i-1}t)\|_\Psi \, \Delta_i(L_i t) \\ &\le \max_{t\in T_m} \sum_{i=k+1}^{m} K_0 d(L_i t, L_{i-1}t)\mathbb{M}_i \qquad \text{if <2> holds.} \end{aligned} \tag{14}$$

10.4 Covering and packing numbers

For a metric space (T, d), a subset S of T is called ***totally bounded*** if: for each $\delta > 0$ there exists a finite set of closed balls of radius δ whose union contains S. Equivalently, there exists a finite subset T_δ of T for which

$$d(s, T_\delta) := \min_{t\in T_\delta} d(s, t) \le \delta \qquad \text{for each } s \text{ in } S. \tag{15}$$

Remark. A metric space is compact if and only if it is both complete and totally bounded (Dudley, 2003, Section 2.3). Thus $\mathbb{R}$ is not totally bounded under the usual ℓ^2 metric.

The set T_δ is sometimes called a ***δ-net for S***. At the cost of some extra factors of 2 we could assume that $T_\delta \subset S$. Indeed, if $t \in T_{\delta/2}$ then we may assume that there exists a point $\sigma(t)$ in S for which $d(\sigma(t), t) \le \delta/2$; if not, then that t could be discarded from T_δ. Replace $T_{\delta/2}$ by $S_\delta := \{\sigma(t) : t \in T_{\delta/2}\}$. Then for each $s \in S$ we have $d(s, \sigma(t)) \le d(s, t) + d(t, \sigma(t)$, which is $\le \delta$ for some t.

It is easy to construct a chaining framework for a totally bounded S from such T_δ's. Starting with some sequence $\{\delta_i\}$ that decreases towards zero, choose finite subsets T_i of S for which $d(s, T_i) \le \delta_i$ for each s in S. Then choose $t' := \ell_{i-1}(t) \in T_{i-1}$ for $t \in T_i$ to satisfy $d(t, t') = d(t, T_{i-1})$.

One should not regard this description as a "construction". It is not constructive in the sense of defining an algorithm for selecting the points that make up T_i. Indeed, most of the "constructions" in this Chapter (and the next two) could be criticised on the same grounds. Perhaps they should be called existence theorems.

For a chaining framework to be useful as a means for deriving maximal inequalities we need to have some upper bound on each $|T_i|$, the cardinality of the approximating set T_i. That requirement leads us to the concept of a ***covering number***.

Definition. *For a totally bounded subset S of a metric space (T, d) define* $\text{COVER}_T(\delta, S, d)$ *to be the smallest N for which there exists a set T_δ of carinality N such that <15> holds.* □ <16>

Remark. The set of closed balls $\{t' \in T : d(t',t) \le \delta\}$ for $t \in T_\delta$ is sometimes called a δ-***covering of*** S. The log of the covering number is sometimes called the ***metric entropy***. See Dudley (1973, page 70) for the origin of that name.

The subscript T on COVER serves as a reminder that the approximating set T_δ need not be a subset of S. The closely related concept of ***packing numbers*** captures the same approximation idea without the need for a subscript to denote where the centers of the balls must lie.

<17> **Definition.** *For a totally bounded subset S of a metric space (T,d) define* $\text{PACK}(\delta,S,d)$ *to be the largest N for which there exists a* δ-***separated*** *subset $\{t_1,\ldots,t_N\}$ of S, that is, if $d(t_i,t_j)>\delta$ for $i \ne j$.* □

The packing number of S depends on (T,d) only through the metric obtained by restricting d to the set $S \times S$. Actually we could dispense with T altogether and just think of (S,d) a metric space in its own right.

<18> **Lemma.** *Suppose S is a totally bounded subset of a metric space (T,d). Then for each $\delta > 0$,*

$$\text{COVER}_S(\delta,S,d) \le \text{PACK}(\delta,S,d) \le \text{COVER}_T(\delta/2,S,d).$$

Proof. Suppose $\text{PACK}(\delta,S,d) = N$ with $F = \{t_1,\ldots,t_N\}$ a δ-separated subset of S. If $t \in S\backslash F$ then the larger set $F \cup \{t\}$ cannot be δ-separated, which implies $\min_i d(t,t_i) \le \delta$.

For the second inequality, observe that no closed ball of radius $\delta/2$ can contain two points more than δ apart. In particular, each of the points t_i must lie in a different ball for a $\delta/2$ covering of S. □

Henceforth I'll call a (finite) subset F of S a δ-***packing*** if it is δ-separated and maximal, in the sense that it is a not a proper subset of any other δ-separated set. As in the proof of Lemma<18>, for such an F the maximality implies $\sup_{s\in S} d(s,F) \le \delta$. Also I'll write $\text{PACK}(\epsilon,S)$, instead of $\text{PACK}(\epsilon,S,d)$, when there is no ambiguity regarding the metric d.

It is largely a matter of taste (or convenience) whether one works with covering or packing numbers, provided an occasional extra factor of 2 causes no significant concerns. However, packing does have some slight advantages (Dudley, 2014, §1.2) over covering.

<19> **Example.** Let $\|\cdot\|$ denote some norm on $\mathbb{R}^n$, such as an ℓ^p norm, $|x|_p = \left(\sum_{i\le n}|x_i|^p\right)^{1/p}$ for $p \ge 1$. Let d denote the metric corresponding to $\|\cdot\|$. As usual, write $B[t,r] := \{x \in \mathbb{R}^n : \|x-t\| \le r\}$, the closed ball centered at t with radius r. Note that $\text{diam}B[t,r] \le 2r$.

The covering/packing numbers for such norms share a common geometric bound, a property derived from the fact that

$$\mathfrak{m}_n B[t,r] := \{x \in \mathbb{R}^n : \|x-t\| \le r\} = r^n \Lambda_n \qquad \text{where } \Lambda_n := \mathfrak{m}_n B[0,1],$$

where $\mathfrak{m}_n$ denotes LEBESGUE measure on $\mathcal{B}(\mathbb{R}^n)$.

Let $\{x_1, \ldots, x_N\}$ be any ϵr-packing set of points in $B[0,r]$. The closed balls $B[x_i, \epsilon r/2]$, of radius $\epsilon r/2$ centered at the x_i, are disjoint and their union lies within $B_{r+\epsilon r/2}$. Thus

$$N \le \frac{(r+\epsilon r/2)^n \Lambda_n}{(\epsilon r/2)^n \Lambda_n} = \left(\frac{2+\epsilon}{\epsilon}\right)^n \le (4/\epsilon)^n \qquad \text{if } 0 < \epsilon \le 2.$$

Similarly, $N \ge (1/\epsilon)^n$ because $\cup_{i \le N} B[x_i, r\epsilon] \supset B[0,r]$. To summarize:

$$(1/\epsilon)^n \le \text{PACK}(\epsilon r, B[0,r], d) \le (4/\epsilon)^n \qquad \text{for } 0 < \epsilon \le 2.$$

□ for every norm on $\mathbb{R}^n$.

The next Lemma shows another advantage of packing over covering: it easily generates nested chaining frameworks.

<20> **Lemma.** *Suppose (S,d) is a totally bounded metric space with diameter D. Let $\delta_0 = D > \delta_1 > \delta_2 \ldots$ be a sequence of positive numbers that decreases to zero. Then there exist a nested chaining framework for S with T_i a δ_i-packing of S for each i.*

Proof. We have $\text{PACK}(D,T) = 1$ for the trivial reason that no pair of points in T is greater than D apart. Define $T_0 = \{t_0\}$ for an arbitrarily chosen point t_0 in T. Build T_0 up to a δ_1-separated set: start with $T_1 = T_0$ then loop,

while $R := \{t \in T : d(t, T_1) > \delta_1\} \ne \emptyset$
    $T_1 \leftarrow T_1 \cup \{t\}$ for some t in R # a new δ_1-separated set

This procedure must exit with $R = \emptyset$ after a finite number of steps because $\text{PACK}(\delta_1, T)$ is finite. Similarly, T_1 can be built up to a δ_2-packing set. And so on.

Take ℓ_{i-1} as any map from T_i into T_{i-1} for which $d(t, \ell_{i-1}t) = d(t, T_{i-1})$ □ for each t in T_i.

For future reference, I'll refer to the chaining framework described by Lemma <20> as a ***$\{\delta_i : i \in \mathbb{N}_0\}$-packing framework***, or $\{\delta_i\}$-packing framework, for short.

10.5 A reminder about ORLICZ norms

The following facts from §5.4 provide one way to handle the $\mathbb{M}_i$ term from inequality <14>.

If an ORLICZ function Ψ has no flat spot then $g := \log(1+\Psi)$ is continuous and strictly increasing with $g(0) = 0$, and $\Psi = e^g - 1$. The inverse

functions Ψ^{-1} and g^{-1} are well defined, continuous, and strictly increasing.

Assumption <2> with $\|\cdot\| = \|\cdot\|_\Psi$ becomes

$$\|X_s - X_t\|_\Psi := \inf\{c > 0 : \mathbb{P}\Psi\left(|X_s - X_t|/c\right) \le 1\} \le K_0 d(s,t).$$

If $d(s,t) > 0$, the standardized increment $Y_{s,t} := |X_s - X_t|/K_0 d(s,t)$ has $\|Y_{s,t}\|_\Psi \le 1$, so that $\mathbb{P}\Psi\left(Y_{s,t}\right) \le 1$ and

$$2 \ge 1 + \mathbb{P}\Psi(Y_{s,t}) = \mathbb{P}\exp(g(Y_{s,t})).$$

It then follows that

<21> $$\mathbb{P}\{|X_s - X_t| \ge K_0 d(s,t) r\} \le \mathbb{P}\exp\left(g(Y_{s,t}) - g(r)\right) \le 2e^{-g(r)}.$$

That is, the $\beta(r)$ in <3> can be taken to equal $2e^{-g(r)}$.

For the special case where X is a centered gaussian process it is natural to define the metric by $d(s,t) = \|X_s - X_t\|_2$ and use the ORLICZ function corresponding to $g(x) = x^2$. For those choices (see §5.6) the constant K_0 equals the awkward $\sqrt{8/3}$. Even with the choice $g(x) = x^2/2$, which would produce the tail bound $2\exp(-r^2/2)$, the constant K_0 would be $\sqrt{2}$, only slightly less awkward. I'll mostly ignore such nuisances, perhaps with the suggestion that we could just work with X/K_0.

For many calculations it helps to have a graceful way to bound products of the form $\Psi(x)\Psi(y)$. Suitable bounds come easily if $\Psi \in \mathcal{Y}_{\exp}$, meaning that there exist constant K_0 and x_1 such that $2g(x) \le g(K_0 x)$ for $x \ge x_1$. For such a Ψ there exists a constant C_Ψ for which:

<22> $$\Psi(x)\Psi(y) \le \Psi\left(C_\Psi \max(x,y)\right) \qquad \text{for all } x, y \text{ in } \mathbb{R}^+,$$

<23> $$\Psi^{-1}(ab) \le C_\Psi \max\left(\Psi^{-1}(a), \Psi^{-1}(b)\right) \qquad \text{for all } a, b \text{ in } \mathbb{R}^+.$$

For $1 \le \alpha < \infty$, each of the ORLICZ functions $\Psi_\alpha := e^{g_\alpha} - 1$, with $g_\alpha(x) = x^\alpha$, belongs to $\mathcal{Y}_{\exp}$. Moreover, g_α is convex and $g_a(x) + g_\alpha(y) \le g_\alpha(x+y)$ for all $x, y \in \mathbb{R}^+$, so that <22> actually holds with $C_{g_\alpha} = 2$.

The bounds from §5.4 for handling maxima of finitely many variables will also be particularly useful in the present Chapter. Here is a slight restatement of the main result:

<24> **Theorem.** *Let $W_1, \ldots, W_N$ be random variables, **not necessarily independent**, but all defined on the same probability space. Define $M := \max_{i \le N} |W_i|$. Suppose $\max_{i \le N} \|W_i\|_\Psi \le 1$ for some* ORLICZ *function Ψ. Then:*

(i) $\mathbb{P}M \le \Psi^{-1}(N)$.

(ii) $\mathbb{P}_B M \le \Psi^{-1}(N/\mathbb{P}B)$ *where $\mathbb{P}_B$ denotes conditional expectation given an event B with $\mathbb{P}B > 0$.*

(iii) If $\Psi \in \mathcal{Y}_{\exp}$ then $\|M\|_\Psi \le 2C\Psi^{-1}(N)$ for some constant C.

□

10.6 Chaining with packing numbers

This Section presents several ways to obtain probabilistic bounds involving $\sup_{t\in T}|X_t|$, where $\{X_t : t \in T\}$ for (T,d) a totally bounded metric space. Without loss of generality assume T is at worst countably infinite.

Remark. Either of inequalities <2> and <3> would ensure that any countable dense subset of T could be used as approximating set for a DOOB-separable version, in which case $\sup_{t\in S}|X_t| = \sup_{t\in T}|X_t|$ almost surely for each countable dense S.

From now on I'll omit the annoying constant K_0 from <2> and just assume

<25> $$\|X_s - X_t\|_\Psi \le d(s,t) \qquad \text{for all } s,t \in T\ .$$

If you find this simplification troubling, you could work with X/K_0.

The inequalities in the rest of the Section usually involve various constants, all denoted by C (or something similar), which depend only on Ψ (and the abolished K_0). Here I am following the lead of Talagrand (2021, page 24), to avoid having to keep track of how a host of different constants are interrelated. I leave for you the instructive, if somewhat tedious, exercise of filling in the details.

Consider any $\{\delta_i\}$-packing framework, constructed using the method described in Lemma <20> with $\delta_i = D/2^i$, for $D := \operatorname{diam}(T)$. That is, we have an increasing sequence of finite subsets $T_0 = \{t_0\} \subset T_1 \subset \dots$ with $N_i := |T_i| \le \text{PACK}(\delta_i, T)$ for each i. The ℓ_{i-1} maps each t in T_i to its closest point in T_{i-1}. In particular, $d(s, \ell_{i-1}s) \le \delta_{i-1}$ for each s in T_i. The set $T_\infty = \cup_{i\in\mathbb{N}_0} T_i$ is countable and dense in T.

In each case the starting point is an inequality derived from <14> (with K_0 equal to 1): for $0 \le k < m$,

<26> $$\mathcal{M}_{k,m} := \max_{t\in T_m}|X(t) - X(L_k t)| \le \sum\nolimits_{i=k+1}^{m} \delta_{i-1}\mathcal{M}_i,$$

with $\mathcal{M}_i$ a maximum of $\le N_i$ random variables, each with Ψ-norm ≤ 1.

As m increases to ∞ the set T_m expands up to T_∞, which implies that $\mathcal{M}_{k,m}$ increases to $\mathcal{M}_{k,\infty} := \sup_{t\in T_\infty}|X(t) - X(L_k t)|$. Probabilistic bounds for $\mathcal{M}_{k,\infty}$ can be obtained by taking limits from bounds for $\mathcal{M}_{k,m}$.

The following four subsections present four types of chaining argument, to illustrate the different ways that packing numbers lead to different sorts of probability bound. For the first subsection Ψ can be any ORLICZ function; for the other three Ψ is assumed to belong to $\mathcal{Y}_{\exp}$, making it easier to simplify a few products. The fourth subsection devotes some attention to the question of how the sequences involved in bounding tail probabilities should be chosen. The other three subsections use the maximal inequalities from Theorem <24>.

Expected values

Take expected values of both sides of <26>, invoking <24>(i) for $\mathbb{P}\mathcal{M}_i$ to get

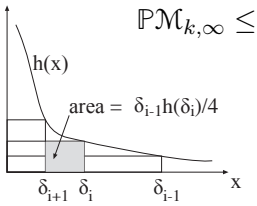
$$\mathbb{P}\mathcal{M}_{k,m} \le \sum_{i=k+1}^{m} \delta_{i-1}\Psi^{-1}(N_i) \le \sum_{i=k+1}^{\infty} \delta_{i-1}\Psi^{-1}(\text{PACK}(\delta_i, T)).$$

Invoke Monotone Convergence as m increases to deduce that

<27> $$\mathbb{P}\mathcal{M}_{k,\infty} := \mathbb{P}\sup_{t\in T_\infty} |X(t) - X(L_k t)| \le \sum_{i=k+1}^{\infty} \delta_{i-1}\Psi^{-1}(\text{PACK}(\delta_i, T)).$$

It is traditional to exploit the geometric rate of decrease of the δ_i's to bound such a sum using an integral if the decreasing function $h(x) := \Psi^{-1}(\text{PACK}(x, T))$ is integrable on the interval $(0, \text{diam}(T)]$. For $J(x) := \int_0^x h(r)\,dr$ and $\delta_i := D/2^i$ then, as in §2.7,

$$\mathbb{P}\mathcal{M}_{k,\infty} \le \sum_{i=k+1}^{\infty} 4(\delta_i - \delta_{i+1})h(\delta_i) \le \sum_{i=k+1}^{\infty} 4\int_{\delta_{i+1}}^{\delta_i} h(r)\,dr = 4J(\delta_{k+1}).$$



In practice, the step function $\text{PACK}(\,\cdot\,, T)$ is usually not known exactly but can be bounded from above by a smooth function, as in the picture.

Orlicz norms

Suppose $\Psi \in \mathcal{Y}_{\exp}$. This time take Ψ-norms of both sides of <26>, invoking <24>(iii) for $\|M_i\|_\Psi$ to get

$$\|\mathcal{M}_{k,m}\|_\Psi \le C_0 \sum_{i=k+1}^{m} \delta_{i-1}\Psi^{-1}(N_i).$$

Then take a limit, as in Section 10.6, to deduce that

<28> $$\|\mathcal{M}_{k,\infty}\|_\Psi \le C_1 \int_0^{\delta_{k+1}} \Psi^{-1}\left(\text{PACK}(x, T)\right)\,dx$$

for a constant C_1 (depending only on Ψ and K_0).

Remark. For processes with subgaussian increments the inequality $\|W\|_p \le C_p \|W\|_{\Psi_2}$, for $p \ge 1$, leads to upper bounds for $\|\mathcal{M}_{k,\infty}\|_p$ similar to those derived by Pollard (1989, Section 3). Those bounds became the main technical tool for the cube-root asymptotic theory developed by Kim and Pollard (1990).

Conditional expected values

The bound based on <24>(ii) is the most surprising. It uses a delightful trick (apparently due to Fernique, 1983) that I learned from Ledoux and Talagrand (1991, Section 11.1). For a fixed event B with $\mathbb{P}B > 0$, the argument parallels the one for expected values until we get to

$$\mathbb{P}_B \mathcal{M}_{k,\infty} \le \sum_{i=k+1}^{\infty} \delta_{i-1} \Psi^{-1}(N_i/\mathbb{P}B) \qquad \text{with } N_i \le \text{PACK}(\delta_i, T).$$

To disentangle the N_i from the $\mathbb{P}B$ use <23> to split the $\Psi^{-1}(N_i/\mathbb{P}B)$:

$$\begin{aligned}\mathbb{P}_B \mathcal{M}_{k,\infty} &\le C_0 \sum_{i=k+1}^{\infty} \delta_{i-1} \Psi^{-1}(N_i) + C_0 \Psi^{-1}(1/\mathbb{P}B) \sum_{i=k+1}^{\infty} \delta_{i-1} \\ &\le 4CJ(\delta_{k+1}) + 4C\delta_{k+1} \Psi^{-1}(1/\mathbb{P}B),\end{aligned}$$

with $J(r) = \int_0^r \Psi^{-1}(\text{PACK})x, T)\, dx$, as in Section 10.6.

Now comes the very clever part. Choose

$$B = \{\mathcal{M}_{k,\infty} \ge 4CJ(\delta_{k+1}) + 4C\delta_{k+1} r\} \qquad \text{with } r > 0,$$

so that $\mathbb{P}_B \mathcal{M}_{k,\infty} \ge 4CJ(\delta_{k+1}) + 4C\delta_{k+1} r$. After some cancellations it then follows that $r \le \Psi^{-1}(1/\mathbb{P}B)$, which rearranges to

<29> $$\mathbb{P}\{\mathcal{M}_{k,\infty} \ge 4CJ(\delta_{k+1}) + 4C\delta_{k+1} r\} \le 1/\Psi(r) \qquad \text{for } r > 0.$$

Remark. For the subgaussian case we have $1/\Psi_2(r) \le 2\exp(-r^2/2)$ if $r \ge \sqrt{\log 2} \approx 0.83$. Thus <29> is a subgaussian tail bound for deviations above $4CJ(\delta_{k+1})$, which is an upper bound for $\mathbb{P}\mathcal{M}_{k,\infty}$.

Tail probabilities

Finally we come to the case involving some nonobvious choices. Inequality <26> again provides the starting point. Suppose $\{\eta_i\}$ is a sequence of strictly positive numbers. Define $G_i := \{\mathbb{M}_i \le \eta_i\}$. On the set $\cap_{i=k+1}^m G_i$ we have

$$|X(t) - X(L_k t)| \le \sum_{i=k+1}^{m} \delta_{i-1} \eta_i \qquad \text{for } t \in T_m;$$

and a union bound based on inequality <21> gives $\mathbb{P}G_i^c \le 2N_i e^{-g(\eta_i)}$. Thus

<30> $$\mathbb{P}\{\mathcal{M}_{k,m} > \sum_{i=k+1}^{m} \delta_{i-1}\eta_i\} \le \sum_{i=k+1}^{m} \mathbb{P}G_i^c \le 2\sum_{i=k+1}^{m} N_i e^{-g(\eta_i)}.$$

So much for the easy bit.

Now we have to figure out a good choice for the η_i's. If I hadn't fixed $\delta_i := \text{diam}(T)/2^i$ then we would be facing something like an optimization over a pair of sequences.

Consider first the special case where $\Psi_1(x) := e^x - 1$ and $g(x) = x$. Inequality <30> is not of much use unless $\sum_{i \ge k+1} \exp(\log N_i - \eta_i)$ is small. That suggests that η_i should be big enough to kill off the $\log N_i$, leaving a remainder γ_i for which $\sum_{i \ge k+1} e^{-\gamma_i}$ is small. If we replaced $\log(N_i)$ by the slightly larger $\log(1 + N_i) = \Psi_1^{-1}(N_i)$ then we would have a term $\sum_{i=k+1}^m \delta_{i-1} \Psi_1^{-1}(N_i)$, an upper bound for $\mathbb{P}\mathcal{M}_{k,m}$ and a contribution to the sum inside the left side of <30>. It seems that we might be headed for a bound like inequality <29> for deviations beyond the expected value. That suggests we choose $\eta_i = \Psi_1^{-1}(N_i) + \gamma_i + r$, with γ_i large enough to make both $\sum_i e^{-\gamma_i}$ and $\sum_{i \ge k+1} \delta_{i-1}\gamma_i$ not too large. For example, the choice $\gamma_i = \log(1/\delta_i)$ would result in sums not much larger than their first terms. The r would contribute a $2\delta_k r$ to the left side and $2e^{-r}$ to the right side of the tail bound. For universal constants c_1 and c_2 we would have

$$\mathbb{P}\{\mathcal{M}_{k,m} \ge \sum_{i=k+1}^m \delta_{i-1}\Psi_1^{-1}(N_i) + c_1\delta_k \log(1/\delta_k) + 2\delta_k r\} \le c_2 e^{-r},$$

which looks a lot like <29>. In fact, if $\text{PACK}(\delta, T)$ were increasing at least like $1/\delta$, which usually happens in nontrivial cases, the contribution of $\{\gamma_i\}$ to $\sum_i \delta_{i-1}\eta_i$ could be absorbed into the $\Psi_1^{-1}(N_i)$ contribution.

The lessons I draw from the special case are: the main role of η_i is to kill off the N_i from the union bound with a little bit extra tacked on to ensures a nicely convergent bounding sum. Here is one way to apply these lessons for a more general $\Psi := e^g - 1$: choose η_i so that

$$g(\eta_i) = \log(1 + N_i/\delta_i) + g(r).$$

Take g^{-1} of both sides, invoking inequality <22>:

$$\eta_i \le C_g^{-1}(\log(1 + N_i/\delta_i)) + C_g r = C_g \Psi^{-1}(N_i/\delta_i) + C_g r$$

and

$$\begin{aligned} &\mathbb{P}\{\mathcal{M}_{k,m} \ge C_g \sum_{i=k+1}^m \delta_{i-1}\Psi^{-1}(N_i/\delta_i) + 2C_g\delta_k r\} \\ &\qquad \le 2\sum_{i=k+1}^m \frac{N_i}{1 + N_i/\delta_i} e^{-g(r)} \le 2\delta_k e^{-g(r)}. \end{aligned} \tag{<31>}$$

Again we have a bound that involves deviations beyond a multiple of $\mathbb{P}\mathcal{M}_{k,m}$ with a tail that decreases like the tail for a single random variable with Ψ-norm at most 1. (And if $g(x) = x^\alpha$ with $\alpha \ge 1$ we would have $C_g = 1$ and $x_0 = 0$.)

As in subsection 10.6, the sum of a series can be replaced by an integral. Also, if $\text{PACK}(x, t)$ is bounded from below by a positive constant (as is usually the case) one of the N_i/δ_i's can be replaced by N_i at the cost of a larger constant. In short, inequality <31> can be replaced by the inequality

$$\mathbb{P}\{\mathcal{M}_{k,m} \ge CJ_k(r)\} \le 2\delta_k e^{-g(r)} \qquad \text{for } r \ge 0$$

$$\text{where } J_k(r) := \int_0^{\delta_k} \Psi^{-1}(\text{PACK}(x, T)\, dx + \delta_k r$$

for some constant C that depends only on Ψ and $\delta_k = \text{diam}(T)/2^k$.

You might ask why I went to so much trouble with the tail probabilities just to get an inequality comparable with the one obtained using conditional expectations. The reason is that the first three chaining methods depend on the existence of a deterministic upper bound, δ_{i-1}, for the link lengths $\max_{s\in S_i} d(s, \ell_{i-1}s)$; it was essential that $\mathbb{P}$, $\mathbb{P}_B$, and $\|\cdot\|_\Psi$ could be moved inside the sum in inequality <26> to act directly on the $\mathbb{M}_i$. Chaining with tail probabilities would also work without the uniform bound on the link lengths, which will be essential for Talagrand's approach to chaining.

10.7 Oscillation and continuity of sample paths

As in Section 10.1, suppose $X = \{X_t : t \in T\}$ is a stochastic process whose index set is a totally bounded metric space (T, d). Remember that a sample path X_ω is d-uniformly continuous if

$$\begin{aligned}\text{OSC}(\delta, X_\omega, T) &:= \sup\{|X(\omega, s) - X(\omega, t)| : s, t \in T \text{ and } d(s, t) < \delta\} \\ &\to 0 \qquad \text{as } \delta \to 0.\end{aligned}$$

If the process is DOOB-separable (or it is replaced by a DOOB-separable version) then we can replace T by any countable, dense subset T_∞ of T.

To control the oscillation it is enough to show that to each $\epsilon > 0$ there exists a $\delta > 0$ such that $\mathbb{P}\text{OSC}(\delta, X, T_\infty) \le \epsilon$ Indeed, it is then just a matter of finding the $\{\delta_j\}$ correponding to $\epsilon_j := 2^{-j}$ then deducing from $\sum_j \mathbb{P}\text{OSC}(\delta_j, X, T_\infty) < \infty$ that $\sum_j \text{OSC}(\delta_j, X_\omega, T_\infty) < \infty$ for almost all ω. If you wanted a version with all (not just almost all) sample paths uniformly continuous then you could redefine X_ω to be the zero function on a suitable negligible set.

If $\{F_j\}$ is a sequence of finite subsets of T that increase to the countable, dense T_∞ then $\text{OSC}(\delta, X_\omega, F_j)$ increases to $\text{OSC}(\delta, X_\omega, T_\infty)$ as $j \to \infty$. Thus it suffices to show that for each $\epsilon > 0$ there exists a $\delta > 0$ such that

<32> $$\mathbb{P}\text{OSC}(\delta, X, F) \le \epsilon \qquad \text{for every finite subset } F \text{ of } T.$$

That is, everything comes down to proving an inequality that holds uniformly over all finite subsets of T.

Remark. Almost all of the stochastic process maximal inequalities that I know about seem to have this form: some sort of limit of inequalities involving finitely many random variables, $\{X_t : t \in F\}$ where the size of the index set F does not enter the bounds.

Instead of hard-coding in assumptions about packing numbers and chaining frameworks let us start from a more primitive requirement, which captures the idea that the behavior of the process is controlled by an approximation property for finite subsets of T.

<33> **Theorem.** *Let $\{X_t : t \in T\}$ be a stochastic process indexed by a metric space (T, d) for which:*

(i) $\|X_s - X_t\|_\Psi \le K_0 d(s,t)$ *for all s, t in T, for some* ORLICZ *norm.*

(ii) For each $\epsilon > 0$ there exists a finite $N(\epsilon)$ such that: to each finite subset F of T there corresponds an $F_0 \subset T$ with $|F_0| \le N(\epsilon)$ and a map $\gamma : F \to F_0$ such that

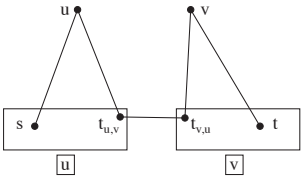
$$\mathbb{P}\mathcal{M}_F \le \epsilon \qquad \text{where } \mathcal{M}_F := \max_{s\in F} |X(s) - X(\gamma s)|.$$

Then there is a version of the process with d-uniformly continuous sample paths.

Remark. Problem [3] shows how a typical chaining argument leads to approximation property (ii). The $\mathbb{P}\mathcal{M}_F$ could be replaced by a tail probability, $\mathbb{P}\{\mathcal{M}_F > \eta\}$, at the cost of some juggling of constants.

Proof. As before, assume $K_0 = 1$. Fix an $\epsilon > 0$ and a finite F. Choose F_0 as in (ii).

The following up-down-across-up-down idea comes from Ledoux and Talagrand (1991, Theorem 11.6). For each u in F_0 define $\boxed{u} := \gamma^{-1}(u) = \{s \in F : \gamma s = u\}$. The sets $\{\boxed{u} : u \in F_0\}$ partition F into at most $N := N(\epsilon)$ equivalence classes. For each distinct pair u, v in F_0 choose points $t_{u,v} \in \boxed{u}$ and $t_{v,u} \in \boxed{v}$ such that



$$d(\boxed{u}, \boxed{v}) := \min(d(s,t) : s \in \boxed{u}, t \in \boxed{v}\} = d(t_{u,v}, t_{v,u}).$$

Define $\mathbb{M}_\square := \max_{u\ne v} |X(t_{u,v}) - X(t_{v,u})|/d(t_{u,v}, t_{v,u})$, a maximum of at most $n_\epsilon := \binom{N}{2}$ increments standardized to have Ψ-norm at most 1. By <24>(i), we have $\mathbb{P}\mathbb{M}_\square \le \Psi^{-1}(n_\epsilon)$.

Consider a pair $s, t \in T_m$ with $d(s,t) < \delta$. If s, t belong to the same $\boxed{u}$ then $|X_s - X_t| \le |X_s - X_u| + |X_u - X_t| \le 2\mathcal{M}_F$. If $s \in \boxed{u}$ and $t \in \boxed{v}$ for distinct u and v then

$$\begin{aligned} |X(s) - X(t)| &\le |X(s) - X(t_{u,v})| + |X(t_{u,v}) - X(t_{v,u})| + |X(t_{v,u}) - X(t)| \\ &\le 2\mathcal{M}_F + d(t_{u,v}, t_{v,u})\mathbb{M}_\square + 2\mathcal{M}_F. \end{aligned}$$

Note that $d(t_{u,v}, t_{v,u}) = d(\boxed{u}, \boxed{v}) < \delta$ if $s \in \boxed{u}$ and $t \in \boxed{v}$ and $d(s,t) < \delta$. Take a maximum over all pairs $\{s, t\}$ with $d(s,t) < \delta$ to deduce that

<34> $$\mathbb{P}\text{OSC}(\delta, X, F) \le 4\mathbb{P}\mathcal{M}_F + \delta\mathbb{P}\mathbb{M}_\square \le 4\epsilon + \delta\Psi^{-1}(n_\epsilon),$$

□ which is $\le 5\epsilon$ if δ is chosen small enough.

Remark. If analogous inequalities hold for a whole sequence of processes $\{X_n(t,\omega) : t \in T\}$ with the same δ and ϵ for each n (or, sometimes, just for all n large enough) then we have an example of ***stochastic equicontinuity***.In its tail probability version, this condition asserts: for each $\eta > 0$ and $\epsilon > 0$ there exists a $\delta > 0$ such that $\limsup_n \mathbb{P}\{\text{OSC}(\delta, X_n, T) > \eta\} < \epsilon$. Such a property plays a central role in the theory of convergence in distribution of random processes. See §9.7 for details.

*10.8 An example of a classical chaining argument

The chaining arguments used to derive maximal inequalities in well-known texts, such as Billingsley (1968, Sections 12, 13), usually involved the index set $[0,1]$. They focussed on processes with sample paths either in the space $C[0,1]$ of continuous real functions or in the space $D[0,1]$ of right-continuous functions with left limits at each point of $(0,1]$, the famous ***cadlag*** functions. It was natural to exploit the ordering of the index set, as in the notion of a limit from the left or from the right. However, particularly for the theory of empirical processes, the reliance on ordering properties leads into a chaining cul-de-sac when one seeks generalizations to more complicated index sets.

Even more importantly from my point of view, the special features of $[0,1]$ allowed for some clever tricks that, to me, made chaining seem more mysterious than it really was. To illustrate this point I'll reinterpret the presentation by Parthasarathy (1967, page 216) of a technique apparently first introduced by Kolmogorov (see Notes). The subject is a stochastic process $\{X_t : 0 \le t \le 1\}$ for which

<35> $$\mathbb{P}|X_s - X_t|^\alpha \le d(s,t)^{1+r} \qquad \text{for all } s,t \in [0,1],$$

where $d(s,t) = |s-t|_2$ (the usual Eucliden distance) and both α and r are strictly positive constants. The aim is to show that there is a version of the process with continuous sample paths.

The moment assumption <35> implies a tail bound,

<36> $$\mathbb{P}\{|X_s - X_t| \ge \eta\} \le d(s,t)^{1+r}/\eta^\alpha \qquad \text{for } \eta > 0.$$

Note that neither <35> nor <36> fits neatly into either of my schemes <2> or <3> for how a metric should control increments. Nevertheless, chaining arguments very similar to those in Section 10.6 will provide the necessary maximal inequalities.

The dyadic rationals provide a very natural chaining framework for $[0,1]$, namely $T_0 = \{0\}$ and $T_i = \{j/2^i : j = 0, 1, \ldots, 2^i\}$, with ℓ_{i-1} the map that rounds down to an integer multiple of $(1/2)^{i-1}$ when $t \in S_i = T_i \backslash T_{i-1}$. Thus $d(t, \ell_{i-1}t) \le \delta_i := (1/2)^i$ for each t in T_i. Instead of working

with the scaled maxima $\mathbb{M}_i$, as in Section 10.6, this time use the unscaled random variables $M_i := \max_{s\in S_i} |X(s) - X(\ell_{i-1}s)|$, so that

$$\mathcal{M}_{k,m} := \max_{t\in T_m} |X(t) - X(L_k t)| \le \sum_{i=k+1}^{m} M_i.$$

Inequality <36> gives

$$\mathbb{P}\{M_i \ge \eta_i\} \le \sum_{s\in S_i} \mathbb{P}\{|X_s - X_{\ell_{i-1}s}| \ge \eta_i\} \le 2^i \delta_i^{1+r}/\eta_i^\alpha = \delta_i^r/\eta_i^\alpha.$$

Notice that 2^i is effectively the covering number for a δ_i-approximation. (My calculation differs a little from Parthsarathy's at this point: he was working with a continuous approximation obtained by linear interpolation between the values $\{X_s : s \in T_i\}$.) The 1 in the $1 + r$ exponent has been used to cancel out the size of the set T_i, a slightly subtle idea. Even more subtle is the choice $\eta_i = \delta_i^\theta$ where $0 < \theta < r/\alpha$ so that an analog of my inequality <30> gives

$$\mathbb{P}\{\mathcal{M}_{k,m} > \sum_{i=k+1}^{m} C\delta_i^\theta\} \le C \sum_{i=k+1}^{m} \delta_i^{r-\theta\alpha}.$$

The rest of the argument could then follow the method described in Section 10.7.

10.9 Problems

[1] Suppose $Y_1, \ldots, Y_n$ are independent random variables. For a fixed $y > 0$ define $M = \mathbb{P}\{\max_{i\le n} Y_i \ge y\}$ and $S = \sum_{i\le n} \mathbb{P}\{Y_i \ge y\}$. Show that

$$S \ge M \ge 1 - e^{-S} = S + o(S)$$

The union bound can be quite sharp when dealing with tails of independent random variables.

[2] This problem offers some insight into the cost of chaining, using the heuristics from Section 10.6. Suppose $\{B_t : 0 \le t \le 1\}$ is a standard Brownian motion process. The increment $B_1 - B_0$ has a $N(0, 1)$ distribution. The usual subgaussian tail bound gives $\mathbb{P}\{|B_1 - B_0| > r\} \le 2\exp(-r^2/2)$ for $r \in \mathbb{R}^+$. What sort of bound can we get using

$$|B(t_m) - B(t_0)| \le \sum_{i=1}^{m} |B(t_i) - B(t_{i-1})| \qquad \text{where } t_i = 2^{-i}?$$

(i) The natural metric is $d(s, t) = \|B_s - B_t\|_2 = \sqrt{|s - t|}$. Show that the increment $B(t_i) - B(t_{i-1})$ has a $N(0, \delta_{i-1}^2)$, where $\delta_i = \sqrt{(1/2)^{i+1}}$.

(ii) Define $\Delta_i := |B(t_i) - B(t_{i-1})|/\delta_{i-1}$. Show that $\mathbb{P}\{\Delta_i > \eta_i\} \le 2\exp(-\eta_i^2/2)$ for $\eta_i \in \mathbb{R}^+$.

(iii) The heuristic from Section 10.6 suggests we choose η_i so that $\eta_i^2/2 = r^2/2 + \gamma_i$, with $\{\gamma_i\}$ increasing rapidly enough to make $\sum_{i\in\mathbb{N}} \exp(-\gamma_i) < \infty$. Show that this choice leads to

$$\mathbb{P}\{|B_1 - B_0| > \sum\nolimits_{i\in\mathbb{N}} \delta_{i-1}(r + \sqrt{2\gamma_i})\} \le 2 \sum\nolimits_{i\in\mathbb{N}} e^{-\gamma_i} e^{-r^2/2}.$$

Experiment with some γ_i values. With $\gamma_i = \log(i \log(2+i)^2)$ I got

$$\sum\nolimits_{i\in\mathbb{N}} \delta_{i-1} \approx 2.4 \qquad \sum\nolimits_{i\in\mathbb{N}} \delta_{i-1}\sqrt{2\gamma_i} \approx 4.2 \qquad \sum\nolimits_{i\in\mathbb{N}} 2e^{-\gamma_i} \approx 3.7.$$

[3] As in Section 10.2, suppose $\{X_t : t \in T\}$ is a stochastic process and $\Psi = e^g - 1$ is an ORLICZ function for which:

(a) $\|X_s - X_t\|_\Psi \le K_0 d(s,t)$ for all $s, t \in T$.

(b) $\{(T_i, \ell_i) : i \in \mathbb{N}_0\}$ is a chaining framework, not necessarily nested, on T for which to each $\epsilon > 0$ there exists a finite $k(\epsilon)$ such that $\mathbb{P}\mathcal{M}_{k,m} \le \epsilon$ if $m > k \ge k(\epsilon)$.

Show that there is a finite $N(\epsilon)$ for which: to each finite subset F of T there exists an $F_0 \subset T$ with $|F_0| \le N(\epsilon)$ and a map $\gamma : F \to F_0$ such that

$$\mathbb{P} \max_{s\in F} |X(s) - X(\gamma s)| \le \epsilon.$$

Argue as follows. Consider any fixed, finite F.

(i) Deduce from Definition <9>(i) that $\delta_{m,F} := \max\{d(s, T_m) : s \in F\} \to 0$ as $m \to \infty$. Choose $\pi_m : F \to T_m$ so that $d(s, \pi_m s) = \delta_{m,F}$. Define $\gamma := L_k \circ \pi_m$

(ii) For $m > k \ge k(\epsilon/2)$ show that

$$\begin{aligned}\mathbb{P} \max_{s\in F} |X(s) - X(\gamma s)| &\le \mathbb{P} \max_{s\in F} |X(s) - X(\pi_m s)| + \mathbb{P}\mathcal{M}_{k,m} \\ &\le K_0 \delta_{m,F} \Psi^{-1}(|F|) + \epsilon/2,\end{aligned}$$

which is smaller than ϵ if m is large enough.

[4] The following statistical example is based loosely on results from Ibragimov and Has'minskii (1981, Section 1.5). It involves a set of probability measures $\{P_\theta : 0 \le \theta \le 1\}$ all defined on the same $(\mathcal{X}, \mathcal{F})$, with densities $p(x,\theta) = dP_\theta/d\mu$ for some dominating measure μ. For simplicity assume $\theta \mapsto p(x,\theta)$ is continuous. Define $\mathbb{P}_{\theta,n} = P_\theta^n$, the probability measure on the product sigma-field on $\mathcal{X}^n$ under which the coordinate maps $x_1, \ldots, x_n$ represent independent observations from P_θ. As a regularity condition assume existence of positive constants C_i for which

<37> $$C_1|\theta - \theta'|_2 \le h(P_\theta, P_{\theta'}) \le C_2|\theta - \theta'|_2 \qquad \text{for all } \theta, \theta' \in [0,1],$$

where $h(\theta,\theta') = \left(\mu|\sqrt{p(x,\theta)} - \sqrt{p(x,\theta')}|\right)^{1/2}$ denotes the Hellinger distance between P_θ and $P_{\theta'}$. See UGMTP Problem 4.18 for useful properties of Hellinger distance, such as $h^2(P^n,Q^n) \le nh^2(P,Q)$.

The maximum likelihood estimator $\widehat{\theta}_n$ is defined to maximize the likelihood function

$$\mathcal{L}_n(\theta) = \mathcal{L}_n(\theta, x_1, \ldots, x_n) := \prod\nolimits_{i\le n} p(x_i,\theta).$$

Show that there exists constants K_1 and K_2 for which

$$\mathbb{P}_{\theta_0,n}\{\sqrt{n}|\widehat{\theta}_n - \theta_0| \ge y\} \le K_1 \exp(-K_2 y^2) \qquad \text{all } y \ge 0, \text{ all } \theta_0 \in [0,1], \text{ all } n.$$

Argue as follows. (If issues of uniqueness worry you, feel free to generalize.) To simplify notation, for a fixed θ_0 abbreviate $\mathbb{P}_{\theta_0,n}$ to $\mathbb{P}$, write $\widehat{t}_n$ for $\sqrt{n}(\widehat{\theta}_n - \theta_0)$, and define

$$\xi(z,t) = \sqrt{p(z,\theta_0 + t/\sqrt{n})/p(z,\theta_0)}\,.$$

(Add some indicator functions if you worry about $p(z,\theta_0) = 0$.)

(i) Show that $\widehat{t}_n$ maximizes the process

$$Z_n(t) := \sqrt{\mathcal{L}_n(\theta_0 + t/\sqrt{n})/\mathcal{L}_n(\theta_0)} = \prod\nolimits_{i\le n} \xi(x_i,t).$$

(ii) Show that $\mathbb{P}|Z_n(t_1) - Z_n(t_2)|^2 \le C_2^2|t_1 - t_2|^2$ and $\mathbb{P}Z_n(t) \le \exp(-\frac{1}{2}C_1^2t^2)$.

(iii) Use the fact that $Z_n(\widehat{t}_n) \ge Z_n(0) = 1$ to show that

$$\mathbb{P}\{|\widehat{t}_n| \ge y\} \le \mathbb{P}\{\sup_{|t|\ge y} Z_n(t) \ge 1\} \le \mathbb{P}\sup_{|t|\ge y} Z_n(t).$$

(iv) Split the set $\{t : |t| \ge y \text{ and } \theta_0 + t/\sqrt{n} \in [0,1]\}$ into a union of intervals like $J = [y_k, y_k + 1]$ where $y_k = y + k$ for an integer k. For a suitable δ, let S_J be a δ-packing set. Use the chaining bound from Section 10.6 to show that

$$\mathbb{P}\sup_{t\in J} Z_n(t) \le (1/\delta)e^{-\frac{1}{2}C_1^2y_k^2} + C_2\sqrt{8}\delta^{1/2}.$$

Minimize over δ.

(v) Sum over intervals like J to get a suitable bound for $\mathbb{P}\sup_{|t|\ge y} Z_n(t)$.

10.10 Notes

Credit for the idea of chaining as a method of successive approximations seems to belong to Kolmogorov, at least for the case of a one-dimensional index set. For example, the start of the paper of Chentsov (1956) commented that "In 1934 A. N. Kolmogorov proved [the result mentioned in

my Section 10.2] ... A generalization of this theorem is the following proposition which was suggested to the author by A. N. Kolmogorov". He added the footnote: "This theorem was first published in a paper by E. E. Slutskii" (= Slutsky, 1937), a paper that I have not read. See Billingsley (1968, Section 12) for a small generalization—with credit to Kolmogorov, via Slutsky, and Chentsov—and a chaining proof.

The Billingsley book was my main source of information when I first learned about stochastic processes with well behaved sample paths. It prepared me for the study of general empirical processes, which I first encountered in the ground-breaking paper of Dudley (1978). To grapple with the host of ideas in that paper I needed to study the earlier work by Dudley (1967, 1973) on gaussian processes, which became my textbooks for general chaining methods. Only later did I stumble on the French work expounded so convincingly by Ledoux and Talagrand (1991).

Dudley (1973) used chaining with covering numbers and exponential tail bounds to establish various probabilistic bounds for gaussian processes. Dudley (1978) adapted the methods using Bernstein's inequality for the increments and metric entropy plus inclusion assumptions (now called bracketing) to extend the gaussian techniques to empirical processes indexed by collections of sets. He also derived bounds for processes indexed by VC classes of sets (see Chapter 14) via symmetrization (see Chapter 13) arguments.

Dudley's 1978 paper became the basis for his famous St. Flour lectures (Dudley, 1984), which were widely circulated in note form. Those notes turned into the 1999 first edition of the definitive Dudley (2014) text. See Dudley (1973, Section 1) and Dudley (2014, Section 1.2 and Notes) for more about packing and covering.

Dudley (2016) took pains to give V.N. Sudakov credit for an earlier use of covering numbers to bound the expected supremum of a Gaussian process. (A scholar less scrupulous than Dudley might have been contented to accept Sudakov's generous comment that "The idea of using ϵ-entropy characteristics here is due independently to several authors.")

Pisier (1983) is usually credited for realizing that the entropy methods used for Gaussian processes could also be extended to nongaussian processes with ORLICZ norm control of the increments. However, as Pisier (page 127) remarked:

> For the proof of this theorem, we follow essentially [10]; I have included a slight improvement over [10] which was kindly pointed out to me by X. Fernique. Moreover, I should mention that N. Kôno [6] proved a result which is very close to the above; at the time of [10], I was not aware of Kôno's paper [6].

Here [10] = Pisier (1980) and [6] = Kôno (1980). The earlier paper [10] included extensive discussion of other precursors for the idea. See also the Notes to Section 2.8 of Dudley (2014).

Using a method like the "expected value" approach in my Section 10.6, Pisier (1983) proved existence of continuous versions of stochastic processes $\{X_t : t \in T\}$ with $\|X_s - X_t\|_\Psi \le d(s,t)$, adapting arguments from Pisier (1980) for the case $\Psi(x) = x^p$. In his July 1981 Saint-Flour lectures, Fernique (1983) gave his own proofs of Pisier's result and generalizations.

Using methods like those in Section 10.6, Nolan and Pollard (1988) proved a functional central limit for the U-statistic analog of the empirical process. Kim and Pollard (1990) and Pollard (1990) proved limit theorems for a variety of statistical estimators using second moment control for suprema of empirical processes. See also Pollard (1985) for one way to use a form of oscillation bound (under the name *stochastic differentiability*) to establish central limit theorems for M-estimators. Pakes and Pollard (1989, Lemma 2.17) used a property more easily recognized as oscillation around a fixed index point.

My analysis in Problem [4] is based on arguments of Ibragimov and Has'minskii (1981, Section 1.5), with the chaining bound replacing their method for deriving maximal inequalities. The analysis could be extended to unbounded subsets of $\mathbb{R}$ by similar adaptations of their arguments for unbounded sets.

References

Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley.

Chentsov, N. N. (1956). Weak convergence of stochastic processes whose trajectories have no discontinuities of the second kind and the "heuristic" approach to the Kolmogorov-Smirnov tests. *Theory of Probability and Its Applications 1*(1), 140–144.

Dudley, R. M. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis 1*(3), 290–330.

Dudley, R. M. (1973). Sample functions of the Gaussian process. *Annals of Probability 1*, 66–103.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability 6*, 899–929.

Dudley, R. M. (1984). A course on empirical processes. *Springer Lecture Notes in Mathematics 1097*, 1–142. École d'Été de Probabilités de St-Flour XII , 1982.

Dudley, R. M. (2003). *Real Analysis and Probability* (2nd ed.), Volume 74 of *Cambridge studies in advanced mathematics*. Cambridge University Press.

Dudley, R. M. (2014). *Uniform Central Limit Theorems* (2nd ed.), Volume 142 of *Cambridge studies in advanced mathematics*. Cambridge University Press. (First edition, 1999).

Dudley, R. M. (2016). V. N. Sudakov's work on expected suprema of Gaussian processes. In *High Dimensional Probability VII*, pp. 37–43. Springer.

Fernique, X. (1983). Regularité de fonctions aléatoires non gaussiennes. *Springer Lecture Notes in Mathematics 976*, 1–74. École d'Été de Probabilités de St-Flour XI, 1981.

Ibragimov, I. A. and R. Z. Has'minskii (1981). *Statistical Estimation: Asymptotic Theory*. New York: Springer. (English translation from 1979 Russian edition).

Kim, J. and D. Pollard (1990). Cube root asymptotics. *Annals of Statistics 18*, 191–219.

Kôno, N. (1980). Sample path properties of stochastic processes. *J. Math. Kyoto Univ. 20*(2), 295–313.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. New York: Springer.

Nolan, D. and D. Pollard (1988). Functional limit theorems for U-processes. *Annals of Probability 16*, 1291–1298.

Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica 57*, 1027–1058.

Parthasarathy, K. R. (1967). *Probability Measures on Metric Spaces*. New York: Academic.

Pisier, G. (1980). Conditions d'entropie assurant la continuité de certains processus et applications à l'analyse harmonique. In *Séminaire d'analyse fonctionnelle, 1979-80*, pp. 1–41. École Polytechnique Palaiseau. Available from http://archive.numdam.org/.

Pisier, G. (1983). Some applications of the metric entropy condition to harmonic analysis. *Springer Lecture Notes in Mathematics 995*, 123–154. (A collection of lecture notes from the 1980-81 Special Year in Analysis at the University of Connecticut Department of Mathematics).

Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory 1*, 295–314.

Pollard, D. (1989). Asymptotics via empirical processes (with discussion). *Statistical Science 4*, 341–366.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*, Volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Hayward, CA: Institute of Mathematical Statistics.

Slutsky, E. (1937). Qualche proposizione relativa alla teoria delle funzioni aleatorie. *Giorn. Ist. Ital. Attuari 8*, 183–199.

Talagrand, M. (2021). *Upper and Lower Bounds for Stochastic Processes: Decomposition Theorems* (Second ed.), Volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag.

Chapter 11
Majorizing measures and partitions

SECTION *11.1 explains why you might want to read this Chapter.*
SECTION *11.2 shows how to construct chaining frameworks from a sequence of finite partitions of the index set T. The differences between nested and non-nested partitions are discussed.*
SECTION *11.3 descibes a small handful of majorizing measures tricks that are helpful for getting some feel for the way a probability measure can tell us something about a sample path.*
SECTION *11.4 describes how majorizing measures were originally used to smooth sample paths.*
SECTION *11.5 describes a method, apparently due to Fernique, that uses weighted frameworks to develop chaining bounds.*
SECTION *11.6 establishes a partial equivalence between weighted partitions and majorizing measures.*
SECTION *11.7 presents a minor extension of Talagrand's method for deriving upper bounds using admissible sequence of partitions.*
SECTION *11.8 establishes an equivalence between admissible partitions and majorizing measures.*
SECTION *11.9 presents a slight modification of an example, due to Fernique, that illustrates the differences between the various methods described in previous Sections. It shows how majorizing measure ideas can improve on packing numbers as a tool for chaining.*
SECTION *11.10 offers an explanation for why the majorizing measure approaches are superior to packing numbers in the previous Section.*

11.1 Overview

The methods (based on packing/covering numbers) described in §10.6 have long been identified with the name 'chaining'. But, as noted at the end of §10.2, there are some more delicate examples where that approach fails.

Beginning in the 1970's, slightly different approaches began joining the collection of chaining tools. Initially the change was driven by the work of Xavier Fernique on gaussian processes indexed by a general metric space (T, d), not just subsets of euclidean spaces. (The metric was supplied by covariances of the process.) He focussed on the use of ***majoriz-***

ing measures (MM's), following a line of development that seems to have been started by Garsia, Rodemich, and Rumsey (1970). See the Notes for some history.

<1> **Definition.** *A majorizing measure with respect to an* ORLICZ *function* Ψ *is a* BOREL *probability measure* μ *defined on a metric space* (T, d) *for which* $\sup_{t\in T} \mathbb{F}_\mu(t, \Psi) < \infty$, *where the functional* $\mathbb{F}_\mu$ *is defined by*

$$\mathbb{F}_\mu(t, \Psi) := \int_0^D \Psi^{-1}\left(1/\mu B[t, r]\right)\, dr \qquad \text{with } D := \operatorname{diam}(T). \tag{2}$$

□ *Here, as usual,* $B[t, r]$ *denotes the closed ball of radius* r *and center* t.

Of course for gaussian processes the natural ORLICZ function is $\Psi_2(x) := \exp(x^2) - 1$. For other processes, different Ψ in $\mathcal{Y}_{\exp}$ seem useful. (It helped me a lot to recognize that some strange quantities in the literature were essentially just $\Psi^{-1}(\,\cdot\,)$ for a suitable Ψ.)

For a centered, gaussian process $\{X_t : t \in T\}$, Fernique (1975, Theorem 6.1.1) showed that the sample paths are bounded if there exists a MM μ with respect to $\Psi_2(x)$. He also commented (page 69) that he didn't know of any examples of gaussian processes with bounded sample paths for which no MM existed ("Je ne connais pas d'exemple de fonctions aléatoires gaussiennes à trajectoires presque sûrement majorées et ne possédant pas de mesure majorante."). Fernique (1981, first page) subsequently turned this suggestive remark into a formal conjecture: modulo a few measurability qualifiers, if a centered gaussian process has bounded sample paths then a MM (with respect to Ψ_2) must exist.

In his 1975 paper Fernique viewed the MM μ as a tool to smooth the sample paths of X, an approach that seems to have fallen into disfavor in recent years. Section 11.4 contains a slight modification of Fernique's argument.

Subsequently, Michel Talagrand proved Fernique's conjecture to be correct by transforming it into an equivalent form involving what he called ***admissible sequences of partitions***. For a centered gaussian process $\{X_t : t \in T\}$, he was able to produce upper and lower bounds for $\mathbb{P}\sup_t |X_t|$ that differed by only constant factors. In this important sense the new methods were claimed to be optimal. Subsequently, Talagrand and various other authors extended this success to other, nongaussian, settings. Much of this progress has been described by Talagrand (2021).

With Talagrand's approach the argument starts from a sequence of finite partitions $\mathbb{A} = \{\mathcal{A}_i : i \in \mathbb{N}_0\}$ of the index set T for a stochastic process $X = \{X_t : t \in T\}$. A chaining framework is easily constructed from $\mathbb{A}$. The effort needed to then derive upper bounds for X is not much greater than the effort needed for the packing/covering methods described in §10.6. In some cases the partitions can be 'nested', which means that each $\mathcal{A}_{i+1}$ is obtained by further partitioning of each A in $\mathcal{A}_i$, starting

from $\mathcal{A}_0 = \{T\}$. The 'nested' assumption is a great convenience but it does come at some cost, as you will see.

Even though Talagrand has argued that his admissible partition method makes other approaches obsolete, I still feel it worthwhile to study the precursors. Maybe the original majorizing measure idea will turn out to be useful in some future variation on the chaining method.

To relate $\mathbb{F}_\mu$ to the sample paths of a general process $\{X_t : t \in T\}$ we of course need to assume something like

$$\|X_s - X_t\|_\Psi \le K_0 d(s,t) \qquad \text{for all } s,t \in T. \tag{<3>}$$

Such a constraint is particularly useful when $\Psi \in \mathcal{Y}_{\exp}$, that is, when there exist a constant K_0 such that $2g(x) \le g(K_0 x)$ for all x large enough, where $g := \log(1+\Psi)$. As shown in §5.3, this inequality implies existence of a constant C_g for which $2g(x) \le g(C_g x)$ for all x and

$$g^{-1}(a+b) \le C_g \max\left(g^{-1}(a), g^{-1}(b)\right) \qquad \text{for all } a,b \in \mathbb{R}^+ \tag{<4>}$$

$$\Psi(x)\Psi(y) \le \Psi\left(C_g \max(x,y)\right) \qquad \text{for all } x,y \in \mathbb{R}^+, \tag{<5>}$$

$$\Psi^{-1}(ab) \le C_g \max\left(\Psi^{-1}(a), \Psi^{-1}(b)\right) \qquad \text{for all } a,b \text{ in } \mathbb{R}^+. \tag{<6>}$$

$$g(x) \ge A_g x^\alpha \qquad \text{if } x \ge x_0, \text{ for constants } x_0 \ge 0,\ \alpha > 0, \text{ and } A_g > 0, \tag{<7>}$$

$$\mathbb{P}\{|Y| \ge r\rho\} \le 2e^{-g(r)} \qquad \text{for } r \ge 0, \text{ if } \rho := \|Y\|_\Psi < \infty. \tag{<8>}$$

This Chapter will explain the general relationship between MM's and partitions, even though the various equivalences currently seem to be regarded as good reasons to ignore the MM altogether. If you subscribe to that view then you'll probably want to skip Sections 11.6 and 11.8. However, the equivalences do bring out some ideas that pop up in some currently popular ways of thinking about chaining.

The Chapter will focus on upper bounds for processes satisfying <3>. For the gaussian lower bound result see Chapter 12.

Remark. I must confess to understanding only a tiny part of Talagrand's 2021 book and, except for some dabbling with bracketing arguments, have never had to use any of the partioning ideas in this Chapter to solve an interesting statistical problem. Nevertheless, I do regard partitioning ideas to be central to the chaining method.

11.2 Partitions, nested and not nested

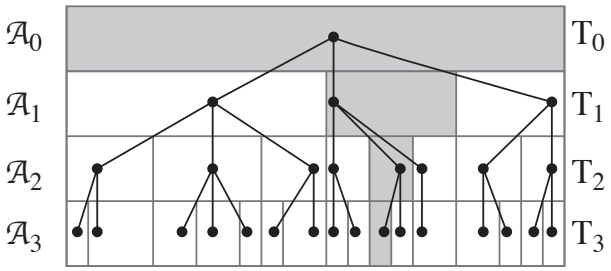
Suppose $\mathbb{A} = \{\mathcal{A}_i : i \in \mathbb{N}_0\}$ is a set of finite partitions of a metric space (T,d), where T will soon become the index set of a stochastic process. This Section describes two slightly different ways of constructing chaining frameworks from $\mathbb{A}$, followed by a short reminder of how the frameworks lead to maximal inequalities.

Case: Nested partitions

Suppose $\mathcal{A}_i$ is a refinement of $\mathcal{A}_{i-1}$ for each i. That is, each B in $\mathcal{A}_{i-1}$ is a disjoint union of sets A from $\mathcal{A}_i$.

To build a nested chaining framework on top of $\mathbb{A}$ we need to choose a single point t_A from each partition set A. The choices should be made recursively, starting with a single t_0 from T then working down through successive levels. If $t_B \in B \in \mathcal{A}_{i-1}$ has been chosen and if $t_B \in A \in \mathcal{A}_i$ then we should make sure to choose t_A equal to t_B. That precaution ensures that the sets $T_i := \{t_A : A \in \mathcal{A}_i\}$ for $i \in \mathbb{N}_0$ are nested: $T_{i-1} \subset T_i$.

If $A \in \mathcal{A}_i$ and $A \subset B \in \mathcal{A}_{i-1}$ define $\ell_{i-1}t_A = t_B$. We then have a nested chaining framework.

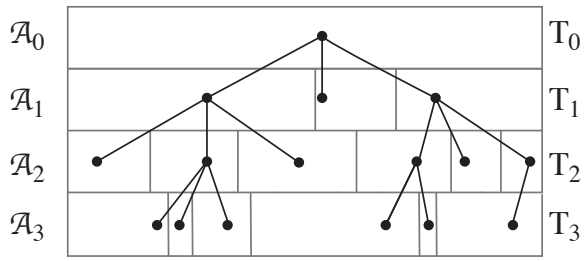


For each t in T there are sets for which $T = A_0(t) \supset A_1(t) \supset A_2(t) \supset \ldots$ with $t \in A_i(t) \in \mathcal{A}_i$. (Compare with the shaded sets in the picture.) If t is a member of $T_\infty := \cup_{i\in\mathbb{N}_0} T_i$ then $L_i t \in A_i(t)$ for every i. Usually there will be constraints on $\mathbb{A}$ that ensure $\operatorname{diam}(A_i(t)) \downarrow 0$ as $i \uparrow \infty$. In that case, $\cap_{i\in\mathbb{N}_0} A_i(t) = \{t\}$ for each t in T.

Case: Non-nested partitions

In this case there is little hope of constructing a nested framework. Accordingly, it suffices to construct sets $T_i := \{t_A : A \in \mathcal{A}_i\}$ with any point $t_A \in A$ chosen from A. If $A \in \mathcal{A}_i$ and $t_A \in B \in \mathcal{A}_{i-1}$ define $\ell_{i-1}t_A = t_B$, which ensures that $d(t_A, \ell_{i-1}t_A) \le \operatorname{diam}(B)$. Note that some members of T_{i-1} might have no links with T_i.

Remark. It is possible that some set A might belong to two $\mathcal{A}_i$'s. To guard against such ambiguity, we could add an extra subscript, writing $\{t_{A,i} : A \in \mathcal{A}_i\}$. Or we could just rely on context to know which $\mathcal{A}_i$ is meant. I prefer the latter approach.



The sets $A_i(t)$ are not as useful as for nested partitions. For example, it need not be true that $L_i t$ is a point of $A_i(t)$. We will need stronger control over the diameters of sets in each $\mathcal{A}_i$ to make up for this deficiency.

Chaining bounds

In both cases, nested and non-nested, as in §10.3 we have a decomposition of the difference $X(t) - X(t_0)$ into a sum of increments, with

$$\begin{aligned} |X(t) - X(L_k t)| &\le \sum_{i=k+1}^{m} |X(L_i t) - X(L_{i-1} t)| \\ &\le \sum_{i=k+1}^{m} \|X(L_i t) - X(L_{i-1} t)\|_\Psi \, \Delta_i(L_i t) \end{aligned} \tag{<9>}$$

where, for each s in T_i,

$$\Delta_i(s) := \begin{cases} |X(s) - X(\ell_{i-1} s)| / \|X(s) - X(\ell_{i-1} s)\|_\Psi & \text{if } s \ne \ell_{i-1} s \\ 0 & \text{otherwise} \end{cases} . \tag{<10>}$$

If <3> holds then

$$\begin{aligned} \mathcal{M}_{k,m} &:= \max_{t \in T_m} |X(t) - X(L_k t)| \\ &\le \max_{t \in T_m} \sum_{i=k+1}^{m} K_0 d(L_i t, L_{i-1} t) \Delta_i(L_i t). \end{aligned} \tag{<11>}$$

Notice that the $\max_{\dots}$ has stayed outside the sum, unlike what happened in Chapter 10 for chaining arguments with packing numbers. The methods in this Chapter handle the maximum in a different way. As before, bounds on things like $\mathbb{P}\mathcal{M}_{k,m}$, if uniform over m, will lead to useful maximal inequalities involving the supremum or oscillation of a sample path.

11.3 Small facts about majorizing measures

When I first encountered MMs I was most puzzled by how such a μ could possibly tell us anything about a sample path. It took me a long time to gain even a slight understanding. It helped me to see how such a μ could be constructed by traditional packing number arguments.

<12> **Example.** Suppose $\Psi = e^g - 1 \in \mathcal{Y}_{\exp}$ and $\int_0^D \Psi^{-1}(\text{PACK}(r,T))\,dr$ is finite. Construct the packing framework as in §10.4: with $\delta_i := D/2^i$, build nested δ_i-packing sets $T_0 = \{t_0\} \subset T_1 \subset \ldots$. For $i \in \mathbb{N}$ we then have $N_i := |T_i| \le \text{PACK}(\delta_i, T)$. Define μ_i to be the uniform distribution on T_i and $\mu := \sum_{i\in\mathbb{N}_0}(1/2)^{i+1}\mu_i$.

To see that μ is a MM, consider any t in T. If $r \ge \delta_j$ then the ball $B[t,r]$ must contain at least one member of the δ_j-packing set T_j, which implies $\mu B[t,r] \ge (1/2)^{j+1}N_j^{-1}$ and

$$\int_{\delta_j}^{\delta_{j-1}} \Psi^{-1}(1/\mu B[t,r])\,dr \le (\delta_{j-1}-\delta_j)\,\Psi^{-1}\left(2^{j+1}N_j, T\right)\right)$$
$$\le \delta_j C_g \max\left[\Psi^{-1}(2^{j+1}), \Psi^{-1}(\text{PACK}(\delta_j,T))\right].$$

□ Sum over j. And so on. (Inequality <7> takes care of the 2^{j+1}.)

Actually, the previous Example has very little to do with packing; it is a special case of a more general result. Again for a Ψ in $\mathcal{Y}_{\exp}$, suppose there exist subsets T_i of T with $|T_i| = N_i < \infty$ for which

<13> $$\sup_{t\in T}\sum_{i\in\mathbb{N}_0}\Psi^{-1}\left(\max(2^{i+1},N_i)\right)d(t,T_i) < \infty.$$

Let μ_i denote the uniform distribution on T_i. Then $\mu := \sum_{i\in\mathbb{N}_0}(1/2)^{i+1}\mu_i$ is a MM. The proof is almost the same as in the previous Example: for each fixed t, just replace δ_i by $\min_{j\le i} d(t,T_j)$.

It also helped my intuition to see why existence of a MM implies that T is totally bounded.

<14> **Example.** Temporarily write β for $\sup_{t\in T}\mathbb{F}_\mu(t,\Psi)$, which is assumed finite. Then for each $t \in T$ and $0 < \delta \le \text{diam}(T)$ we have

$$\beta \ge \int_0^\delta \Psi^{-1}\left(1/\mu B[t,r]\right)\,dr \ge \delta\Psi^{-1}\left(1/\mu B[t,\delta]\right),$$

so that $\mu B[t,\delta] \ge \epsilon := 1/\Psi(\beta/\delta)$. If $\{t_1,\ldots,t_k\}$ is a 2δ-separated subset of T then the balls $B[t_i,\delta]$ for $i = 1,\ldots,k$ are disjoint and each has μ measure at least ϵ. It follows that $1 \ge k\epsilon$, that is, $\text{PACK}(2\delta,T) \le \Psi(\beta/\delta)$.

Notice the similarity to the argument in §10.4, where LEBESGUE measure on $\mathbb{R}^k$ played an analogous role in bounding covering numbers in euclidean space.

Notice also that the bound on the packing numbers i s not quite enough to imply integrability of $\Psi^{-1}(\text{PACK}(x,T))$; the calculations with majorizing measures are not special cases of the calculations in §10.6. See Section 11.9 for a specific example. □

Finally, it will be helpful to know that $\mathbb{F}_\mu$ is controlled by its values on a countable dense subset of T.

<15> **Example.** Suppose S is a countable, dense subset of T and μ is a probability measure on S for which $\mathbb{F}_\mu(S) := \sup_{s\in S} \mathbb{F}_\mu(s)$ is finite. We can also think of μ as a probability measure on the BOREL sigma-field $\mathcal{B}(T)$. I claim that $\mathbb{F}_\mu(t) \le \mathbb{F}_\mu(S)$ for every t in T. That is, each MM on S also defines a MM on T.

To see why this assertion is true, consider any t in T and any $\delta > 0$. There must exist an s in S for which $d(s,t) < \delta$ because S is dense in T. If $r > \delta$ then $B[t,r] \supset B[s, t-\delta]$, which implies

$$\int_\delta^D \Psi^{-1}(1/\mu B[t,r])\, dr \le \int_\delta^D \Psi^{-1}(1/\mu B[s,r-\delta])\, dr \le \mathbb{F}_\mu(s) \le \mathbb{F}_\mu(S).$$

Let δ decrease to 0 to deduce that $\mathbb{F}_\mu(t) \le \mathbb{F}_\mu(S)$.

Problem [3] provides a converse if $\Psi \in \mathcal{Y}_{\exp}$: from each MM μ on $\mathcal{B}(T)$ and each countable, dense subset S of T we can build a MM ν on S.

Similarly, to show that a probability measure μ on $\mathcal{B}(T)$ is a MM it suffices to consider $\mathbb{F}_\mu(t,\Psi)$ only for t in the countable, dense subset T_∞, such as the T_∞ for a chaining framework. □

The previous Example justifies a typical simplification when working with a MM μ: without loss of generality we may assume that μ concentrates on a (given) countable dense subset S of T. For example, if $X = \{X_t : t \in T\}$ is a DOOB-separable process of interest then we can reduce a MM argument to an analysis of the process $\{X_t : t \in S\}$ for any countable, dense approximating subset S. If A is any BOREL subset of $\mathbb{R}$ then

$$\{(s,\omega) \in S \times \Omega : X(s,\omega) \in A\} = \cup_{s\in S} (\{s\} \times \{\omega \in \Omega : X(s,\omega) \in A\})$$

is measurable. That is, X is product-measurable as a map from $S \times \Omega$ into $\mathbb{R}$. See the next Section for an illustration of how this simplification can be helpful.

*11.4 Majorizing measures as smoothers

The following discussion modifies the smoothing argument presented by Fernique (1975, §6), extending it to nongaussian processes and other ORLICZ functions Ψ in $\mathcal{Y}_{\exp}$. It also includes some explanations regarding a few details that troubled me.

Once again consider a (DOOB-separable, to avoid measurabilty issue) stochastic process $\{X_t : t \in T\}$ indexed by a totally bounded metric space (T,d). Actually, by Example <15>, we may as well assume T is countable. Suppose the increments are controlled by an ORLICZ norm for some $\Psi = e^g - 1 \in \mathcal{Y}_{\exp}$, as in <3>, with the $K_0 = 1$ to simplify notation:

$$\|X_s - X_t\|_{\Psi,\mathbb{P}} \le d(s,t) \qquad \text{for } s,t \in T$$

The added $\mathbb{P}$ on the norm symbol serves to distinguish it from another ORLICZ norm that plays a key role in the proof: $\|\cdot\|_{\Psi,\mu\otimes\mu}$ for functions on $T \times T$, where μ is a majorizing measure on T, as in <2>.

The following discussion shows how these assumptions imply that almost all sample paths $X(\cdot,\omega)$ are bounded as functions on T.

Define $\Delta(s,t,\omega)$ to be $|X(s,\omega) - X(t,\omega)|/d(s,t)$ if $s \neq t$ and 0 otherwise. From <3> we have

$$\mathbb{P}^\omega \Psi\left(\Delta(s,t,\omega) \leq 1 \right) \qquad \text{for all } s,t, \text{ in } T.$$

Integrate over s,t with the product measure $\mu \otimes \mu$ then switch the order of integration to deduce that $\mathbb{P}H \leq 1$ where $H(\omega) := \mu^s\mu^t\Psi\left(\Delta(s,t,\omega)\right.$. Consequently $H(\omega) < \infty$ ae[$\mathbb{P}$]. I'll ignore the bad $\mathbb{P}$-neglible set and just assume $H(\omega) < \infty$ for all ω, which implies that $\Delta(\cdot,\cdot,\omega) \in \mathcal{L}^\Psi(\mu \otimes \mu)$. Define

$$Y(\omega) = \|\Delta(\cdot,\cdot,\omega)\|_{\Psi,\mu\otimes\mu}\,.$$

I'll also ignore the issue of measurability (which is needed to justify integration arguments) and the possibility that some $Y(\omega)$ values might be zero.

Remark. We could also take care of both problems by redefining $Y(\omega)$ to equal the smallest m in $\mathbb{N}$ for which $\mu^s\mu^t\Psi(\Delta(s,t,\omega)/m) \leq 1$.

The main idea is to approximate $X(\tau,\omega)$, for each fixed τ in T, by a sequence of random variables $X_n(\tau,\omega)$ for which

<16> $$\sum\nolimits_{n=0}^{\infty} |X_{n+1}(\tau,\omega) - X_n(\tau,\omega)| \leq \beta(\omega) := 3C_g Y(\omega)\mathbb{F}_\mu(T),$$

<17> $$X_n(\omega,\tau) \to X(\omega,\tau) \qquad \text{in } \mathbb{P} \text{ probability as } n \to \infty.$$

It will turn out that $X_0(\tau,\omega) = \mu^s X(s,\omega)$, which is the same for every τ. Thus there is no harm in writing it as $X_0(\omega)$.

The expression on the right-hand side of <16> is finite for each ω, which forces $\{X_n(\tau,\omega) : n \in \mathbb{N}\}$ to be a CAUCHY sequence of real numbers: for $n < N$,

$$|X_N(\tau,\omega) - X_n(\tau,\omega)| \leq \sum\nolimits_{i=n}^{\infty} |X_{i+1}(\tau,\omega) - X_i(\tau,\omega)| \to 0 \qquad \text{as } n \to \infty.$$

Accordingly, there is a value $X_\infty(\tau,\omega)$ such that $X_n(\tau,\omega) \to X_\infty(\tau,\omega)$ as $n \to \infty$, for each ω. The convergence <17> then forces $X_\infty(\tau,\omega) = X(\tau,\omega)$ ae[$\mathbb{P}$] and

<18> $$|X(\tau,\omega) - X_0(\omega)| \leq \beta(\omega) \qquad \text{ae}[\mathbb{P}].$$

The boundedness of almost all sample paths $X(\cdot,\omega)$ follows.

It remains only to construct the X_n's that satisfy <16> and <17>. Each of these variables will be defined by a local 'smoothing' argument using the MM μ.

Proof (of <16>). (Compare with the proof by Garsia 1972, Lemma 1.) For each Borel subset A for which $\mu A > 0$, write μ_A for the conditional distribution $\mu(\cdot \mid A)$ and $X_A(\omega) := \mu_A^s X(s,\omega)$.

Remark. I worry slightly whether $X(\cdot,\omega)$ is μ-integrable for (almost all) ω. Perhaps it would be prudent to replace X_t by $X_t - X_{t_0}$ for a fixed t_0, which would ensure that $X_{t_0} = 0$. The increments would not be affected by such a centering. Then by JENSEN and FUBINI we have

$$\mathbb{P}^\omega \Psi(\mu^s |X(s,\omega)|/D) \le \mu^s \mathbb{P}\Psi\left(|X(s,\omega) - X(t_0,\omega)|/D\right) \le 1,$$

which implies $\mu^s|X(s,\omega)| < \infty$ ae$[\mathbb{P}]$.

For subsets A and B, define $\rho(A,B) := \sup\{d(s,t) : s \in A, t \in B\}$. Then we have

$$\begin{aligned}
\Psi\left(\frac{|X_A(\omega) - X_B(\omega)|}{Y(\omega)\rho(A,B)}\right) &= \Psi\left(\frac{|\mu_A^s \mu_B^t (X(s,\omega) - X(t,\omega))|}{Y(\omega)\rho(A,B)}\right) \\
&\le \Psi\left(\mu_A^s \mu_B^t \Delta(s,t,\omega)/Y(\omega)\right) \\
&\le \mu_A^s \mu_B^t \Psi\left(\Delta(s,t,\omega)/Y(\omega)\right) \qquad \text{by JENSEN} \\
&= \frac{1}{\mu A\, \mu B} \mu^s \mu^t \{s \in A, t \in B\} \Psi\left(\Delta(s,t,\omega)/Y(\omega)\right).
\end{aligned}$$

The first inequality comes from the fact that $\mu_A\{s \in A\} = \mu_B\{t \in B\} = 1$ so that $\rho(A,B) \ge d(s,t)$ for all (s,t) pairs involved in the μ_A^s and μ_B^t integrations. If we discard the $\{s \in A, t \in B\}$ in the final expression we are left with $\mu^s\mu^t\Psi\left(\Delta(s,t,\omega)/Y(\omega)\right)$, which is ≤ 1 by definition of Y. Some minor rearrangement then gives

<19> $$|X_A(\omega) - X_B(\omega)| \le Y(\omega)\rho(A,B)\Psi^{-1}\left(\frac{1}{\mu A\, \mu B}\right).$$

For a fixed τ in T define $X_n(\tau,\omega) := X_{B_n}(\omega)$ for $B_n := B[\tau,\delta_n]$, where $\delta_n := D/2^n$. Note that $B_0 = T$ because $\delta_0 = D$. Thus $X_0(\omega) = \mu^s X(s,\omega)$, which doesn't depend on τ. For all other n the ball B_n changes with τ; the corresponding X_n's all depend on τ. From <19> we get

$$\begin{aligned}
&|X_{n+1}(\tau,\omega) - X_n(\tau,\omega)| \\
&\qquad \le Y(\omega)\left(\delta_n + \delta_{n+1}\right)\Psi^{-1}\left(1/\mu B_n\, \mu B_{n+1}\right) \\
&\qquad \le Y(\omega)(3\delta_{n+1})C_g\Psi^{-1}\left(1/B[\tau,\delta_{n+1}]\right) \qquad \text{by <6>} \\
&\qquad \le Y(\omega)C_g(3\delta_{n+1})/\delta_{n+1}\int_{\delta_{n+1}}^{\delta_n} \Psi^{-1}\left(1/\mu B[\tau,r]\right)\, dr,
\end{aligned}$$

the last inequality coming from the monotonicity of the integrand. When summed over $n = 0, 1, \ldots$ the integrals contribute a quantity smaller than $\mathbb{F}_\mu(\tau)$. When this term is bounded by $\mathbb{F}_\mu(T)$ we get <16>. □

Proof (of <17>). It suffices to show that $\|X(\tau,\omega) - X_n(\omega)\|_{\Psi,\mathbb{P}} \to 0$ as $n \to \infty$. The argument is similar to the one leading to <19>.

$$\begin{aligned}\mathbb{P}\Psi\left(|X(\tau,\omega) - X_n(\tau,\omega)|/\delta_n\right) &= \mathbb{P}\Psi\left(\mu^s_{B_n}|X(\tau,\omega) - X(s,\omega)|/\delta_n\right)\\ &\le \mu^s_{B_n}\mathbb{P}\Psi\left(|X(\tau,\omega) - X(s,\omega)|/\delta_n\right)\\ &\le 1\end{aligned}$$

because $\|X(\tau,\omega) - X(s,\omega)\|_{\Psi,\mathbb{P}} \le \delta_n$ for each s in the ball $B[\tau,\delta_n]$ on which μ_{B_n} concentrates. It follows that $\|X(\tau,\omega) - X_n(\omega)\|_{\Psi,\mathbb{P}} \le \delta_n$. □

11.5 Partitions with weights

Apparently, Fernique also developed another approach, which captures the spirit of a MM by means of a type of weighted chaining framework: Ledoux (1996, Theorem 6.2) presented a proof (using the Ψ_2-norm) for a centered gaussian process $\{X_t : t \in T\}$, which he attributed to Fernique (1971, 1975).

Ledoux wrote the argument using nested sequences of finite partitions, $\mathbb{A} = \{\mathcal{A}_i : i \in \mathbb{N}_0\}$, of T. He imposed a bound on the diameters of each of the partitioning sets, which, in my notation, corresponds to assuming that $\max_{A\in\mathcal{A}_i} \operatorname{diam}(A) = O(\theta^i)$, for some fixed θ in $(0, 1/2]$. He assumed existence of weight functions $w_i : \mathcal{A}_i \to (0,1]$ for which $\sum_{A\in\mathcal{A}_i} w_i(A) \le 1$. He gave conditions under which, for some universal constant C,

$$\mathbb{P}\sup_{t\in T} X_t \le C \sup_{t\in T} \sum\nolimits_{i\in\mathbb{N}} \theta^i \sqrt{\log(1/w_i(A_i(t)))},$$

where $A_i(t)$ denotes the member of $\mathcal{A}_i$ that contains t. He also noted that a similar argument would show

<20> $$\mathbb{P}\sup_{t\in T} |X(t) - X(L_k t)| \le C \sup_{t\in T} \sum\nolimits_{i=k+1}^{\infty} \theta^i \sqrt{\log(1/w_i(A_i(t)))},$$

with the constant C not depending on k.

Remarks.

(i) You should recognize $\sqrt{\log(x)}$ as a surrogate for $\Psi_2^{-1}(x) = \sqrt{\log(1+x)}$.

(ii) Ledoux did not assume DOOB-separabilty for the X process. Instead he used the convention that the $\mathbb{P}\sup_{t\in T}\ldots$ should be interpreted as $\sup_{\text{finite } F\subset T} \mathbb{P}\max_{t\in F}\ldots$.

As explained in Section 11.2, the nesting of the $\mathcal{A}_i$ partitions ensures that $A_1(t) \supset A_2(t) \supset \cdots \supset \{t\}$ and $L_i t \in A_i(t)$, for each t in T, a property that implies $d(L_i t, L_{i-1} t) \le \operatorname{diam}(A_{i-1}(t))$.

It is not difficult to extend inequality <20> beyond the gaussian setting by using a different ORLICZ norm. Also, to avoid the minor task of building a chaining framework from a sequence of partitions, I'll regard

w_i as a function on the T_i of a chaining framework. (It could also be regarded a function on the edges joining T_i to T_{i-1}.) With this formulation there is no need to assume explicitly that the framework is nested. It also sidesteps the issue of whether $\{L_i t, L_{i-1} t\} \subset A_{i-1}(t)$ if the partitions are not nested. Another version of the next Theorem that uses nested partitions will appear in §12.2.

<21> **Theorem.** *Let $\{X_t : t \in T\}$ be a* DOOB*-separable process indexed by a totally bounded metric space (T, d) for which:*

(i) $\|X_s - X_t\|_\Psi \le K_0 d(s,t)$ *where* $\Psi = e^g - 1$ *is an* ORLICZ *function in* $\mathcal{Y}_{\exp}$.

(ii) There is a chaining framework $\{(T_i, \ell_i) : i \in \mathbb{N}_0\}$ *on* T *and positive constants* $\{\delta_i\}$ *such that* $d(t, \ell_{i-1} t) \le \delta_i$ *for each* t *in* T_i.

(iii) For each i *there exists a weight function* $w_i : T_i \to (0, 1]$.

Then $\mathbb{P}\mathcal{M}_{k,m} \le C\beta_{k,m}$ *for all* m *larger than* k, *where*

$$\beta_{k,m} := \max_{t \in T_m} \sum_{i=k+1}^{m} \delta_i \Psi^{-1}\left(1/w_i(L_i t)\right) + \sum_{i=k+1}^{m} \delta_i w_i(T_i)$$

where $w_i(T_i) := \sum_{s \in T_i} w_i(s)$ *and* C *is a constant that depends only on* K_0 *and* Ψ. □

Remark. Of course the asserted maximal inequality is useful for oscillation purposes only if $\sup\{\beta_{k,m} : m > k\}$ can be made small by choosing k large. Compare with §10.7. The literature usually requires $w_i(T_i) \le 1$ but there are cases (cf. §12.2) where $w_i(T_i)$ should be allowed to grow larger. If $w_i(T_i) = O(R^i)$ for some R and $\delta_i = O(\theta^i)$ for some θ with $R\theta < 1$ then $\sum_i \delta_i w_i(T_i)$ converges, taking care of one contribution to $\beta_{k,m}$.

Proof. The argument again starts from the chaining inequality <9>:

<22> $$|X(t) - X(L_k t)| \le \sum_{i=k+1}^{m} K_0 d(L_i t, L_{i-1} t) \Delta_i(L_i t) \qquad \text{for } t \text{ in } T_m.$$

The main trick is to split $\Delta_i(s)$ into C_g times $b_i(s) := \Psi^{-1}(1/w_i(s))$ plus a remainder term, $R_i(s) := \Delta_i(s)\{\Delta_i(s) > C_g b_i(s)\}$, where C_g is the constant from the definition of $\mathcal{Y}_{\exp}$. We thereby bound $|X(t) - X(L_k t)|$ by a non-random function of t plus a sum of random term that turns out to be easily controlled.

The key to the splitting argument is the convexity fact that $\Psi(x)/x$ increases with x together with the growth condition imposed by $\mathcal{Y}_{\exp}$. For values x, b with $x/C_g \ge b \ge \kappa := \Psi^{-1}(1)$ we have

$$\frac{\Psi(x)}{x} \ge \frac{\Psi(C_g b)}{C_g b} \ge \frac{\Psi(b)^2}{C_g b} = \frac{\Psi(b)}{C_g} \frac{\Psi(b)}{b} \ge \frac{\Psi(b)}{C_g} \frac{\Psi(\kappa)}{\kappa} = \frac{\Psi(b)}{C_g \kappa}.$$

Put another way,

<23> $$x \le C_g \kappa \Psi(x)/\Psi(b) \qquad \text{if } x \ge C_g b \text{ and } \Psi(b) \ge 1.$$

Take x equal to $\Delta_i(s)$ for an s in T_i and argue only on the set where $\Delta_i(s) > C_g b_i(s)$ to get

<24> $$R_i(s) := \Delta_i(s)\{\Delta_i(s) > C_g b_i(s)\} \le C_g \kappa \frac{\Psi(\Delta_i(s))}{\Psi(b_i(s))} = C_g \kappa \Psi(\Delta_i(s)) w_i(s).$$

From the fact that $\mathbb{P}\Psi(\Delta_i(s)) \le 1$ it follows that

$$\mathbb{P} \max_{s \in T_i} R_i(s) \le \mathbb{P} \sum\nolimits_{s \in T_i} C_g \kappa \Psi(\Delta_i(s)) w_i(s) \le C_g \kappa w(T_i).$$

From <22>, we can then bound $\mathbb{P}\mathcal{M}_{k,m}$ by

$$\begin{aligned} &\mathbb{P} \max_{t \in T_m} \sum\nolimits_{i=k+1}^{m} K_0 \delta_i \Big(C_g b_i(L_i t) + R_i(L_i t) \Big) \\ &\quad \le K_0 C_g \max_t \sum\nolimits_{i=k+1}^{m} \delta_i b_i(L_i t) + K_0 \sum\nolimits_{i=k+1}^{m} \delta_i \mathbb{P} \max_{s \in T_i} R_i(s), \end{aligned}$$

□ which implies the asserted inequality.

Remark. If $w_i(s) = 1/|T_i|$ for each s in T_i and each i, conclusion of the Theorem becomes $\mathbb{P}\mathcal{M}_{k,m} \le C \sum_{i=k+1}^{m} \delta_i \left(1 + \Psi^{-1}(|T_i|\right)$, which is essentially the first bound (based on packing numbers) from §10.6.

*11.6 From majorizing measures to weighted frameworks

Problem [4] shows how to construct a majorizing measure from the sort of weighted framework appearing in Theorem <21>. This Section partially complements that result. Even though the rough equivalence between MM's and weighted partitions is not essential to the story for gaussian processes told in Chapter 12, it does give some insight into the role of various greedy arguments for constructing finite partitions of an index set.

The proof of the main result in this Section will work by first constructing a nested sequence of partitions on T. A similar approach will be taken in Section 11.8 for Talagrand's alternative to weighted partitions. There the initial part of the argument will ignore the nesting requirement, first constructing a (probably non-nested) sequence $\{\widetilde{\mathcal{A}}_i\}$ then subsequently working with the finer partition $\mathcal{A}_i$ defined as the common refinement of $\widetilde{\mathcal{A}}_1, \ldots, \widetilde{\mathcal{A}}_i$.

Unfortunately the same idea does not seem to work with weighted partions: there is no obvious way (known to me) to combine the weight functions for such $\widetilde{\mathcal{A}}_i$'s into suitable weight functions for the $\mathcal{A}_i$'s. (However the constructions used by Andersen et al., 1988, §2 might offer a clue.)

Instead, the partitions will be constructed from the bottom up, starting from a partition built around a finite subset of T then taking successive unions to construct a sequence of successively coarser, nested partitions. This approach is based on L&T = Ledoux and Talagrand (1991, Lemma 11.10, Rem. 11.11), which Ledoux (1996, p 239) described as "a somewhat technical and actually nontrivial discretization lemma".

Initially the constraint of starting from a finite subset of T seemed very restrictive to me. That concern led me to rethink some of my earlier chaining arguments. I tried to recast them as results about finite subsets of T. The resulting rewrite of §10.7 was one encouraging success. It became clear that to derive useful facts (such as existence of a version with uniformly continous sample paths) about a process $\{X_t : t \in T\}$ it is not essential to work with the same chaining framework on T for each finite subset F of T; it suffices to construct a separate chaining framework for each F provided the resulting probabilistic bounds hold in some uniform sense over a sufficiently large collection of finite F's.

<25> **Theorem.** *Suppose F is a finite subset of a metric space (T, d) on which there is a* MM *μ. Then there is a nested chaining framework $\{(T_i, \ell_i) : 0 \le i \le m\}$ on T with $F \subset T_m$, weight functions $w_i : T_i \to (0, 1]$, and a universal constant C for which:*

(i) $\sum_{t \in T_i} w_i(t) \le 1$ *for each i for $0 \le i \le m$.*

(ii) $\max_{s \in T_i} d(s, \ell_{i-1} s) \le 6 D_i$ *for $0 \le i \le m$, where $D_i := \operatorname{diam}(T)/3^i$.*

(iii) For each t in T_m and $0 \le k \le m$,

$$\sum_{i=k}^{m} D_i \Psi^{-1}\left(1/w_i(L_i t)\right) \le C \int_0^{D_k/2} \Psi^{-1}\left(1/\mu B[t, r]\right)\, dr.$$

□

Remark. If $\{X_t : t \in T\}$ is DOOB-separable process and $\|X_s - X_t\|_\Psi \le K_0 d(s, t)$ for all s, t in T, then by Theorem <21> there is a constant C such that

<26> $$\mathbb{P} \max_{t \in F} |X(t) - X(L_k t)| \le C \int_0^{D_k/2} \Psi^{-1}\left(1/\mu B[t, r]\right)\, dr.$$

Consequently, if $\sup_{t \in T} \int_0^\delta \Psi^{-1}\left(1/\mu B[t, r]\right)\, dr \to 0$ as $\delta \to 0$ then §10.7 shows that there there is a version of the process with uniformly continuous sample paths.

The main argument in the Proof will consist of repeated appeals to the following Lemma. The process $\{X_t : t \in T\}$ will play no role; only the MM on T is used.

<27> **Lemma.** *Let μ be a* MM *on T and S be a finite subset of T. Suppose $\mathcal{E}$ is a finite partition of T such that that each E in $\mathcal{E}$ contains at least one point of S and $\max_{E \in \mathcal{E}} \operatorname{diam}(E) \le \rho$. Then for each $R > 0$ there is a coarser partition $\mathcal{E}^*$ of T and a weight function $W : \mathcal{E}^* \to (0, 1]$ such that*

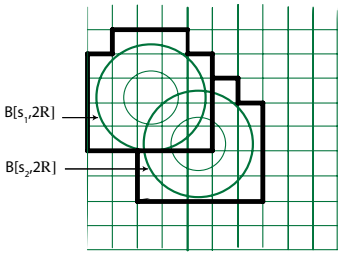
(i) $\sum_{E^* \in \mathcal{E}^*} W(E^*) \le 1$.

(ii) *If* $t \in S \cap E^*$ *with* $E^* \in \mathcal{E}^*$ *then* $W(E^*) \ge \mu B[t, R]$

(iii) $\max_{E* \in \mathcal{E}^*} \operatorname{diam}(E^*) \le 4R + 2\rho$.

Proof (of Lemma). The partition $\mathcal{E}^*$ will be created by repeated appeals to a greedy procedure that assigns a positive integer as a label to each set E in $\mathcal{E}$. Each point of $S \cap E$ inherits the label applied to E. Eventually E_j^* will be the union of all E's with label j, for $j = 1, 2, \ldots$.

1: **procedure** COARSEN($\mathcal{E}$,R)
2: Initialize: $j = 0$. No labels on $\mathcal{E}$ or S.
3: **loop**
4: Exit the loop if all members of S have labels.
5: Increment j by 1.
6: Find an unlabeled s_j in S that maximizes $\mu B[s_j, R]$ over all unlabeled members of S.
7: Define E_j^* to be the union of all the unlabeled members of $\mathcal{E}$ that have a nonempty intersection with the ball $B[s_j, 2R]$. Attach label j to all those E's and to all the members of $E_j^* \cap S$. Set $W(E_j^*)$ equal to $\mu B[s_j, R]$.
8: **end loop**
9: Return the partition $\mathcal{E}^* = \{E_1^*, E_2^*, \ldots\}$ and the weight function W.
10: **end procedure**



The small squares denote the members of $\mathcal{E}$. *The two outlined regions represent* E_1^* *and* E_2^*, *defined by the balls* $B[s_1, 2R]$ *and* $B[s_2, 2R]$. *Notice how some of the* $\mathcal{E}$ *sets that intersect* $B[s_2, 2R]$ *have already been allocated to* E_1^*.

Suppose it takes k passes through the loop before all members of $\mathcal{E}$ are labeled. Then $\mathcal{E}^* := \{E_j^* : 1 \le j \le k\}$ provides the desired coarser partition of F. The $2R$ padding ensures that the $B[\xi_j, R]$ balls are disjoint, implying $\sum_{j \le k} W(E_j^*) \le 1$.

If $t \in E_j^*$ then t must have been one of the unlabeled candidates for the center ξ_j, which implies $W(E_j^*) = \mu B[\xi_j, R] \ge \mu B[t, R]$.

If t and t' are two members of $S \cap E_j^*$ then there are $E, E' \in \mathcal{E}$ with $t \in E$ and $t \in E'$. Choose points $s \in E \cap B[s_j, 2R]$ and $s' \in E' \cap B[s_j, 2R]$. Then

$$d(t, t') \le d(t, s) + d(s, \xi_j) + d(\xi_j, s') + d(s', t) \le \rho + 2R + 2R + \rho,$$

□ as required for (iii).

Remark. I'll write $\{\mathcal{E}_i : i = 0, 1, \dots\}$ instead of $\{\mathcal{A}_i : i = 0, 1, \dots\}$ for the partitions to emphasize the 'bottom up' idea. Each $\mathcal{E}_i$ will be coarser than the preceding $\mathcal{E}_{i-1}$ Similarly, I'll write $W_0, W_1, \dots$ for the weight function and $\rho_0, \rho_1, \dots$ for the upper bound on $\max\{\text{diam}(E) : E \in \mathcal{E}_i\}$. Ony near the end of the proof will these quantities be translated to 'top down' notation.

Proof (of Theorem <25>). Find the smallest m for which F is D_m-separated and then build F up to a D_m-packing set $T_m \supset F$. (Except in trivial cases there exists such an m in $\mathbb{N}$.) Define $R_0 = D_m/2$.

Start by defining a partition $\mathcal{E}_0 := \{E_s : s \in T_m\}$ of T by taking E_s to consist of those t in T for which s is the nearest point in F, with some suitable tie-breaking rule. Because T_m is a D_m-packing and $R_0 = D_m/2$ we have $d(t, T_m) \le D_m$ for each t in T and, for each $s \in T_m$,

<28> $$B[s, R_0] \subset E_s \subset B[s, 2R_0] \quad \text{AND} \quad \text{diam}(E_s) \le \rho_0 := 4R_0.$$

Define $W_0(E_s) := \mu B[s, R_0]$. The balls $\{B[s, R_0] : s \in T_m\}$ are disjoint; the sum of the weights is ≤ 1.

For $i = 1, 2, \dots$, define $R_i := 3^i R_0 = D_{m-i}$. Successively define

$$(\mathcal{E}_i, W_i) := \text{COARSEN}(\mathcal{E}_{i-1}, R_i) \qquad \text{for } i = 1, 2, \dots, m.$$

Notice that the E_1^* constructed by Lemma <27> with R equal to R_m contains a closed ball of radius $2R_m = \text{diam}(T)$. Thus $\mathcal{E}_m = \{T\}$.

By construction the $\{\mathcal{E}_i\}$ partitions are nested and satisfy the inequalities

$$\rho_i := \max\{\text{diam}(E) : E \in \mathcal{E}_i\} \le 4R_i + 2\rho_{i-1}.$$

More specifically,

$$\begin{aligned}
\rho_1 &= 4R_1 + 2\rho_0 = 2^2 R_1 + 2^3 R_0, \\
\rho_2 &= 4R_2 + 2\rho_1 = 2^2 R_2 + 2^3 R_1 + 2^4 R_0 \\
&\dots \\
\rho_i &= 4R_i + 2\rho_{i-1} = 2^2 R_i + 2^3 R_{i-1} + 2^4 R_{i-2} + \cdots + 2^{i+1} R_1 + 2^{i+2} R_0 \\
&\le 4R_i \left(1 + 2/3 + 2^2/3^2 + \cdots + 2^i/3^i\right) \le 12 R_i = 6 D_{m-i}.
\end{aligned}$$

As before, let $E_i(t)$ denote the member of $\mathcal{E}_i$ that contains a given t in T. If $t \in T_m$ then Lemma <27>(iii) ensures that $W_i(E_i(t)) \ge \mu B[t, R_i]$, so that

$$\begin{aligned}
{}^2\!/_3 R_i \Psi^{-1}\left(1/W_i(E_i(t))\right) &\le (R_i - R_{i-1}) \Psi^{-1}\left(1/\mu B[t, R_i]\right)) \\
&\le \int_{R_{i-1}}^{R_i} \Psi^{-1}\left(1/\mu B[t, r]\right)\right) \, dr \qquad \text{for } i \ge 1.
\end{aligned}$$

Similarly, $R_0 W_0(E_0(t)) \le \int_0^{R_0} \Psi^{-1}(1/\mu B[t,r]))\, dr$. Summation over i gives, for $0 \le k \le m$,

<29> $$\sum_{i=0}^{m-k} D_{m-i}\Psi^{-1}\left(\frac{1}{W_i(E_i(t))}\right) \le 3\int_0^{R_{m-k}} \Psi^{-1}\left(\frac{1}{\mu B[t,r]}\right) dr.$$

Now let me translate into 'top down' notation. Define $\mathcal{A}_i := \mathcal{E}_{m-i}$, a nested sequence of partitions of T such that $\mathcal{A}_0 = \{T\} \preccurlyeq \mathcal{A}_1 \preccurlyeq \ldots \preccurlyeq \mathcal{A}_m$, where $\preccurlyeq$ indicates that the partition on the right is a refinement of the partition on the left. Each A in $\mathcal{A}_m$ contains exactly one point from T_m and $A_j(t) = E_{m-j}(t)$ with

$$\max\{\operatorname{diam}(A) : A \in \mathcal{A}_j\} = \rho_{m-j} \le 6D_j \qquad \text{for } 0 \le j \le m.$$

To construct a nested framework on T_m choose a point $t_A \in A \cap T_m$ for each A in $\mathcal{A}_j$, following the procedure described in Section 11.2 to ensure that $T_j \subset T_{j+1}$ for $0 \le j < m$. Transfer the weights from the members of the partitions to the elements of the framework: $w_j\{t_A\} = W_{m-j}(A)$ for A in $\mathcal{A}_j$. In particular, if $A = A_j(t)$ for t in T_m then $L_j t = t_A$ and inequality <29> translates into

□ $$\sum_{j=k}^{m} D_j \Psi^{-1}\left(\frac{1}{w_j(L_j t)}\right) \le 3\int_0^{D_k/2} \Psi^{-1}\left(\frac{1}{\mu B[t,r]}\right) dr \qquad \text{for } 0 \le k \le m.$$

11.7 Talagrand and admissible partitions

In a truly amazing feat, Talagrand (1987) verified Fernique's gaussian conjecture. (Read Section 2.16 of Talagrand (2021) for a fascinating account of how he succeeded.) Unfortunately, I was unable to follow the details of his 1987 argument. Fortunately, Talagrand (1992) came up with a much simpler argument, which is described in my Chapter 12. Eventually—compare with Talagrand (1996, 2001)—he decided to completely eliminate majorizing measures from the story, replacing them by ***admissible partitions***. He showed that boundedness of sample paths of a centered gaussian process implies the existence of *at least one* nested sequence $\mathbb{A} = \{\mathcal{A}_i : i \in \mathbb{N}_0\}$ of finite partitions of T (in the sense of Section 11.2) for which:

(Tal.i) $|\mathcal{A}_i|$, the number of sets in $\mathcal{A}_i$, is at most $n_i := 2^{2^i}$ for each i in $\mathbb{N}$. He referred to this constraint as ***admissibility***.

(Tal.ii) If $A_i(t)$ denotes the member of $\mathcal{A}_i$ that contains a given point t of T then, for some universal constant C,

<30> $$\sup_{t\in T} \sum_{i\in\mathbb{N}} 2^{i/2}\operatorname{diam}(A_i(t)) \le C\mathbb{P}\sup_{t\in T} X_t.$$

See Chapter 12 for the delicate details.

Remarks.

(i) The $2^{i/2}$ in inequality <30> corresponds to $\Psi_2^{-1}(n_i) = \sqrt{\log(1+n_i)} \approx 2^{i/2}\sqrt{\log 2}$, especially for larger i.

(ii) The sequence $n_i = 2^{2^i}$ increases very rapidly because $n_{i+1} = n_i^2$, a relationship that greatly simplifies some calculations. For example,

$$(n_k + n_{k-1} + \cdots + n_1)/n_k = 1 + n_k^{-1/2} + n_k^{-3/4} + \cdots \le 2,$$

so that $\sum_{i=1}^k n_i \le 2n_k$.

(iii) The concept of an admissible sequence of finite partitions lies at the center of what Talagrand (2021) called his ***generic chaining*** method.

(iv) Most of the magic of admissible partitions comes from the fact that $\Psi_2^{-1}\left(2^{2^i}\right)$ increases at a geometric rate. In general, it seems that the key idea is that $\Psi^{-1}(|\mathcal{A}_i|)$ should increase slowly enough that it doesn't overwhelm $\operatorname{diam}(A_i(t))$. A geometric rate of increase does simplify the argument.

It is relatively easy to establish upper bound analogs of <30>, even without a gaussianity assumption, using more general ORLICZ functions and a nested sequence of finite partitions $\mathbb{A} = \{\mathcal{A}_i : i \in \mathbb{N}_0\}$ of T. To shed more light on the role played by the n_i's, I'll initially impose only the mild constraint that $|\mathcal{A}_i| \ge 2^i$ (not $\le 2^{2^i}$), which excludes some trivial cases. Other constraints will be added later. The main argument will be very similar to the one for tail probabilities in §10.6. I'll leave some assertions in a slightly raw state to simplify comparisons with a theorem of Talagrand (2021, page 43).

Section 11.2 showed how to build a nested chaining framework on $\mathbb{A}$, with sets T_i, consisting of a single point from each A in $\mathcal{A}_i$, and maps $\ell_{i-1} : T_i \to T_{i-1}$. If $A_i(t)$ denotes the A in $\mathcal{A}_i$ for which $t \in A$ then $A_0(t) \supset A_1(t) \supset \ldots$ and $L_i t \in A_i(t)$ for each i. Consequently,

<31> $$d(L_i t, L_{i-1} t) \le \delta_i(t) := \operatorname{diam}(A_{i-1}(t)) \qquad \text{for each } i \le m \text{ and } t \text{ in } T_m.$$

<32> **Theorem.** *Let $\{X_t : t \in T\}$ be a* DOOB*-separable stochastic process indexed by a metric space (T, d), for which $\|X_s - X_t\|_\Psi \le K_0 d(s,t)$ for all $s, t \in T$, for some* ORLICZ *function $\Psi = e^g - 1$ in $\mathcal{Y}_{\exp}$. Suppose T is equipped with a nested sequence of finite partitions $\mathbb{A} = \{\mathcal{A}_i : i \in \mathbb{N}_0\}$, and corresponding nested chaining framework, with $N_i := |\mathcal{A}_i| = |T_i| \ge 2^i$ for each i. Then:*

(i) There exists a constant $C_0 = C_0(\Psi, K_0)$ and an event G_k with $\mathbb{P}G_k^c \le 2e^{-g(r)}/2^k$ on which

$$|X(t) - X(L_k t)| \le C_0 \sum_{k < i \le m} \delta_i(t) \left(\Psi^{-1}(N_i) + r\right) \qquad \text{for all } t \in T_m,$$

for each m that is $\ge k$ and each $r \ge 0$.

(ii) For another constant $C_1 = C_1(\Psi, K_0)$,

$$\mathbb{P} \max_{t\in T_m} |X(t) - X(L_k t)| \le C_1 \beta_{k,m}$$

□ *where* $\beta_{k,m} := \max_{t\in T_m} \sum_{i=k+1}^{m} \delta_i(t) \Psi^{-1}(N_i)$.

Remark. If $\sup_m \beta_{k,m} \to 0$ as $k \to \infty$ then X has uniformly continuous sample paths.

Proof. Again let me assume $K_0 = 1$. From Section 11.2 we have

$$\begin{aligned}\mathcal{M}_{k,m} &:= \max_{t\in T_m} |X(t) - X(L_k t)| \\ &\le \max_{t\in T_m} \sum_{i=k+1}^{m} \delta_i(t) \mathbb{M}_i \qquad \text{where } \mathbb{M}_i := \max_{s\in T_i} \Delta_i(s).\end{aligned}$$

Define $G_k := \cap_{i>k} \{\mathbb{M}_i \le \eta_i\}$, for some positive η_i (to be defined soon). Because $\mathbb{M}_i$ is a maximum of at most N_i random variables, each with Ψ-norm at most 1, inequality <8> gives

$$\mathbb{P} G_k^c \le \sum\nolimits_{i>k} \mathbb{P}\{\mathbb{M}_i > \eta_i\} \le \sum\nolimits_{i>k} 2N_i e^{-g(\eta_i)} \qquad \text{for } \eta_i \ge 0.$$

If we choose η_i so that $g(\eta_i) = \log(1 + N_i) + \log(2^i) + g(r)$ then

$$\sum\nolimits_{i>k} 2N_i e^{-g(\eta_i)} \le \sum\nolimits_{i>k} 2 \frac{N_i}{(1+N_i)2^i} e^{-g(r)} \le 2e^{-g(r)}/2^k.$$

Multiple appeals to inequality <4> show that $\eta_i \le C_0 \left(\Psi^{-1}(N_i) + r\right)$ so that the inequality asserted by (i) holds on G_k.

For assertion (ii), first note that

$$\epsilon_{k,m} := \max_{t\in T_m} \sum\nolimits_{i>k} \delta_i(t) \le \beta_{k,m} / \min\{\Psi^{-1}(N_i) : i > k\},$$

which is bounded above by $\beta_{k,m}/\Psi^{-1}(2^{k+1})$. The bound from (i) gives

$$\mathbb{P}\{\mathcal{M}_{k,m} \ge C_0 (\beta_{k,m} + r\epsilon_{k,m})\} \le 2e^{-g(r)}/2^k \qquad \text{for each } r \ge 0.$$

Integrate.

$$\begin{aligned}&\mathbb{P} \left(\mathcal{M}_{k,m} - C_0\beta_{k,m}\right)^+ / (C_0\epsilon_{k,m}) \\ &\quad = \int_0^\infty \mathbb{P}\{\left(\mathcal{M}_{k,m} - C_0\beta_{k,m}\right)^+ / (C_0\epsilon_{k,m}) \ge r\}\, dr \\ &\quad \le \int_0^\infty 2e^{-g(r)}/2^k\, dr \le \text{const},\end{aligned}$$

which rearranges to give

□ $$\mathbb{P}\mathcal{M}_{k,m} \le C_0\beta_{k,m} + \mathbb{P}\left(\mathcal{M}_{k,m} - C_0\beta_{k,m}\right)^+ \le C_1\beta_{k,m}.$$

Remark. Assertion (ii) could also be proved by a method similar to the one used to prove Theorem <21>: see Ledoux (2014).

*11.8 From majorizing measures to admissible partitions

The equivalence between MM's and partitions holds in greater generality than suggested in Section 11.7. It has nothing to do with stochastic processes, it works for more general ORLICZ functions in $\mathcal{Y}_{\exp}$, and the sizes of the partitions need not be controlled by the magical 2^{2^i}.

It is easy (see Problem [5]) to construct a majorizing measure, supported by a countable subset of T, from the framework described in Theorem <32>.

It takes more effort to argue in the other direction, to construct a suitable nested sequence of partitions from a majorizing measure. The next Theorem, which I write in a form that emphasizes the role of $\Psi^{-1}(|\mathcal{A}_i|)$, is a minor extension of an argument due to Talagrand (2001, Theorem 2.2).

To simplify notation, let me assume that the metric on T is rescaled so that $D := \operatorname{diam}(T) = 1$. It also helps to work with a slightly different functional, obtained by inserting an extra factor of r in the definition <2> of $\mathbb{F}_\mu(t,\Psi)$:

<33> $$\widetilde{\mathbb{F}}_\mu(t,\Psi) := \int_0^1 \Psi^{-1}(1/G_t(r))\,dr \qquad \text{where } G_t(r) := r\mu B[t,r] \text{ for } t \in T.$$

It is an easy exercise with $\mathcal{Y}_{\exp}$ to show that $C_\Psi := \int_0^1 \Psi^{-1}(1/r)dr$ is finite (Problem [1]) and for some other constant C, which also depends only on Ψ,

$$\mathbb{F}_\mu(t,\Psi) \le \widetilde{\mathbb{F}}_\mu(t,\Psi) \le C_g\left(\mathbb{F}_\mu(t,\Psi) + C_\Psi\right).$$

It is largely a matter of taste whether one works with the original $\mathbb{F}_\mu$ or its modification $\widetilde{\mathbb{F}}_\mu$ when μ is a MM.

<34> **Theorem.** *Suppose $\Psi = e^g - 1 \in \mathcal{Y}_{\exp}$ and μ is a* MM *on T. If $\lambda \ge C_g$ there exists a nested sequence $\mathbb{A} = \{\mathcal{A}_i : i \in \mathbb{N}_0\}$ of finite partitions of T with $\mathcal{A}_0 = \{T\}$, and constants $C_1 = C_1(\lambda,\Psi)$ and $C_2 = C_2(\lambda,\Psi)$ such that:*

(i) $\Psi^{-1}\left(|\mathcal{A}_i|\right) \le C_1\lambda^i$ *for $i \in \mathbb{N}$.*

(ii) $\sum_{i\in\mathbb{N}} \operatorname{diam}(A_i(t))\lambda^i \le C_2\widetilde{\mathbb{F}}_\mu(t,\Psi)$ *for each $t \in T$.*

Remark. Of course, if $\widetilde{\mathbb{F}}_\mu(t,\Psi)$ is uniformly bounded over T then so is $\sum_{i\in\mathbb{N}}\lambda^i \text{diam}(A_i(t))$. I prefer the statement used in the Theorem because it emphasizes the pointwise nature of the argument.

Proof. Most of the argument will work for any λ that is > 1. Only at the end will it be useful to require $\lambda \geq C_g$ to ensure that $g(\kappa\lambda^i)$, with $\kappa := \Psi^{-1}(1)$, increases at a geometric rate, making it easy to create nested partitions.

The main idea is to initially construct partitions $\widetilde{\mathcal{A}}_0 = \{T\}, \widetilde{\mathcal{A}}_1, \widetilde{\mathcal{A}}_2, \ldots$ which are not necessarily nested, but for which $|\widetilde{\mathcal{A}}_i| \leq 1/y_i$ for cunningly chosen constants $1 = y_0 > y_1 \ldots \downarrow 0$. The procedure is the same for each $i \geq 1$. Thus it simplifies notation to consider the construction of a generic finite partition $\widetilde{\mathcal{A}}$ from a generic y in $(0,1)$. Later everything will acquire an extra subscript i.

Construction of a generic $\widetilde{\mathcal{A}}$ for a fixed y in $(0,1)$.

The partition is built in two stages: first slice T into regions $F_1, F_2, \ldots, F_\ell$ for which $G_t(\delta_j) \approx y$ if $t \in F_j$, where $\delta_j := 2^{-j}$ for $j = 1, 2, \ldots$. (Don't forget that $\text{diam}(T) = 1$.) The number of regions created in the second step is controlled by a trick similar to the one used in Example <14>.

For the first step, it is helpful that $G_t(r) := r\mu B[t,r] \leq r$ is a strictly increasing, right-continuous function of r with $G_t(0) = 0$ and $G_t(1) = 1$. Thus $G_t(\cdot)$ is a probability distribution function on $[0,1]$. As G_t can have jumps it need not have an inverse. The next best thing is its quantile function (UGMTP p41):

$$Q_t(y) := \inf\{r \in [0,1] : G_t(r) \geq y\} \qquad \text{for } 0 \leq y \leq 1.$$

The function Q_t is increasing, with $Q_t(0) = 0$ and $Q_t(1) = 1$ and, by right-continuity of G_t, we have

<35> $$G_t(r) \geq y \qquad \text{iff} \qquad r \geq Q_t(y) \qquad \text{for } 0 \leq r, y \leq 1,$$

an equivalence that will be important for the construction of the $\widetilde{\mathcal{A}}_i$'s.

These properties of G_t imply

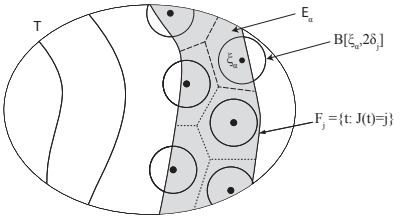
$$1 = G_t(2\delta_1) > G_t(\delta_1) = G_t(2\delta_2) > G_t(\delta_2) = G_t(2\delta_3) > \ldots \downarrow 0$$

Accordingly

$$J(t) := \text{the } j \text{ in } \mathbb{N} \text{ for which } G_t(2\delta_j) \geq y > G_t(\delta_j).$$

is a well-defined map from T into $\mathbb{N}$. If ℓ denotes the largest positive integer for which $2\delta_\ell \geq y$ then the sets $F_j := \{t \in T : J(t) = j\}$ for $1 \leq j \leq \ell$ define a partition of T.

As (T, d) is totally bounded (Example <14>), the set F_j has a finite $4\delta_j$-packing set $S_j := \{\xi_\alpha : \alpha = 1, \ldots, k\}$ and a corresponding partition $\{E_\alpha : \alpha = 1, \ldots, k\}$ of F_j for which $d(t, \xi_\alpha) = d(t, S_j) \le 4\delta_j$ for each t in E_α. By this construction, $\operatorname{diam}(E_\alpha) \le 8\delta_j$.



Remark. You might prefer to write $E_{\alpha,j}$ for $\alpha = 1, \ldots, k_j$. However, that would lead to an orgy of subscripts in the next segment of the Proof.

Now we can use the neat MM trick. The balls $B[\xi_\alpha, 2\delta_j]$ are disjoint. (It doesn't matter that they might extend beyond F_j.) Thus

$$1 \ge \sum_{\alpha=1}^{k} \mu B[\xi_\alpha, 2\delta_j] = \sum_{\alpha=1}^{k} G_{\xi_\alpha}(2\delta_j)/(2\delta_j) \ge ky/(2\delta_j),$$

implying $k \le 2\delta_j/y$. The partition $\widetilde{\mathcal{A}}$ of T made up of all the E_α's from all the nonempty F_j's consists of fewer than $\sum_{j\in\mathbb{N}} 2\delta_j/y = 2/y$ sets. Most importantly, if $t \in T$ and $\widetilde{A}(t)$ denotes the member of $\widetilde{\mathcal{A}}$ that contains t then

$$\begin{aligned}
\operatorname{diam}(\widetilde{A}(t)) &\le \sum_{j\in\mathbb{N}} 8\delta_j \{G_t(2\delta_j) \ge y > G_t(\delta_j)\} \\
&= \sum_{j\in\mathbb{N}} 8\delta_j \{2\delta_j \ge Q_t(y) > \delta_j\} \qquad \text{by <35>} \\
&\le 8Q_t(y).
\end{aligned}$$

Construction of $\widetilde{\mathcal{A}}_i$ using a cunningly chosen y_i.

Requirement (i) suggests that we generate $\widetilde{\mathcal{A}}_i$ for $i \in \mathbb{N}$ from the sequence $y_i = 1/\Psi(\kappa\lambda^i)$, where $\kappa := \Psi^{-1}(1)$ to ensure that $y_0 = 1 > y_1 > \ldots$. Everything now gets an extra subscript i: we have $|\widetilde{\mathcal{A}}_i| \le 2\Psi(\kappa\lambda^i)$ and $\operatorname{diam}(A_i(t)) \le 8Q_t(y_i)$.

For each positive real number R, if $\kappa\lambda^m < R \le \kappa\lambda^{m+1}$ then

$$\sum_{i\in\mathbb{N}} \kappa\lambda^i \{R > \kappa\lambda^i\} = \sum_{i=1}^{m} \kappa\lambda^i \le \kappa\lambda^{m+1}/(\lambda - 1) \le \lambda R/(\lambda - 1).$$

When applied with $R = \Psi^{-1}(1/G_t(r))$ this elementary inequality gives

$$\begin{aligned}\sum_{i\in\mathbb{N}} \kappa\lambda^i \operatorname{diam}(\widetilde{A}_i(t)) &\le 8 \sum_{i\in\mathbb{N}} \kappa\lambda^i \int_0^1 \{r < Q_t(y_i)\}\, dr \\ &= 8 \sum_{i\in\mathbb{N}} \kappa\lambda^i \int_0^1 \{G_t(r) < y_i\}\, dr \\ &= 8 \int_0^1 \sum_{i\in\mathbb{N}} \kappa\lambda^i \{\Psi^{-1}(1/G_t(r)) > \kappa\lambda^i\}\, dr \\ &\le \frac{8\lambda}{\lambda-1} \int_0^1 \Psi^{-1}(1/G_t(r))\, dr.\end{aligned}$$

That is, except for the annoying fact that the $\{\widetilde{\mathcal{A}}_i\}$ need not be nested, we have (ii). The final step in the argument will eliminate this deficiency.

Creating nested partitions from the $\widetilde{\mathcal{A}}_i$'s.

Consider the finer partition $\mathcal{A}_i$ that consists of all nonempty subsets of T expressible as $\cap_{\ell\le i}\widetilde{A}_\ell$ with $\widetilde{A}_\ell \in \mathcal{A}_\ell$. That is, $\mathcal{A}_i$ is the coarsest partition that is a refinement of each $\widetilde{\mathcal{A}}_\ell$, for $\ell \le i$. And, of course, $\mathcal{A}_{i+1}$ is a refinement of $\mathcal{A}_i$, so that $\mathbb{A} = \{\mathcal{A}_i : i \in \mathbb{N}_0\}$ is a nested sequence.

Assertion (ii) of the Theorem is undisturbed if $\mathbb{A}$ replaces $\widetilde{\mathbb{A}}$, because $A_i(t)$ is a subset of $\widetilde{A}_i(t)$. However, we do need to worry about assertion (i). Fortunately, the exponential growth in the ORLICZ function $\Psi = e^g - 1$ will come to the rescue. By $2g(x) \le g(C_g x)$ for all $x \ge 0$.

The exponential rate of growth of g controls the size of $\mathcal{A}_i$:

$$|\mathcal{A}_i| \le \prod_{\ell\le i} |\widetilde{\mathcal{A}}_\ell| \le 2^i \prod_{\ell\le i} \Psi(\kappa\lambda^\ell) \le 2^i \exp\left(g(\kappa\lambda) + \cdots + g(\kappa\lambda^i)\right).$$

If we choose $\lambda \ge C_g$ then repeated appeals to the bound for $2g(x)$ give $g(\kappa\lambda^\ell) \le g(\kappa\lambda^{i+1})/2^{i-\ell}$ for $1 \le \ell \le i$. The sum in the exponent is less than $g(\kappa\lambda^{i+1})$. Thus

$$\Psi^{-1}\left(|\mathcal{A}_i|\right) \le C_g \max\left(g^{-1}\left(\log(1+2^i)\right), g^{-1}\log\left(1 + e^{g(\kappa\lambda^{i+1})}\right)\right) = O(\lambda^i).$$

□ I leave to you the pleasure of finding a suitable C_1.

<36> **Example.** Suppose $\Psi(x) = \exp(x^\alpha) - 1$ for some $\alpha \ge 1$. That is, $g(x) = x^\alpha$ and $g(\lambda x) = \lambda^\alpha g(x)$ for all $x \ge 0$. Theorem <34> provides nested partitions for which there is a constant C such that

$$|\mathcal{A}_n| \le \Psi(C\lambda^n) \le \exp\left((C\lambda^n)^\alpha\right) = 2^{(\log_2 e) C^\alpha \lambda^{n\alpha}}.$$

If we choose $\lambda = 2^{1/\alpha}$ then there exists some positive integer p for which $\log_2 |\mathcal{A}_n| \le 2^{n+p}$. It takes only a little relabelling to get an admissible sequence in Talagrand's sense: define $\mathcal{A}_n$ to be the trivial partition $\{T\}$ for $n \le p$ and to be the old $\mathcal{A}_{n-p}$ for $n > p$.

□

11.9 An illuminating example

The following example is based on an analogous construction by Fernique (1975, Exemple 5.5.2). It highlights some subtle differences between the packing/covering condition,

$$\int_0^D \Psi^{-1}\text{PACK}(x,T)\,dx < \infty \qquad \text{with } D = \operatorname{diam}(T), \tag{37}$$

and the more refined majorizing measure type of condition.

Suppose $T = \mathbb{N}_0$ and $\{X_t : t \in T\}$ is a set of independent random variables with $X_t \sim N(0,\sigma_t^2)$. Specify the variances by partitioning T into consecutive blocks $E_0 = \{0\}, E_1, E_2, \ldots$ on each of which the σ_t^2's are constant: make $|E_j| = n_j := 2^{2^j}$ for $j \geq 1$, then define $\sigma_0^2 = \tau_0^2 = 0$ and $\sigma_t^2 := \tau_j^2 := 2^{-j}$ when $t \in E_j$ for $j \geq 1$. Equip T with the metric $d(s,t) := \|X_s - X_t\|_2$. That is, for $s \neq t$,

$$d(s,t) = \sqrt{\sigma_s^2 + \sigma_t^2} = \sqrt{\tau_j^2 + \tau_k^2} \qquad \text{if } s \in E_j \text{ and } t \in E_k.$$

We then have $\operatorname{diam}(T) = 1$ and $\operatorname{diam}(E_j) = \sqrt{2}\tau_j$.

Remark. The choice $\sigma_0 = 0$ ensures that $0 \in B[t,\sigma_t]$ for every t, which greatly simplifies MM calculations.

Slightly inconveniently, with the natural choice $\Psi_2(x) = \exp(x^2) - 1$ for ORLICZ function we then have $\|X_s - X_t\|_{\Psi_2} = \sqrt{8/3}\,d(s,t)$.

To avoid a plethora of $\sqrt{2}$ factors in what follows, let me define $\lambda := \sqrt{2}$. so that, for $j \geq 1$ we have $\tau_j = \lambda^{-j}$ and $\operatorname{diam}(E_j) = \lambda^{1-j}$ and $n_j = \exp(c_1\lambda^{2j})$ where $c_1 := \log 2$, implying

$$\Psi_2^{-1}(n_j) = \sqrt{\log(1+n_j)} \approx \sqrt{c_1}\lambda^j.$$

Regarding control of sample paths, the short version of the story is: $\mathbb{P}\sup_t X_t$ is finite, the majorizing measure, weighted partition, and admissible sequence approaches work, but the packing condition <37> fails. Keep reading for the explanations.

Paths bounded but not continuous

Define $Z_t = X_t/\sigma_t$ for $t \geq 1$. You know from §3.4 that $Y_j := \max_{t\in E_j}|Z_t|$ is highly concentrated around its expected value, which is close to

$$\sqrt{2\log n_j} = \sqrt{2\times 2^j \log 2} = c_2/\tau_j \qquad \text{where } c_2 := \sqrt{2\log 2}.$$

Each random variable $\max_{t\in E_j}|X_t| = \tau_j Y_j$ should be highly concentrated around c_2, which suggests $\sup_{t\in T}|X_t|$ might also be concentrated

around c_2 (cf. §6.1). With a little luck the supremum over T should be integrable. A simple integration of tail probabilities confirms this suspicion:

$$\begin{aligned}\mathbb{P}\sup_t |X_t| &= \int_0^\infty \mathbb{P}\{\sup_t |X_t| > x\}\,dx \\ &\le c_2 + \int_{c_2}^\infty \sum\nolimits_t \mathbb{P}\{|Z_t| > x/\sigma_t\}\,dx \\ &\le c_2 + \sum\nolimits_{j\in\mathbb{N}} n_j \int_{c_2}^\infty 2\exp\left(-x^2/(2\tau_j^2)\right)\,dx.\end{aligned} \tag{38}$$

For the jth term in the sum, make the change of variable $x = c_2 + \tau_j y$ and note that $n_j = \exp\left(c_2^2/(2\tau_j^2)\right)$ to bound the term by

$$\tau_j \int_0^\infty \exp\left(\frac{c_2^2}{2\tau_j^2} - \frac{(c_2+\tau_j y)^2}{2\tau_j^2}\right)\,dy \le \tau_j \int_0^\infty \exp(-y^2/2)\,dy.$$

The sum in the bound for $\mathbb{P}\sup_{t\in T}|X_t|$ is less than a constant multiple of the sum of a geometric series. That is, the expected value is finite.

The X process does not have continuous sample paths. For example, $d(0,E_n)\to 0$ as $n\to\infty$ but $\sup_{t\in E_n}|X_t|$ concentrates near c_2, which does not converge to $|X_0| = 0$.

Failure of the packing argument

If s and t are distinct points in that E_j then $d(s,t) = \lambda^{1-j} = \tau_{j-1}$. Thus $\text{PACK}(r,T) \ge |E_j| = n_j$ for $r < \tau_{j-1}$ and

$$\int_{\tau_j}^{\tau_{j-1}} \Psi_2^{-1}(\text{PACK}(r,T))\,dr \ge (\tau_{j-1}-\tau_j)\,\Psi_2^{-1}(n_j) \ge \lambda^{-j}(\lambda-1)\sqrt{c_1}\lambda^j.$$

The sum over j is divergent. Inequality <37> does not hold.

Majorizing measure

With μ_i denoting the uniform distribution on E_i, define $\mu := \sum_{i\in\mathbb{N}_0}(1/2)^{i+1}\mu_i$. We need to show that $\mathbb{F}_\mu(t) := \int_0^1 \Psi_2^{-1}(1/\mu B[t,r])\,dr$ is uniformly bounded.

If $t \in E_0$ there is no difficulty because $\mu\{0\} = 1/2$ implies $\mathbb{F}_\mu(0) \le \Psi_2^{-1}(2)$. The calculation becomes more interesting for a t in some E_j with $j \ge 1$. By inequality <6>, there is a constant C for which

$$\Psi_2^{-1}(1/\mu B[t,r]) \le \Psi_2^{-1}(1/\mu\{t\}) = \Psi_2^{-1}\left(2^{j+1}n_j\right) \le C2^{j/2} = C\lambda^j.$$

And for $r > \tau_j = \lambda^{-j}$, the ball $B[t,r]$ contains 0. Thus

$$\mathbb{F}_\mu(t) \le \int_0^{\tau_j} C\lambda^j dr + \int_{\tau_j}^1 \Psi_2^{-1}(1/\mu\{0\})\,dr \le C + \Psi_2^{-1}(2),$$

implying that μ is a majorizing measure.

Weighted partitions

As you might expect from Section 11.6, a suitable set of partitions should bear a strong similarity to the majorizing measure μ. Indeed, we can take $\mathcal{A}_i$ to consist of all singleton sets $\{t\}$ for $t \in G_i := \cup_{j\le i} E_j$ together with the set G_i^c, then take $w_i(A) = \mu A$ for each A in $\mathcal{A}_i$. As the singleton sets all have zero diameter, we have

$$\delta_i := \max_{A\in\mathcal{A}_i} \operatorname{diam}(A) = \operatorname{diam}(G_i^c) = \operatorname{diam}(E_{i+1}) = \lambda^{-i}.$$

Now consider the sum $\sum_{i=1}^{\infty} \delta_i \Psi^{-1}(1/w_i\{A_i(t)\})$ from Theorem <21>. There is no point in replacing "$i = 1$" by "$i = k + 1$" because there is no hope for getting small oscillations. Suppose $t \in E_j$. Then

$$A_i(t) = \begin{cases} \{t\} & \text{if } j \le i \\ G_i^c & \text{if } j > i \end{cases} \qquad \text{with} \quad \mu A_i(t) = \begin{cases} 1/(2^{j+1}n_j) & \text{if } j \le i \\ (1/2)^i & \text{if } j > i \end{cases}.$$

For some constant C,

$$\Psi_2^{-1}(2^{j+1}n_j) = \sqrt{\log(1+2^{j+1}n_j)} \le C2^{j/2} = C\lambda^j \quad \text{AND} \quad \Psi_2^{-1}(2^i) \le C\sqrt{i}.$$

Hence

$$\begin{aligned}\sum_{i\in\mathbb{N}} \delta_i \Psi_2^{-1}\left(1/\mu A_i(t)\right) &= \sum_{i\in\mathbb{N}} \lambda^{-i} \Psi_2^{-1}\Big(\{j\le i\}/\mu\{t\} + \{j > i\}/\mu G_i^c\Big) \\ &\le C\sum_{i=1}^{j-1} \lambda^{-i}\sqrt{i} + C\lambda^j \sum_{i=j}^{\infty} \lambda^{-i},\end{aligned}$$

which is uniformly bounded over j in $\mathbb{N}_0$.

Admissible partitions

It is less obvious how the admissible sequence of partitions should be chosen. Again write G_i for $\cup_{j\le i} E_j$. Define $\mathcal{A}_0 = \{T\}$ and, for $i \in \mathbb{N}$, define $\mathcal{A}_i$ to consist of all the singleton sets t for $t \in G_{i-1}$, together with each of the sets E_j for $i \le j \le 2i$, together with G_{2i}^c. That is,

$$\mathcal{A}_i = \{\{t\} : t \in G_{i-1}\} \cup \{E_i, \ldots, E_{2i}, G_{2i}^c\} \qquad \text{for } i \in \mathbb{N},$$

which obeys the size constraint: $|\mathcal{A}_i| = (1+n_1+\cdots+n_{i-1})+(i+1)+1 \le n_i$ with equality only at $i = 1$.

Remember that $\Psi_2^{-1}(n_i) = O(\lambda^i)$. I claim that these partitions satisify the condition

$$\sup_{t\in T} \sum_{i\in\mathbb{N}} \lambda^i \operatorname{diam}(A_i(t)) < \infty,$$

from Theorem <32>.

The inclusion of the $E_i, \ldots, E_{2i}$ provides a small "buffer region" between two desirable forms for the set $A_i(t)$. For each fixed t in E_j,

$$A_i(t) = \begin{cases} \{t\} \text{ with diameter } 0 & \text{if } j < i \\ E_j \text{ with diameter } \lambda^{1-j} & \text{if } i \le j \le 2i \\ G_{2i}^c \text{ with diameter } \lambda^{-2i} & \text{if } 2i < j \end{cases} .$$

Written another way,

$$\operatorname{diam}(A_i(t)) = 0\{j < i\} + \lambda^{1-j}\{j/2 \le i \le j\} + \lambda^{-2i}\{i < j/2\}.$$

It follows, for this particular t in E_j, that

$$\begin{aligned} \sum\nolimits_{i\in\mathbb{N}} \lambda^i \operatorname{diam}(A_i(t)) &= \lambda^{1-j} \sum\nolimits_{i\in\mathbb{N}} \{j/2 \le i \le j\}\lambda^i + \sum\nolimits_{i\in\mathbb{N}} \lambda^{-i}\{i < j/2\} \\ &\le \lambda^{1-j}\lambda^{j+1}/(\lambda-1) + 1/(\lambda-1), \end{aligned}$$

which is uniformly bounded over j in $\mathbb{N}_0$.

11.10 A comparison of chaining methods

Why did the chaining with weights (or admissible partitions) in Section 11.9 work but the approach based on packing/covering numbers fail? Each of the chaining arguments effectively make use of chaining frameworks $\{(T_i, \ell_i)\}$ via an inequality

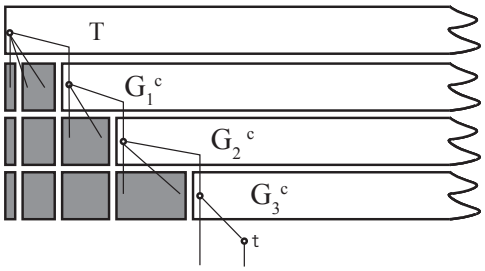
$$\mathcal{M}_{k,m}/K_0 \le \max_{t\in T_m} \sum\nolimits_{i=k+1}^{m} d(L_i t, L_{i-1}t)\Delta_i(L_i t).$$

They differ in the way they bound the expression on the right-hand side:

$$\begin{aligned} \text{packing:} \quad & \sum\nolimits_{i=k+1}^{m} \delta_i \max\nolimits_{s\in T_i} \Delta_i(s), \\ \text{weighted:} \quad & \max_{t\in T_m} \sum_{i=k+1}^{m} \delta_i \Psi^{-1}\left(\frac{1}{w_i(L_i t)}\right) + \sum_{i=k+1}^{m} \delta_i \max\nolimits_{s\in T_i} R_i(s), \\ \text{admissible:} \quad & \max\nolimits_{t\in T_m} \sum\nolimits_{i=k+1}^{m} \operatorname{diam}(A_i(t)) \max\nolimits_{s\in T_i} \Delta_i(s). \end{aligned}$$

For the last two cases, the frameworks are only implicit; they would be obtained by applying the method from Section 11.2 to the nested sequence of partitions. For weighted partitions the only ambiguity comes from the choice of a point f_i from G_i^c. For concreteness, let me choose it to be the smallest t (under the usual ordering on $\mathbb{N}_0$) in G_i^c. Thus $T_i = G_i \cup \{f_i\}$ for $i \ge 1$. Similarly, we could use $T_i = G_{i-1} \cup \{f_i, \ldots, f_{2i}\}$ for $i \ge 1$.

Consider the weighted case. (The analysis for the admissible case is similar.) The sketch indicates the path from a t in E_4 up to 0. (Shading of a box corresponding to a single E_j indicates that all the singletons belong to the T_i corresponding to that row in the picture. Also, there is not enough room to show all the edges in the tree.)



In general, if $t \in E_j$ with $j \le m$ then for $i \le m$ we have

$$\begin{aligned} L_i t &= f_i\{i<j\} + t\{i=j\} + t\{i>j\}, \\ d(L_i t, L_{i-1}t) &= d(f_{i-1}, f_i)\{i<j\} + d(t, f_{i-1})\{i=j\} + 0\{i>j\} \\ &\le \delta_i := \lambda^{1-i}, \\ \Psi_2^{-1}\left(1/w_i(L_i t)\right) &= \sqrt{\log(1+2^i)}\,\{i<j\} + \sqrt{\log(1+2^{j+1}n_j)}\,\{i\ge j\}. \end{aligned}$$

As in Section 11.9 these calculations show that the terms making the important contributions to $\sum_i \delta_i \Psi_2^{-1}\left(1/w_i(L_i t)\right)$ depend on t.

Now notice that $\sup_{t\in T} d(t, T_i) \le \delta_i$ and $|T_i| \le n_{i+1}$ when $T_i = G_i \cup \{f_i\}$. That is, T_i provides a δ_i-covering for T. We could use the method from §10.6 to bound $\mathbb{P}\max_{t\in T_m}|X(t) - X(0)|$:

$$\begin{aligned} \mathbb{P}\mathcal{M}_{0,m} &\le \sum_{i=1}^m \mathbb{P}\max_{s\in T_i}|X(s) - X(\ell_{i-1}s)| \\ &\le \sum_{i=1}^m D_i \Psi_2^{-1}(|T_i|) = \sum_{i=1}^m \lambda^{-i}\sqrt{\log(1+n_i)} \\ &\le \sum_{i=1}^m \lambda^{-i} C\lambda^{i+1} \to \infty \qquad \text{as } m\to\infty, \end{aligned}$$

the same sort of packing failure as before. In effect, the method has failed because it ignores the fact that the links that make the most important contributions to the bound change from one t to the next. Put another way, it failed because it was giving the same weight, $1/|T_i|$, to each node in T_i.

11.11 Problems

[1] Suppose $\Psi = e^g - 1 \in \mathcal{Y}_{\exp}$. Show that $\int_0^1 \Psi^{-1}(1/r)\,dr < \infty$. Hint: Change of variable $r = e^{-x}$ then use the fact that $g(x) \ge (c_0 x)^\alpha$ for all x large enough.

[2] Ledoux and Talagrand (1991, Lemma 11.9) Suppose μ is a majorizing measure on a metric space (T, d) and $F = \{\xi_1, \ldots, \xi_k\}$ is a finite subset of T. Argue as follows to show that there is another probability measure μ_0, with $\mu_0(F) = 1$, such that

$$\int_0^\delta \Psi^{-1}\left(1/\mu_0 B[\xi_i, r]\right)\,dr \le 2\int_0^{\delta/2} \Psi^{-1}\left(1/\mu B[\xi_i, r]\right)\,dr \qquad \text{for each } \xi_i \text{ in } F.$$

for each ξ_i in F and each $\delta > 0$. That is, μ_0 is a majorizing measure for the metric space (F, d).

(i) Show that there is a measurable map $\pi : T \to F$ for which $d(t, \pi(t)) = \min_i d(t, \xi_i)$ for each t in T. Define μ_0 to be the image of μ under the map π. (That is, $\mu_0 A = \mu\pi^{-1}A$ for each subset A of F.)

(ii) Consider a fixed ξ_i in F. For $t \in B[\xi_i, r]$ show that

$$d(\xi_i, \pi(t)) \le d(\xi_i, t) + d(t, \pi(t)) + \le 2d(\xi_i, t) \le 2r.$$

Deduce that $B[\xi_i, r] \subset \pi^{-1}B[\xi_i, 2r]$ and hence $\mu B[\xi_i, r] \le \mu_0 B[\xi_i, 2r]$.

(iii) Make a change of variable to obtain the asserted inequality.

[3] Suppose μ is a MM on (T, d) and $\Psi = e^g - 1 \in \mathcal{Y}_{\exp}$. For each $i \in \mathbb{N}_0$ let T_i be an δ_i-packing set for T, where $\delta_i = D/2^i$. Use Problem [2] to find a probability measure ν_i on T_i for which $\mu B[\xi, r] \le \nu_i B[\xi, 2r]$ for each $r > 0$ and ξ in T_i. Define $\nu = \sum_{i\in\mathbb{N}} 2^{-i}\nu_i$. Show that ν is a MM that concentrates on the countable set $T_\infty := \cup_i T_i$ by the following steps.

(i) If $t \in T$ and $r > \delta_i$ show that there is a ξ in T_i with $d(t, \xi) < r$ and

$$\nu_i B[t, 5r] \ge \nu_i B[\xi, 4r] \ge \mu B[\xi, 2r] \ge \mu B[t, r].$$

(ii) Deduce that

$$\begin{aligned}\int_{\delta_i}^{\delta_{i-1}} \Psi^{-1}\left(1/\nu B[t, 5r]\right)\, dr &\le \int_{\delta_i}^{\delta_{i-1}} \Psi^{-1}\left(2^i/\nu_i B[t, 5r]\right)\, dr \\ &\le \int_{\delta_i}^{\delta_{i-1}} \Psi^{-1}\left(2^i/\mu B[t, r]\right)\, dr \\ &\le C_g\delta_i\Psi^{-1}(2^i) + C_g \int_{\delta_i}^{\delta_{i-1}} \Psi^{-1}\left(1/\mu B[t, r]\right)\, dr\end{aligned}$$

(iii) By summing over i and making a change of variable, show that

$$1/5 \int_{\delta_i}^{5\delta_k} \Psi^{-1}\left(1/\nu B[t, r]\right)\, dr \le C_k + C_g \int_0^{\delta_k} \Psi^{-1}\left(1/\nu B[t, r]\right)\, dr$$

where $C_k := \sum_{i=k+1}^{\infty} \delta_i\Psi^{-1}(2^i)$.

(iv) Show $C_0 > C_1 > \cdots \to 0$ as $k \to \infty$.

(v) Deduce that ν is a MM on T.

[4] Suppose $\{(T_i, \ell_i) : i \in \mathbb{N}_0\}$ is a chaining framework and $w_i : T_i \to (0, 1]$ is a weight function with $0 < w_i(t) \le 1$ for each t in T_i. Assume

(a) $\max_{s\in T_i} d(s, \ell_{i-1}s) \le \delta_i$ for each i.

(b) $\Psi \in \mathcal{Y}_{\exp}$ and $\sum_i \delta_i\Psi^{-1}(2^i w_i(T_i)) < \infty$, where $w_i(T_i) := \sum_{s\in T_i} w_i(s)$.

(c) $\beta_k := \sup_{m>k} \max_{t\in T_m} \sum_{i=k+1}^{m} \delta_i \Psi^{-1}\left(1/w_i(L_i t)\right) < \infty$.

Let ν_i be the measure that puts mass $w_i(s)$ each s in T_i. Argue as follows to show that $\mu := \sum_{i\in\mathbb{N}} 2^{-i}\nu_i/\nu_i T_i$ is a MM concentrating on the countable dense set $T_\infty = \cup_i T_i$.

(i) Show that $d(t, L_k t) \le D_k := \sum_{i\ge k} \delta_i < \infty$ if $t \in T_m$ and $k < m$.

(ii) Deduce that $\mu B[t, D_k] \ge 2^{-k} w_k(L_k t)/w_k(T_k)$ so that

$$\Psi^{-1}\left(1/\mu B[t, D_k]\right) \le C_g\left(\Psi^{-1}(2^k w_k(T_k)) + \Psi^{-1}\left(1/w_k(L_k t)\right)\right) \qquad \text{for } t \in T_m.$$

(iii) If $t \in T_m$ and $k < m$ deduce that

$$\int_{D_k}^{D_{k-1}} \Psi^{-1}\left(1/\mu B[t, r]\right)\, dr \le \delta_{k-1} C_g\left(\Psi^{-1}(2^k \nu_k T_k) + \Psi^{-1}\left(1/\nu_k\{L_k t\}\right)\right).$$

(iv) Sum over k to complete the proof. (In fact you should have proved something stronger than the MM assertion.)

[5] Suppose (T, d) is a totally bounded metric space equipped with a nested sequence of finite partitions, $\mathbb{A} = \{\mathcal{A}_i : i \in \mathbb{N}_0\}$, with corresponding chaining framework $\{(T_i, \ell_{i-1}) : i \in \mathbb{N}\}$. As in Theorem <34>, assume $N_i := |\mathcal{A}_i| \ge 2^i$ and $\Psi^{-1}(N_i) \le C\lambda^i$ for some ORLICZ function $\Psi = e^g - 1$ in $\mathcal{Y}_{\exp}$. Suppose also that

$$\sup_{t\in T} \sum\nolimits_{i\in\mathbb{N}} \delta_i(t)\lambda^i < \infty \qquad \text{where } \delta_i(t) := \operatorname{diam}\left(A_{i-1}(t)\right).$$

Let μ_i be the uniform distribution on T_i. Argue as follows to show that $\mu := \sum_{i\in\mathbb{N}} 2^{-i}\mu_i$ is a MM that concentrates on the countable dense set $T_\infty = \cup_i T_i$.

(i) By Example <15> it suffices to consider t in some T_m. For $r > \delta_{i+1}(t)$ show that $\mu B[t, r] \ge 2^{-i}/N_i \ge N_i^{-2}$. Deduce that, for some constant C,

$$\int_{\delta_{i+1}(t)}^{\delta_i(t)} \Psi^{-1}(1/\mu B[t, r])\, dr \le \delta_i(t)\Psi^{-1}(N_i^2) \le C\delta_i(t)\lambda^i.$$

(ii) Sum over i then complete the argument.

[6] As in §10.7, suppose $\{X_t : t \in T\}$ is a stochastic process and $\Psi = e^g - 1$ is an ORLICZ function in $\mathcal{Y}_{\exp}$ for which:

(a) $\|X_s - X_t\|_\Psi \le K_0 d(s, t)$ for all $s, t \in T$.

(b) $\{(T_i, \ell_i) : i \in \mathbb{N}_0\}$ is a chaining framework, not necessarily nested, on T for which to each $\epsilon > 0$ there exists a finite $k(\epsilon)$ such that $\mathbb{P}\mathcal{M}_{k,m} \le \epsilon$ if $m > k \ge k(\epsilon)$.

Show that there is a finite $N(\epsilon)$ for which: to each finite subset F of T there exists an $F_0 \subset T$ with $|F_0| \le N(\epsilon)$ and a map $\gamma : F \to F_0$ such that

$$\mathbb{P} \max_{s\in F} |X(s) - X(\gamma s)| \le \epsilon.$$

Argue as follows. Consider any fixed, finite F.

(i) Show that $\delta_{m,F} := \max\{d(s, T_m) : s \in F\} \to 0$ as $m \to \infty$. Choose $\pi_m : F \to T_m$ so that $d(s, \pi_m s) = \delta_{m,F}$. Define $\gamma := L_k \circ \pi_m$

(ii) For $m > k \ge k(\epsilon/2)$ show that

$$\begin{aligned}\mathbb{P} \max_{s\in F} |X(s) - X(\gamma s)| &\le \mathbb{P} \max_{s\in F} |X(s) - X(\pi_m s)| + \mathbb{P}\mathcal{M}_{k,m} \\ &\le K_0 \delta_{m,F} \Psi^{-1}(|F|) + \epsilon/2,\end{aligned}$$

which is smaller than ϵ if m is large enough.

Remark. Similar results can be proved using other measures of closeness, such as the one corresponding to convergence in probability.

[7] (based on Fernique, 1975, Ex 5.2.2) Suppose $\{X_t : t \in \mathbb{N}\}$ is set of independent random variables with $X_t \sim N(0, \sigma_t^2)$ with $\sigma_t \ne 0$. (Similar to the setting in Section 11.9 except that the variances are not specified and that the $X_0 \equiv 0$ is not included.) Define $M_n := \max_{t\le n} X_t$ and $M_\infty := \sup_t X_t$. Suppose $\mathbb{P}M_\infty < \infty$.

(i) For each $R > 0$ and each $n \in \mathbb{N}$ show that

$$\begin{aligned}\mathbb{P}\{M_\infty \le R\} \le \mathbb{P}\{M_n \le R\} &= \prod\nolimits_{t\le n} \left(1 - \bar{\Phi}(R/\sigma_t)\right) \\ &\le \exp\left(-\sum\nolimits_{t\le n} \bar{\Phi}(R/\sigma_t)\right).\end{aligned}$$

(Here, as in Chapter 3, $\bar{\Phi}(x) := \mathbb{P}\{Z \ge x\}$ for $Z \sim N(0, 1)$. You might find it useful to know that $\bar{\Phi}(x) \ge \exp(-x^2)/3$ for $x \ge 0$.) Deduce that there exists an R for which $\sum_{t\in\mathbb{N}} \bar{\Phi}(R/\sigma_t) \le 1$.

(ii) Define a sub-probability measure μ on $\mathbb{N}_0$ by $\mu\{0\} = 1/2$ and $\mu\{t\} = \bar{\Phi}(R/\sigma_t)/2$ for $t \ge 1$. Show that there exists a MM ν on $\mathbb{N}_0$ with $\mu \le \nu$. Hint: The MM argument in Section 11.9.

(iii) Does it really matter that ν lives on $\mathbb{N}_0$ rather than on $\mathbb{N}$? For example, could ν be used a smoother as in Section 11.4? Note that $\max(0, M_\infty) = M_\infty$ almost surely.

11.12 Notes

While writing this Chapter I benefitted greatly from reading the expositions by Ledoux (1996, Chap 6) and van Handel (2016).

Garsia, Rodemich, and Rumsey (1970) developed a smoothing argument for functions on the real line, with LEBESGUE measure acting as the

smoother. Garsia (1972) simplified the argument and extended it to establish a sufficient condition for a gaussian process indexed by the unit hypercube in multidimensional euclidean space to have continuous sample paths. Preston (1972) extended that method to gaussian processes indexed by compact metric spaces. (According to mathgenealogy.org, Garsia was Preston's advisor.) I have not tried to work through all the earlier papers cited by these authors, even though I suspect the smoothing idea has a longer history. Compare with Bednorz (2006, 2007).

See the most informative book review by Ledoux (1999) for an account of the early history regarding majorizing measures and gaussian processes, with particular reference to the role played by Xavier Fernique.

For Section 11.9, the idea to use constant variances over large blocks came from remarks made by John Hartigan during a seminar.

References

Andersen, N. T., E. Giné, M. Ossiander, and J. Zinn (1988). The central limit theorem and the law of the iterated logarithm for empirical processes under local conditions. *Probability Theory and Related Fields 77*, 271–306.

Bednorz, W. (2006). A theorem on majorizing measures. *Annals of Probability 34*, 1771–1781.

Bednorz, W. (2007). Hölder continuity of random processes. *Journal of Theoretical Probability 20*, 917–934.

Fernique, X. (1971). Régularité de processus gaussiens. *Inventiones mathematicae 12*, 304– 320.

Fernique, X. (1975). Regularité des trajectoires des fonctions aléatoires gaussiennes. *Springer Lecture Notes in Mathematics 480*, 1–97. École d'Été de Probabilités de Saint-Flour IV, 1974.

Fernique, X. (1981). Mesures majorantes, majoration et continuité de fonctions aléatoires, exemples de construction. *Springer Lecture Notes in Mathematics 860*, 124–137. Probability in Banach Spaces III, Medford 1980.

Garsia, A. M. (1972). Continuity properties of Gaussian processes with multidimensional time parameter. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, pp. 369–375. University of California Press.

Garsia, A. M., E. Rodemich, and H. Rumsey (1970). A real variable lemma and the continuity of paths of some Gaussian processes. *Indiana University Mathematics Journal 20*(6), 565–578.

Ledoux, M. (1996). Isoperimetry and Gaussian analysis. *Springer Lecture Notes in Mathematics 1648*, 165–294. École d'Été de Probabilités de St-Flour XXIV , 1994.

Ledoux, M. (1999). Review of 'Fonctions aléatoires gaussiennes, vecteurs aléatoires gaussiens' by Xavier Fernique. *Annals of Probability 27*(4), 2109–2114. (Book published by CRM, Montreal, 1997).

Ledoux, M. (2014). Appendix to Chapter 6 of Ledoux's 1994 St. Flour lectures on "Isoperimetry and Gaussian Analysis". Available at https://perso.math.univ-toulouse.fr/ledoux/books/.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. New York: Springer.

Preston, C. (1972). Continuity properties of some Gaussian processes. *The Annals of Mathematical Statistics 1972*(1), 285–292.

Talagrand, M. (1987). Regularity of Gaussian processes. *Acta Mathematica 159*, 99–149.

Talagrand, M. (1992). A simple proof of the majorizing measure theorem. *Geometric and Functional Analysis 2*(1), 118–125.

Talagrand, M. (1996). Majorizing measures: The generic chaining. *Annals of Probability 24*, 1049–1103.

Talagrand, M. (2001). Majorizing measures without measures. *The Annals of Probability 29*(1), 411–417.

Talagrand, M. (2021). *Upper and Lower Bounds for Stochastic Processes: Decomposition Theorems* (Second ed.), Volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag.

van Handel, R. (2016, December). Probability in high dimension. APC 550 Lecture Notes, Princeton University.

Chapter 12
Majorizing measures for gaussian processes

SECTION *12.1 introduces a functional that underlies the remarkable properties of centered gaussian processes with bounded sample paths. Using special properties of that functional, it then derives a strengthened version of the* SUDAKOV *minoration, which plays a key role for developing lower bounds to complement the chaining upper bounds.*

SECTION *12.2 constructs nested partitions with weights for centered gaussian processes with bounded sample paths.*

SECTION *12.3 describes the effect on the majorizing measure lower bound if a centered gaussian processes has uniformly continuous sample paths.*

SECTION *12.4 constructs admissible partitions for* DOOB*-separable gaussian processes with bounded sample paths.*

SECTION *12.5 briefly describes why a thorough understanding of the gaussian case is a good starting point for further exploration of Talagrand's amazing discoveries.*

12.1 Introduction

Broadly speaking, Chapter 11 established the equivalence of three methods (majorizing measures, weighted frameworks, admissible partitions) to establish maximal inequalities for a stochastic process whose increments are controlled by an ORLICZ norm. This Chapter will focus on the companion lower bounds for gaussian processes.

Throughout the Chapter, $X := \{X_t : t \in T\}$ will always be a centered gaussian process with bounded sample paths, which is at least DOOB-separable. As explained in §9.6, there is no loss of generality in assuming T to be countably infinite, an assumption that ***will be implicit for the whole of this Chapter***. The index set T will always by equipped with the natural metric, $d(s,t) := \|X_s - X_t\|_2$. The diameter of T, the supremum of $d(s,t)$ over all pairs of points s and t in T, will always be denoted by $\operatorname{diam}(T)$ or abbreviated to D.

For such a T, Problem [1] shows that boundedness of sample paths is equivalent to finiteness of $\mathbb{P}\sup_{t\in T} X_t$. Consequently, the functional

$$\mathfrak{F}(A) := \mathbb{P}\sup_{t\in A} X_t \qquad \text{for } A \subset T \tag{1}$$

is well-defined and finite.

Remark. As explained in §9.7, there is no loss of generality in assuming $d(s,t) > 0$ for $s \neq t$.

The functional $\mathfrak{F}$ is the key to the whole analysis. It works its magic through the following strengthened form of the SUDAKOV minoration, which encapsulates everything we need to know about gaussianity.

<2> **Lemma.** *Let $\{X_t : t \in T\}$ be a centered gaussian process with $\mathfrak{F}(T) < \infty$. If $\{\xi_1, \ldots, \xi_k\}$ is a δ-separated subset of T and $B_j \subseteq B[\xi_j, \theta\delta]$ for each j, then*

<3> $$\mathfrak{F}\left(\bigcup_{j\le k} B_j\right) \ge C_\theta \delta \mathcal{L}(k) + \min_{j\le k} \mathfrak{F}(B_j) \qquad \text{where } \mathcal{L}(k) := \sqrt{\log k},$$

□ *with $C_\theta := C_{sud} - 4\theta$, which is > 0 if θ is small enough.*

Remark. I write ξ_j, instead of the more natural t_j, to avoid subsequent confusion with the points in a chaining framework.

Using just this Lemma we will be able to recursively construct two slightly different nested sequences of finite partitions $\mathbb{A} = \{\mathcal{A}_i : i \in \mathbb{N}_0\}$ of T. The easier construction, in Section 12.2, leads to weighted frameworks like those in §11.5; the other, in Section 12.4, involves some extra subtleties in building admissible sequences like those in §11.7.

Of course there is some redundancy in proving two separate results equivalent to the existence of a majorizing measure (MM). However, I think it worthwhile to see both constructions as a way of comparing the advantages of (and the technical difficulties involved in) both approaches.

The Proof of the Lemma requires two gaussian facts from §6.1:

GSUD (SUDAKOV minoration) Suppose $W := (W_1, W_2, \ldots, W_k)$ has a centered MVN/ distribution with $\mathbb{P}|W_i - W_j|^2 \ge \delta^2$ for all distinct i and j. Then $\mathbb{P}\max_{j\le k} W_j \ge C_{\text{sud}}\delta\mathcal{L}(k)$ with C_{sud} a universal (positive) constant and $\mathcal{L}(k) := (\log k)^{1/2}$.

GCONC (SUBGAUSSIAN max) Suppose $M := \sup_{s\in S} Y_s$, where $\{Y_s : s \in S\}$ is a centered gaussian process with S at worst countably infinite. If both $\mathbb{P}M$ and $\sigma^2 := \sup_{s\in B} \operatorname{var}(Y_s)$ are finite then $M - \mathbb{P}M$ has a subgaussian distribution:

$$\mathbb{P}e^{\lambda(M-\mathbb{P}M)} \le \exp(\sigma^2\lambda^2/2) \qquad \text{for all real } \lambda.$$

Proof (of Lemma <2>). Inequality <3> is trivial if k equals 1, so assume $k \ge 2$.

Define $W_j := X_{\xi_j}$ and $Y_t := X_t - W_j$ for $t \in B_j$ and

$$M_j := \sup_{t\in B_j} Y_t = \sup_{t\in B_j} X_t - W_j.$$

Notice that $\mathbb{P}M_j = \mathbb{P}\sup_{t\in B_j} X_t - \mathbb{P}W_j = \mathfrak{F}(B_j)$ and

$$\operatorname{var}(Y_t) = \|X_t - W_j\|_2^2 = d(t,\xi_j)^2 \le (\theta\delta)^2 \qquad \text{for all } t \text{ in } B_j.$$

Thus $\mathbb{P}e^{\lambda(M_j - \mathfrak{F}(B_j))} \le e^{(\lambda\theta\delta)^2/2}$ for all real λ, by GCONC. The usual JENSEN trick then leads to a maximal ineualty for $\mathbb{M} := \max_j |M_j - \mathfrak{F}(B_j)|$:

$$e^{\lambda\mathbb{P}\mathbb{M}} \le \mathbb{P}e^{\lambda\mathbb{M}} \le \sum\nolimits_j \mathbb{P}\Big(e^{\lambda(M_j-\mathfrak{F}(B_j))} + e^{-\lambda(M_j-\mathfrak{F}(B_j))}\Big) \le 2ke^{(\lambda\theta\delta)^2/2}.$$

Take logs then (assuming $\lambda > 0$) divide through by λ. The choice $\lambda = \sqrt{2}\,\mathcal{L}(2k)/(\theta\delta)$ minimizes, leaving

<4> $$\mathbb{P}\mathbb{M} \le 2\theta\delta\sqrt{2\log(2k)} \le 4\theta\delta\mathcal{L}(k).$$

Similarly, from $\mathbb{P}|W_i - W_j|^2 = d(\xi_i,\xi_j)^2 > \delta^2$ for $i \ne j$ we get

<5> $$\mathbb{P}\max_{i\le k} W_j \ge C_{\text{sud}}\delta\mathcal{L}(k) \qquad \text{by GSUD.}$$

Roughly speaking, $\sup_{t\in B_j} X_t = W_j + M_j$ should be close to the corresponding $W_j + \mathfrak{F}(B_j)$ for each j, by virtue of GCONC. The supremum of X_t over $\cup_j B_j$ should then be close to the maximum of $W_j + \mathfrak{F}(B_j)$. More precisely,

$$\begin{aligned}\sup_{t\in\cup_j B_j} X_t &= \max_j \sup_{t\in B_j} X_t \\ &= \max_j\big(W_j + \mathfrak{F}(B_j) + M_j - \mathfrak{F}(B_j)\big) \\ &\ge \max_j\big(W_j + \min_j \mathfrak{F}(B_j) - \mathbb{M}\big) = \max_j W_j + \min_j \mathfrak{F}(B_j) - \mathbb{M}.\end{aligned}$$

Take expected values.

$$\mathfrak{F}\Big(\bigcup\nolimits_{j\le k} B_j\Big) = \mathbb{P}\sup_{t\in\cup_j B_j} X_t \ge \max_j W_j + \min_j \mathfrak{F}(B_j) - \mathbb{P}\mathbb{M}.$$

□ Inequalities <4> and <5> then deliver the asserted inequality.

12.2 From functional to nested partitions with weights

The Proof of the following Theorem comes from Talagrand (1992) via the exposition by Ledoux (1996, Chapter 6). Once again, the index set T is assumed countable and equipped with the metric $d(s,t) := \|X_s - X_t\|_2$.

<6> **Theorem.** *Suppose $\{X_t : t \in T\}$ is centered gaussian process for which $\mathfrak{F}(T) := \mathbb{P}\sup_{t\in T} X_t$ is finite. Define $\theta := \min(C_{sud}/5, 1/3)$ and $\delta_i := D\theta^i$. Then there exists a nested sequence of finite partitions $\mathbb{A} = \{\mathcal{A}_i : i \in \mathbb{N}_0\}$ with $\mathcal{A}_0 = \{T\}$ and $\max_{A\in\mathcal{A}_i} \operatorname{diam}(A) \le 2\delta_i$, and weight functions $w_i : \mathcal{A}_i \to (0,1]$ with $w_i(\mathcal{A}_i) := \sum_{A\in\mathcal{A}_i} w_i(A) \le 2^i$, for which*

$$C\mathfrak{F}(T) \ge \sum\nolimits_{i\in\mathbb{N}}^{\infty} \delta_i\sqrt{\log(1/w_i(A_i(t))} \qquad \text{for each } t \in T,$$

where $A_i(t)$ denotes the member of $\mathcal{A}_i$ that contains t and C is a constant
□ *that depends on the choice for θ.*

This Theorem complements the main result from §11.5, which was written using a weighted chaining framework. You might find it instructive to rewrite that result using weighted partitions.

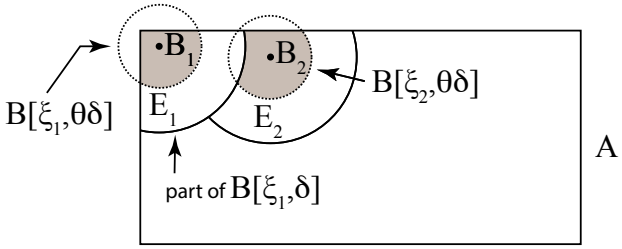
Remark. As noted at the start of §11.6, the existence of such a weighting scheme implies existence of a MM μ on T for which $C_1\mathfrak{F}(T) \geq \sup_{t\in T}\int_0^{\operatorname{diam}(T)} \Psi_2^{-1}(1/B[t,r])\,dr$ with C_1 a universal constant.

The proof of Theorem <6> will consist of recursive appeals to a greedy procedure (WPARTITION, with the leading W to suggest weights) that partitions a set into smaller pieces (of diameter at most 2δ) using a localized version of the functional,

<7> $$f(S,r) := \sup_{s\in S} \mathfrak{F}(S\cap B[s,r]) \qquad \text{for each subset } S \text{ of } T.$$

The starting point is a set A with weight W. The η provides some wiggle room when the supremum is not achieved at some s in S.

```
1:   procedure WPARTITION(A, δ, W, η)
2:       Initialize: j = 1, S_1 = A.
3:       loop
4:           Find a ξ_j in S_j for which the set B_j := S_j ∩ B[ξ_j, θδ]
                 satisfies 𝔉(B_j) ≥ f(S_j, θδ) − η.
5:           Define E_j := S_j ∩ B[ξ_j, δ].        % The jth child of A.
6:           Attach weight w(E_j) := W/j^2 to the set E_j.
7:           Define S_{j+1} := S_j \ E_j.
8:           If S_{j+1} is empty then
9:               Exit the loop.
10:          else
11:              Increase j by 1 (and continue with the next iteration).
12:          end If
13:      end loop
14:  end procedure
```



After the loop exits we are left with a δ-separated set of points $\xi_1,\ldots,\xi_J$ from A (so that $J \leq \text{PACK}(\delta, A) < \infty$), a weighted partition $E_1,\ldots,E_J$ of A, and sets $S_{j+1} = S_j\backslash E_j$ with $A = S_1 \supset S_2 \supset \cdots \supset S_J = E_J$ for which

$$\xi_j \in B_j := S_j \cap B[\xi_j,\theta\delta] \subset E_j := S_j \cap B[\xi_j,\delta].$$

The greed ensures that

$$\eta + \mathfrak{F}(B_j) \geq f(S_j, \theta\delta) \geq \mathfrak{F}(S_j \cap B[s, \theta\delta]) \qquad \text{for each } s \text{ in } S_j.$$

Most importantly, for $1 \leq \alpha \leq J$, Lemma <2> gives

$$\begin{aligned} \mathfrak{F}\left(\cup_{j\leq\alpha} B_j\right) &\geq C_\theta\delta\sqrt{\log\alpha} + \min_{j\leq\alpha}\mathfrak{F}(B_j) \\ &\geq 2^{-1/2}C_\theta\delta\sqrt{\log\alpha^2} + \min_{j\leq\alpha} f(S_j, \theta\delta) - \eta \\ &= C_1\delta\sqrt{\log(W/w(E_\alpha))} + \sup_{\xi\in S_\alpha}\mathfrak{F}\left(S_\alpha \cap B[\xi, \theta\delta]\right) - \eta \end{aligned} \qquad \text{<8>}$$

where $C_1 := 2^{-1/2}C_\theta$. This inequality will provide an upper bound for $\mathfrak{F}(E'')$ for a partitioning set E'' generated in a later iteration of WPARTITION.

Remark. For <8> to hold when α equals 1 then we must have $w(E_1) = W$. The choice $w(E_j) = W/j^2$ then gives a bound $\sum_j w(E_j) \leq 2W$ (Compare with $\sum_{j\in\mathbb{N}} j^{-2} \approx 1.65$.) That was my reason for allowing $w_i(T_i) > 1$ in §11.5 and for inserting the 1/3 into the definition of θ. Talagrand (1992, p. 122)'s argument seems to corresponds to weights $w(E_j) = W/(j+1)^2$ and Ledoux (1996, p. 243)'s to $w(E_j) = W/(2j^2)$, presumably to ensure $\sum_j w(E_j) \leq W$.

Proof (of Theorem <6>, based on Ledoux 1996, pp. 242–244). For reasons that will become clearer towards the end of the proof, define $\eta_i := \mathfrak{F}(T)/2^i$ for $i \in \mathbb{N}$.

The construction consists of recursive appeals to WPARTITION, starting from $\mathcal{A}_0 = \{T\}$ and $w_0(T) = 1$.

(i) Let $\mathcal{A}_1$ consist of the sets $E_1, E_2, \ldots, E_J$, with weights $w_1(E_j) = j^{-2}$, generated by WPARTITION$(T, \delta_1, w_0(T), \eta_1)$.

(ii) Then partition each A in $\mathcal{A}_1$ by WPARTITION$(A, \delta_2, w_1(A), \eta_2)$.

(iii) And so on. In general, WPARTITION$(A, \delta_{i+1}, w_i(A), \eta_{i+1})$ applied to each A in $\mathcal{A}_i$ generates $\mathcal{A}_{i+1}$.

By construction, $w_{i+1}(\mathcal{A}_{i+1}) \leq 2w_i(\mathcal{A}_i)$ and hence $w_i(\mathcal{A}_i) \leq 2^i$ for each i.

Consider a fixed t in T. To simplify notation write A_j instead of $A_j(t)$ for $j = 0, 1, \ldots$. Suppose $A_{i+1} = E'_\alpha$ (the αth child of A_i) and $A_{i+2} = E''_\beta$ (the βth child of A_{i+1}) as generated by

$$\begin{aligned} &\text{WPARTITION}(A_i, \delta_{i+1}, w_i(A_i), \eta_{i+1}) \qquad \text{with } \delta_{i+1} = D\theta^{i+1} = \theta\delta_i, \\ &\text{WPARTITION}(A_{i+1}, \delta_{i+2}, w_{i+1}(A_{i+1}), \eta_{i+2}) \qquad \text{with } \delta_{i+2} = \theta^2\delta_i. \end{aligned}$$

The primes identify quantities involved in the construction of $\mathcal{A}_{i+1}$ and the double primes identify quantities involved in the construction of $\mathcal{A}_{i+2}$. (Some such notation is needed, for example, to distinguish between the E_1 that is a member of $\mathcal{A}_{i+1}$ and the E_1 that is a member of $\mathcal{A}_{i+1}$.)

The construction gives the following pattern of inclusions and inequalities:

$$\begin{array}{lllll} A_i \supset S'_\alpha & \supset & A_{i+1} = & E'_\alpha & \supset \quad B'_\alpha = S'_\alpha \cap B[\xi'_\alpha, \theta^2\delta_i] \\ & & & \cup & \\ & & & S''_\beta & \supset \quad A_{i+2} = E''_\beta \quad = S''_\beta \cap B[\xi''_\beta, \theta^2\delta_i]; \end{array}$$

and

$$\begin{aligned} &\mathfrak{F}A_i \geq C_1\delta_{i+1}\sqrt{\log\left(w_i(A_i)/w_{i+1}(E'_\alpha)\right)} + \mathfrak{F}B'_\alpha \qquad \text{by } <8>; \\ &\eta_{i+1} + \mathfrak{F}(B'_\alpha) \geq \sup_{\xi\in S'_\alpha} \mathfrak{F}(S'_\alpha \cap B[\xi, \theta^2\delta_i]) \geq \mathfrak{F}E''_\beta \qquad \text{because } \xi''_\beta \in S''_\beta \subset S'_\alpha. \end{aligned}$$

It follows that

$$\eta_{i+1} + \mathfrak{F}\left(A_i\right) \geq C_1\delta_{i+1}\sqrt{\rho_{i+1}} + \mathfrak{F}\left(A_{i+2}\right)$$

<9> $$\text{where } \rho_{i+1} := \log\left(w_i(A_i)/w_{i+1}(A_{i+1})\right).$$

If you write out <9> for $i = 0, 2, \ldots, 2m$, add, cancel a bunch of terms, then let m tend to infinity (and remember that $A_0(t) = T$) you should get

$$\mathfrak{F}(T) + \sum\nolimits_{j \text{ odd}} \eta_j \geq C_1 \sum\nolimits_{j \text{ odd}} \delta_j\sqrt{\rho_j}.$$

A similar summation over $i = 1, 3, 5, \ldots$ gives

$$\mathfrak{F}(A_1) + \sum\nolimits_{j \text{ even}} \eta_j \geq C_1 \sum\nolimits_{j \text{ even}} \delta_j\sqrt{\rho_j}.$$

Replace $\mathfrak{F}(A_1)$ by the larger $\mathfrak{F}(T)$ then add the sums for even and odd j to deduce (remembering that $\eta_i = \mathfrak{F}(T)/2^i$) that

$$3\mathfrak{F}(T) = 2\mathfrak{F}(T) + \sum\nolimits_{i\in\mathbb{N}} \eta_j \geq C_1 \sum\nolimits_{i\in\mathbb{N}} \delta_i\sqrt{\rho_i}.$$

This inequality does not quite match the inequality asserted by Theorem <6>, which has $\sqrt{\log(1/w_i(A_i(t)))}$ in place of $\sqrt{\rho_i(t)}$. That problem is easily fixed because $w_0(A_0) = w_0T = 1$ and

$$\sum\nolimits_{j=1}^{i} \rho_j(t) = \log\left(\frac{w_0(T)}{w_1(A_1)}\frac{w_1(A_1)}{w_2(A_2)}\cdots\frac{w_{i-1}(A_{i-1})}{w_i(A_i)}\right) = \log\left(\frac{1}{w_i(A_i)}\right).$$

Using the inequality $\left(\sum_i a_i\right)^{1/2} \leq \sum_i a_i^{1/2}$ for $a_i \in \mathbb{R}^+$ and making an interchange in the order of summation we then get

$$\begin{aligned} \sum\nolimits_{i\in\mathbb{N}} \delta_i\sqrt{\log(1/w_i(A_i(t)))} &= \sum\nolimits_{i\in\mathbb{N}} \delta_i\sqrt{\sum\nolimits_{j\in\mathbb{N}} \{j \leq i\}\rho_j(t)} \\ &\leq \sum\nolimits_{j\in\mathbb{N}} \sqrt{\rho_j(t)} \sum\nolimits_{i\in\mathbb{N}_0} \delta_i\{j \leq i\} \\ &= \sum\nolimits_{j\in\mathbb{N}} \sqrt{\rho_j(t)}\,\delta_j/(1-\theta). \end{aligned}$$

□ It is now just a matter of juggling some constants.

What a beautiful argument! Fernique was correct: existence of a MM is both necessary and sufficient for a (DOOB-separable) centered gaussian process to have bounded sample paths.

*12.3 Uniformly continuous gaussian sample paths

As you saw in §10.7, if the increments of a DOOB-separable process are controlled by an ORLICZ norm $\|\cdot\|_\Psi$ then a sufficient condition for the process to have uniformly continuous sample paths is the existence of a MM μ for which

$$\sup_{t\in T}\int_0^\delta \Psi^{-1}(1/\mu B[t,r])\,dr \to 0 \qquad \text{as } \delta\to 0.$$

For gaussian processes this condition, with the ORLICZ function $\Psi(x)$ specialized to $\Psi_2(x) := \exp(x^2)-1$, is also necessary.

<10> **Theorem.** *If $X = \{X_t : t \in T\}$ is a centered gaussian process with bounded, uniformly continuous sample paths (for $d(s,t) := \|X_s - X_t\|_2$) then there exists a MM μ on T for which*

$$\sup_{t\in T}\int_0^\delta \Psi_2^{-1}(1/\mu B[t,r])\,dr \to 0 \qquad \text{as } \delta\to 0.$$

Proof. The argument will combine the virtues of packing numbers and majorizing measures.

The boundedness implies $\mathfrak{F}(T) < \infty$. The SUDAKOV minoration ensures that T is totally bounded. The assumed uniform continuity implies

$$\text{OSC}(\delta, X, T) := \sup\{|X_s - X_t| : d(s,t) < \delta\} \to 0 \qquad \text{as } \delta\to 0.$$

Define $\delta_k := 2^{-k}$. By dominated convergence (with $2\sup_{t\in T}|X_t|$ as the dominating function) we have

$$\beta_k := \mathbb{P}\text{OSC}(\delta_k, X, T) \to 0 \qquad \text{as } k\to\infty.$$

Total boundedness implies that $\text{PACK}(\delta, T)$ is finite for each positive δ. As in §10.4, there exists a nested sequence of finite subsets $\{T_k : k\in\mathbb{N}\}$ with T_k a δ_k-packing set, $|T_k| = N_k \le \text{PACK}(\delta_k, T)$. Also there exists a map $\tau_k : T \to T_k$ with $\sup_{s\in T} d(s,\tau_k(s)) \le \delta_k$. The sets $\{s\in T : \tau_k(s) = t\}$ for $t\in T_k$ form a partition $\mathcal{E}_k$ of T with each E in $\mathcal{E}_k$ containing exactly one point, $\tau_k(E)$, from T_k.

From the fact that $\sup_{s\in E} d(s,\tau_k(E)) \le \delta_k$ we get

$$\mathfrak{F}(E) := \mathbb{P}\sup_{t\in E} X_t = \mathbb{P}\sup_{t\in E}(X_t - X_{\tau_k(E)}) \le \beta_k \qquad \text{for } E\in\mathcal{E}_k.$$

Theorem <6> then provides a probability measure μ_E on E for which

<11> $$\sup_{t\in E}\int_0^{\operatorname{diam}(E)} \Psi_2^{-1}(1/\mu_E B[t,r])\,dr \le C\beta_k,$$

where C is a constant that doesn't depend on E or k.

I claim that the probability measure

$$\mu := \sum\nolimits_{k\ge 1} \frac{1}{2^k N_k} \sum\nolimits_{E\in\mathcal{E}_k} \mu_E,$$

is a MM with the property asserted by the Theorem. The proof is easy.

For any given positive ϵ first choose $k = k(\epsilon)$ so that $\beta_k < \epsilon$ then choose $\delta := \epsilon/\Psi_2^{-1}(2^k N_k)$. Consider any point t in T. To simplify notation write E_k for the member of $\mathcal{E}_k$ that contains t and abbreviate μ_{E_k} to μ_k. For $r \le \delta$ use the fact that $\mu B[t,r] \ge (2^k N_k)^{-1}\mu_k B[t,r]$ and the inequality $\Psi_2^{-1}(uv) \le \Psi_2^{-1}(u) + \Psi_2^{-1}(v)$ for $u, v \ge 0$ to deduce that

$$\begin{aligned}
\int_0^\delta \Psi_2^{-1}(1/\mu B[t,r])\,dr &\le \int_0^\delta \Psi_2^{-1}\left(\frac{2^k N_k}{\mu_k B[t,r]}\right) dr \\
&\le \int_0^\delta \Psi_2^{-1}(2^k N_k)\,dr + \int_0^\delta \Psi_2^{-1}(1/\mu_k B[t,r])\,dr. \\
&\le \delta\Psi_2^{-1}(2^k N_k) + C\beta_k + \delta\Psi_2^{-1}(1) \le (2+C)\epsilon.
\end{aligned}$$

For the final bound it helps to consider two cases: if $\delta \le \operatorname{diam}(E)$ then <11> bounds the second integral by $C\beta_k$; if $\delta > \operatorname{diam}(E)$ then $\int_0^{\operatorname{diam}(E)} \ldots dr \le C\beta_k$ and the integrand for $\int_{\operatorname{diam}(E)}^\delta \ldots dr$ is less than $\Psi_2^{-1}(1)$ because over that range we have $B[t,r] \supset E$, implying $\mu_k B[t,r] = 1$.

To check that μ is actually a MM just note that $\int_0^D \ldots dr$ is smaller than D/δ times $\int_0^\delta \ldots dr$ when the integrand is a decreasing function of r. □

*12.4 From functional to admissible partitions

This Section shows how to construct suitable admissible partitions from the functional $\mathfrak{F}$. I have cobbled together a proof from several different arguments used by Talagrand (see Notes), rearranged to emphasize the similarities with the proof in my Section 12.2. My treatment corresponds roughly to Talagrand (2021, pp 47–58).

Remark. The Section has gone through many versions, as each fix led to new, more subtle errors. I would be very pleasantly surprised if it were now error-free, even though I am confident that the underlying idea is sound.

<12> **Theorem.** *Let $\{X_t : t \in T\}$ be a centered gaussian process for which $\mathfrak{F}(T) := \mathbb{P} \sup_{t\in T} X_t$ is finite. Then there exists a nested sequence $\mathbb{A} = \{\mathcal{A}_i : i \in \mathbb{N}_0\}$ of finite partitions, with $\mathcal{A}_0 = \{T\}$ and $|\mathcal{A}_i| \le n_i := 2^{2^i}$ such that, for some universal constant C and $\lambda := \sqrt{2}$,*

$$C\mathfrak{F}(T) \ge \sum\nolimits_{i\in\mathbb{N}_0}^{\infty} \operatorname{diam}(A_i(t))\, \lambda^i \qquad \text{for each } t \in T,$$

□ *where $A_i(t)$ denotes the member of $\mathcal{A}_i$ that contains t.*

The construction in Section 12.2 gave a very tight control over the diameters of the individual members of $\mathcal{A}_i$, with the weights providing only an indirect control for the $|\mathcal{A}_i|$, the number of sets in $\mathcal{A}_i$. By contrast, Talagrand's generic chaining involves the strict bound $n_i = 2^{2^i}$ on the size $|\mathcal{A}_i|$ but only indirect control over the diameters of the members of the partitions.

Despite these differences, the two partitioning methods have much in common. Brutally speaking, the new procedure could be implemented by:

(i) Generate the sets $E_1, \ldots, E_J$ using WPARTITION$(A, \ldots)$.

(ii) Discard all the weights.

(iii) Replace E_m by $\cup_{j\ge m} E_j$ if J is larger than some magic number m. and label the new E_m as BIG. (The sets that are not BIG are called SMALL, of course.) Then discard the individual E_j sets for $j > m$.

Needless to say, it makes more sense to start from a clean definition, although the brutal description does explain why some of the arguments in this Section share similarities with the arguments in Section 12.2. The main differences comes in:

(a) The key role that is played by the BIG sets.

(b) The choice $\theta := \min\left(C_{\text{sud}}/5, (2\lambda)^{-1}\right)$. The $C_{\text{sud}}/5$ term will let us invoke Lemma <2> and the $(2\lambda)^{-1}$ will ensure that $\sum_i (\lambda\theta)^i$ converges.

(c) The use of a quantity $\rho(A) \ge \operatorname{diam}(A)/2$ for the each set in each $\mathcal{A}_i$, so that the desired inequality becomes $C\mathfrak{F}(T) \ge \sup_{t\in T} \sum_i \rho(A_i(t))\, \lambda^i$.

Again the localized functional, $f(S, r) := \sup_{s\in S} \mathfrak{F}(S \cap B[s, r])$ will guide the greedy construction of the partitions using the following procedure, which assumes that we have already assigned the value $\rho(A)$ for the the set that is to be partitioned. The parameter m plays the role of a generic n_i. Again the small η provides wiggle room below a supremum.

```
1:  procedure APARTITION(A, m, η)
2:      Initialize: j = 1, S_1 = A, and define ρ := ρ(A).
```

```
    loop
        Find ξ_j in S_j for which the set B_j := S_j ∩ B[ξ_j, ρθ²]
              satisfies 𝔉(B_j) ≥ f(S_j, ρθ²) − η.
        If j is equal to m then
            Define E_m := S_m and ρ(E_m) := ρ and label E_m as BIG.
            Exit the loop.
        end If
        Define E_j := S_j ∩ B[ξ_j, ρθ] and ρ(E_j) := ρθ.
        Define S_{j+1} := S_j \ E_j.
        If S_{j+1} is empty then
            Exit the loop.
        end If
        Increase j by 1 (and continue with the next iteration).
    end loop
end procedure
```

For future reference, here are the most important facts about the output from APARTITION summarized in the form of a Lemma.

<13> **Lemma.** *The subset A of T, with $\rho := \rho(A)$, can be partitioned into subsets $E_1, \ldots, E_J$, for $J \le m$, such that:*

(i) If $j < m$ then E_j is SMALL *with $\rho(E_j) = \rho\theta$ and $E_j \subset A \cap B[\xi_j, \rho\theta]$ for some ξ_j in A.*

(ii) If $J = m$ then E_m is BIG *with $\rho(E_m) = \rho$ and*

$$\mathfrak{F}(A) \ge C_\theta \rho\theta \sqrt{\log m}, \tag{14}$$

$$\eta + \mathfrak{F}(A) \ge C_\theta \rho\theta \sqrt{\log m} + f(E_m, \rho\theta^2), \tag{15}$$

where $f(E_m, r) := \sup\{\mathfrak{F}(E_m \cap B[\xi, r]) : \xi \in E_m\}$. □

Proof. The facts about the the $\rho\theta$ separated ξ_j's and the greedily chosen B_j's are hidden inside the two inequalities for the BIG sets. The first is effectively just the SUDAKOV minoration (GSUD). The second comes from Lemma <2>:

$$\begin{aligned}
\mathfrak{F}(A) &\ge C_\theta(\rho\theta)\sqrt{\log m} + \min_{j\le m} \mathfrak{F}(B_j) \qquad \text{which is} \ge C_\theta \rho\theta \sqrt{\log m} \\
&\ge C_\theta \rho\theta \sqrt{\log m} + \min_{j\le m} \big(f(S_j, \rho\theta^2) - \eta\big) \\
&= C_\theta \rho\theta \sqrt{\log m} + f(E_m, \rho\theta^2) - \eta.
\end{aligned}$$

□

Remark. Lemma <2> is invoked only when E_m is BIG. Notice that the $\mathfrak{F}(B_j)$'s need not decrease monotonely but the the $f(S_j, \rho\theta^2)$'s do. I have separated out the role for GSUD in order to emphasize the places where the full force of Lemma <2> is needed.

Proof (of Theorem <12>). Again define $\eta_i := \mathfrak{F}(T)/2^{i+1}$ for $i \in \mathbb{N}_0$ and choose $\theta := \min\left(C_{\text{sud}}/5, (2\lambda)^{-1}\right)$ as in (b) on the previous page. Make recursive appeals to procedure APARTITION, starting from $\mathcal{A}_0 = \{T\}$ and $\rho(T) := D$. For the sake of a uniform notation, define $n_0 := n_1$. (The first step doesn't fit neatly into the general pattern because $|\mathcal{A}_0| = 1$.)

- Let $\mathcal{A}_1$ consist of the sets $E_1, E_2, \ldots, E_J$, with $J \le n_0 = n_1$, generated by APARTITION(T, n_0, η_0).
- Then partition each A in $\mathcal{A}_1$ into the new sets $E_1, E_2, \ldots$ generated by APARTITION(A, n_1, η_1). Each of the (at most) n_1 sets in $\mathcal{A}_1$ is partitioned into (at most) n_1 subsets to create an $\mathcal{A}_2$ consisting of at most $n_1^2 = n_2$ subsets, as required by admissibility.
- And so on: in general, APARTITION(A, n_i, η_i) applied to each A in $\mathcal{A}_i$ generates the sets for $\mathcal{A}_{i+1}$ and $|\mathcal{A}_{i+1}| \le n_i|\mathcal{A}_i| \le n_i^2 = n_{i+1}$.

As before, consider a fixed t in T and the sets $A_i := A_i(t) \in \mathcal{A}_i$ for which $t \in A_i(t)$. The sequence $\{A_i : i \in \mathbb{N}_0\}$ can consist of both SMALL sets and BIG sets. To simplify notation I'll treat $A_0 = T$ as small. Lemma <13> gives:

- If A_{i+1} is SMALL then $\rho(A_{i+1}) = \rho(A_i)\theta$ and $A_{i+1} \subset A_i \cap B[\xi, \rho(A_i)\theta]$ for some ξ in A_i.
- If A_{i+1} is BIG then $\rho(A_{i+1}) = \rho(A_i)$ and

$$\mathfrak{F}(A_i) \ge C_1 \rho(A_{i+1})\lambda^{i+1}, \tag{16}$$

$$\eta_i + \mathfrak{F}(A_i) \ge C_1 \rho(A_{i+1})\lambda^{i+1} + f(A_{i+1}, \rho(A_{i+1})\theta^2), \tag{17}$$

where $C_1 := \theta\lambda^{-1} C_\theta \sqrt{\log 2}$.

If every A_i is SMALL then $\rho(A_i) \le D\theta^i$ for every i so that

$$\sum\nolimits_{i\in\mathbb{N}_0}^{\infty} \lambda^i \rho(A_i) \le \sum\nolimits_{i\in\mathbb{N}_0} D(\theta\lambda)^i \le 2D \le 2\sqrt{2\pi}\mathfrak{F}(T),$$

the final inequality coming from Problem [2].

Things become more interesting if some of the sets are BIG. The argument then involves a close cousin of the odd/even subsequencing idea used in the Proof of Theorem <6>. To make it easier to spot any errors I had to break the argument into a short sequence of claims followed by proofs and consequences.

Claim (a):
There can be no infinite subsequence of consecutive BIG A_j's.

Proof of (a):
Suppose A_j is big for $\alpha \le j \le \omega$, so that $\rho := \rho(A_{\alpha-1}) = \rho(A_\alpha) = \cdots = \rho(A_\omega)$. Inequality <16> with $i = \omega - 1$ and monotonicity of $\mathfrak{F}$ then imply

<18> $$\mathfrak{F}(T) \ge \mathfrak{F}(A_{\omega-1}) \ge C_1\lambda^\omega \rho\,(A_\omega) = C_1\lambda^\omega\rho,$$

□ an inequality that prevents ω from being too large.

Remark. Note that argument would not work if we replaced $\rho(A_i)$ by the smaller $\operatorname{diam}(A_i)/2$. It was important that $\rho(A_i) = \rho(A_\omega)$ for $\alpha - 1 \le i \le \omega$.

Consequence (a):
It follows that each stretch of consecutive BIG A_j's must be finite and separated from the next stretch of BIG's by at least one SMALL set. Moreover, one stretch of SMALL A_j's might be infinite, in which case there are no more BIG sets. The pattern could look like:

small_0	big_1	small_1	big_2	small_2	big_3	small_3	big_4	
sss	*bbb*	*sssss*	*bb*	*sss*	*b*	*sss*	*bb*	...
	↓		↓		↓		↓	
	$\omega[1]$		$\omega[2]$		$\omega[3]$		$\omega[4]$	

The index $\omega[\ell]$ gives the position of the last BIG A_j in big_ℓ, the ℓth block
□ of BIG sets.

The contribution from small_0 can be handled in the same way as the case where all A_j's are SMALL.

Claim (b):
If A_j is BIG for $\alpha \le j \le \omega$ and SMALL for $\omega < j \le \beta$ then $\sum_{j=\alpha}^{\beta} \lambda^j \rho(A_j) \le C_0\lambda^\omega\rho(A_\omega)$ for some constant C_0.

Proof of (b):

$$\sum\nolimits_{\alpha\le j\le\omega} \lambda^j\rho(A_j) = \rho(A_\omega)\sum\nolimits_{\alpha\le j\le\omega}\lambda^j \le \rho(A_\omega)\lambda^{\omega+1}/(\lambda-1),$$
$$\sum\nolimits_{\omega< j\le\beta} \lambda^j\rho(A_j) \le \sum\nolimits_{\omega< j\le\beta}\lambda^\omega\lambda^{j-\omega}\rho(A_\omega)\theta^{j-\omega}.$$

□ The last sum is $\le \lambda^\omega\rho(A_\omega)$ because we chose θ to make $\theta\lambda \le 1/2$.

Consequence (b):
The contributions from any big_ℓ plus the following small_ℓ are bounded
□ above by $C_0\lambda^\omega\rho(A_\omega)$ with $\omega = \omega[\ell]$.

Thus it remains only to control the sum

<19> $$\sum_{\ell\in\mathbb{L}} \lambda^{\omega[\ell]}\rho(A_\omega[\ell]) \qquad \text{where } \mathbb{L} := \{\ell : \text{big}_\ell \neq \emptyset\}.$$

If $|\mathbb{L}| \le 3$ then three appeals to inequality <18> would suffice. If $|\mathbb{L}| \ge 4$ we can use a trick that is similar to the odd/even decomposition used to prove Theorem <6>.

Claim (c):
If $\ell, \ell+1, \ell+2, \ell+3 \in \mathbb{L}$ then

<20> $$\eta_{\omega[\ell]} + \mathfrak{F}(A_{\omega[\ell]}) \ge C_1\lambda^{\omega[\ell+1]}\rho(A_{\omega[\ell+1]}) + \mathfrak{F}(A_{\omega[\ell+3]}).$$

Proof of (c):
To slightly simplify notation I'll assume $\ell = 1$. The proof for general ℓ is similar.

It is just a matter of pulling together the various inequalities derived in the previous few pages. Define $i := \omega[2] - 1$ and $i' := \omega[3]$.

$$\begin{array}{ccccccccc} s\ldots & \underline{bbb} & s\ldots & & \underline{bb} & s\ldots & & \underline{bb} & s\ldots & \underline{bb} & \ldots \\ & \omega[1] & & \omega[2]=i+1 & & & \omega[3]=i' & & & \omega[4] & \end{array}$$

Because there is at least one SMALL between each big block we must have $\omega[1] + 2 \le \omega[2] = i + 1 \le \omega[3] - 2$ and $\omega[3] = i' \le \omega[4] - 2$. Most importantly, $A_{i'+1}$ is a SMALL set contained within a ball of radius at most $\rho(A_{i'})\theta \le \rho(A_{i+1})\theta^2$ with center in $A_{i'}$, a subset of A_{i+1}. Thus

$$\begin{aligned} &\eta_{\omega[1]} + \mathfrak{F}(A_{\omega[1]}) \\ &\quad\ge \eta_i + \mathfrak{F}(A_i) \qquad \text{by monotonicity of } \mathfrak{F} \text{ and } \eta_{\omega[1]} > \eta_i \\ &\quad\ge C_1\rho(A_{i+1})\lambda^{i+1} + \sup\{\mathfrak{F}(A_{i+1}\cap B[\xi,\rho(A_{i+1})\theta^2] : \xi \in A_{i+1}\} \\ &\quad\ge C_1\rho(A_{\omega[2]})\lambda^{\omega[2]} + \mathfrak{F}\left(A_{i'+1}\right). \end{aligned}$$

The second inequality comes from <17> with $f(A_{i+1}, \rho(A_{i+1})\theta^2)$ replaced by its definition. Monotonicity gives $\mathfrak{F}\left(A_{i'+1}\right) \ge \mathfrak{F}\left(A_{\omega[4]}\right)$, completing the argument □

Consequence (c):
The case where $\mathbb{L}$ is infinite is the most interesting. The argument for finite $\mathbb{L}$ is similar.

Successive appeals to inequality <20> gives

$$\begin{aligned} \eta_{\omega[1]} + \mathfrak{F}(A_{\omega[1]}) - \mathfrak{F}(A_{\omega[4]}) &\ge C_1\lambda^{\omega[2]}\rho(A_{\omega[2]}), \\ \eta_{\omega[4]} + \mathfrak{F}(A_{\omega[4]}) - \mathfrak{F}(A_{\omega[7]}) &\ge C_1\lambda^{\omega[5]}\rho(A_{\omega[5]}), \\ \eta_{\omega[7]} + \mathfrak{F}(A_{\omega[7]}) - \mathfrak{F}(A_{\omega[10]}) &\ge C_1\lambda^{\omega[8]}\rho(A_{\omega[8]}), \end{aligned}$$

and so on. If we add these inequalities the sum on the left-hand side telescopes, leaving

$$\mathfrak{F}(A_{\omega[1]}) + \left(\eta_{\omega[1]} + \eta_{\omega[4]} + \dots\right) \geq C_1 \left(\lambda^{\omega[2]}\rho(A_{\omega[2]}) + \lambda^{\omega[5]}\rho(A_{\omega[5]}) + \dots\right)$$

It follows that, for some constant C_2,

$$C_2\mathfrak{F}(T) \geq \sum\nolimits_{\ell\in\mathbb{L}_2} \lambda^{\omega[\ell]}\rho(A_{\omega[\ell]}) \qquad \text{where } \mathbb{L}_2 := \{\ell \in \mathbb{L} : \ell \equiv 2(\text{mod } 3)\ \}.$$

Similar arguments bound the sums over the similarly defined subsequences $\mathbb{L}_0$ and $\mathbb{L}_1$. After combining the three bounds then invoking inequality <16> to take care of $\rho(A_{\omega[1]})\lambda^{\omega[1]}$ we arrive at the inequality asserted by Theorem <12>. □

12.5 Beyond gaussian

Even though this Chapter has established lower bounds only in the gaussian case, I believe it does point the way to other applications of Talagrand's general method. As explained by Talagrand (2021, pp. 47–49), the key requirement of his approach is existence of an analog of Lemma <2> for analogs of the functional $\mathfrak{F}$.

12.6 Problems

[1] (repeated from Chapter 6) Suppose $\{X_i : i \in \mathbb{N}\}$ is a gaussian process with $X_i \sim N(0, \sigma_i^2)$. Define $M_n := \max_{i\leq n} X_i$ and $M_\infty := \sup_{i\in\mathbb{N}} X_i$. Show that $M_\infty < \infty$ almost surely if and only if $\mathbb{P}M_\infty < \infty$. Argue as follows for the nontrivial implication. Suppose $\mathbb{P}\{M_\infty < \infty\} = 1$.

(i) Show that there exists an $R \in \mathbb{R}$ for which

$$\bar{\Phi}(1) > \mathbb{P}\{M_\infty > R\} \geq \mathbb{P}\{M_n > R\} \geq \mathbb{P}\{X_n > R\} = \bar{\Phi}(R/\sigma_n).$$

Deduce that $\sigma := \sup_n \sigma_n \leq R$.

(ii) Choose any constant r_0 for which $1 > 2e^{-r_0^2/2} + \bar{\Phi}(1)$. Write m_n for $\mathbb{P}M_n$. Use the concentration inequality from §6.1 to show that

$$\mathbb{P}\{M_n \geq m_n - \sigma r_0\} \geq 1 - 2e^{-r_0^2/2} > \bar{\Phi}(1) > \mathbb{P}\{M_n > R\}.$$

Deduce that $m_n - \sigma r_0 \leq R$ for all n.

(iii) Deduce that $\mathbb{P}M_\infty \leq R + \sigma r_0 < \infty$. Hint: $0 \leq |X_1| + M_n \nearrow |X_1| + M_\infty$.

(iv) Extend the argument to a two-sided equivalence: $\mathbb{P}\{\sup_i |X_i| < \infty\} = 1$ iff $\mathbb{P}\sup_i |X_i| < \infty$. Hint: $\sup_i |X_i| = \max(\sup_i X_i, \sup_i(-X_i))$.

[2] Suppose (X_1, X_2) has a centered bivariate normal distribution for which $\|X_1 - X_2\|_2 = \delta > 0$. Show that $\mathbb{P}\max(X_1, X_2) = \delta/\sqrt{2\pi}$. Hint: First

show that $2\max(X_1, X_2) - X_1 - X_2 = |X_1 - X_2|$. Also note that $\mathbb{P}|Z| = 2/\sqrt{2\pi}$ if $Z \sim N(0,1)$.

[3] Suppose $\{X_i : i \in \mathbb{N}\}$ is a centered gaussian process with $F := \mathbb{P}\sup_{i\in\mathbb{N}} X_i$ finite.

(i) Show that $\mathbb{P}\sup_{i,j}|X_i - X_j| = \mathbb{P}\sup_{i,j}(X_i - X_j) = 2F$. Hint: The first equality comes from the fact that both $X_i - X_j$ and $X_j - X_i$ appear in $\sup_{i,j}(X_i - X_j)$. Then note that $\sup_{i,j}|X_i - X_j|$ is an increasing limit of

$$\max_{(i,j)\in[\![n]\!]^2}|X_i - X_j| = \max_{(i,j)\in[\![n]\!]^2}(X_i - X_j) = \max_{i\in[\![n]\!]} X_i + \max_{j\in[\![n]\!]}(-X_j).$$

Use the fact that $\{(-X_j) : j \in \mathbb{N}\}$ has the same distribution as X.

(ii) Define $D := \sup_{i,j}\|X_i - X_j\|_2$. Show that $F \geq D/\sqrt{2\pi}$. Hint: Problem [2].

[4] Under the conditions of Theorem <10>, prove that $\delta\Psi_2^{-1}(\text{PACK}(\delta, T, d)) \to 0$ as $\delta \to 0$.

12.7 Notes

My exposition is based largely on papers by Talagrand (1992, 1996, 2001) and the three versions of his book on generic chaining: Talagrand (2005, Sections 1.3 and 2.1), Talagrand (2014, Chapter 2), and Talagrand (2021, Chapter 2). In particular, Lemma <2> comes from Talagrand (1992) via Talagrand (2021, Proposition 2.10.8).

The highly informative book of Ledoux and Talagrand (1991) provided welcome backup, particularly for the insights it gave into earlier ways of handling chaining ideas. And again I also benefitted greatly from reading the expositions by Ledoux (1996, Chap 6) and van Handel (2016). See Chapter 11 for further references.

The proof for Theorem <10> is based on Ledoux and Talagrand (1991, Thm 12.9), a version of a more involved result presented by Talagrand (1987, page 122-124). My rearrangement of their argument avoids the need for an appeal to Ledoux and Talagrand (1991, Cor 3.19).

References

Ledoux, M. (1996). Isoperimetry and Gaussian analysis. *Springer Lecture Notes in Mathematics 1648*, 165–294. École d'Été de Probabilités de St-Flour XXIV , 1994.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. New York: Springer.

Talagrand, M. (1987). Regularity of Gaussian processes. *Acta Mathematica 159*, 99–149.

Talagrand, M. (1992). A simple proof of the majorizing measure theorem. *Geometric and Functional Analysis 2*(1), 118–125.

Talagrand, M. (1996). Majorizing measures: The generic chaining. *Annals of Probability 24*, 1049–1103.

Talagrand, M. (2001). Majorizing measures without measures. *The Annals of Probability 29*(1), 411–417.

Talagrand, M. (2005). *The Generic Chaining: Upper and lower bounds of stochastic processes*. Springer-Verlag.

Talagrand, M. (2014). *Upper and Lower Bounds for Stochastic Processes: Modern methods and classical problems*. Springer-Verlag.

Talagrand, M. (2021). *Upper and Lower Bounds for Stochastic Processes: Decomposition Theorems* (Second ed.), Volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag.

van Handel, R. (2016, December). Probability in high dimension. APC 550 Lecture Notes, Princeton University.

Chapter 13
Symmetrization

SECTION *13.1 presents a collection of examples, some classical and some of more recent provenance, that introduces several of the key ideas behind the symmetrization method.*

SECTION *13.2 explains why it helps to create "independent copies" of random processes by means of explicit product space constructions.*

SECTION *13.3 explains why empirical processes (my main motivating application) should be thought of as sums of independent stochastic processes.*

SECTION *13.4 describes a symmetrization method for bounding the expected value for the supremum of an empirical processes, introducing the idea of controlling a centered sum of processes $X_i(t) - \mathbb{P}X_i(t)$ by a sum of processes $\mathfrak{s}_i X_i(t)$ with $\mathfrak{s}_i$ a random ± 1-valued sign variable.*

SECTION *13.5 extends the ideas from § 13.4 to control oscillations for empirical processes. The argument involves a small calculation known as the "square-root-trick".*

SECTION *13.6 shows how to strengthen the results from Section 13.5 to obtain oscillation control for empirical processes.*

13.1 Symmetry?

In general, the word 'symmetrize' refers to some operation that replaces an object by a more symmetric object. For example, a classical method due to Steiner (Billingsley, 1968, §19) can be used to prove an isoperimetric inequality for measurable subsets of $\mathbb{R}^k$, by using a succession of transformations that bring the set ever closer to a ball.

A different sort of symmetrization has long been a useful probability tool. For example, as you saw for the PISIER-MAUREY method in §6.3, if X and Y are independent with each $N(0, I_n)$ distributed then, under mild assumptions on the function f,

$$\mathbb{P}e^{\lambda(f(X)-\mathbb{P}f(X))} \le \mathbb{P}e^{\lambda(f(X)-f(Y))} \qquad \text{for all real } \lambda.$$

The distribution of $f(X) - f(Y)$ is symmetric around the origin. However, symmetry was not the main reason for making the transformation; it was merely a step that put the problem into a form that was easier to analyze by means of a path argument.

In fact many arguments traditionally referred to as 'symmetrizations' do not depend on symmetry for their effectiveness. They actually involve

a more general idea. Suppose, for example we wish to bound $\mathbb{P}W$ for a random variable W defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We can sometimes argue using some combination of two general strategies:

(i) RANDOMIZE: Inject more randomness into the problem, to create a new random variable W^*, which is easier to analyze than W and for which $\mathbb{P}W \le \mathbb{P}W^*$.

(ii) CONDITION: For some random object Y to decompose the analyis into two or more steps by conditioning on Y: if Y has distribution Q then $\mathbb{P}W^* = Q^y \mathbb{P}(W^* \mid Y = y)$. The hope is that the analysis of the conditional expectation will be easier than the original problem.

Typically the new W^* is created on a larger probability space, often by taking a product of the original $\mathbb{P}$ with some new probability measure. In such situations the conditioning argument can usually be recast as an appeal to a FUBINI theorem, a simplification that avoids dubious arguments involving abstract KOLMOGOROV-style conditioning on sub-sigma-fields. The next Section will explain.

The following Examples illustrate how these ideas can solve some nontrivial problems. Even if you are not interested in these particular applications, I urge you to at least take note of how the same ideas seem to pop up in apparently unrelated problems.

The first Example does not involve any RANDOMIZE step because the problem already comes in a form amenable to analysis. However the analysis does preview an idea that appears in many RANDOMIZE arguments.

<1> **Example.** A famous result of Cramér (1936) asserts that a sum $X + Y$ of independent random variables can have a normal distribution only if each of X and Y also has a normal distribution (possibly degenerate). As a first step in the proof, one needs to show that X has tails that decrease rapidly.

With no loss of generality, assume that Y has a zero median, that is,

$$\mathbb{P}\{Y \ge 0\} \ge 1/2 \qquad \text{and} \qquad \mathbb{P}\{Y \le 0\} \ge 1/2.$$

Also assume, for simplicity, that $X + Y$ has a $N(0,1)$ distribution. Then, for $t \ge 0$,

$$\begin{aligned} \exp(-t^2/2) \ge \mathbb{P}\{X + Y \ge t\} &\ge \mathbb{P}\{X \ge t, Y \ge 0\} \\ &\ge {}^1\!/_2 \mathbb{P}\{X \ge t\} \qquad \text{by independence.} \end{aligned}$$

A similar argument gives a similar bound for the lower tail. If you read the previous display in reverse order it will look like an instance of RANDOMIZE.

See UGMTP §8.8 for an explanation of how this tail bound leads to Cramér's result. □

Remark. Cramér proved his result in response to a conjecture by Paul Lévy, asserting (roughly speaking) that a sum of independent random variables is approximately normally distributed if and only each summand either contributes a very small amount to the sum or (if not small) is itself approximately normally distributed. See Problem [1] for a result along these lines. See Le Cam (1986, §2) for a most entertaining account of the history.

In the statistical theory of experimental design, randomization is actually a key ingredient in the statistical interpretation and not just a device for bounding one expected value by another. Nevertheless, it does point the way for the use of a RANDOMIZE step in other fields.

<2> **Example.** Fisher (1935, §§13–21) discussed in detail the analysis of an experiment on plant growth made by Charles Darwin. The data consisted of observed heights for 15 pairs of plants, one self-fertilized the other cross-fertilized, each pair being grown under conditions that Darwin had tried to make as similar as possible. Nevertheless, there were probably small differences between those conditions.

Fisher criticized Darwin's procedure because he had not randomly assigned ("as by tossing a coin") the locations where each member of the pair had been planted. He commented (§20):

> Randomisation properly carried out, in which each pair of plants are assigned their positions independently at random, ensures that the estimates of error will take proper care of all such causes of different growth rates, and relieves the experimenter from the anxiety of considering and estimating the innumerable causes by which his data may be disturbed. The one flaw in Darwin's procedure was the absence of randomisation.

Such an experimental design is now referred to as the "method of paired comparisons". Fisher also asserted (§21) that

> . . . the physical act of randomisation, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied.

In other words, the randomization turns some systematic effects into random noise, so that the probabilistic analysis leads to conclusions similar to those obtained under a model with normal errors. He didn't seem to give any theoretical justification for that claim.

Subsequently, several authors tried to justify Fisher's assertions. Suppose D_i denotes the difference in heights (cross- minus self-) within the ith

pair, for $i \in [[n]]$ with $n = 15$. The typical statistical model treats the D_i's as independent $N(\mu, \sigma^2)$ random variables, with μ representing the hypothetical mean difference between treatments. A formal test of the null hypothesis $\mu = 0$ would compare the observed value

$$T := \frac{\sum_i D_i/\sqrt{n}}{\left(\sum_i (D_i - \overline{D})^2/(n-1)\right)^{1/2}}$$

with the percentiles of a standard t-distribution on $n-1$ degrees of freedom.

If there were literally no difference between the treatments, the magnitude $y_i := |D_i|$ would be determined by the locations within pots, and the sign would be determined by the randomization. We would then find ourselves in the situation where the observations had the same distribution as $\mathfrak{s}_1 y_1, \ldots, \mathfrak{s}_n y_n$ for independent random variables for which $\mathbb{P}\{\mathfrak{s}_i = +1\} = 1/2 = \mathbb{P}\{\mathfrak{s}_i = -1\}$. Any statistical test based on T could just as easily be based on the one-to-one transformation

$$S = \frac{\sqrt{n}T}{\sqrt{n-1+T^2}} = \frac{\sum_i \mathfrak{s}_i y_i}{\sqrt{\sum_i y_i^2}} = \sum\nolimits_i \mathfrak{s}_i w_i \qquad \text{where } w_i := y_i/(\sum\nolimits_i y_i^2)^{1/2}.$$

By construction, $\sum_i w_i^2 = 1$. If the normality assumption held, S would have a symmmetric distribution with S^2 distributed $\text{BETA}(1/2, (n-1)/2)$. Normal approximation for T is essentially equivalent to BETA approximation to the distribution of S^2 conditional on the $\{w_i\}$ weights.

As one measure of similarity between the distributions one can compare the moments of S^2 with the corresponding moments of a beta distribution. For a most informative account of this idea see the discussion paper by Box and Andersen (1955). □

In §4.3 you saw how a result of Isaac Newton implies log-concavity of the Poisson-Binomial distribution. The next Example shows how the same fact can be established by a coupling argument involving RANDOMIZE and CONDITION. Notice the similarity to the method of paired comparisons in the previous Example.

<3> **Example.** Suppose $S = X_1 + \cdots + X_n$ for independent $\{0,1\}$-valued X_i's with $\mathbb{P}\{X_i = 1\} = p_i$ for $0 < p_i < 1$. That is, $X_i \sim \text{BER}(p_i)$. In §4.5 it was shown that the distribution of S is log-concave (and hence unimodal):

<4> $$(\mathbb{P}\{S = k\})^2 > \mathbb{P}\{S = k-1\}\mathbb{P}\{S = k+1\} \qquad \text{for } k = 0, 1, \ldots, n.$$

The RANDOMIZE/CONDITION method will give another simple proof of this result.

For the RANDOMIZE step make an "independent copy" of the X_i's. That is, create Y_i's such that $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ are mutually independent and $Y_i \sim \text{BER}(p_i)$. Define $T = Y_1 + \cdots + Y_n$. By independence, inequality <4> is equivalent to

<5> $$\mathbb{P}\{S = k, T = k\} > \mathbb{P}\{S = k-1, T = k+1\} \qquad \text{for } k = 0, 1, \ldots, n.$$

The verification of <5> will be greatly simplified by conditioning on the random variables $W_i := X_i + Y_i$. The triples (X_i, Y_i, W_i) for $i = 1, \ldots, n$ are independent with $W_i \sim \text{BIN}(2, p_i)$ and

$$\mathbb{P}\{X_i = Y_i = 1 \mid W_i = 2\} = 1 = \mathbb{P}\{X_i = Y_i = 0 \mid W_i = 0\},$$
$$\mathbb{P}\{X_i = 1, Y_i = 0 \mid W_i = 1\} = 1/2 = \mathbb{P}\{X_i = 0, Y_i = 1 \mid W_i = 1\}.$$

Remark. The random variables S and T are said to be ***coupled***, That is, we have defined them on a new probability space where a conditioning argument can simplify the analysis.

For each w in $\{0, 1, 2\}^n$ write $\mathbb{P}_w$ for the conditional probability measure $\mathbb{P}(\cdot \mid W = w)$. It will suffice if we can show

<6> $$\mathbb{P}_w\{S = k, T = k\} \geq \mathbb{P}_w\{S = k-1, T = k+1\} \qquad \text{for each } w,$$

with strict inequality for at least one w with $\mathbb{P}\{W = w\} > 0$. To that end define $N_\alpha := \sum_{i \leq n} \{W_i = \alpha\}$ for $\alpha \in \{0, 1, 2\}$ and $Z := \sum_{i \leq n} X_i \{W_i = 1\}$. Under $\mathbb{P}_w$ this Z has a $\text{BIN}(N_1, 1/2)$ distribution and we have $S = N_2 + Z$ and $T = N_2 + (N_1 - Z)$.

For the quantity on the right-hand side of <6> to be nonzero we must have $2k = 2N_2 + N_1$ and $N_1 > 0$, which forces $N_1 = 2\ell$ for some positive integer ℓ. The desired inequality becomes

$$\mathbb{P}_w\{Z = \ell\} \geq \mathbb{P}_w\{Z = \ell - 1\} \qquad \text{with } > \text{ for at least one } \ell.$$

The strict inequality holds for each ℓ in $\mathbb{N}$ because the $\text{BIN}(2\ell, 1/2)$ distribution has a unique mode at ℓ. □

Conditional symmetry has long been a standard tool in the study of empirical distribution functions.

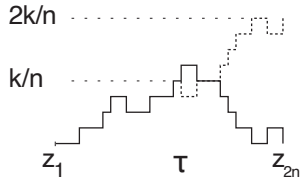
<7> **Example.** Suppose F_n is the empirical distribution function based on a sample $x_1, \ldots, x_n$ from a distribution function F on the real line, and G_n is the empirical distribution function based on a sample $y_1, \ldots, y_n$ from a distribution function G. For a one-sided test of the null hypothesis that $F = G$ one could use the SMIRNOV statistic, $S_n^+ = \max_{x \in \mathbb{R}} (F_n(x) - G_n(x))$.

At least when F has no jumps—which rules out ties in the observations—the exact distribution of S_n^+ can be calculated by means of a conditioning argument (Gnedenko, 1968, Section 68). Write $z_1 \leq \cdots \leq z_{2n}$ for the

ordered values of the combined sample of x_i's and y_i's. Calculate the conditional probability

$$p_k(\boldsymbol{z}) := \mathbb{P}_{\boldsymbol{z}}\{F_n(x) - G_n(x) \geq k/n \text{ for some } x\}.$$

Given $\boldsymbol{z} = (z_1, \ldots, z_{2n})$, the step function $F_n - G_n$ has jumps of size $1/n$ at each z_i, with n jumps up and n jumps down. The sample path starts from $(0, 0)$ and ends at $(z_{2n}, 0)$. All $\binom{2n}{n}$ orderings of jumps are equally likely; the conditional distribution of the x_i sample given $\boldsymbol{z}$ puts equal weight on each subset of size n. To calculate p_k we have merely to count how many orderings will give a path that reaches the level k/n. Write $\tau = \tau(\boldsymbol{z})$ for the first i on such a path for which $F_n(z_i) - G_n(z_i) = k/n$.



A reflection of the path segment from τ to z_{2n} about the horizontal line through $(z_\tau, k/n)$ defines a one-one correspondence between paths that reach k/n and the set of all step functions from $(0, 0)$ to $(z_{2n}, 2k/n)$ with jumps of size $\pm 1/n$ at the z_i. There are exactly $\binom{2n}{n-k}$ such step functions, one for each choice of the $n - k$ positions of the $-1/n$ jumps.

Thus,

$$p_k(\boldsymbol{z}) = \binom{2n}{n-k} \Big/ \binom{2n}{n} \qquad \text{for every } \boldsymbol{z} \text{ and } k = 0, 1, \ldots, n.$$

The conditional probability does not depend on the configuration of the z_i's. When we average out over the distribution of $\boldsymbol{z}$, whatever it might be, we get the same ratio for the unconditional probability,

$$\mathbb{P}\{F_n(x) - G_n(x) \geq k/n \text{ for some } x\} = \binom{2n}{n-k} \Big/ \binom{2n}{n}.$$

The analogous one-sample problem, involving $F_n - F$ instead of $F_n - G_n$, does not lend itself to such a simple combinatorial analysis. However, as you will see in Example <8>, it is possible to use the two-sample process to derive bounds for $F_n - F$. □

The combinatorial idea in the previous Example has a much more powerful analog in the modern theory for more general empirical measures.

<8> **Example.** In an exceedingly famous paper, VC71=Vapnik and Chervonenkis (1971) described a method for establishing an analog of the classical GLIVENKO-CANTELLI theorem. What follows is a rewrite of their argument using the notation that later became popular in the empirical process literature.

Suppose $(\mathbb{A}, \mathcal{A}, P)$ is a probability space and $\xi_1, \xi_2, \ldots$ are independent random elements of $\mathbb{A}$, each with distribution P. Define the empirical measure $P_n = P_{n,\omega}$ on S by putting mass n^{-1} at each of the points $\xi_1(\omega), \ldots, \xi_n(\omega)$. By the strong law of large numbers, for each A in $\mathcal{A}$,

$$P_n A := n^{-1} \sum\nolimits_1 \{\xi_i(\omega) \in A\} \to PA \qquad \text{almost surely as } n \to \infty.$$

VC71 were able to extend this result to a convergence assertion holding uniformly over a subcollection $\mathcal{D}$ of $\mathcal{A}$.

To avoid measurability difficulties of the kind discussed in Chapter 9, let me assume that $\mathcal{D}$ is countable: $\mathcal{D} = \{D_j : j \in \mathbb{N}\}$.

Remark. I would argue that existence of a DOOB-separable version (see §9.6) of the process $\{P_n D : D \in \mathcal{D}\}$ justified the simplifying assumption that $\mathcal{D}$ is countable. However authors such as Dudley (2014, p267) would not accept this argument. Similar measurability issues greatly complicated the development of limit theory for general empirical processes. See Giné and Zinn (1984, §2) for one particularly careful attempt to spell out the measurability assumptions.

More recently, Ledoux and Talagrand (1991, pp44–45) have eliminated the difficulty by defining

$$\mathbb{P} \sup_{t\in T} |X_t|^p := \sup\{\mathbb{P} \max_{t\in F} |X_t|^p : F \text{ a finite subset of } T\},$$

which is equivalent to working with an essential supremum (see §9.4), a close cousin of DOOB-separability. That is, a maximal inequality over an infinite set T should be interpreted as a uniform bound that holds for all finite subsets of T, a bound that does not blow up as the size of F increases.

Define $W(\omega) := \sup_{D\in\mathcal{D}} |P_{n,\omega} D - PD|$. VC71 used a symmetrization argument to bound $\mathbb{P}\{W > \epsilon\}$. For the RANDOMIZE step, generate independent observations $\xi_1, \xi_2, \ldots, \xi_{2n}$ on P, with $P_n = P_{n,\omega}$ putting mass $1/n$ at $\xi_1(\omega), \ldots, \xi_n(\omega)$ and a second empirical measure $\widetilde{P}_n = \widetilde{P}_{n,\omega}$ putting mass n^{-1} at $\xi_{n+1}(\omega), \ldots, \xi_{2n}(\omega)$. Define $P^\circ_{n,\omega} := P_{n,\omega} - \widetilde{P}_{n,\omega}$. The random measures P_n and $\widetilde{P}_n$ are mutually independent. A simple argument will show that, for all n large enough,

<9> $$\mathbb{P}\{W > \epsilon\} \le 2\mathbb{P}\{W^\circ > \epsilon/2\} \qquad \text{where } W^\circ(\omega) := \sup_{D\in\mathcal{D}} |P^\circ_{n,\omega} D|.$$

It is just a matter of keeping track of the first D_j that makes W bigger than ϵ. Define

$$\tau(\omega) := \inf\{j \in \mathbb{N} : |P_{n,\omega} D_j - PD_j| > \epsilon\},$$

with the usual convention that $\tau(\omega) = \infty$ when $W(\omega) \le \epsilon$. The event $\{W > \epsilon\}$ is a disjoint union of the events $\{\tau = j\}$ for $j \in \mathbb{N}$. For each

such j, the event $\{\tau = j\}$ (which depends only on $\xi_1, \ldots, \xi_n$) is independent of the event $B_j := \{\omega : |\widetilde{P}_{n,\omega} D_j - P D_j| \le \epsilon/2\}$ and on the intersection $\{\tau = j\}B_j$ we have $|P^\circ_{n,\omega} D_j| > \epsilon/2$. Note that $\operatorname{var}(\widetilde{P}_{n,\omega} D_j) = n^{-1}(PD_j)(PD_j^c) \le (4n)^{-1}$, so that

$$\mathbb{P}B_j^c \le (1/4n)/(\epsilon/2)^2 < 1/2 \qquad \text{when } n > 2/\epsilon^2, \text{ by CHEBYSHEV.}$$

For such an n we then have

$$\begin{aligned}
\mathbb{P}\{W > \epsilon\} &= \sum\nolimits_{j\in\mathbb{N}} \mathbb{P}\{\tau = j\} \le \sum\nolimits_{j\in\mathbb{N}} \mathbb{P}\{\tau = j\} 2\mathbb{P}B_j \\
&\le 2 \sum\nolimits_{j\in\mathbb{N}} \mathbb{P}\{\tau = j, |P_{n,\omega} D_j - P D_j| > \epsilon\} B_j \qquad \text{by independence} \\
&\le 2 \sum\nolimits_{j\in\mathbb{N}} \mathbb{P}\{\tau = j, W^\circ > \epsilon/2\} \le 2\mathbb{P}\{W^\circ > \epsilon/2\}.
\end{aligned}$$

The process $P^\circ_n := P_n - \widetilde{P}_n$ is easier to analyze than $P_n - P$ because it has simple conditional distributions. The VC71 method effectively conditions on $P_n + \widetilde{P}_n$, in the sense that we treat the locations of the observations in the double sample as given but without their identification as members of the P_n or the $\widetilde{P}_n$ sample. That is, we work with the conditional distribution under which each of the $\binom{2n}{n}$ subsets of $\{\xi_i : i \in [\![2n]\!]\}$ is equally likely to correspond to $\{\xi_1, \ldots, \xi_n\}$. After conditioning, the analysis reduces to a combinatorial problem similar to taking a sample of size n from an urn containing n red balls and n black balls.

The key VC71 insight was similar in spirit to the simplification for the classical case described in Example <7>. For a given ω, if D_j and D_k are two members of $\mathcal{D}$ for which

$$\text{pattern}_j := \{i \in [\![2n]\!] : \xi_i(\omega) \in D_j\} = \{i \in [\![2n]\!] : \xi_i(\omega) \in D_k\} =: \text{pattern}_k.$$

then we must have $P^\circ_{n,\omega} D_j = P^\circ_{n,\omega} D_k$ for every possible partition into two subsets of size n. When we are taking the supremum over all members of $\mathcal{D}$ we need not consider both D_j and D_k. Indeed,

$$\sup\nolimits_{D\in\mathcal{D}} |P^\circ_{n,\omega} D| = \sup\nolimits_{D\in\mathcal{D}_\omega} |P^\circ_{n,\omega} D|$$

for any $\mathcal{D}_\omega$ that picks out all possible patterns from $\{\xi_i(\omega) : i \in [\![2n]\!]\}$. (The subset $\mathcal{D}_\omega$ is allowed to depend on ω, of course.)

VC71 had noticed the possibility that, for some $\mathcal{D}$, it is possible that there might always exist a $\mathcal{D}_\omega$ whose size is bounded by $p(n)$, for a fixed polynomial p. (They cited the classical case of intervals $(-\infty, a]$ on the real line as an example.) Such a $\mathcal{D}$ is called a ***VC-class of sets***. When combined with a hypergeometric tail bound for the conditional distribution, the polynomial bound on $|\mathcal{D}_\omega|$ leads to a most satisfactory upper bound for $\mathbb{P}\{W > \epsilon\}$. □

Remark. Admittedly the distribution of each $P_n^\circ A$ in the previous Example is symmetric but that is not the whole reason for the method's success. For example, Devroye (1982) was able to sharpen the VC inequalities by constructing $\widetilde{P}_n$ from an independent sample of size n^2, so that the variability of $P_n - \widetilde{P}_n$ was only slightly greater than the variability of $P_n - P$. For that construction the distribution of $P_n f - \widetilde{P}_n f$ is no longer symmetric, but it is still more tractable than the distribution of $P_n f - P f$. Massart (1986) took the idea even further by taking the second sample to be of size mn, where m was a parameter that could be optimized over.

<10> **Example.** VC81=Vapnik and Chervonenkis (1981) extended their result for sets to the analogous results for uniformly bounded collections of $\mathcal{A}$-measurable, real-valued functions, $\{f_t(\cdot) : t \in T\}$ on $\mathbb{A}$. The following discussion translates some of their arguments into the more familiar notation.

Again they considered samples of observations $\xi_1, \ldots, \xi_n$ from a fixed probability measure P on $\mathbb{A}$. They assumed $0 \le f_t(a) \le 1$ for each t. For each realization $(\xi_i(\omega) : i \in [\![n]\!])$, each t defines a point in the unit cube:

$$z_{t,\omega} := (f_t(\xi_i(\omega)) : i \in [\![n]\!]) \in C_n := [0,1]^n.$$

Thus $\mathbb{A}_{\omega,n} := \{z_{t,\omega} : t \in T\}$ can be thought of as a random subset of C_n. They equipped the cube with its ℓ^∞ metric, $\rho(x,y) := \max_{i \le n} |x_i - y_i|$. By using a symmetrization argument similar to the one for VC71 they showed that $\sup_t |P_n f - P f| \to 0$ in probability iff $\mathbb{P} \log_2 \text{COVER}(\epsilon, \mathbb{A}_{\omega,n}, \rho)/n \to 0$ for each $\epsilon > 0$, where $\text{COVER}(\epsilon, \mathbb{A}_{\omega,n}, \rho)$ denotes the smallest number of ρ balls whose union constains $\mathbb{A}_{\omega,n}$. See Pollard (1984, §II.5) for an exposition of a slight variant of the VC81 result. □

13.2 Symmetrization for general stochastic processes

Consider a stochastic process $X = \{X_t : t \in T\}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. That is, for each t in T we have an $\mathcal{F}\backslash\mathcal{B}(\mathbb{R})$-measurable map from Ω into $\mathbb{R}$. As in Example <8>, to avoid measurability difficulties, assume T to be at worst countably infinite. (Again such an assumption could be justified by an appeal to the results from §9.6.) Suppose we want to bound $\mathbb{P}\Psi(\sup_t |X_t - \mathbb{P}X_t)$ for some ORLICZ function Ψ (see Chapter 5.)

Symmetrize by means of an ***independent copy*** $\{\widetilde{X}_t : t \in T\}$. That is, $\widetilde{X}$ has the same distribution as X and it is independent of X.

Remark. As explained in Section 2 of 9, the distribution of X could be defined as the probability measure P on the cylinder sigma-field $\mathcal{C}(\mathbb{R}^T)$ by $PB := \mathbb{P}\{\omega : X(\cdot, \omega) \in B\}$. As noted in that same Chapter, this definition could also be applied with uncountable T but then other measurability problems arise: even for processes as simple as the empirical distribution function for sampling from UNIF$[0,1]$, the cylinder

sigma-field is not large enough for the traditional analysis of many interesting functionals. It can be difficult to pin down the notion of "identical distributions" for random objects that don't have distributions that live on BOREL sigma-fields.

We might try to apply a JENSEN inequality,

$$\begin{aligned}\mathbb{P}\Psi\left(\sup_t |X(t,\omega) - \mathbb{P}X_t|\right) &= \mathbb{P}\Psi\left(\sup_t |X(t,\omega) - \mathbb{P}\widetilde{X}(t,\omega)|\right)\\ &\overset{?}{\le} \mathbb{P}\Psi\left(\sup_t |X(t,\omega) - \widetilde{X}(t,\omega)|\right),\end{aligned} \tag{11}$$

by arguing that independence of X and $\widetilde{X}$ means that the $\mathbb{P}\widetilde{X}(t,\omega)$ inside the Ψ acts only on "the part of ω that controls the randomness of $\widetilde{X}_t$ but really has no effect on the X_t part" and hence "it can be pulled outside the $\sup_t |\dots|$ then outside the Ψ and then combined with the first $\mathbb{P}$", whatever that means. To eliminate the handwaving many authors would adorn the innermost $\mathbb{P}$ (or the more traditional $\mathbb{E}$) with some subscript to suggest that it should be treated as an expectation conditional on the value of $X(\cdot,\omega)$. Such an argument seems to me a particularly clumsy way of invoking independence.

Alternatively, to manufacture an independent copy $\widetilde{X}$ of X we could work with the product space $\mathbb{R}^T \times \mathbb{R}^T$ equipped with its product sigma-field $\mathcal{C}(\mathbb{R}^T) \otimes \mathcal{C}(\mathbb{R}^T)$. Under the product measure $P \otimes P$ the coordinate maps X and $\widetilde{X}$ are independent, each with distribution P. We could then replace <11> by

$$\begin{aligned}&\mathbb{P}\Psi\left(\sup_t |X(t,\omega) - \mathbb{P}X_t|\right)\\ &\quad= P^x\Psi\left(\sup_t |x(t) - P^{\widetilde{x}}\widetilde{x}(t)|\right) \le P^x\Psi\left(P^{\widetilde{x}}\sup_t |x(t) - \widetilde{x}(t)|\right)\\ &\quad\le P^xP^{\widetilde{x}}\Psi\left(\sup_t |x(t) - \widetilde{x}(t)|\right) \qquad \text{by JENSEN}\\ &\quad= \mathbb{P}\Psi\left(\sup_t |X(t,\omega) - \widetilde{X}(t,\omega)|\right).\end{aligned} \tag{12}$$

An appeal to FUBINI is usually much cleaner than a hand-waving conditioning argument.

Inequality <12> has an analog for tail probabilities, whose proof requires a litle more work: it replaces the appeal to a JENSEN inequality by a direct independence argument analogous to the one described in Example <8>.

<13> **Theorem.** *Let $\{X_t(\omega) : t \in T\}$ and $\{\widetilde{X}_t : t \in T\}$ be independent stochastic processes with T countable. Let α, R_1, and R_2 be positive constants.*

(i) If $\alpha\mathbb{P}\{\widetilde{X}_t \le R_2\} \ge 1$ for each fixed t in T then

$$\mathbb{P}\{\sup_{t\in T} X_t > R_1 + R_2\} \le \alpha\mathbb{P}\{\sup_{t\in T}(X_t - \widetilde{X}_t) > R_1\}.$$

(ii) If $\alpha\mathbb{P}\{|\widetilde{X}_t| \le R_2\} \ge 1$ *for each* t *in* T *then*

$$\mathbb{P}\{\sup_t |X_t| > R_1 + R_2\} \le \alpha\mathbb{P}\{\sup_t |X_t - \widetilde{X}_t| > R_1\}.$$

Proof. For asssertion (i) suppose $T = \{t_j : j \in \mathbb{N}\}$, as in Example <8>, and define $\tau(\omega) := \inf\{j \in \mathbb{N} : X(t_j,\omega) > R_1 + R_2\}$ with $\tau(\omega) = \infty$ if $\sup_{j\in\mathbb{N}} X(t_j,\omega) \le R_1 + R_2$. Then the events $\{\tau = j\}$ and $\{\widetilde{X}(t_j) \le R_2\}$ are independent and $X(t_j,\omega) - \widetilde{X}(t_j,\omega) > R_1$ on their intersection. Thus

$$\begin{aligned}
\mathbb{P}\{\sup_{t\in T} X_t > R_1 + R_2\} &= \mathbb{P}\{\tau < \infty\} = \sum\nolimits_{j\in\mathbb{N}} \mathbb{P}\{\tau = j\} \\
&\le \alpha \sum\nolimits_{j\in\mathbb{N}} \mathbb{P}\{\tau = j, \widetilde{X}(t_j) \le R_2\} \qquad \text{by independence} \\
&\le \alpha \sum\nolimits_j \mathbb{P}\{\tau = j,\ X(t_j) - \widetilde{X}(t_j) > R_1\} \\
&\le \alpha \sum\nolimits_j \mathbb{P}\{\tau = j,\ \sup_{t\in T}(X_t - \widetilde{X}_t) > R_1\} \\
&\le \alpha\mathbb{P}\{\sup_{t\in T}(X_t - \widetilde{X}_t) > R_1\} \qquad \text{by } \textstyle\sum_j\{\tau(\omega) = j\} \le 1 \text{ for each } \omega.
\end{aligned}$$

The argument for assertion (ii) is almost the same except that we should define $\tau(\omega) := \inf\{j \in \mathbb{N} : |X(t_j,\omega)| > R_1 + R_2\}$ and use the inequality $|X_t| - |\widetilde{X}_t| \le |X_t - \widetilde{X}_t|$. □

Remark. Assertion (ii) of the Theorem is also valid if X_t takes values in some $\mathbb{R}^k$, or even if it takes values in some general normed linear space (provided you take care of a few measurability issues).

13.3 Empirical processes?

My interest in symmetrization began when I first tried to understand the argument used by VC71=Vapnik and Chervonenkis (1971). Their results would now be called ***uniform strong laws of large numbers*** (or USLLNs for short).

Amongst other innovations, Dudley (1978) extended the VC71 approach to establish a ***uniform central limit theorem*** (UCLT)for an ***empirical process*** ν_n based on a sample $\xi_1, \ldots, \xi_n$ from a probability measure P on $(\mathbb{A}, \mathcal{A})$ and a collection of sets $\mathcal{D}$ of $\mathbb{A}$:

$$\nu_n(D) := \sqrt{n}(P_nD - PD) := n^{-1/2} \sum\nolimits_i \Big(\{\xi_i(\omega) \in D\} - PD\Big) \qquad \text{for } D \in \mathcal{D}.$$

That is, he proved convergence in ditribution (in the sense described in Chapter 9) for ν_n as a random element of a space of bounded real functions on $\mathcal{D}$. He subsequently (Dudley, 1981) generalized his UCLT by regarding ν_n a stochastic process indexed by various subsets $\mathbb{F}$ of $\mathcal{L}^2(\mathbb{A}, \mathcal{A}, P)$:

<14> $$\nu_n(f) := n^{-1/2} \sum\nolimits_i \Big(f(\xi_i) - Pf\Big) \qquad \text{for } f \in \mathbb{F}.$$

Empirical process ideas soon became popular as tools for analyzing asymptotic problems in Statistics. For example, the ideas could be applied to a regression problem with observations $y_i = \langle x_i, \theta_0 \rangle + u_i$, for $i \in [\![n]\!]$, with observed vectors x_i in some euclidean space, an unknown θ_0, and unobserved errors u_i. They also worked (Pollard, 1990, §11) for an even more complicated problem with censored observations and an estimator $\widehat{\theta}$ chosen to minimize $\sum_{i \le n} |y_i^+ - \langle x_i, \theta \rangle^+|$, a problem very difficult to analyze using classical methods but easily handled by an empirical process analysis with $\mathbb{F}$ replaced by a set of functions $\{f_i(\cdot, \theta) : i \in [\![n]\!], \theta \in \Theta\}$ and a normalization different from the $n^{-1/2}$ for Dudley's theorems.

The customary $n^{-1/2}$ standardization also turned out to be irrelevant for a problem analyzed by Kim and Pollard (1990), which involved a process $\{G_n(t) - \mathbb{P}G_n(t) : t \in T\}$ with $T \subset \mathbb{R}^k$ and

<15> $$G_n(t) := n^{-1/3} \sum\nolimits_{i \le n} f\left(\xi_i(\omega), \theta_0 + t n^{-1/3}\right),$$

for a collection of functions $\{f(\cdot, \theta) : \theta \in \Theta\}$ indexed by a subset $\Theta \subseteq \mathbb{R}^k$.

Such examples persuaded me to follow the leads of Alexander (1987) and BLM=Boucheron, Lugosi, and Massart (2013, Chap 11) in focussing on general sums of independent processes $X_i := \{X_i(t, \omega) : t \in T\}$ for $i \in [\![n]\!]$. That is, any standardization (such as an $n^{-1/2}$ factor) is absorbed into the definition of the X_i process. For example, $X_i(t, \omega) := n^{-1/3} f\left(\xi_i(\omega), \theta_0 + t n^{-1/3}\right)$ for <15>; and for Dudley's UCLT (for functions) we would have $X_i(t, \omega) := n^{-1/2} f\left(\xi_i(\omega), t\right)$ where $\{f(\cdot, t) : t \in T\}$ is a (countable) collection of measurable maps into $\mathbb{A}$.

Following BLM I'll also use the name "empirical process" for any (centered) sum $\sum_i X_i(t, \omega)$ of independent stochastic processes, despite the fact that there need be no underlying "empirical measure" P_n constructed from observed (a.k.a. "empirical") data. When I feel the need for an explicit P_n I'll add the qualifier "traditional". Despite the risk of confusion, the generalization seems worthwhile because the traditional techniques largely carry over to the independent process setting. To avoid measurability difficulties I'll again assume that the index set T is countable.

Typically one chooses the standardization to control $\sum_i \operatorname{var}(X_i(t))$ for each t, with some sort of central limit effect in mind, which then has the side effect of making each individual $X_i(t)$ small in some probabilistic sense.

As Alexander (1987, §2) noted, the limit theory takes a slightly simpler form (particularly for processes centered to have zero expected values) if the ***envelope*** functions

<16> $$E_i(\omega) := \sup_{t \in T} |X_i(t, \omega)|$$

satsify something like a LINDEBERG condition or some other condition that makes $\max_i E_i$ small in some probabilistic sense.

<17> **Example.** Consider the traditional case where $\xi_1, \xi_2, \ldots$ are independent observations from some distribution P on a set $\mathbb{A}$ and $\mathbb{F} := \{f_t : t \in T\}$ is a collection of measurable functions on $\mathbb{A}$ for which $F(\cdot) := \sup_{t\in T} |f_t(\cdot)|$ belongs to $\mathcal{L}^2(P)$. The traditional empirical process

$$\nu_n(t,\omega) := n^{-1/2} \sum\nolimits_{i\le n} \Big(f_t(\xi_i(\omega)) - P f_t \Big)$$

corresponds to $X_i(t,\omega) = n^{-1/2} f_t(\xi_i(\omega))$, with envelope $E_i \le n^{-1/2} 2F(\xi_i(\omega))$. In that case

$$\mathbb{P}\{\max_{i\le n} E_i > \eta\} \le nP\{2F > n^{1/2}\eta\} \le PF^2\{2F > n^{1/2}\eta\}/\eta^2,$$

□ which tends to zero for each fixed $\eta > 0$ as $n \to \infty$.

Remark. If $\{\xi_{n,i} : i \in [\![n]\!],\ n \in \mathbb{N}\}$ is a triangular array of random variables, independent within each row and satisfying the mild regularity assumption that $\max_i \mathbb{P}\{|\xi_{n,i}| > \epsilon\} \to 0$ for each $\epsilon > 0$, then classical theory (see Petrov, 1975, §IV.4 and Le Cam, 1986) tells us that $\sum_i \xi_{n,i} \rightsquigarrow N(0,1)$ iff for each $\epsilon > 0$ we have:

(a) $\mathbb{P}\{\max_i |\xi_{n,i}| > \epsilon\} \to 0$ as $n \to \infty$.

(b) $\sum_i \mathbb{P}\left(\xi_{n,i}\{|\xi_{n,i}| \le \epsilon\}\right) \to 0$.

(c) $\sum_i \operatorname{var}\left(\xi_{n,i}\{|\xi_{n,i}| \le \epsilon\}\right) \to 1$.

13.4 Maximal inequalities for empirical processes: norm-like

Let me start with a simple example that expands on the symmetrization idea introduced in Section 13.2. The conditioning appears in the thin disguise of an upper bound expressed as a product integral. The following Theorem involves a general ORLICZ function Ψ. On a first reading you might find it easier to assume $\Psi(r) = r$.

<18> **Theorem.** *Suppose:*

(i) $S(t,\omega) := \sum_{i=1}^n X_i(t,\omega)$ *where* $\{X_i(t,\omega) : i \in [\![n]\!],\ t \in T\}$ *is a collection of independent stochastic processes with* T *countable.*

(ii) $X_i(t) \in \mathcal{L}^\Psi(\mathbb{P})$, *for each* i *in* $[\![n]\!]$ *and each* t *in* T, *for a given* ORLICZ *function* Ψ.

Then, for every probability measure $\mathbb{Q}$ *on* $\mathbb{B}_n := \{-1,+1\}^n$,

<19> $$\mathbb{P}\Psi\left(\sup_t \Big| S(t,\omega) - \mathbb{P}S(t) \Big|\right) \le \mathbb{P}^\omega \mathbb{Q}^{\mathfrak{s}} \Psi\left(2 \sup_t \Big| \sum\nolimits_i \mathfrak{s}_i X_i(t,\omega) \Big|\right)$$

□ *where* $S_t^\circ := S^\circ(t,\omega,\mathfrak{s}) := \sum_i \mathfrak{s}_i X_i(t,\omega)$.

Remark. If you worry about integrability issues you could initially replace T by a finite set. The general case for countable T would then follow by taking a supremum over an increasing sequence of finite subsets of T.

Proof. Make independent copies $\{\widetilde{X}_i(t,\widetilde{\omega}) : i \in [\![n]\!], t \in T\}$ on a new probability space $(\widetilde{\Omega},\widetilde{\mathcal{F}},\widetilde{\mathbb{P}})$ and define $\widetilde{S}_n(t,\widetilde{\omega}) := \sum_i \widetilde{X}_i(t,\widetilde{\omega})$. Then $\mathbb{P}\widetilde{S}(t) = \mathbb{P}S(t)$ for each t and the expression on the left-hand side of <19> equals

$$\begin{aligned}(\star) &:= \mathbb{P}^\omega \Psi\left(\sup_{t\in T}\left|S(t,\omega) - \widetilde{\mathbb{P}}^{\widetilde{\omega}}\widetilde{S}(t,\widetilde{\omega})\right|\right)\\ &\le \mathbb{P}^\omega \sup_t \Psi\left(\widetilde{\mathbb{P}}^{\widetilde{\omega}}\left|\sum_i X_i(t,\omega) - \widetilde{X}_i(t,\widetilde{\omega})\right|\right) \qquad \text{because } \Psi \text{ is increasing}\\ &\le \mathbb{P}^\omega\widetilde{\mathbb{P}}^{\widetilde{\omega}} \sup_t \Psi\left(\left|\sum_i X_i(t,\omega) - \widetilde{X}_i(t,\widetilde{\omega})\right|\right), \qquad <20>\end{aligned}$$

the final inequality coming from the JENSEN inequality and the fact that $\sup_t \widetilde{\mathbb{P}}(\dots) \le \widetilde{\mathbb{P}}\sup_t(\dots)$.

The final bound would be unchanged if we swapped some X_i with the corresponding $\widetilde{X}_i$. For example, under $\mathbb{P}\otimes\widetilde{\mathbb{P}}$ the joint distribution of

$$X_1, X_2, X_3, \dots, X_{n-1}, X_n, \widetilde{X}_1, \widetilde{X}_2, \widetilde{X}_3 \dots, \widetilde{X}_{n-1}, \widetilde{X}_n$$

is the same as the joint distribution of

$$X_1, \widetilde{X}_2, \widetilde{X}_3, \dots, \widetilde{X}_{n-1}, X_n, \widetilde{X}_1, X_2, X_3, \dots, X_{n-1}, \widetilde{X}_n.$$

Equivalently, in <20> we can just replace $X_2 - \widetilde{X}_2$ by $-(X_2 - \widetilde{X}_2)$, and $X_3 - \widetilde{X}_3$ by $-(X_3 - \widetilde{X}_3)$, and so on. In general we have

$$(\star) \le \mathbb{P}^\omega\widetilde{\mathbb{P}}^{\widetilde{\omega}} \sup_{t\in T} \Psi\left(\left|\sum_i \mathfrak{s}_i\Big(X_i(t,\omega) - \widetilde{X}_i(t,\widetilde{\omega})\Big)\right|\right)$$

for every choice of $\mathfrak{s} := (\mathfrak{s}_1,\dots,\mathfrak{s}_n)$ from $\mathbb{B}_n$. Integrate with respect to $\mathbb{Q}$ to get

$$\begin{aligned}(\star) &\le \mathbb{Q}^{\mathfrak{s}}\mathbb{P}^\omega\widetilde{\mathbb{P}}^{\widetilde{\omega}} \sup_{t\in T} \Psi\left(\left|\sum_i \mathfrak{s}_i\Big(X_i(t,\omega) - \widetilde{X}_i(t,\widetilde{\omega})\Big)\right|\right)\\ &\le \mathbb{P}\widetilde{\mathbb{P}}\mathbb{Q}\Psi\left(\sup_t\left|\sum_i \mathfrak{s}_i X_i(t,\omega)\right| + \sup_t\left|\sum_i \mathfrak{s}_i \widetilde{X}_i(t,\widetilde{\omega})\right|\right).\end{aligned}$$

Invoke the convexity inequality $\Psi(A+B) \le \frac{1}{2}\Psi(2A) + \frac{1}{2}\Psi(2B)$ to split the upper bound into a sum of two terms then discard unnecessary $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ to end up with the asserted inequality. □

Remark. The split into two equal contributions from the X_i and $\widetilde{X}_i$ processes is not essential; it just makes for a neater-looking bound.

The simplest choice for $\mathbb{Q}$ in the previous Theorem is the uniform distribution over $\mathbb{B}_n$, in which case the coordinates $\mathfrak{s}_1,\dots,\mathfrak{s}_n$ become independent random variables with $\mathbb{Q}\{\mathfrak{s}_i = +1\} = 1/2 = \mathbb{Q}\{\mathfrak{s}_i = -1\}$. The random variable $\mathfrak{s}_i X_i(t,\omega)$ has a symmetric distribution under $\mathbb{P}\otimes\mathbb{Q}$. From now on, with $\mathbb{Q}$ uniform on $\mathbb{B}_n$, I'll refer to $\mathfrak{s}_1,\dots,\mathfrak{s}_n$ as ***sign variables***, as in §7.1.

Remark. Maybe someone will one day come up with a clever trick involving a different $\mathbb{Q}$. For example, if i were replaced by a double indexing i, j, it might be helpful to have $\mathbb{Q}$ invariant under a fancier group of transformations of the index set.

Theorem <18> is useful because it lets us break the analysis of the empirical process into two steps, the first—effectively a conditioning step—involving only the randomness in S° provided by the sign variables under $\mathbb{Q}$. Moreover, for each fixed ω, the S° process involves the $\{X_i\}$'s only through the set of n-dimensional vectors $X(t,\omega) := (X_i(t,\omega) : i \in [\![n]\!])$, that is, $S^\circ(t,\omega,\mathfrak{s}) = \langle X(t,\omega), \mathfrak{s}\rangle$ where $\langle\cdot,\cdot\rangle$ denotes the usual ℓ^2 inner product.

With $\mathbb{Q}$ uniform, for each pair of vectors x, y in $\mathbb{R}^n$ and each real λ,

$$\mathbb{Q}e^{\langle x,\mathfrak{s}\rangle - \langle y,\mathfrak{s}\rangle} \le e^{\lambda^2 \sum_i (x_i - y_i)^2/2} = e^{\lambda^2 |x-y|_2^2/2},$$

where $|\cdot|_2$ denotes the usual ℓ^2 norm on $\mathbb{R}^n$. That is, under $\mathbb{Q}$, the increment has a subgaussian distribution, in the sense of §7.1. Using the ORLICZ norm characterization of subgaussianity from §7.2 we also have

$$\|\langle x,\mathfrak{s}\rangle - \langle y,\mathfrak{s}\rangle\|_{\Psi_2,\mathbb{Q}} \le K_0 |x-y|_2 \qquad \text{where } \Psi_2(r) := \exp(r^2) - 1, \tag{21}$$

for K_0 a universal constant (smaller than $\sqrt{6}$).

Remark. I write $\|\ldots\|_{\Psi_2,\mathbb{Q}}$ instead of $\|\ldots\|_{\Psi_2}$, as a reminder that the calculation takes place in $\mathcal{L}^{\psi_2}(\mathbb{Q})$ with ω held fixed.

In particular, under $\mathbb{Q}$ the process $\{S_t^\circ : t \in T\}$ has subgaussian increments controlled by a random semimetric $\rho_\omega(s,t) := |X(s,\omega) - X(t,\omega)|_2$ for $s, t \in T$:

$$S_s^\circ - S_t^\circ \in \text{SUBG}(\rho_\omega(s,t)^2). \tag{22}$$

As t runs over the index set T the $X(t,\omega)$ vectors trace out a subset $\mathcal{X}_\omega := \{X(t,\omega) : t \in T\}$ of R^n. The $\sup_t \left|S^\circ(t,\omega,\mathfrak{s})\right|$ in <19> could be rewritten as $\sup_{x\in\mathcal{X}_\omega} |\langle x,\mathfrak{s}\rangle|$. We might reasonably hope to control this process for fixed ω using the chaining methods described in Chapters 10 and 11, leaving a function of ω that can be bounded more easily (we also hope) than the expression on the left-hand side of <19>.

For example, suppose $\mathcal{X}_\omega$ has ℓ^2-diameter $D(\omega)$ and, for simplicity, assume $0 \in \mathcal{X}_\omega$. With $\delta_i := D(\omega)/2^i$ for $i = 0, 1, \ldots$ suppose we have a $\{\delta_i\}$-packing framework, in the sense described in §10.4, rooted at 0. Then the chaining argument from §10.6 gives

$$\left\|\sup_{x\in\mathcal{X}_\omega} |\langle x,\mathfrak{s}\rangle|\right\|_{\Psi_2,\mathbb{Q}} \le J(\omega) := C\int_0^{D(\omega)} \Psi_2^{-1}\Big(\text{PACK}(r,\mathcal{X}_\omega,\rho_2)\Big)\, dr, \tag{23}$$

with C a universal constant. As shown in §7.2, if the upper bound in <23> is finite then it implies a host of weaker (but more easily interpreted) inequalities for S° as an element of $\mathcal{L}^p(\mathbb{Q})$: for each $p \geq 1$ there is a constant C_p for which

$$<24> \qquad \left\| \sup_{x \in \mathcal{X}_\omega} |\langle x, \mathfrak{s} \rangle| \right\|_{p,\mathbb{Q}} \leq C_p J(\omega),$$

from which we get

$$\mathbb{P} \sup_t |S_t(\omega) - \mathbb{P} S_t|^p \leq C_p' \, \mathbb{P} \, (C_p J(\omega))^p$$

for some new constant C_p'. Of course the inequality is useful only if J belongs to $\mathcal{L}^p(\mathbb{P})$. See Pollard (1989) and Kim and Pollard (1990), where very similar bounds were applied to derive limit theorems for statistical models.

13.5 Maximal inequalities for empirical processes: tail-probabilty

The symmetrization method used to prove Theorem <18> can also be applied to tail probability bounds. This time the $\mathfrak{s}_i$'s will be built into the whole construction—which might remind you of the paired comparison method described in Example <2>—and not just added in as an afterthought justified by the symmetry.

Again, as a measure-theoretic precaution, assume that T is at worst countably infinite. and start from a set of independent processes $X_i := \{X_i(t,\omega) : t \in T\}$ for $i \in [\![n]\!]$, living on $(\Omega, \mathcal{F}, \mathbb{P})$, with independent copies $\widetilde{X}_i := \{\widetilde{X}_i(t,\omega) : t \in T\}$ for $i \in [\![n]\!]$, living on some $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$. For each $\mathfrak{s}$ in $\mathbb{B}_n := \{-1, +1\}^n$ and each i in $[\![n]\!]$ define a pair of new processes by

$$\begin{aligned} \left(Y_i, \widetilde{Y}_i\right) &= \left(Y_i(t,\omega,\widetilde{\omega},\mathfrak{s}), \widetilde{Y}_i(t,\omega,\widetilde{\omega},\mathfrak{s}) : t \in T\right) \\ <25> \qquad &:= \{\mathfrak{s}_i = +1\}\left(X_i, \widetilde{X}_i\right) + \{\mathfrak{s}_i = -1\}\left(\widetilde{X}_i, X_i\right). \end{aligned}$$

The processes Y_i and $\widetilde{Y}_i$ each have the same distribution as X_i and all $2n$ processes $Y_1, \ldots, Y_n, \widetilde{Y}_1, \ldots, \widetilde{Y}_n$ are independent under $\mathbb{Q} \otimes \mathbb{P} \otimes \widetilde{\mathbb{P}}$. Most importantly, for each fixed $\mathfrak{s}$ the sum processes

$$<26> \qquad \mathcal{S}(t,\omega,\widetilde{\omega},\mathfrak{s}) := \sum\nolimits_i Y_i(t,\omega,\widetilde{\omega},\mathfrak{s}) \quad \text{AND} \quad \widetilde{\mathcal{S}}(t,\omega,\widetilde{\omega},\mathfrak{s}) := \sum\nolimits_i \widetilde{Y}_i(t,\omega,\widetilde{\omega},\mathfrak{s})$$

are also independent, with each having the same distribution as the process $S(t,\omega) := \sum_i X_i(t,\omega)$ under $\mathbb{P}$. That is, $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are independent copies of S for which

$$<27> \qquad \mathcal{S}(t) - \widetilde{\mathcal{S}}(t) = \sum\nolimits_i \mathfrak{s}_i \left(X_i(t,\omega) - \widetilde{X}_i(t,\widetilde{\omega})\right).$$

<28> **Theorem.** *Suppose* $S(t,\omega) := \sum_{i=1}^n X_i(t,\omega)$ *for independent processes* $\{X_i(t,\omega) : i \in [\![n]\!], t \in T\}$ *with* T *countable. Define* $S^\circ(t,\omega,\mathfrak{s}) := \sum_i \mathfrak{s}_i X_i(t,\omega)$. *Let* α, R_1, R_2 *be positive constants and* $\mathbb{Q}$ *be the uniform distribution on* $\mathbb{B}_n := \{-1,+1\}^n$. *Then:*

(i) If $\alpha\mathbb{P}\{S(t,\omega) - \mathbb{P}S(t) \le R_2\} \ge 1$ *for each* $t \in T$ *then*

$$\mathbb{P}\{\exists t \in T : S(t,\omega) - \mathbb{P}S(t) > R_1 + R_2\} \le 2\alpha\mathbb{P}^\omega\mathbb{Q}^{\mathfrak{s}}\{\exists t : S^\circ(t,\omega,\mathfrak{s}) > R_1/2\}.$$

(ii) If $\alpha\mathbb{P}\{|S(t,\omega) - \mu(t)| \le R_2\} \ge 1$ *for each* t *then*

$$\mathbb{P}\{\sup_t |S(t,\omega) - \mathbb{P}S(t)| > R_1 + R_2\} \le 2\alpha\mathbb{P}^\omega\mathbb{Q}^{\mathfrak{s}}\{\sup_t |S^\circ(t,\omega,\mathfrak{s})| > R_1/2\}.$$

Proof. For a fixed $\mathfrak{s}$, build independent copies as in <25> through <27> then invoke Theorem <13> with X_t replaced by $S_t - \mathbb{P}S_t$ and $\widetilde{X}_t$ replaced by $\widetilde{S}_t - \widetilde{\mathbb{P}}\widetilde{S}_t$ to get

$$\begin{aligned}
&\mathbb{P}\{\exists t : S(t,\omega) - \mathbb{P}S(t) > R_1 + R_2\} \\
&\quad \le \alpha\mathbb{P}^\omega\widetilde{\mathbb{P}}^{\widetilde{\omega}}\{\sup_t : S(t,\omega) - \widetilde{S}(t,\widetilde{\omega}) > R_1\} \\
&\quad = \alpha\mathbb{P}^\omega\widetilde{\mathbb{P}}^{\widetilde{\omega}}\{\sup_t : \mathbb{S}(t,\omega,\widetilde{\omega},\mathfrak{s}) - \widetilde{\mathbb{S}}(t,\omega,\widetilde{\omega},\mathfrak{s}) > R_1\} \\
&\quad = \alpha\mathbb{P}^\omega\widetilde{\mathbb{P}}^{\widetilde{\omega}}\{\sup_t : \sum\nolimits_i \mathfrak{s}_i\Big(X_i(t,\omega) - \widetilde{X}_i(t,\widetilde{\omega})\Big) > R_1\} \qquad \text{by <27>.}
\end{aligned}$$

Bound the event in the last line by

$$\{\sup_t \sum\nolimits_i \mathfrak{s}_i X_i(t,\omega) > R_1/2\} + \{\sup_t \sum\nolimits_i (-\mathfrak{s}_i)\widetilde{X}_i(t,\omega) > R_1/2\},$$

then integrate through with respect to $\mathbb{Q}$ (using the fact that $-\mathfrak{s}$ has the same uniform distribution as $\mathfrak{s}$ under $\mathbb{Q}$) to arrive at (i).

□ The argument for (ii) is similar.

13.6 Oscillation control for an empirical process

From now on, $\mathbb{Q}$ *will always denote the uniform probability measure on the discrete cube* $\mathbb{B}_n := \{-1,+1\}^n$. *Under* $\mathbb{Q}$, *the coordinates* $\mathfrak{s}_1, \ldots, \mathfrak{s}_n$ *become independent sign variables.*

The main Theorem in §9.8 established a result slightly stronger than the following assertion:

Let $\{\nu_t^{(n)} : t \in T,\ n \in \mathbb{N}\}$ be a sequence of stochastic processes defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and indexed by a totally bounded semimetric space (T, ρ). Let μ be a tight probability measure that concentrates on $\text{UC}(T, \rho)$, the set of all bounded, uniformly ρ-continuous real functions on T. Then $\nu^{(n)} \overset{*}{\rightsquigarrow} \mu$ (in the sense defined by §9.7) if:

(OSC.i) $\pi_J \nu^{(n)} \rightsquigarrow \pi_J \mu$ *for each finite subset J of T (fidi convergence).*

(OSC.ii) For each $\eta > 0$ and $\epsilon > 0$ there exists a $\delta > 0$ such that $\limsup_n \mathbb{P}^*\{\text{OSC}(\delta, \nu^{(n)}, T, \rho) \geq \eta\} < \epsilon$.

Here $\text{OSC}(\delta, x, T, \rho) := \sup\{|x(t_1) - x(t_2)| : t_1, t_2 \in T \text{ and } \rho(t_1, t_2) < \delta\}$ for each bounded, real-valued function x on T. In Chapter 9 the use of ρ was implicit in the definition of the oscillation of a function; in this Section we must specify that ρ is being used because there will be several semimetrics floating around.

Symmetrization helps with verification of (OSC.ii) for a centered empirical processes $\nu^{(n)}(t, \omega) := S(t, \omega) - \mathbb{P}S_t$, where $S(t, \omega) := \sum_{i \in [\![n]\!]} X_i(t, \omega)$. Consider just a single n. As before, let $X(t, \omega)$ be the random element of $\mathbb{R}^n$ with ith component $X_i(t, \omega)$. Define new processes indexed by $\mathbb{T} := T \times T$:

<29> $$V(\tau, \omega) = X(t_1, \omega) - X(t_2, \omega) \qquad \text{for } \tau := (t_1, t_2) \in \mathbb{T},$$

<30> $$U(\tau, \omega) := \sum\nolimits_i V_i(\tau, \omega) \quad \text{AND} \quad U^\circ(\tau, \omega, \mathfrak{s}) := \sum\nolimits_i \mathfrak{s}_i V_i(\tau, \omega),$$

so that $\nu^{(n)}(t_1, \omega) - \nu^{(n)}(t_2, \omega) = U(\tau, \omega) - \mathbb{P}U_\tau$ for $\tau := (t_1, t_2) \in \mathbb{T}$ and

$$\text{OSC}(\delta, \nu^{(n)}, T, \rho) = \sup\{|U(\tau, \omega) - \mathbb{P}U_\tau| : \tau = (t_1, t_2) \text{ with } \rho(t_1, t_2) < \delta\}.$$

That is, the oscillation condition (OSC.ii) involves a maximal inequality for a new empirical process

<31> $$\{U(\tau, \omega) : \tau := (t_1, t_2) \in \mathbb{T}(\delta)\} \qquad \text{for } \mathbb{T}(\delta) := \{(t_1, t_2) \in \mathbb{T} : \rho(t_1, t_2) < \delta\}.$$

For each given $\eta > 0$ and $\epsilon > 0$ we need to find a $\delta > 0$ for which $\mathbb{P}\{\sup_{\mathbb{T}(\delta)} |U_\tau - \mathbb{P}U_\tau| \geq \eta\} < \epsilon$.

Remark. Verification of conditions for $\overset{*}{\rightsquigarrow}$ from §9.7 involves a whole sequence of maximal inequalities for different $\nu^{(n)}$'s: in that setting we need to find a single δ that works for all n large enough.

Under mild assumptions Theorem <28>(ii) will reduce the problem to an analogous problem involving the symmetrized process U°. First note that, for each fixed τ in $\mathbb{T}(\delta)$,

$$\left(\tfrac{\eta}{5}\right)^2 \mathbb{P}\{|U_\tau - \mathbb{P}U_\tau| > \eta/5\} \leq \text{var}(U_\tau) = \sum\nolimits_i \text{var}(V_i(\tau)) \leq \mathbb{P}|V_\tau|_2^2.$$

If we assume that

<32> $$\mathbb{P}|V_\tau|_2^2 := \sum\nolimits_i \mathbb{P}|V_i(t_1,t_2)|^2 \le \rho(t_1,t_2)^2 \qquad \text{for all } \tau=(t_1,t_2)\in\mathbb{T}$$

then any δ less than $\eta/\sqrt{50}$ will make the tail probability less than $1/2$, in which case (by Theorem <28>)

<33> $$\mathbb{P}\{\sup\nolimits_{\mathbb{T}(\delta)}|U_\tau-\mathbb{P}U_\tau|\ge\eta\}\le 4\mathbb{P}\mathbb{Q}\{\sup\nolimits_{\mathbb{T}(\delta)}|U^\circ(\tau,\omega,\mathfrak{s})|>2\eta/5\}.$$

Remark. Clearly the weaker assumption $\operatorname{var}(U_\tau)\le\rho(t_1,t_2)^2$, with $\tau=(t_1,t_2)$, would also suffice for <33>. The whole symmetrization argument would also work if we replaced the X_i's by the centered processes, $X_i(t,\omega)-\mathbb{P}X_i(t)$, in which case $\operatorname{var}(U_\tau)=\mathbb{P}|V_\tau|_2^2$. However, centering is not harmless: it would interfere with the sorts of combinatorial arguments that you will learn about in Chapters 14 and 15.

In the early empirical process literature, where $X_i(t,\omega))=n^{-1/2}f_t(\xi_i(\omega))$ for independent observations on P, the usual choice for the semimetric ρ was $\|f_{t_1}-f_{t_2}\|_{2,P}$, the $\mathcal{L}^2(P)$ distance, so that <32> was built in. Giné and Zinn (1984) seemed to prefer $\|f_{t_1}-f_{t_2}-(Pt_1-Pt_2)\|_{2,P}$, although they often required some sort of comparability with the $\mathcal{L}^2(P)$ distance. See their Theorem 3.1 and Remark 3.3 for example.

Under $\mathbb{Q}$ the process $\{U_\tau^\circ:\tau\in\mathbb{T}\}$ has subgaussian increments: for all real λ and $\sigma:=(s_1,s_2)$ and $\tau:=(t_1,t_2)$ in $\mathbb{T}$,

<34> $$\mathbb{Q}e^{\lambda(U_\sigma^\circ-U_\tau^\circ)}\le e^{\lambda^2 d_\omega(\sigma,\tau)^2/2} \qquad \text{for } d_\omega(\sigma,\tau):=|V(\sigma,\omega)-V(\tau,\omega)|_2.$$

If we wanted to control $\sup\{|U_\tau^\circ|:\tau\in\mathbb{T}(\delta)\}$ by one of the chaining arguments from Chapters 10 and 11 we would need to bound the diameter of the subset $\mathcal{V}_\omega:=\{V(\tau,\omega):\tau\in\mathbb{T}(\delta)\}$ of $\mathbb{R}^n$:

$$D(\omega):=\operatorname{diam}(\mathcal{V}_\omega):=\sup\{d_\omega(\sigma,\tau):\sigma,\tau\in\mathbb{T}(\delta)\}.$$

Remarks. There is no avoiding the task of bounding $D(\omega)$ because $\mathbb{Q}\{\sup_{v\in\mathcal{V}_\omega}|\langle v,\mathfrak{s}\rangle|\ge D(\omega)/4\}\ge c_0$ for some positive universal constant c_0. See Problem [2].

As the next Theorem will show, under mild assumptions, we can do even better than bounding $D(\omega)$: we can ensure that $\mathcal{V}_\omega$ lies inside a ball $\{v\in\mathbb{R}^n:|v|_2\le C\delta\}$ for suitably large C, with high probability. The key new idea will be another symmetrization argument, slightly different from the one used for Theorem <28>. This time we symmetrize norms of sums using Theorem <13> together with a simple inequality that somehow acquired the name ***square-root trick***: for $A=(a_1,\dots,a_n)$ and $B=(b_1,\dots,b_n)$ in $\mathbb{R}^n$,

<35> $$\big|\,|A|_2-|B|_2\,\big|=\frac{\big||A|_2^2-|B|_2^2\big|}{|A|_2+|B|_2}\le|\sum\nolimits_i a_i^2-b_i^2|/\sqrt{\sum\nolimits_i a_i^2+b_i^2}\,.$$

The $\sum_i a_i^2+b_i^2$ contribution makes the trick work.

<36> **Theorem.** *Let $\{X_i(t) : t \in T\}$ for $i \in [\![n]\!]$ be independent processes with T countable and envelope functions $E_i(\omega) := \sup_{t\in T} |X_i(t,\omega)|$. Define processes $\{V_\tau : \tau \in \mathbb{T}(\delta)\}$ and $\mathbb{T}(\delta)$ as in <29> and <31>. Then $\mathbb{P}\{\sup_{t\in\mathbb{T}(\delta)} |V(t,\omega)|_2 > C\delta\}$ is less than*

<37> $$\mathbb{P}\{\max_i E_i(\omega) > \theta\} + \mathbb{P}\{\text{PACK}(\delta, T, \rho_\omega) > m\} + 4\exp\left(\log m - \tfrac{1}{8}(R\delta/\theta)^2\right)$$

□ *for constants $R > 0$, each $\theta > 0$, and each $m \in \mathbb{N}$, if $C := 2(\sqrt{2} + 4 + R)$.*

Remark. The choice of the constants will depend on the details for particular applications. For example, to establish convergence in distribution for Doob's heuristic (§9.1) I claim that $R = 1$, $\theta = n^{-1/2}$, and $m = \lceil 2/\delta^2 \rceil$ would work for all n large enough. You might find it helpful to use the results from §10.6 and Theorem <36> to provide a complete justification for Doob's idea. For something more challenging, try to prove Giné and Zinn (1984, Thm 8.10).

Proof. The calculations will involve constants $C = 2(\sqrt{2} + 4 + R)$, $C_1 = \sqrt{2}+4+R$, $C_2 = 4+R$, $C_3 = R$. The reasons for these choices will become clearer in the five main steps of the argument.

Use the envelope functions to split each $X_i(t,\omega)$ into a bounded part, $Y_i(t,\omega) := X_i(t,\omega)\{E_i(\omega) \le \theta\}$ plus a remainder for which a crude bound will suffice. Define $W(\tau,\omega) := Y(t_1,\omega) - Y(t_2,\omega)$ for $\tau := (t_1,t_2) \in \mathbb{T}(\delta)$. Then $|Y_i(t,\omega)| \le 2\theta$ and $W_\tau = V_\tau$ on the event $\{\max_i E_i(\omega) \le \theta\}$, so that

$$\mathbb{P}\{\sup_\tau |V(\tau,\omega)|_2 > C\delta\} \le \mathbb{P}\{\sup_\tau |W(\tau,\omega)|_2 > C\delta\} + \mathbb{P}\{\max_i E_i(\omega) > \theta\}.$$

The rest of the argument can then focus on W and Y. Equivalently, to cut down on notation I'll just forget about the first term in <37> and assume straight off that $|X_i(t,\omega)| \le \theta$ for all t in T.

Remark. As shown by Talagrand (2021, §6.8), it sometimes pays to split $X_i(t,\omega)$ at levels that depend on t.

For each $\mathfrak{s}$ in $\mathbb{B}_n$ split the vector $V(\tau,\omega)$ into another sum of two random n-vectors, $V^\oplus(\tau,\omega,\mathfrak{s})$ and $V^\ominus(\tau,\omega,\mathfrak{s})$, defined by

$$V_i^\oplus(\tau) := V_i^\oplus(\tau,\omega,\mathfrak{s}) = V_i(\tau,\omega)\{\mathfrak{s}_i = +1\} \text{ for each } i \text{ in } [\![n]\!],$$
$$V_i^\ominus(\tau) := V_i^\ominus(\tau,\omega,\mathfrak{s}) = V_i(\tau,\omega)\{\mathfrak{s}_i = -1\} \text{ for each } i \text{ in } [\![n]\!].$$

Remark. For each fixed $\mathfrak{s}$ the processes $V^\oplus(\,\cdot\,,\omega,\mathfrak{s})$ and $V^\ominus(\,\cdot\,,\omega,\mathfrak{s})$ depend on different blocks of $V_i(\,\cdot\,,\omega)$'s. Replacement of $\mathfrak{s}$ by $-\mathfrak{s}$ (which has the same distribution under $\mathbb{Q}$ as $\mathfrak{s}$) would interchange the roles of $V^\oplus$ and $V^\ominus$. These facts underlie the two important properties <41> and <42> stated below.

To avoid notational confusion with expressions like $||\dots|_2 - |\dots|_2|$, define random variables

$$\mathcal{N}_\tau := |V(\tau,\omega)|_2, \quad \mathcal{N}_{\tau,\oplus} := |V^\oplus(\tau,\omega,\mathfrak{s})|_2, \quad \mathcal{N}_{\tau,\ominus} := |V^\ominus(\tau,\omega,\mathfrak{s})|_2.$$

By assumption <32> we have $\mathbb{P}\mathcal{N}_\tau^2 \le \delta^2$ for all τ in $\mathbb{T}(\delta)$ and by construction, for all ω and $\mathfrak{s}$,

<38> $$V(\tau,\omega) = V^\oplus(\tau,\omega,\mathfrak{s}) + V^\ominus(\tau,\omega,\mathfrak{s}),$$

<39> $$\mathcal{N}_\tau^2 = \mathcal{N}_{\tau,\oplus}^2 + \mathcal{N}_{\tau,\ominus}^2 \text{ and } \max\Big(\mathcal{N}_{\tau,\oplus}, \mathcal{N}_{\tau,\ominus}\Big) \le \mathcal{N}_\tau \le \mathcal{N}_{\tau,\oplus} + \mathcal{N}_{\tau,\ominus},$$

<40> $$\sum\nolimits_i \mathfrak{s}_i V_i(\tau,\omega)^2 = \mathcal{N}_{\tau,\oplus}^2 - \mathcal{N}_{\tau,\ominus}^2,$$

and:

<41> for each fixed $\mathfrak{s}$, the processes $V^\oplus$ and $V^\ominus$ are independent under $\mathbb{P}$;

<42> $|V^\oplus(\cdot,\omega,\mathfrak{s})|_2$ has the same distribution as $|V^\ominus(\cdot,\omega,\mathfrak{s})|_2$ under $\mathbb{P}\otimes\mathbb{Q}$.

The main argument will involve several steps using facts <38> through <42>.

(i) Control V by $V^\oplus$: By <38>, for each fixed $\mathfrak{s}$ in $\mathbb{B}_n$,

$$\mathbb{P}\{\sup_\tau \mathcal{N}_\tau > C\delta\} \le \mathbb{P}\{\sup_\tau \mathcal{N}_{\tau,\oplus} > C\delta/2\} + \mathbb{P}\{\sup_\tau \mathcal{N}_{\tau,\ominus} > C\delta/2\}.$$

Notice that only the right-hand side depends on $\mathfrak{s}$. Integrate with respect to $\mathbb{Q}$ on both sides of the inequality then invoke <42> to deduce that

$$\mathbb{P}\{\sup_\tau \mathcal{N}_\tau > C\delta\} \le 2\mathbb{Q}\mathbb{P}\{\sup_\tau \mathcal{N}_{\tau,\oplus} > C_1\delta\} \qquad \text{where } C_1 := C/2.$$

(ii) Symmetrize $V^\oplus$: Fix $\mathfrak{s}$. Note that

$$\mathbb{P}\{\mathcal{N}_{\tau,\ominus} > \delta\sqrt{2}\} \le \mathbb{P}\mathcal{N}_{\tau,\ominus}^2/(2\delta^2) \le \mathbb{P}\mathcal{N}_\tau^2/(2\delta^2) \le 1/2.$$

Thus, by <41> (with $\mathfrak{s}$ still fixed) and the symmetrization bound from Theorem <13>(i) with $C_2 := C_1 - \sqrt{2}$:

$$\mathbb{P}\{\sup_\tau \mathcal{N}_{\tau,\oplus} > C_1\delta\} \le 2\mathbb{P}\{\sup_\tau \Big(\mathcal{N}_{\tau,\oplus} - \mathcal{N}_{\tau,\ominus}\Big) > C_2\delta\}.$$

(iii) Approximate by a ρ_ω packing set: For each fixed ω let S_ω be a δ-packing set for T with cardinality $\text{PACK}(\delta, T, \rho_\omega)$. Then (see §10.4)

$$\min\{\rho_\omega(t,s) : s \in S_\omega\} \le \delta \qquad \text{for each } t \text{ in } T.$$

Consequently, for each $\tau = (t_1,t_2)$ in $\mathbb{T}$ there exists a $\sigma_\tau := \sigma(\tau,\omega) = (s_1,s_2)$ in $\mathbb{S}_\omega := S_\omega \times S_\omega$ for which

$$|V_\tau - V_{\sigma_\tau}|_2 \le |X_{t_1} - X_{s_1}|_2 + |X_{t_2} - X_{s_2}|_2 \le 2\delta.$$

For each $\mathfrak{s}$ the approximating V_{σ_τ} also has a decomposition, $V^{\oplus}(\sigma_\tau,\omega,\mathfrak{s}) + V^{\ominus}(\sigma_\tau,\omega,\mathfrak{s})$. As the same $\mathfrak{s}$ is used to split both V_τ and V_σ we get

$$\begin{aligned}|V_\tau^{\oplus} - V_{\sigma_\tau}^{\oplus}|_2^2 + |V_\tau^{\ominus} - V_{\sigma_\tau}^{\ominus}|_2^2 &= |V_\tau^{\oplus} + V_\tau^{\ominus} - V_{\sigma_\tau}^{\oplus} - V_{\sigma_\tau}^{\ominus}|_2^2 \\ &= |V_\tau - V_{\sigma_\tau}|_2^2 \le 4\delta^2.\end{aligned}$$

Take square roots to deduce that

$$\max\big(\,|V_\tau^{\oplus} - V_{\sigma_\tau}^{\oplus}|_2, |V_\tau^{\ominus} - V_{\sigma_\tau}^{\ominus}|_2\,\big) \le 2\delta,$$

implying $\max\left(|\mathcal{N}_{\tau,\oplus} - \mathcal{N}_{\sigma_\tau,\oplus}|, |\mathcal{N}_{\tau,\ominus} - \mathcal{N}_{\sigma_\tau,\ominus}|\right) \le 2\delta$ for each τ in $\mathbb{T}$.

As τ ranges over $\mathbb{T}(\delta)$ the approximating $\sigma(\tau,\omega)$'s range over a finite subset $\mathbb{S}_\omega(\delta)$ of $S_\omega \times S_\omega$. Thus, for each $\mathfrak{s}$,

$$\sup_{\tau\in\mathbb{T}(\delta)} |\mathcal{N}_{\tau,\oplus} - \mathcal{N}_{,\ominus}\tau| \le 4\delta + \max_{\sigma\in\mathbb{S}_\omega(\delta)} |\mathcal{N}_{\sigma,\oplus} - \mathcal{N}_{\sigma,\ominus}|,$$

<43> $$\{\omega : \sup_{\mathbb{T}(\delta)} |\mathcal{N}_{\tau,\oplus} - \mathcal{N}_{\tau,\ominus}| > C_2\delta\} \subseteq \{\omega : \max_{\sigma\in\mathbb{S}_\omega} |\mathcal{N}_{\sigma,\oplus} - \mathcal{N}_{\sigma,\ominus}| > C_3\delta\}.$$

where $C_3 := C_2 - 4$.

(iv) Use the square-root trick: For each σ, with ω and $\mathfrak{s}$ held fixed, inequality <35> followed by <39> and <40> give

$$\left|\mathcal{N}_{\sigma,\oplus} - \mathcal{N}_{\sigma,\ominus}\right| \le \frac{|\sum_i V_i^{\oplus}(\sigma)^2 - V_i^{\ominus}(\sigma)^2|}{\sqrt{\sum_i V_i^{\oplus}(\sigma)^2 + V_i^{\ominus}(\sigma)^2}} = \sum\nolimits_i \mathfrak{s}_i V_i(\sigma)^2/\mathcal{N}_\sigma.$$

For each σ in S_ω, the previous inequality together with the HOEFFDING inequality and the fact that $\sum_i V_i(\sigma,\omega)^4 \le 4\theta^2 \sum_i V_i(\sigma,\omega)^2 = 4\theta^2\mathcal{N}_\sigma^2$ give

$$\begin{aligned}\mathbb{Q}\{\big|\ \mathcal{N}_{\sigma,\oplus} - \mathcal{N}_{\sigma,\ominus}\ \big| > C_3\delta\} &\le \mathbb{Q}\{|\sum\nolimits_i \mathfrak{s}_i V_i(\sigma,\omega)^2| > C_3\delta\mathcal{N}_s\} \\ &\le 2\exp\left(-\tfrac{1}{2}(C_3\delta\mathcal{N}_\sigma)^2/\sum\nolimits_i V_i(s,\omega)^4\right) \le 2\exp(-\tfrac{1}{8}(C_3\delta/\theta)^2).\end{aligned}$$

(v) Combine (i)—(iv): A simple union bound on the set where the size of S_ω is at most m will now complete the Proof:

$$\begin{aligned}\mathbb{P}\{\sup_{\tau\in\mathbb{T}(\delta)} \mathcal{N}_\tau > C\delta\} &\le 2\mathbb{Q}\mathbb{P}\{\sup_\tau \mathcal{N}_{t,\oplus} > C_1\delta\} \qquad \text{by (i)} \\ &\le 2\mathbb{Q}\mathbb{P}\{\sup_\tau\big(\mathcal{N}_{\tau,\oplus} - \mathcal{N}_{\tau,\oplus}\big) > C_2\delta\} \qquad \text{by (ii)} \\ &\le 2\mathbb{P}\mathbb{Q}\{\max_{\sigma\in\mathbb{S}_\omega} |\mathcal{N}_{\sigma,\oplus} - \mathcal{N}_{\sigma,\ominus}| > C_3\delta\} \qquad \text{by (iii)} \\ &\le 2\mathbb{P}\mathbb{Q}\left(\{|S_\omega| > m\} + \{|S_\omega| \le m\}\sum\nolimits_{s\in S_\omega}\{|\mathcal{N}_{s,\oplus} - \mathcal{N}_{s,\ominus}| > C_3\delta\}\right) \\ &\le 2\mathbb{P}\{|S_\omega| > m\} + 4m^2\mathbb{P}\{|S_\omega| \le m\}\exp(-\tfrac{1}{8}(C_3\delta/\theta)^2) \qquad \text{by (iv).}\end{aligned}$$

□ The asserted inequality follows.

13.7 Problems

[1] Suppose $\{X_n\}$ and $\{Y_n\}$ are sequences of random variables for which X_n is independent of Y_n and $X_n + Y_n \rightsquigarrow N(0,1)$.

(i) Show that there is a sequence of real numbers $\{a_n\}$ for which the sequence $\{X_n + a_n\}$ is tight, that is, for each $\epsilon > 0$ there exists a constant K_ϵ for which

$$\limsup \mathbb{P}\{X_n + a_n \notin [-K_\epsilon, K_\epsilon]\} < \epsilon$$

Hint: Let a_n be a median of Y_n, that is $\mathbb{P}\{Y_n \geq a_n\} \geq 1/2$ and $\mathbb{P}\{Y_n \leq a_n\} \geq 1/2$. Show that

$$\mathbb{P}\{X_n+a_n > K\} \leq 2\mathbb{P}\{X_n-a_n > K, Y_n-a_n \leq 0\} \leq 2\mathbb{P}\{X_n+Y_n \geq K\} \to \bar{\Phi}(K).$$

Argue similarly for the lower tail.

(ii) Let $\mathbb{N}_1 \subset \mathbb{N}$ be a subsequence (Billingsley, 1968, §29) along which $X_n + a_n$ converges in distribution to some X. Show that $Y_n - a_n$ converges in distribution to some Y along the same subsequence. Hint: $\mathbb{P}e^{it(X_n+a_n)}\mathbb{P}e^{it(Y_n-a_n)} \to e^{-t^2/2}$ along $\mathbb{N}_1$.

(iii) Use Example <1> to deduce that X has a normal distribution, possibly degenerate.

[2] Suppose $W := \sum_i \mathfrak{s}_i a_i$ for a fixed vector $a = (a_1, \ldots, a_n)$ in $\mathbb{R}^n$ and independent sign variables $\mathfrak{s}_1, \ldots, \mathfrak{s}_n$. Argue as follows to prove that there exists a universal constant $c_0 > 0$ (depending neither on a nor on n) for which $\mathbb{Q}\{|W| \geq |a|_2/2\} \geq c_0$ for each a in $\mathbb{R}^n$.

(i) Show that $\mathbb{Q}W^2 = |a|_2^2$ and $\mathbb{Q}W^4 = \sum_i a_i^4 + 6\sum_{1\leq i<j\leq n} a_i^2 a_j^2 \leq 3|a|_2^4$.

(ii) Without loss of generality suppose $|a|_2 = 1$. Show that

$$\sqrt{(\mathbb{Q}W^4)\,\mathbb{Q}\{|W| \geq 1/2\}} \geq \mathbb{Q}W^2\{|W| \geq 1/2\} = \mathbb{Q}W^2 - \mathbb{Q}W^2\{W^2 < 1/4\} \geq 3/4,$$

which rearranges to $\mathbb{Q}\{|W| \geq 1/2\} \geq (3/4)^2/3$.

(iii) Suppose $\mathcal{X}$ is a bounded subset of $\mathbb{R}^n$ with ℓ^2 diameter R. Deduce from part (i) that $\mathbb{Q}\{\sup_{x\in\mathcal{X}} |\langle x, \mathfrak{s}\rangle| \geq R/4\} \geq c_0$ Hint: If $x_1, x_2 \in \mathcal{X}$ and $W_j := \langle x_j, \mathfrak{s}\rangle$ for $j = 1, 2$, show that $\max(|W_1|, |W_2|) \geq |W_1 - W_2|/2$.

[3] As in Theorem <28>, suppose $X_1, \ldots, X_n$ are independent stochastic processes, each indexed by a countable set $T = \{t_j : j \in \mathbb{N}\}$, but now also assume that $\mathbb{P}X_i(t) = 0$ for $i \in [\![n]\!]$ and $t \in T$. Define $S_i(t,\omega) := \sum_{j\leq i} X_j(t,\omega)$ and $M_i(\omega) := \sup_t |S_i(t,\omega)|$ for $i \in [\![n]\!]$. For some positive constants η and α suppose $\alpha\mathbb{P}\{|S_n(t_j) - S_i(t_j)| \leq \eta\} \geq 1$ for each $i \in [\![n]\!]$ and $j \in \mathbb{N}$. Argue as follows to show that

$$\mathbb{P}\{\max_i \sup_t |X_i(t)| > 4\eta\} < 2\alpha\mathbb{P}\{\max_t |S_n(t)| > \eta\}.$$

(i) Define $\theta(\omega) := \inf\{i \in [\![n]\!] : M_i(\omega) > 2\eta\}$ and, for each i in $[\![n]\!]$, define $\tau_i(\omega) := \inf\{j \in \mathbb{N} : |S_i(t_j,\omega)| > \eta\}$, with $\inf\emptyset = +\infty$. Show that $\mathbb{P}\{\max_i M_i > 2\eta\}$ is less than

$$\begin{aligned}&\alpha\mathbb{P}\sum\nolimits_{i\in[\![n]\!],j\in\mathbb{N}}\{\theta=i,\tau_i=j\}\{|S_i(t_j)|>2\eta\}\{|S_n(t_j)-S_i(t_j)|\le\eta\}\\&\quad\le\alpha\mathbb{P}\{\max_t|S_n(t)|>\eta\}\end{aligned}$$

(ii) Define $S_0(t,\omega)=0$. Show that $\sup_t|X_i(t)|\le M_i+M_{i-1}$ for each i.

[4] Suppose $\{X_i(t,\omega) : i \in [\![n]\!], t \in T\}$ is a collection of independent stochastic processes with T countable and Ψ is an ORLICZ function. Define $\mu_i(t) := \mathbb{P}X_i(t)$. Starting from inequality <19> show that

$$\mathbb{P}\Psi\left(\sup_t\left|\sum\nolimits_i X_i(t,\omega)-\mu_i(t)\right|\right)\le\mathbb{P}^\omega\gamma_n^g\Psi\left(2\sup_t\left|\sum\nolimits_i g_iX_i(t,\omega)\right|/\kappa\right)$$

where $\gamma_n := N(0,I_n)$ and $\kappa := \gamma_n|g_i| = \sqrt{2/\pi}$. Hint: Start with a factor $\gamma_n|g_i|/\kappa$ inside the sum. Note that $\mathfrak{s}_i|g_i| \sim N(0,1)$ under $\mathbb{Q}\otimes\gamma_n$.

Remark. This trick gains us very little if we just use the subgaussian tail bound for the increments of the process $W := \{\langle g,x\rangle : x\in\mathcal{X}_\omega\}$. In a similar setting Giné and Zinn (1984, Prop 3.4) did reap some benefits by assuming nice limiting properties for a sequence of empical processes then invoking a gaussian comparison inequality.

Talagrand (2021, p. 176) has pointed out that the subgaussian process $S^\circ := \{\langle\mathfrak{s},x\rangle : x\in\mathcal{X}_\omega\}$ has "better tails" than W. It helps greatly to apply some truncations to the x's before gaussifying. More precisely, it pays to apply the chaining method described in §11.7 to the truncated S° process.

[5] Let $B = (b_{i,j})$ be an $n\times n$ symmetric (real) matrix with zeros on its diagonal and $\mathfrak{s} = (\mathfrak{s}_1,\ldots,\mathfrak{s}_n)$ be a vector of independent sign variables under the uniform distribution $\mathbb{Q}$ on $\{-1,+1\}^{[\![n]\!]}$. Define $W := \mathfrak{s}'B\mathfrak{s} = \sum_{i,j}\mathfrak{s}_i\mathfrak{s}_jb_ib_j$. Show that there exists a universal constant C for which $\|W\|_{\Psi_1,\mathbb{Q}} \le C\,\|B\|_{\mathrm{F}}$, where $\|B\|_{\mathrm{F}} := (\sum_{i,j}b_{i,j}^2)^{1/2}$ and $\Psi_1(r) := e^r-1$, by the following steps.

(i) With $\gamma_n := N(0,I_n)$ and $\kappa=\gamma_n|g_i|$, as in the previous Problem, show that

$$\begin{aligned}\mathbb{Q}e^{|W|/C}&=\mathbb{Q}\exp\left(\left|\sum\nolimits_{i,j}(\gamma_n|g_i|/\kappa)\mathfrak{s}_i(\gamma_n|g_j|/\kappa)\mathfrak{s}_jb_{i,j}/C\right|\right)\\&\le\gamma_n\exp\left(\left|\sum\nolimits_{i,j}g_ig_jb_{i,j}/C_1\right|\right)\qquad\text{where } C_1:=C\kappa^2\\&\le\gamma_n\exp\left(g'Bg/C_1\right)+\gamma_n\exp\left(-g'Bg/C_1\right).\end{aligned}$$

(ii) Let B have spectral representation $B = L\Theta L'$, with L orthogonal and $\Theta := \text{diag}(\theta_1, \ldots, \theta_n)$. (The columns of L are the eigenvectors corresponding to the eigenvalues $\{\theta_i\}$.) Show that

$$\sum\nolimits_i \theta_i = \text{trace}B = 0 \quad \text{AND} \quad \sum\nolimits_i \theta_i^2 = \text{trace}(B'B) = \|B\|_{\text{F}}^2 .$$

(iii) Show that, provided $\max |2\theta_i/C_1| \le 1/2$,

$$\begin{aligned}
&\gamma_n \exp\left(g'Bg/C_1\right) = \gamma_n \exp\left(g'\Theta g/C_1\right) = \prod\nolimits_i \gamma_n \exp(\theta_i g_i^2/C_1) \\
&= \exp\left(-\tfrac{1}{2}\sum\nolimits_i \log(1 - 2\theta_i/C_1)\right) \\
&\le \exp\left(\tfrac{1}{2}\sum\nolimits_i 2\theta_i/C_1 + 4\theta_i^2/C_1^2\right) = \exp\left(2\,\|B\|_{\text{F}}^2/C_1^2\right).
\end{aligned}$$

(iv) Take a peek at §5.1 then choose C_1 as a suitable multiple of $\|B\|_{\text{F}}$ to complete the argument. Note: $\|-B\|_{\text{F}} = \|B\|_{\text{F}}$.

13.8 Notes

Judging by the many results I have seen attributed to Paul Lévy, he deserves a lot of the credit for the use of symmetrization ideas in probability theory. See, for example, the discussions of Lévy's work by Le Cam (1972, p. xvii) and Loève (1973, p. 4). Unfortunately I do not know his work well enough to give precise citations.

There is a story behind the use of sign variables to symmetrize an empirical process. When I first tried to understand the VC71 technique (as described in my Example <8>) I got stuck in their section 4, where they stated an exponential bound for tails of a HYPERGEOMETRIC distribution, with the comment "This estimate can be derived by a simple but long computation and so we omit the proof". I knew that the HYPERGEOMETRIC came from sampling without replacement, which I thought might be replaced by another sampling method: include each member of the population into the sample with probability 1/2, independently across members, thereby replacing the HYPERGEOMETRIC by a $\text{Bin}(2n, 1/2)$ distribution at the cost of creating samples of random size. Unfortunately, at that time I didn't know about the standard inequalities described in Chapter 3. In particular, I didn't know how to bound Binomial tails until I read Dudley (1978), when I became aware of the BERNSTEIN inequality. That discovery made Pollard (1981) possible. Subsequently (Pollard, 1982) I realized that the sampling idea could be taken further, which led me to the method described in Section 13.5. Some of my friends later informed me that I had reinvented "Rademacher" variables, which were well known in the BANACH-space literature (cf. Marcus and Pisier, 1981, page 3 and Kahane, 1968, §1.7). Also I learned that Koltchinskii (1981, Lemma 2) had

come up with an equivalent symmetrization idea. The obvious similarity to the paired-comparsion method, described in Example <2>, only dawned on me many years later. Belatedly I came to understand how much modern theory relies on classical exponential tail bounds (such as those described in Chapters 3 through 8).

The original idea for the square root trick, for traditional empirical processes indexed by collections of sets, came from Le Cam (1983, §3). Giné and Zinn (1984, Lemma 5.2) extended the idea to traditional empirical processes indexed by uniformly bounded collections of functions. Alexander (1987, §VII) realized that the result could be extended to unbounded processes by means of a truncation of the envelope variables.

References

Alexander, K. S. (1987). Central limit theorems for stochastic processes under random entropy conditions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 75*, 351–378.

Billingsley, P. (1968). *Probability Measures* (2nd ed.). New York: Wiley.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Box, G. E. P. and S. L. Andersen (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society, Series B 17*, 1–34.

Cramér, H. (1936). Über eine Eigenschaft der normalen Verteilungsfunktion. *Mathematische Zeitung 41*, 405–414.

Devroye, L. (1982). Bounds for the uniform deviation of empirical measures. *Journal of Multivariate Analysis 12*, 72–79.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability 6*, 899–929.

Dudley, R. M. (1981). Donsker classes of functions. In M. Csörgő, D. A. Dawson, J. N. K. Rao, and A. K. M. E. Saleh (Eds.), *Statistics and Related Topics*, pp. 341–352. Amsterdam: North-Holland.

Dudley, R. M. (2014). *Uniform Central Limit Theorems* (2nd ed.), Volume 142 of *Cambridge studies in advanced mathematics*. Cambridge University Press. (First edition, 1999).

Fisher, R. A. (1935). *The Design of Experiments* (first ed.). Oxford University Press. The quotes are taken from the 8th edition (Hafner) of 1966, which was reprinted in 1991 by Oxford University Press as part of

a collection of three of Fisher's books, under the title "Statistical Methods, Experimental Design, and Scientific Inference".

Giné, E. and J. Zinn (1984). Some limit theorems for empirical processes. *Annals of Probability 12*, 929–989. (plus 9 pages of contributed discussion).

Gnedenko, B. V. (1968). *The Theory of Probability* (Fourth ed.). New York: Chelsea.

Kahane, J.-P. (1968). *Some Random Series of Functions*. Heath. Second edition: Cambridge University Press 1985.

Kim, J. and D. Pollard (1990). Cube root asymptotics. *Annals of Statistics 18*, 191–219.

Koltchinskii, V. I. (1981). On the central limit theorem for empirical measures. *Theory of probability and mathematical statistics 24*, 71–82.

Le Cam, L. (1972). Paul lévy. In L. Le Cam, J. Neyman, and E. L. Scott (Eds.), *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Volume III, Berkeley, pp. xiv–xx. University of California Press.

Le Cam, L. (1983). A remark on empirical measures. In Bickel, Doksum, and Hodges (Eds.), *A festschrift for Erich Lehmann in honor of his sixty-fifth birthday*, pp. 305–327. Wadsworth. Available online from https://www.stat.berkeley.edu/users/rice/LeCam/papers/.

Le Cam, L. (1986). The central limit theorem around 1935. *Statistical Science 1*, 78–96.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. New York: Springer.

Loève, M. (1973). Paul Lévy, 1886-1971. *Annals of Probability 1*, 1–18. Includes a list of Lévy's publications.

Marcus, M. and G. Pisier (1981). *Random Fourier Series with Applications to Harmonic Analysis*. Princeton NJ: Princeton University Press.

Massart, P. (1986). Rates of convergence in the central limit theorem for empirical processes. *Annales de l'Institut Henri Poincaré 22*, 381–423.

Petrov, V. V. (1975). *Sums of Independent Random Variables*. Springer-Verlag. Enlish translation from 1972 Russian edition.

Pollard, D. (1981). Limit theorems for empirical processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 57*, 181–195.

Pollard, D. (1982). A central limit theorem for empirical processes. *J. Austral. Math. Soc. Ser. A 33*, 235–248.

Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer.

Pollard, D. (1989). Asymptotics via empirical processes (with discussion). *Statistical Science 4*, 341–366.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*, Volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Hayward, CA: Institute of Mathematical Statistics.

Talagrand, M. (2021). *Upper and Lower Bounds for Stochastic Processes: Decomposition Theorems* (Second ed.), Volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag.

Vapnik, V. N. and A. Ya. Chervonenkis (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications 16*, 264–280. Russian original received by the editors on May 7, 1969.

Vapnik, V. N. and A. Ya. Chervonenkis (1981). Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and Its Applications 26*, 532–553. Russian original received by the editors on July 28, 1978.

Chapter 14
VC Sets

SECTION *14.1 presents a simple concrete example to illustrate the use of the combinatorial method for deriving bounds on packing numbers.*

SECTION *14.2 defines the concept of a VC-class of subsets $\mathcal{D}$ of some set $\mathbb{A}$, which leads to a polynomial bound for the numbers of subsets of the form $F \cap D$ for $D \in \mathcal{D}$, for each finite subset F of $\mathbb{A}$. The slightly sharper concept of shatter-dimension is obtained by restricting attention to a single F. Several Examples illustrate how shatter-dimension can be bounded and used, with control of packing numbers as the prime application.*

SECTION *14.3 derives the basic polynomial bound using the downshift method.*

*SECTION *14.4 interprets subsets of the discrete unit cube as the vertex sets of graphs. The shatter dimension gives a surprising upper bound for the number of edges of the graph. The method of proof again relies on downshifting.*

*SECTION *14.5 presents Haussler's refinement of the arguments from Section 14.2, leading to a sharper bound for packing numbers.*

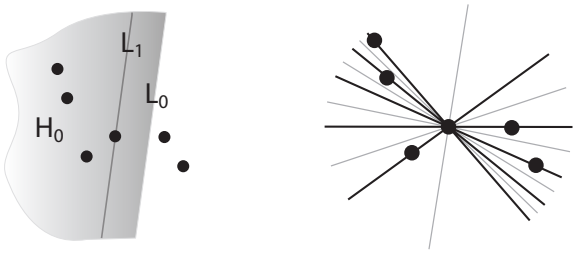
14.1 An introductory example

For my purposes, the combinatorial argument often referred to as the VC method—in honor of the important contributions of Vapnik and Chervonenkis (1971, 1981)—is of use mainly as a step towards calculation of packing/covering numbers. The method is elegant and leads to results not easily obtained in other ways. The basic calculations occupy only a few pages. Nevertheless, the ideas are subtle enough to appear beyond the comfortable reach of many would-be users. With that fact in mind, I offer a more concrete preliminary example, in the hope that the combinatorial ideas might then seem less mysterious.

<1> **Example.** Consider the set $\mathcal{H}$ of all closed half-spaces in $\mathbb{R}^2$. Let F be a set of n points in $\mathbb{R}^2$. How many distinct subsets can $\mathcal{H}$ pick out from F? That is, how large can the set $\mathcal{H}_F := \{F \cap H : H \in \mathcal{H}\}$ be? Certainly

its size (= cardinality) can be no larger than 2^n because F has only that many subsets. Even better, a simple argument will show that

$$|\mathcal{H}_F| := \text{cardinality of } \mathcal{H}_F \le p(n) := 1 + 6n(n-1).$$



Indeed, consider a particular nonempty subset F_0 of F picked out by a particular half-space H_0. (That is, $F_0 := F \cap H_0$.) There is no loss of generality in assuming that at least one point of F_0 (call it x_0) lies on L_0, the boundary of H_0: otherwise, without affecting F_0, we could replace H_0 by a smaller H_1 whose boundary runs parallel to L_0 through a point x_0 of F_0 that is closest to L_0.

For each x in F define $\mathcal{H}(x) = \{H \in \mathcal{H} : x \text{ lies on boundary of } H\}$, so that $\mathcal{H}_F \backslash \{\emptyset\} = \cup_{x \in F} \mathcal{H}(x)$ and $|\mathcal{H}_F| \le |F| \times \max_{x \in F} |\mathcal{H}(x)|$.

Consider any $\mathcal{H}(x_0)$. As seen from x_0, the other $n-1$ points of F all lie on a set $\mathcal{L}(x_0)$ of at most $n-1$ lines through x_0. Augment $\mathcal{L}(x_0)$ by another set $\mathcal{L}'(x_0)$ of at most $2(n-1)$ lines through x_0, one in each angle between two lines from $\mathcal{L}(x_0)$. The lines in $\mathcal{L}(x_0) \cup \mathcal{L}'(x_0)$ define a collection of at most $6(n-1)$ closed half-spaces, each with x_0 on its boundary, that accounts for all possible nonempty subsets of F picked out by $\mathcal{H}(x_0)$. The extra 1 takes care of the empty set. □

Remark. Apparently (Dudley, 1978, page 921) the bound can be reduced to $n^2 - n + 2$, which is sharp in the sense that it is achieved whenever the n points are in general position, that is, no straight line runs through more than 2 of the points. Dudley also described several precursors for the key combinatorial bound (my Theorem <4>). For my purposes the sharper bound is not needed. Indeed any polynomial would suffice for the consequences described below.

The slow increase in $|\mathcal{H}_F|$, at an $O(n^2)$ rate rather than a rapid 2^n rate, has a useful consequence for the packing numbers when $\mathcal{H}$ is equipped with an $\mathcal{L}^1(P)$ metric, for some probability measure P on the plane. The $\mathcal{L}^1(P)$ distance between two Borel sets B and B' is defined as $P|B - B'| = P(B \Delta B')$, the measure of the symmetric difference. The two sets are said to be ϵ-separated (in $\mathcal{L}^1(P)$) if $P(B \Delta B') > \epsilon$. As in Chapter 10, the packing number $\text{PACK}(\epsilon, \mathcal{H}, P)$—or $\text{PACK}(\epsilon, \mathcal{H}, \mathcal{L}^1(P))$ if there is any ambiguity about the norm being used—is defined as the largest N for which there exists a collection of N closed half-spaces, each of the $\binom{N}{2}$ pairs $> \epsilon$ apart.

<2> **Example.** Here is an argument, due to Dudley (1978, Lemma 7.13), to show that the packing numbers $\text{PACK}(\epsilon, \mathcal{H}, P)$ are bounded uniformly over P by a polynomial in $1/\epsilon$, for $0 < \epsilon \leq 1$. The result is surprising because it makes no regularity assumptions about the probability measure P.

Suppose the half-spaces $H_1, H_2, \ldots, H_N$ are ϵ-separated in $\mathcal{L}^1(P)$. By means of a cunningly chosen F, the polynomial bound on $|\mathcal{H}_F|$ will lead to an upper bound for N. The trick is to find a set $F_m := \{\xi_1, \ldots, \xi_m\} \subset \mathbb{R}^2$, with $m = \lceil 2 \log N/\epsilon \rceil$, from which each H_α picks out a different subset. Then $\mathcal{H}$ will pick out at least N subsets from F_m, implying that

$$N \leq p(m) \leq 1 + 6m(m-1) \leq 48\epsilon^{-2} (\log N)^2 .$$

Bounding the $\log N$ by a constant multiple of $N^{1/4}$ and solving the resulting inequality for N, we get an upper bound $N \leq O(1/\epsilon)^4$. With a smaller power in the bound for $\log N$ we would bring the power of $1/\epsilon$ arbitrarily close to 2.

Remark. As shown in Example <14>, the bound for the packing numbers for $\mathcal{H}$ can be sharpened to $C(\epsilon^{-1} \log(1/\epsilon))^2$, with C a universal constant, at least when $0 < \epsilon \leq 1/2$. With a lot more work, even the log term can be removed—see Section 14.5. At this stage there is little point in struggling to get the best bound in ϵ. For many applications, the qualitative consequences of a polynomial bound in $1/\epsilon$ are the same, no matter what the degree of the polynomial.

How do we find a suitable F_m? We need to place at least one point of F_m in each of the $\binom{N}{2}$ symmetric differences $H_i \Delta H_j$. It might seem we are faced with a delicate task involving consideration of all possible configurations of the symmetric differences, but here probability theory comes to the rescue.

As described in the Preface of the wonderful little book by Alon and Spencer (2000), the ***probabilistic method*** can prove existence by artificially introducing a probability measure into a problem:

> In order to prove the existence of a combinatorial structure with certain properties, we construct an appropriate probability space and show that a randomly chosen element in this space has the desired properties with positive probability. This method was initiated by Paul Erdős, who contributed so much to its development over the last fifty years, that is seems appropriate to call it "The Erdős Method."

Generate F_m as a random sample of size m from P. If $m \geq 2 \log N/\epsilon$, then there will be a strictly positive probability that the sample has the

desired property. Indeed, for fixed α, β with $1 \le \alpha < \beta \le 1$,

$$\begin{aligned}
&\mathbb{P}\{H_\alpha \text{ and } H_\beta \text{ pick out same points from } F_m\} \\
&\quad = \mathbb{P}\{\text{no points of sample in } H_\alpha \Delta H_\beta\} \\
&\quad = (1 - P(H_\alpha \Delta H_\beta))^m \\
&\quad \le (1-\epsilon)^m \le \exp(-m\epsilon).
\end{aligned}$$

Add up $\binom{N}{2}$ such probability bounds to get a conservative estimate,

$$\mathbb{P}\{\text{some pair } H_\alpha, H_\beta \text{ pick same subset from } F_m\} \le \tbinom{N}{2} \exp(-m\epsilon).$$

When $m \ge 2 \log N/\epsilon$ the last bound is strictly less than 1, as desired.

Notice how probability theory has been used to prove an existence result, which gives a bound for a packing number, which will be used to derive probabilistic consequences—all based ultimately on the existence of the polynomial bound for $|\mathcal{H}_F|$. □

Remark. The argument in the Example implicitly assumed that the ξ_i's sampled from P were all distinct, which need not be true if P has atoms. Thus $|F_m|$, the size of F_m, might be smaller than m. To be more precise I could have written $N \le p(|F_m|) \le p(m) \le \ldots$. The added rigor is hardly worth the trouble; usually ties cause only notational difficulties. Section 14.2 will eventually avoid the issue altogether by working with the n-tuple $(\xi_1, \ldots, \xi_n)$, which might contain repeats, instead of the set $\{\xi_1, \ldots, \xi_n\}$.

14.2 VC-dimension, shatter-dimension

The argument in the previous Section had little to do with the choice of $\mathcal{H}$ as the set of all closed half-spaces in a particular Euclidean space. It would apply to any collection $\mathcal{D}$ of (measurable) subsets of any probability space $(\mathbb{A}, \mathcal{A}, P)$ for which the size of $\mathcal{D}_F := \{F \cap D : D \in \mathcal{D}\}$ is bounded above by a fixed polynomial in $|F|$.

Unfortunately, for more complicated $\mathcal{D}$'s it can be difficult to derive such a polynomial bound directly although it is easier to prove existence of a finite k such that $|\mathcal{D}_F| < 2^{|F|}$ for every finite subset F of $\mathbb{A}$ with $|F| > k$. (Traditional terminology refers to this situation by saying that "$\mathcal{D}$ cannot shatter F".) Such a $\mathcal{D}$ is usually called a ***VC-class*** of sets in honor of Vapnik & Červonenkis. This property is equivalent to the finiteness of the ***VC-dimension*** for $\mathcal{D}$, defined as follows.

<3> **Definition.** *A collection $\mathcal{D}$ of subsets of a set $\mathbb{A}$, with $|\mathcal{D}| \ge 2$, is said to have VC-dimension d (written* $\text{VCDIM}(\mathcal{D}) = d$*) if both the following conditions hold.*

(i) There exists at least one subset F_0 of $\mathbb{A}$ for which

$$|F_0| = d \quad \text{AND} \quad |\{F_0 \cap D : D \in \mathcal{D}\}| = 2^d.$$

□ *(ii)* $|\{F \cap D : D \in \mathcal{D}\}| < 2^{|F|}$ *for every finite subset F of $\mathbb{A}$ with $|F| > d$.*

Remarks.

(i) If $D_1, D_2 \in \mathcal{D}$ with $D_1 \neq D_2$ then there exists an a in $\mathbb{A}$ for which $\{a\} \cap D_1 \neq \{a\} \cap D_2$. Thus $\text{VCDIM}(\mathcal{D}) \geq 1$ if $|\mathcal{D}| \geq 2$.

(ii) Empirical evidence shows that it can be quite difficult to rapidly convey the meaning of "VC-class" to a seminar audience.

(iii) Some authors use the term ***shatter dimension*** instead of VC-dimension. Motivated by the extension of the concept to classes of functions, and beyond, I feel a better name would be ***surround dimension***. In fact that is the term used in the next Chapter. I chose the name SDIM, which will appear in Definition <8>, because the initial 's' could be interpreted as a reference to either 'shatter' or 'surround'.

The key fact about VC-classes is often called the VC Lemma or the Sauer/Shelah Lemma, although credit should be spread more widely. (See the Notes in Section 14.7.)

<4> **Theorem.** *If $\mathcal{D}$ is VC-class of subsets of some $\mathbb{A}$ with $\text{VCDIM}(\mathcal{D}) = d$ then, for each finite subset F of $\mathbb{A}$ with $|F| = n$,*

□ $$|\mathcal{D}_F| := |\{F \cap D : D \in \mathcal{D}\}| \leq \beta(n, d) := \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}.$$

This Theorem asserts an upper bound that holds uniformly over all subsets F of size n. It will follow from a sharper result (Theorem <9>) that provides an upper bound for each individual F.

Remark. Typically Theorem <4> is invoked to provide an upper bound for the size of $\mathcal{D}_F$ by showing that $\mathcal{D}$ is not large enough to shatter subsets of size greater than some (smallish) d. That is, part (ii) of Definition <3> provides the means for controlling $|\mathcal{D}_F|$.

Clearly $\beta(n, d)$ is a polynomial of degree d in n. More precisely, it equals 2^n if $n \leq d$ and, if $n \geq d$,

$$\begin{aligned} \beta(n, d) &= \sum_{k=0}^{d} n(n-1)\ldots(n-k+1)/k! \\ &\leq \sum_{k=0}^{d} (n/d)^k d^k/k! \\ &\leq (en/d)^d \qquad \text{because } (n/d)^k \leq (n/d)^d \text{ when } n \geq d \geq k. \end{aligned} \tag{<5>}$$

The upper bound can be improved slightly:

$$\beta(n, d) \leq \beta_1(n, d) := 1.5n^d/d! \qquad \text{if } n \geq d + 2, \tag{<6>}$$

a result apparently due to Vapnik and Červonenkis. As Dudley (2014, Proposition 4.3) noted, the β_1 bound is "not far from optimal" because the $\binom{n}{d}$ contributes an $n^d/d!$ as the leading term in the polynomial $\beta(n,d)$. Compare with: $\beta_1(n,d) \approx 0.6 d^{-1/2}(en/d)^d$ by the STIRLING approximation from §2.5.

Vapnik and Chervonenkis (1971, §6) realized that the concept defined by Definition <3> needs sharpening if we seek necessary and sufficient conditions for uniform laws of large numbers. For example, suppose we have independent random elements $\xi_1, \ldots, \xi_n$ of $\mathbb{A}$ with $\xi_i \sim P_i$ for some probability measure P_i defined on a suitable sigma-field $\mathcal{A}$ on $\mathbb{A}$. Define independent $\{0,1\}$-valued stochastic processes

$$X_i(D,\omega) := \{\xi_i(\omega) \in D\} \qquad \text{for } D \in \mathcal{D} \text{ and } i \in [\![n]\!].$$

Remark. To avoid measurability difficulties you could require that each process $\{X_i(D,\cdot) : d \in \mathcal{D}\}$ be DOOB-separable or, essentially equivalently, just suppose $\mathcal{D}$ to be a countable subset of $\mathcal{A}$

Let $\mathbb{Q}$ denote the uniform distribution on the discrete hypercube $\{-1,1\}^n$. Then, as shown in §13.4, for each ORLICZ function we have

$$\mathbb{P}\Psi\left(\sup_{D\in\mathcal{D}} \left|\sum\nolimits_i X_i(D,\omega) - P_i D\right|\right) \le \mathbb{P}^\omega \mathbb{Q}^{\mathfrak{s}} \Psi\left(2 \sup_{D\in\mathcal{D}} \left|\sum\nolimits_i \mathfrak{s}_i X_i(D,\omega)\right|\right). \tag{7}$$

The $\mathbb{Q}$ integral on the right-hand side of the inequality in <7> can also be written as $\mathbb{Q}\Psi\left(\sup_{x\in\mathcal{X}_\omega} |\langle \mathfrak{s}, x\rangle|\right)$ where $\mathcal{X}_\omega$ denotes the set of all members of another discrete hypercube, $\mathcal{C}_n := \{0,1\}^n$, of the form

$$x(D,\omega) := \Big(X_1(D,\omega), \ldots, X_n(D,\omega)\Big) \qquad \text{for some } D \text{ in } \mathcal{D}.$$

Suitable bounds on the ℓ^2 packing numbers for $\mathcal{X}_\omega$ would lead to empirical process maximal inequalities via chaining arguments.

To make the right-hand side of <7> small we do not need a bound on the $\mathbb{Q}$ integrals that holds uniformly over all ω.

Remark. As $\mathcal{X}_\omega$ is finite and $\mathcal{D}$ need not be, there will typically be many different D's that contribute the same $x(D,\omega)$.

You should also take note of a subtle difference between the set $\{X_i(D,\omega) : i \in [\![n]\!]\}$ and the vector $x(D,\omega)$ for fixed ω: even when the set $\{\xi_i(\omega) : i \in [\![n]\!]\}$ has size $< n$, each $x(D,\omega)$ vector still has length n but with repeated coordinates. The possibility of repeated coordinates caused some confusion in the early empirical process literature.

Corresponding to Definition <3> for VC-classes we have an analogous property for ***shatter-dimension*** for subsets of $\mathcal{C}_n$. Instead of subsets of a finite F we now deal with subsets J of the index set $[\![n]\!]$ and the projections they define on $\mathcal{C}_n$ by

$$x[J] := (x_i : i \in J) \in \{0,1\}^J \qquad \text{for each } x := (x_i : i \in [\![n]\!]) \text{ in } \mathcal{C}_n.$$

Remark. Here I am borrowing from the most convenient notation used by the **R** statistical language. Similarly, if M is an $n \times N$ matrix and $I \subset [\![n]\!]$, $J \subset [\![N]\!]$ then $M[I, J]$ denotes the $|I| \times |J|$ submatrix $(M_{i,j} : i \in I, j \in J)$ and $M[-I, J]$ denotes the $(n-|I|)\times|J|$ submatrix $(M_{i,j} : i \notin I, j \in J)$, and so on. If there is no constraint placed on the rows (or columns) just insert a "$\cdot$" symbol. For example, $M[I, \cdot]$ is shorthand for the $|I| \times N$ matrix $M[I, [\![N]\!]]$. The downside of this notation would be the $[[[\ldots]]]$ mess created if I wanted to maintain my former convention that $(1, 2, 3)$ is a row vector of length 3 and $[1, 2, 3]$ is a column vector of length 3. Accordingly, from now on I'll happily write things like $(1, w)$ for a column vector whose first element equals 1 with the remaining coordinates coming from a vector w.

<8> **Definition.** *A subset $\mathcal{X}$ of the discrete hypercube $\mathcal{C}_n := \{0, 1\}^n$ is said to have **shatter-dimension** d (written $\text{SDIM}(\mathcal{X}) = d$) if both the following conditions hold.*

(i) There exists at least one subset I_0, of size d, of $[\![n]\!]$ that is shattered by $\mathcal{X}$, meaning that

$$|I_0| = d \quad \text{AND} \quad |\{x[I_0] : x \in \mathcal{X}\}| = 2^d.$$

□ *(ii) $|\{x[I] : x \in \mathcal{X}\}| < 2^{|I|}$ for every subset I of $[\![n]\!]$ with $|I| > d$.*

Remark. A collection of subsets $\mathcal{D}$ of $\mathbb{A}$ has $\text{VCDIM}(\mathcal{D}) \le d$ iff for all n and all $\alpha := (\alpha_1, \ldots, \alpha_n)$ in $\mathbb{A}^n$, the collection $\mathcal{X}_{\mathcal{D},\alpha}$ of all vectors in $\{0, 1\}^n$ of the form $(\{\alpha_1 \in D\}, \ldots, \{\alpha_n \in D\})$ for $D \in \mathcal{D}$ has $\text{SDIM}(\mathcal{X}_{\mathcal{D},\alpha}) \le d$.

The SDIM point of view will become useful in Chapter 15, when we deal with empirical processes indexed by collections of functions (or sums of more general independent stochastic processes).

The following result, which mirrors the bound from Theorem <4>, will be proved in Section 14.3 by means of a general combinatorial technique known as shifting or, in this case, ***downshifting***.

<9> **Shatter Theorem.** *If $\mathcal{X} \subseteq \mathcal{C}_n := \{0, 1\}^n$ has $\text{SDIM}(\mathcal{X}) = d \ge 1$ then*

□ $$|\mathcal{X}| \le \beta(n, d) := \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d} \le (en/d)^d.$$

Classical results about VC-dimension—as described, for example, by Dudley (2014, Chap 4) or Pollard (1984, §II.4)—usually have simpler analogs for shatter-dimension. The proofs typically require only minor modifications of the proofs for VC-classes. For example, the next result corresponds to Dudley (1978, Theorem 7.2).

<10> **Example.** Suppose $\mathcal{X} \subset \mathcal{C}_n := \{0,1\}^n$ and $\mathcal{L}$ is a k-dimensional subspace of $\mathbb{R}^n$, with $k < n$. Suppose also to each $x = (x_1, \ldots, x_n)$ in $\mathcal{X}$ there exists an $\ell = (\ell_1, \ldots, \ell_n)$ in $\mathcal{L}$ such that $x_i = 1$ iff $\ell_i \geq 0$, for $i \in [\![n]\!]$. Then a simple piece of linear algebra will show that $\text{SDIM}(\mathcal{X}) \leq k$. Indeed, suppose J is a subset of $[\![n]\!]$ of size $k+1$. Then $\mathcal{L}_J := \{\ell[J] : \ell \in \mathcal{L}\}$ is a subspace of dimension at most k within the $(k+1)$-dimensional vector space $\mathbb{R}^J$. There must exist a nonzero vector z in $\mathbb{R}^J$ that is orthogonal to $\mathcal{L}_J$, that is, $\sum_{i \in J} z_i \ell_i = 0$ for all ℓ in $\mathcal{L}$.

Without loss of generality there is at least one i in J at which $z_i < 0$. (You might need to replace z by $-z$.) The equality

$$0 = \sum\nolimits_{i \in J} \Big(\ell_i z_i \{z_i \geq 0\} + (-\ell_i)(-z_i)\{z_i < 0\} \Big)$$

ensures there can be no ℓ in $\mathcal{L}$ for which $\ell_i \geq 0$ iff $z_i \geq 0$, for $i \in J$. From the assumed relationship between $\mathcal{X}$ and $\mathcal{L}$ it then follows that there can be no x in $\mathcal{X}$ for which $x_i = 1$ iff $z_i \geq 0$, for $i \in J$. □

The natural metric for $\mathcal{C}_n$ is the HAMMING distance, defined as

<11> $$\mathfrak{H}(x,y) := \sum\nolimits_{i \in [\![n]\!]} \{x_i \neq y_i\} = \sum\nolimits_i |x_i - y_i| \qquad \text{for } x, y \in \mathcal{C}_n.$$

For empirical process purposes it sometimes help to work with a weighted version of that metric, with the weights coming from some probability measure μ on $[\![n]\!]$:

<12> $$\mathfrak{H}_\mu(x,y) := \sum\nolimits_{i \in [\![n]\!]} \{x_i \neq y_i\}\mu_i = \mu J(x,y) = \sum\nolimits_i |x_i - y_i|\mu_i$$
$$\text{where } J(x,y) := \{i \in [\![n]\!] : x_i \neq y_i\}.$$

The next Example derives a bound on packing numbers due to Dudley (1978, Lemma 7.13). (He worked with VC-classes of sets.) It involves two key ideas that will be applied many times in this and the next Chapter. The second idea is a small inversion trick that will eliminate many messy calculations. For future reference it is recorded here as a Lemma. Problem [5] steps through the simple proof.

<13> **Lemma.** *The function $G(r) := e^r/r$ for $r > 0$ is convex. It achieves its minimum value of e at $r = 1$. If $G(r) \leq w$ then $e \leq w$ and $e^r \leq c_0 w \log w$ where $c_0 := (1 - 1/e)^{-1} \approx 1.58$.*

<14> **Example.** It is a straightforward exercise to extend the argument from Section 14.1 to a general subset $\mathcal{X}$ of $\mathcal{C}_n := \{0,1\}^n$ with $\text{SDIM}(\mathcal{X}) = d$. This time I'll be slightly more careful when controlling a $\log N$ term, in order to obtain Dudley's bound on $\text{PACK}(\epsilon, \mathcal{X}, \mathfrak{H}_\mu)$. For chaining purposes it suffices to have a bound only for small enough ϵ, say $0 < \epsilon \leq 1/2$.

Suppose $\{x_\alpha : \alpha \in [\![N]\!]\}$, for some $N \geq 2$, is an ϵ-separated subset of $\mathcal{X}$, that is, $\mathfrak{H}_\mu(x_\alpha, x_\beta) = \mu J(x_\alpha, x_\beta) > \epsilon$ for $1 \leq \alpha < \beta \leq N$. Consider

the random projection into $\{0,1\}^m$ defined by a vector $\mathcal{I} := (\mathcal{I}_1, \ldots, \mathcal{I}_m)$ of independent observations on μ. For fixed $\alpha < \beta$,

$$\begin{aligned}\mathbb{P}\{x_\alpha[\mathcal{I}] = x_\beta[\mathcal{I}]\} &= \mathbb{P}\{ \text{ no } \mathcal{I}_j\text{'s land in } J(x_\alpha, x_\beta) \ \} \\ &= (1 - \mu J(x_\alpha, x_\beta))^m \le \exp(-m\epsilon)\end{aligned}$$

As before, if $m := \lceil 2(\log N)/\epsilon \rceil$ then

$$\mathbb{P}\{x_\alpha[\mathcal{I}] = x_\beta[\mathcal{I}] \text{ for some } \alpha < \beta \ \} \le \binom{N}{2} \exp(-m\epsilon) < 1,$$

implying existence of a realization of $\mathcal{I}$ for which all the $x_\alpha[\mathcal{I}]$'s are distinct. It follows (via <5>) that $N \le (em/d)^d$. Thus

$$dN^{1/d}/e \le m \le 1 + \frac{2\log N}{\epsilon} = \frac{\epsilon + 2\log N}{\epsilon} \le \frac{3d \log N^{1/d}}{\epsilon},$$

the final inequality coming from the asumptions $\epsilon \le 1/2$ and $N \ge 2$. Equivalently $G(\log N^{1/d}) \le 3e/\epsilon$, which Lemma <13> converts to the assertion that

$$\text{PACK}(\epsilon, \mathcal{X}, \mathfrak{H}_\mu)^{1/d} \le (c_0(3e/\epsilon)\log(3e/\epsilon)) \qquad \text{for } 0 < \epsilon \le 1/2.$$

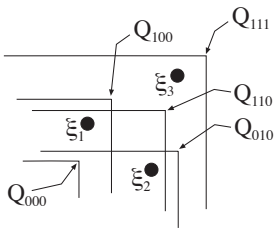
The upper bound comes tantalizingly close to $(C/\epsilon)^{\text{SDIM}(\mathcal{X})}$, which looks a lot like a packing number for a unit ball in an euclidean space of dimension $\text{SDIM}(\mathcal{X})$. As you will see in Section 14.5, such an aesthetically pleasing bound can be established with a lot more work. □

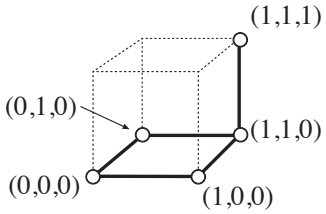
Remark. For each $p \ge 1$, note that $(\sum_i |x_i - y_i|^p \mu_i)^{1/p} = \mathfrak{H}_\mu(x,y)^{1/p}$. Thus the bound derived in the Example also leads to ℓ^p packing numbers for $\mathcal{X}$ that increase like a power of $1/\epsilon$ for smallish ϵ.

A subset $\mathcal{X}$ of $\mathcal{C}_n$ can be hard to visualize for n larger than 2 or 3. However, if $|\mathcal{X}|$ equals N then its elements can be identified with the columns of an $n \times N$ ***binary matrix*** M, a matrix with distinct columns whose elements all belonging to $\{0,1\}$. The arguments in the rest of the Chapter will be much easier to follow when written in matrix form. From now on I'll treat the general $\mathcal{X} \subset \mathcal{C}_n$ and its corresponding $n \times N$ binary matrix M as essentially the same object, writing $\text{SDIM}(M)$ to mean $\text{SDIM}(\mathcal{X})$, and so on.

<15> **Example.** Let $\mathcal{Q}$ be the set of all closed quadrants with a north-east vertex in $\mathbb{R}^2$, that is, sets of the form $\{(u,v) \in \mathbb{R}^2 : u \le a, v \le b\}$ with $a, b \in \mathbb{R}$. You should check that $\text{VCDIM}(\mathcal{Q}) = 2$. (Hint: If a quadrant Q contains the northernmost point and the easternmost point of a finite set F, why must Q contain all of F?) The maximum number of subsets that $\mathcal{Q}$ could pick out from an F of size $|F| = 3$ is seven. (That maximum is

achieved when the three points lie on straight line running from north-west to south-east.)





$$M = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

From the three points ξ_1, ξ_2, ξ_3 of $\mathbb{R}^2$ shown in the picture, $\mathcal{Q}$ picks out only five subsets. The corresponding subset $\mathcal{X}$ consists of five of the eight 'vertices' of the discrete hypercube $\mathcal{C}_3$, which are labelled in the picture to match the columns of the corresponding 3×5 binary matrix M.

For the moment you should ignore the 12 lines (7 dotted and 5 dark) that I added to the picture of $\mathcal{C}_3$ and its subset $\mathcal{X}$. In Section 14.4 the members of $\mathcal{X}$ will become the set of vertices of a graph with edges defined by pairs x, y in $\mathcal{X}$ for which $\mathfrak{H}(x, y) = 1$. (The five dark lines in the middle picture identify the 5 edges of the corresponding graph.) That Section will establish a most surprising relationship between $\text{SDIM}(\mathcal{X})$ and the largest possible number of edges, which will play a key role in Section 14.5. □

Remark. For the $\mathcal{Q}$ in the previous Example there is a polynomial upper bound, of order $O(n^2)$, for the size of $|\mathcal{Q}_F|$ that hold for every F of size n. If we replace $\mathcal{Q}$ by the collection $\mathcal{K}$ of all compact convex subsets of $\mathbb{R}^2$ then the $\text{SDIM}(\cdot)$ situation changes: if F consists of n points spread around a circle then $\mathcal{K}$ picks out all 2^n subsets (consider the convex hull of F_0 for each subset F_0 of F). Thus $\text{VCDIM}(\mathcal{K}) = \infty$. However, if F consists of n points spread along a single line then $\mathcal{K}$ picks out only $\binom{n+1}{2}$ subsets. Less obviously, for observations $\xi_1, \ldots, \xi_n$ from the uniform distribution on the unit square, $\text{SDIM}(\mathcal{X}_\omega)/n$ is small with high probability. See Pollard (1984, p 23).

*14.3 Proof of the Shatter Theorem

The ***downshift method*** offers a particularly clean way to visualize the Proof of Theorem <8> when the subset $\mathcal{X}$ of $\{0,1\}^n$ is represented by the columns of an $n \times N$ binary matrix M_0. (The argument will involve a sequence of binary matrices, $M_0, M_1, \ldots$.)

Remark. Apparently the shifting method has become a standard combinatorial tool (Frankl, 1987).

<16> **Definition.** *Let M be any $n \times N$ binary matrix with distinct columns. A* ***downshift*** *of the ith row is the result of applying the following operation*

to M. For each $j \in [\![N]\!]$, if $M[i,j] = 1$ change it to a 0 unless the resulting column would duplicate an existing column of M. The downshift is said to succeed if it changes the row $M[i,\cdot]$ in at least one position. □

The idea is to cycle through the rows, downshifting as we go, creating new binary matrices $M_1, M_2, \ldots$ until eventually we reach a matrix for which no more successful downshifts (of any row) are possible. As each successful downshift reduces the total number of 1's in the matrix it takes only finitely many steps to reach that goal. The following Example illustrates the procedure for a small M_0.

<17> **Example.** To make it easier to keep track of which downshifts succeed and which are blocked, I'll write beside each M_i the matrix $\overline{M}_i$ consisting of the vectors from $\mathcal{C}_n$ that are missing from M_i. (I found this trick helpful when working through an example on the blackboard.)

Start with an M_0 consisting of 8 of the 16 vectors in $\mathcal{C}_4$.

$$M_0 = \begin{bmatrix} 0&1&0&1&0&1&0&1 \\ 0&0&0&1&0&1&1&1 \\ 0&0&0&0&1&1&1&1 \\ 0&0&1&0&1&0&1&1 \end{bmatrix} \quad \overline{M}_0 = \begin{bmatrix} 0&0&1&1&0&0&1&1 \\ 1&0&0&0&1&1&1&0 \\ 0&1&1&0&1&0&0&1 \\ 0&0&0&1&0&1&1&1 \end{bmatrix}$$

A downshift of row 1 of M_0 changes columns 4 and 6 but column 2 is blocked by column 1 and column 8 is blocked by column 7. The $\boxed{0}$'s in the matrix M_1 indicate which new columns appear and the $\boxed{1}$'s in the matrix $\overline{M}_1$ indicate the columns that disappear from M_0.

$$M_1 = \begin{bmatrix} 0&1&0&\boxed{0}&0&\boxed{0}&0&1 \\ 0&0&0&1&0&1&1&1 \\ 0&0&0&0&1&1&1&1 \\ 0&0&1&0&1&0&1&1 \end{bmatrix} \quad \overline{M}_1 = \begin{bmatrix} \boxed{1}&0&1&1&\boxed{1}&0&1&1 \\ 1&0&0&0&1&1&1&0 \\ 0&1&1&0&1&0&0&1 \\ 0&0&0&1&0&1&1&1 \end{bmatrix}$$

Effectively, the downshifting has swapped two columns from M_0 with two columns from $\overline{M}_0$.

A downshift of the second row of M_1 then swaps two columns between M_1 and $\overline{M}_1$. This time column 1 blocks a downshift of column 4 and column 5 blocks a downshift of column 7:

$$M_2 = \begin{bmatrix} 0&1&0&0&0&0&0&1 \\ 0&0&0&1&0&\boxed{0}&1&\boxed{0} \\ 0&0&0&0&1&1&1&1 \\ 0&0&1&0&1&0&1&1 \end{bmatrix} \quad \overline{M}_2 = \begin{bmatrix} 1&0&1&1&1&0&1&1 \\ 1&\boxed{1}&0&0&1&1&1&\boxed{1} \\ 0&1&1&0&1&0&0&1 \\ 0&0&0&1&0&1&1&1 \end{bmatrix}$$

Then downshift the third row of M_2:

$$M_3 = \begin{bmatrix} 0&1&0&0&0&0&0&1 \\ 0&0&0&1&0&0&1&0 \\ 0&0&0&0&1&1&\boxed{0}&\boxed{0} \\ 0&0&1&0&1&0&1&1 \end{bmatrix} \quad \overline{M}_3 = \begin{bmatrix} 1&0&1&1&1&0&1&1 \\ 1&1&0&0&1&1&1&1 \\ 0&1&1&\boxed{1}&1&\boxed{1}&0&1 \\ 0&0&0&1&0&1&1&1 \end{bmatrix}$$

Every potential downshift of the fourth row of M_3 is blocked. No more downshifts—for any row of M_3—can succeed.

For future reference, note that no column of M_3 contains more than two 1's. □

Now back to the proof of Theorem <9>. First some useful terminology. If M is an $n \times N$ binary matrix with distinct columns and $I \subset [\![n]\!]$, say that M shatters I if $M[I, \cdot]$ contains all members of $\{0,1\}^I$ amongst its columns. Note that if $\text{SDIM}(M) = k$ then M cannot shatter any I with $|I| > k$.

In general, the downshifting operation has two important properties:

(i) No new shattered sets of rows can be created by a downshift.

Proof. For notational convenience, suppose that M^* is created from M by a downshift of the first row and that M^* shatters some I. We may assume $1 \in I$ because if $1 \notin I$ we have $M[I, \cdot] = M^*[I, \cdot]$, ensuring that M also shatters I. Again for notational convenience, we may suppose that $I = [\![k]\!]$ for some k.

The assumption about M^* tells us that for each u in $\{0,1\}^{I\setminus\{1\}}$ there must exist j_0, j_1 in $[\![N]\!]$ for which $M^*[\,\cdot\,, j_1] := (1, u, v_1)$ and $M^*[\,\cdot\,, j_0] := (0, u, v_0)$ for some v_0, v_1. Downshifting creates no new 1's, so $M[\,\cdot\,, j_1] = M^*[\,\cdot\,, j_1]$. And for column j_1 to have survived the downshift the change must have been blocked by the presence of $(0, u, v_1)$ amongst the columns of M. Thus M also shatters I. □

(ii) If no downshifts can succeed for a binary matrix M then it is ***hereditary***. That is, if x is a column of M and if y is obtained from x by changing some of its 1's to 0's then y is also a column of M. Consequently, if $x[i] = 1$ for each i in I, for some $I \subset [\![n]\!]$, then M shatters I.

Proof. Suppose $M[i, j_0] = 1$. The downshift on row i must be blocked from changing $M[i, j_0]$ by another column for which $M[-i, j_1] = M[-i, j_0]$ and $M[i, j_1] = 0$. Then argue that each possible change to $M[,\, \cdot\,, j_1]$ must be blocked. And so on. And so on, □

With these two facts in hand the rest of the argument is straightforward. Suppose we start with M_0 and after a R downshifts end up with an M_R (with all columns distinct) for which no more downshifts can succeed. Suppose for some j the column $\{i \in [\![n]\!] : M_R[i, j] = 1\} = I$. By property (ii), the submatrix $M[I, \cdot]$ must contain all the vectors in $\{0,1\}^I$, that is, M shatters I. And $\text{SDIM}(M_R) \le \text{SDIM}(M_0) = d$ by property (i). Taken together, these facts tell us that $|\{i \in [\![n]\!] : M_R[i, j] = 1\}| \le d$ for each J. All the columns of M_R must belong to $\{x \in \mathcal{C}_n : \sum_i x_i \le d\}$, a set of size $\beta(n, d)$.

*14.4 Subsets of a discrete hypercube as graphs

The results in this Section provide the first steps towards a surprising and elegant result: if $\mathcal{X} \subset \mathcal{C}_n := \{0,1\}^n$ and μ is a probability measure on $[\![n]\!]$ then $\text{PACK}(\epsilon, \mathcal{X}, \mathfrak{H}_\mu)$ increases like $(1/\epsilon)^{\text{SDIM}(\mathcal{X})}$ as $\epsilon \searrow 0$. This slight improvement on Example <14> comes via the trick of regarding $\mathcal{X}$ as the vertex set of a graph with edge set $\mathcal{E}$ (or $\mathcal{E}(\mathcal{X})$ if we wish to distinguish between various graphs) defined as all the pairs $\{x, y\}$ from $\mathcal{X}$ for which

<18> $$\mathfrak{H}(x,y) := \sum\nolimits_{i \in [\![n]\!]} \{x_i \neq y_i\} = 1.$$

For such an edge, write $x \overset{i}{\sim} y$ if $x[i] \neq y[i]$.

There is a most surprising general relationship between the shatter dimension of $\mathcal{X}$ and the number of edges in the corresponding graph.

<19> **Theorem.** *(Haussler, 1995, Lemma 2) If* $\text{SDIM}(\mathcal{X}) \le d$ *then* $|\mathcal{E}(\mathcal{X})| \le d|\mathcal{X}|$.

The proof will make use of the following helpful fact about the effect of a single downshift operation on the set of edges.

<20> **Lemma.** *If* $(\mathcal{X}^*, \mathcal{E}^*)$ *is obtained from* $(\mathcal{X}, \mathcal{E})$ *by means of a single downshift then* $|\mathcal{E}| \le |\mathcal{E}^*|$.

Proof. As before, the members of $\mathcal{X}$ correspond to columns of some binary matrix M. Without loss of generality suppose σ acts on the first row of M. The proof will work by constructing a one-to-one map (also called an injection) $S : \mathcal{E} \backslash \mathcal{E}^* \to \mathcal{E}^* \backslash \mathcal{E}$ to show that $|\mathcal{E} \backslash \mathcal{E}^*| \le |\mathcal{E}^* \backslash \mathcal{E}|$, which will be mostly an exercise in careful bookkeeping.

Suppose $\mathfrak{e} = \{x, y\} \in \mathcal{E} \backslash \mathcal{E}^*$. Then $x \overset{j}{\sim} y$ for some j in $[\![n]\!]$. Without loss of generality suppose $x[j] = 1$ and $y[j] = 0$ (and $x[-j] = y[-j]$ to ensure $\mathfrak{H}(x,y) = 1$). If j were equal to 1 the vector y would block σ from changing x, thereby implying that $\{x, y\} \in \mathcal{E}^*$, contradicting the assumption that $\mathfrak{e} \notin \mathcal{E}^*$. Thus $j \neq 1$.

To simplify notation we may assume $j = 2$, so that $x = (\alpha, 1, u)$ and $y = (\alpha, 0, u)$ for some α in $\{0,1\}$ and u in $\{0,1\}^{n-2}$. If α were equal to 0 then neither x nor y would be affected by σ, again violating the assumption that $\{x, y\} \notin \mathcal{E}^*$. The proof will involve only four members of $\mathcal{C}_n$:

$$\begin{array}{c} x \quad y \quad w \quad z \\ \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ u & u & u & u \end{pmatrix} \end{array}$$

Note that if $w \in \mathcal{X}$ then $\sigma x = x$ and if $w \notin \mathcal{X}$ then $\sigma x = w$. Similarly, if $z \in \mathcal{X}$ then $\sigma y = y$ and if $z \notin \mathcal{X}$ then $\sigma y = z$.

We cannot have $\{w, z\} \subset \mathcal{X}$, for that would force $x = \sigma x \in \mathcal{X} \cap \mathcal{X}^*$ and $y = \sigma y \in \mathcal{X} \cap \mathcal{X}^*$, contradicting the assumption $\mathfrak{e} \notin \mathcal{E}^*$. Thus we have only three cases to consider, in each of which $S(\mathfrak{e}) := \mathfrak{e}^* := \{w, z\}$:

(i) $\sigma x = w$ and $\sigma y = z$, implying $w, z \in \mathcal{X}^* \backslash \mathcal{X}$, so that $\{w, z\} \in \mathcal{E}^* \backslash \mathcal{E}$.

(ii) $\sigma x = w$ and $\sigma y = y$, implying $w \in \mathcal{X}^* \backslash \mathcal{X}$ and $z \in \mathcal{X}^* \cap \mathcal{X}$. Again $\{w, z\} \in \mathcal{E}^* \backslash \mathcal{E}$.

(iii) $\sigma x = x$ and $\sigma y = z$, implying $w \in \mathcal{X} \cap \mathcal{X}^*$ and $z \in \mathcal{X}^* \backslash \mathcal{X}$. And yet again $\{w, z\} \in \mathcal{E}^* \backslash \mathcal{E}$.

The function S is one-to-one because $\mathfrak{e}$ can be uniquely recovered from $\mathfrak{e}^*$ by changing the initial 0's in v and w back to 1's. □

Proof (of Theorem <19>). Let M and $\mathcal{E}$ be the $n \times N$ binary matrix and the edge set corresponding to $\mathcal{X}$. As in the Proof of Theorem <9>, apply a sequence of downshifts until arriving at a hereditary binary matrix $M^\dagger$, with column set $\mathcal{X}^\dagger$ and edge set $\mathcal{E}^\dagger$, for which no more downshifts can succeed. By repeated appeals to Lemma <20> we know that $|\mathcal{E}| \le |\mathcal{E}^\dagger|$ and, from Section 14.3, that $\text{SDIM}(M^\dagger) \le \text{SDIM}(M) \le d$. As $M^\dagger$ is hereditary, it follows that the number of 1's in each column of $M^\dagger$ must be $\le d$.

Consider an edge $\{x, y\}$ from $\mathcal{E}^\dagger$. There must exist some $i \in [[n]]$ for which $x \overset{i}{\sim} y$. Without loss of generality, we may suppose $x[i] = 1$ and $y[i] = 0$. That is, the edge $\mathfrak{e}$ is obtaining by connecting x to some y defined by changing one member of $\{j : x[j] = 1\}$ to a 0. There are at most d ways to choose such a y. Put another way, every edge in $\mathcal{E}^\dagger$ appears on the list of pairs obtained by working through the set of N members of $\mathcal{X}^\dagger$ and pairing each with one of at most d other vertices, which gives $|\mathcal{E}^\dagger| \le Nd$. □

The final paragraph of the previous Proof effectively bounded the size of $\mathcal{E}^\dagger$ by assigning an orientation to each edge: if $\mathfrak{e} = \{x, y\}$ and $\sum_i x[i] = 1 + \sum_i y[i]$ we could define $\text{HEAD}(\mathfrak{e}) := x$ and $\text{TAIL}(\mathfrak{e}) = y$, giving the edge an orientation from tail to the head. The bound $\sum_i x_i \le \text{SDIM}(\mathcal{X}^\dagger)$ then gives

$$\text{indegree}(x) := |\{\mathfrak{e} \in \mathbb{E}^\dagger : \text{HEAD}(\mathfrak{e}) = x\}| \le \text{SDIM}(\mathcal{X}).$$

The proof in the next Section requires a similar ordering for the original graph $(\mathcal{X}, \mathcal{E})$. The existence of such an ordering can be deduced from Theorem <19> via the classical Marriage Lemma (UGMTP Problem 10.5):

<*21*> **Marriage Lemma.** *Suppose for each s in a finite set S we are given a nonempty subset K_s of some finite set $\mathcal{K}$ such that*

$$|\cup_{s \in A} K_s| \ge |A| \qquad \text{for each nonempty subset } A \text{ of } S.$$

Then there exists a one-to-one function ψ mapping S into $\mathcal{K}$ with $\psi(s) \in K_s$ for each $s \in S$.

<22> **Corollary.** *(Based on Haussler (1995, Lemma 3), using Alon and Tarsi (1992, Lemma 3.1) to orient edges.) If* $\text{SDIM}(\mathcal{X}) \le d$. *then there exists an orientation* $\text{TAIL}(\mathfrak{e}) \to \text{HEAD}(\mathfrak{e})$ *of each* $\mathfrak{e}$ *in* $\mathcal{E}(\mathcal{X})$ *for which*

$$|\sum\nolimits_{\mathfrak{e}\in\mathcal{E}(\mathcal{X})}\{\text{HEAD}(\mathfrak{e}) = x\}| \le d \qquad \text{for each } x \text{ in } \mathcal{X}.$$

Proof. Apply the Marriage Lemma with $\mathcal{E}$ playing the role of the set S. We want a map HEAD from $\mathcal{E}$ into the set $\mathcal{X}$ that allows as many as d edges to map to the same vertex. As that Lemma produces a one-to-one map we need $\mathcal{K}$ to contain d copies of each x in $\mathcal{X}$. That can be arranged by defining

$$\mathcal{K} := \mathcal{X} \times [[d]] := \{(x, j) : x \in \mathcal{X}, j \in [[d]]\}.$$

If $\mathfrak{e} = \{x, y\} \in \mathcal{E}(\mathcal{X})$ define

$$K(\mathfrak{e}) := \{(x, j) : j \in [[d]]\} \cup \{(y, j) : j \in [[d]]\}.$$

If $\emptyset \ne A \subseteq \mathcal{E}(\mathcal{X})$ define $\mathcal{X}_0 := \cup A$, the set of all x in $\mathcal{X}$ that appear as a vertex of at least one edge in A, and

$$\mathcal{E}_0 := \mathcal{E}(\mathcal{X}_0) := \{\{x, y\} : x, y \in \mathcal{X}_0 \text{ and } \mathfrak{H}(x, y) = 1\}.$$

Then we have $A \subseteq \mathcal{E}_0$ and $\text{SDIM}(\mathcal{X}_0) \le \text{SDIM}(\mathcal{X}) \le d$, so that $|\mathcal{E}_0| \le d|\mathcal{X}_0|$ by Theorem <19>, and $\cup_{\mathfrak{e}\in A} K(\mathfrak{e}) = \mathcal{X}_0 \times [[d]]$. It follows that

$$|A| \le |\mathcal{E}_0| \le d|\mathcal{X}_0| = |\cup_{\mathfrak{e}\in A} K(\mathfrak{e})|,$$

as required by Lemma <21> to provide the one-to-one map $\psi : \mathcal{E} \to \mathcal{K}$. If $\mathfrak{e}_0 = \{x_0, y_0\} \in \mathcal{E}$ then $\psi(\mathfrak{e})$ equals either (x_0, j) or (y_0, j) for some j in $[[d]]$. Strip off the j to define $\text{HEAD}(\mathfrak{e})$ then take $\text{TAIL}(\mathfrak{e})$ as the other vertex of $\mathfrak{e}$.

Finally, argue for any given x in $\mathcal{X}$ that an edge $\mathfrak{e}$ has $\text{HEAD}(\mathfrak{e}) = x$ iff $\mathfrak{e} = \{x, y\}$ for some $y \in \mathcal{X}$ with $\mathfrak{H}(x, y) = 1$ and $\psi(\mathfrak{e}) = (x, j)$ for some j in $[[d]]$. For each such j there can be at most one edge $\mathfrak{e}$ for which $\psi(\mathfrak{e}) = (x, j)$. Thus $\sum_{\mathfrak{e}\in\mathcal{E}}\{\text{HEAD}(\mathfrak{e}) = x\} \le d$, as asserted. □

Remark. We could also reinterpret HEAD and TAIL from Corollary <22> as functions from $\mathcal{E}$ into $[[N]]$ for which

$$|\sum\nolimits_{\mathfrak{e}\in\mathcal{E}}\{\text{HEAD}(\mathfrak{e}) = j\}| \le \text{SDIM}(M) \qquad \text{for every } j \text{ in } [[N]].$$

*14.5 Haussler's improvement of the packing bound

For the case of a weighted HAMMING distance with uniform weights Haussler (1995) improved the bound on packing numbers from Example <14> by using a more subtle randomization argument to remove the $\log(3e/\epsilon)$ factor. The rearrangement of Haussler's argument presented in this Section resulted from discussions I had in 1994 with Aad van der Vaart and Jon Wellner. For their version see van der Vaart and Wellner (1996, §2.6).

<23> **Theorem.** *Let $\mathcal{X}$ be a subset of $\mathcal{C}_n := \{0,1\}^n$ with $\text{SDIM}(\mathcal{X}) \le d$ and let μ be a probability measure on $[\![n]\!]$. Then*

$$\text{PACK}(\epsilon, \mathcal{X}, \mathfrak{H}_\mu) \le \mathcal{H}_d(\epsilon) := Cd(2e/\epsilon)^d \qquad \text{for } 0 < \epsilon < 1.$$

□ *for a universal constant C.*

Actually it suffices to prove the Theorem for the special case where μ is the uniform distribution: as shown by Problem [6], simple approximation arguments then take care of the general case.

As Haussler's argument is quite subtle, let me sketch the main idea before launching into the Proof of the Theorem. For a fixed ϵ in $(0,1)$ let M be an $n \times N$ binary matrix whose columns form an ϵ-packing of $\mathcal{X}$ under the $\mathfrak{H}_\mu$ metric (for μ uniform on $[\![n]\!]$):

<24> $$n^{-1} \sum\nolimits_{i \in [\![n]\!]} \{M[i,j] \ne M[i,k]\} > \epsilon \qquad \text{for all } j \ne k.$$

We need to show that $N \le Cd(2e/\epsilon)^d$.

As each $M[i,j]$ takes values in $\{0,1\}$ we could also write the indicator $\{M[i,j] \ne M[i,k]\}$ as $|M[i,j] - M[i,k]|$ or as $|M[i,j] - M[i,k]|^2$. For some parts of the argument the quadratic expression will be the most useful.

The central idea involves a small integer m (whose value will be chosen during the Proof of the Theorem) and the strange quantity

<25> $$\mathfrak{D}_M := \sum\nolimits_{|K|=m} \sum\nolimits_{i \in [\![n]\!]} \{i \notin K\} \mathbb{P}\text{var}\Big(M[i,\mathcal{J}] \mid M[K,\mathcal{J}]\Big)$$

<26> $$= \sum\nolimits_{|I|=m+1} \sum\nolimits_{i \in [\![n]\!]} \{i \in I\} \mathbb{P}\text{var}\Big(M[i,\mathcal{J}] \mid M[I\backslash\{i\},\mathcal{J}]\Big),$$

where $\mathcal{J}$ is distributed uniformly on $[\![N]\!]$. The summation $\sum_{|K|=m}$ means that K ranges over all $\binom{n}{m}$ subsets of $[\![n]\!]$ of size m. Similarly, I ranges over all $\binom{n}{m+1}$ subsets of $[\![n]\!]$ of size $m+1$. Of course K corresponds to $I\backslash\{i\}$ and I corresponds to $K \cup \{i\}$.

The key to the Proof will be an inequality

<27> $$\binom{n}{m+1} d \ge \mathfrak{D}_M \ge \binom{n}{m} \tfrac{1}{2} n\epsilon \left(1 - \beta(m,d)/N\right),$$

with the lower bound coming from

<28> $$\sum\nolimits_{i\notin K} \mathbb{P}\mathrm{var}\Big(M[i,\mathcal{J}] \mid M[K,\mathcal{J}]\Big) \geq \tfrac{1}{2}n\epsilon\,\big(1-\beta(m,d)/N\big) \qquad \text{for each } K$$

and the upper bound coming from

<29> $$\sum\nolimits_{i\in I} \mathbb{P}\mathrm{var}\Big(M[i,\mathcal{J}] \mid M[I\backslash\{i\},\mathcal{J}]\Big) \leq d \qquad \text{for each } I.$$

Once we have these inequalities the rest will be easy.

Proof (of Theorem <23>). The argument consists of three steps: prove the lower bound <28>; prove the upper bound <29>; choose an appropriate value for m then invert an inequality to bound N.

Step 1: lower bound for fixed K

Define $\mathcal{Z} := \{M[K,j] : j \in [\![N]\!]\}$, a subset of $\{0,1\}^K$ with $\textsc{sdim}(\mathcal{Z}) \leq d$, and $J(z) := \{j \in [\![N]\!] : M[K,j] = z\}$ for each z in $\mathcal{Z}$. We can regard $Z := M[K,\mathcal{J}]$ as a $\mathcal{Z}$-valued random vector with $\mathbb{P}\{Z=z\} = \theta_z := |J(z)|/N$.

In order to exploit the ϵ-separation property <24> of M we will need a second random variable, $\widetilde{\mathcal{J}}$, with the same distribution as $\mathcal{J}$ and coupled with $\mathcal{J}$ so that $\mathbb{P}\{\widetilde{\mathcal{J}} = k \mid \mathcal{J} \in J(z)\} = 1/J(z)$ for each k in $J(z)$. This can be achieved by generating $(\mathcal{J},\widetilde{\mathcal{J}})$, for each z in $\mathcal{Z}$, as pair of independent observations each uniformly distributed on $J(z)$.

Remark. More formally (and less intuitively), we could define a probability measure $\mathbb{P}$ on $[\![N]\!]\times[\![N]\!]$ by

$$\mathbb{P}\{j,k\} := \sum\nolimits_{z\in\mathcal{Z}} \{j,k\in J(z)\}\frac{|J(z)|}{N}\frac{1}{|J(z)|}\frac{1}{|J(z)|}$$

then take $\mathcal{J}$ and $\widetilde{\mathcal{J}}$ as the coordinate maps on $[\![N]\!]\times[\![N]\!]$.

The random vectors $X := M[\,\cdot\,,\mathcal{J}]$ and $\widetilde{X} := M[\,\cdot\,,\widetilde{\mathcal{J}}]$ both take values in $\{0,1\}^n$ and $X[K] = \widetilde{X}[K] = z$ when $Z = z$. They are coupled: not independent but conditionally independent given Z. (Compare with the Poisson-Binomial Example in §13.1.) For each $i \in [\![n]\!]\backslash K$,

$$\begin{aligned}
&\mathrm{var}\Big(X[i] \mid X[K]=z\Big)\\
&\quad= \mathbb{P}\Big(X[i]^2 \mid Z=z\Big) - \mathbb{P}\Big(X[i] \mid Z=z\Big)\times\mathbb{P}\Big(\widetilde{X}[i] \mid Z=z\Big)\\
&\quad= \tfrac{1}{2}\mathbb{P}\Big(\big(X[i]-\widetilde{X}[i]\big)^2 \mid Z=z\Big) \qquad \text{by conditional independence.}
\end{aligned}$$

Thus the expression on the left-hand side of <28> equals

$$\begin{aligned}
&\sum_{i\notin K}\sum_{z\in\mathcal{Z}}\theta_z \mathrm{var}\Big(X[i] \mid X[K]=z\Big)\\
&\quad=\tfrac{1}{2}\sum_{z\in\mathcal{Z}}\theta_z\mathbb{P}\Big(\sum_{i\notin K}(X[i]-\widetilde{X}[i])^2 \mid Z=z\Big)\\
&\quad=\tfrac{1}{2}\sum_{z\in\mathcal{Z}}\theta_z\mathbb{P}\Big(n\mathfrak{H}_\mu(X,\widetilde{X}) \mid Z=z\Big)\\
&\quad\ge\tfrac{1}{2}\sum_{z\in\mathcal{Z}}n\theta_z\mathbb{P}\Big(\{\mathcal{J}\neq\widetilde{\mathcal{J}}\}\epsilon \mid Z=z\Big) \qquad \text{because } M \text{ is } \epsilon\text{-separated}\\
&\quad=\tfrac{1}{2}n\epsilon\sum_{z\in\mathcal{Z}}(|J(z)|/N)(1-1/|J(z)|)=\tfrac{1}{2}n\epsilon\Big(1-|\mathcal{Z}|/N).
\end{aligned}$$

The second equality comes from the fact that $X[K]=\widetilde{X}[K]=z$ when $Z=z$. As $\mathcal{Z}\subset\{x[K]:x\in\mathcal{X}\}$ and $\textsc{sdim}(\mathcal{X})\le d$, we have $|\mathcal{Z}|\le\beta(m,d)$ by the Shatter Theorem.

Step 2: upper bound for fixed I

The subset $\mathcal{Y}:=\{M[I,j]:j\in[\![N]\!]\}$ of $\{0,1\}^I$ has $\textsc{sdim}(\mathcal{Y})\le d$. We can regard $Y:=M[I,\mathcal{J}]$ as a random element of $\mathcal{Y}$ with distribution $P\{y\}:=\mathbb{P}\{Y=y\}$. (The precise form of P will not matter.) For each i in I, the random vector $Y[-i]$ takes values in $\mathcal{Z}_{-i}:=\{y[-i]:y\in\mathcal{Y}\}$, a subset of $\{0,1\}^{I\setminus\{i\}}$. The left-hand side of <29> equals (in **R** notation)

$$\sum_{i\in I}\sum_{z\in\mathcal{Z}_{-i}}p_{-i}(z)\mathrm{var}\Big(Y[i]\mid Y[-i]=z\Big) \qquad \text{where } p_{-i}(z):=\mathbb{P}\{Y[-i]=z\}.$$

As in Section 14.4, the set $\mathcal{Y}$ can be thought of as the vertices of a graph with edge set $\mathcal{E}$ defined by pairs $\mathfrak{e}=\{y,y'\}$ for which $\mathfrak{H}(y,y')=1$. In particular, $\mathfrak{e}\in\mathcal{E}_i$ when $y[i]\neq y'[i]$. By Corollary <22> the edges can be oriented by maps TAIL and HEAD such that

$$\Big|\sum_{\mathfrak{e}\in\mathcal{E}(\mathcal{X})}\{\textsc{head}(\mathfrak{e})=y\}\Big|\le d \qquad \text{for all } y \text{ in } \mathcal{Y}.$$

If $z\in\mathcal{Z}_{-i}$ then the set $\mathcal{Y}_{-i}(z):=\{y\in\mathcal{Y}:y[-i]=z\}$ contains at most two members. There are ony two cases to consider:

- If $|\mathcal{Y}_i(z)|=1$ the conditional distribution of $Y[i]$ given $Y[-i]=z$ is degenerate, so $\mathrm{var}(Y[i]\mid Y[-i]=z)=0$.
- If $|\mathcal{Y}_i(z)|=2$, say $\mathcal{Y}_i(z)=\mathfrak{e}:=\{y_0,y_1\}$ with $y_0[-i]=y_1[-i]=z$ and $y_0[i]=0$ and $y_1[i]=1$, then $\mathfrak{e}\in\mathcal{E}_i$ with $P\mathfrak{e}=P\{y_0\}+P\{y_1\}$ and

  $$\mathbb{P}\{Y[i]=1\mid Y[-i]=z\}=P\{y_1\}/P\mathfrak{e}.$$

  That is, $Y[i]\mid Y[-i]=z\sim\textsc{Ber}(p_{-i}(z))$, with $\mathrm{var}(Y[i]\mid Y[-i]=z)$ equal to

  $$\frac{P\{y_1\}}{P\mathfrak{e}}\times\frac{P\{y_0\}}{P\mathfrak{e}}\le\min\left(\frac{P\{y_1\}}{P\mathfrak{e}},\frac{P\{y_0\}}{P\mathfrak{e}}\right)\le\frac{P\{\textsc{head}(\mathfrak{e})\}}{P\mathfrak{e}}.$$

In my opinion, the use of HEAD($\mathfrak{e}$) in this way is the cleverest part of Haussler's proof. The expression on the left-hand side of <29> equals

$$\begin{aligned}&\sum_{i\in I}\sum_{z\in\mathcal{Z}_{-i}}\mathbb{P}\{Y[-i]=z\}\operatorname{var}\Big(Y[i]\mid Y[-i]=z\Big)\\&\quad\le\sum_{i\in I}\sum_{\mathfrak{e}\in\mathcal{E}_i}(P\mathfrak{e})\times\big(P\{\text{HEAD}(\mathfrak{e})\}/P\mathfrak{e}\big)\\&\quad=\sum_{\mathfrak{e}\in\mathcal{E}}\mathbb{P}\{Y=\text{HEAD}(\mathfrak{e})\}=\mathbb{P}\sum_{\mathfrak{e}\in\mathcal{E}}\{Y=\text{HEAD}(\mathfrak{e})\}.\end{aligned}$$

The final sum is at most d because $\sum_{\mathfrak{e}\in\mathcal{E}}\{Y=\text{HEAD}(\mathfrak{e})\}\le d$ for every possible realization of Y.

Step 3: choose m

We now know that inequality <27> holds, so that

$$\binom{n}{m}\tfrac{1}{2}n\epsilon\left(1-\beta(m,d)/N\right)\le\binom{n}{m+1}d$$

which rearranges to

$$N\le\beta(m,d)\left(1-\frac{2(n-m)d}{n\epsilon(m+1)}\right)^{-1}\le(em/d)^d\left(1-\frac{2(n-m)d}{n\epsilon(m+1)}\right)^{-1}.$$

We could try to optimize over m immediately, as Haussler (1995, page 225) did, to bound N by a function of ϵ, d, and n, then take a supremum over n. However, it is simpler to note that all the conditions of the Theorem apply if n is replaced by qn, for some positive integer q, and each x in $\mathcal{X}$ is replaced by the concatenation of q copies of x. (The new vectors are the columns of the $(qn)\times N$ binary matrix obtained by stacking q copies of M.) Letting q tend to infinity (with m fixed) eliminates n from the bound, leaving

$$N\le(em/d)^d\left(1-\frac{2d}{\epsilon(m+1)}\right)^{-1}.$$

As a function of m, the first factor is increasing and the second is decreasing. If we treated m as a continuous variable and ignored the difference between m and $m+1$ the minimizing value would be $2(d+1)/\epsilon$. That suggests we choose $m=\lceil 2(d+1)/\epsilon\rceil$, which gives the bound

$$N\le\left(\frac{2e(d+1)}{\epsilon d}\right)^d\frac{2(d+1)}{2(d+1)-2d}\le(2e/\epsilon)^d(1+d^{-1})^{d+1}.$$

□

Remark. For comparison's sake, note that Haussler got

$$N \le e(d+1)\left(2e(n+1)/(n\epsilon+2d+2)\right)^d.$$

14.6 Problems

[1] With points $x_1, x_2, \ldots, x_{n+2}$ in $\mathbb{R}^n$ define $y_i := x_i - x_{n+2}$ for $1 \le i \le n+1$.

(i) Use a linear dependence argument to show that there exist constants $\{\alpha_i\}$, not all zero, for which $\sum_{i=1}^{n+1} \alpha_i y_i = 0$.

(ii) Define $\beta_i := \alpha_i$ for $1 \le i \le n+1$ and $\beta_{n+2} := -\sum_{i\in[\![n+1]\!]} \alpha_i$. Show that $\sum_{i\in[\![n+2]\!]} \beta_i = 0$ and $\sum_{i\in[\![n+2]\!]} \beta_i x_i = 0$.

(iii) Define $A := \{i \in [\![n+2]\!] : \beta_i > 0\}$ and $B := \{i \in [\![n+2]\!] : \beta_i < 0\}$. Show that $c := \sum_{i\in A} \beta_i = -\sum_{i\in B} \beta_i > 0$ and $w := \sum_{i\in A}(\beta_i/c)x_i = \sum_{i\in B}(-\beta_i/c)x_i$. Deduce that $w \in \operatorname{co}(\{) x_i : i \in A\} \cap \operatorname{co}(\{) x_i : i \in B\} \ne \emptyset$.

[2] Let $\mathcal{H}_d$ denote the set of all closed half-spaces in $\mathbb{R}^d$. Show that $\text{VCDIM}(\mathcal{H}_d) = d+1$ by the following arguments.

(i) Show that $\mathcal{H}_d$ shatters the set $F = \{0, e_1, \ldots, e_d\}$, where e_i is the unit vector with $+1$ in the ith position. Hint: Consider closed halfspaces of the form $\{x : \theta \cdot x \ge r\}$ with $\theta \in \{-1, +1\}^d$.

(ii) Suppose $F = \{x_i : i = 0, 1, \ldots, d+1\}$ is a set of $d+2$ distinct points in $\mathbb{R}^d$. If H picks out a subset $F_K := \{x_i : i \in K\}$ from F, show that that the convex hull of F_K is a subset of H, which is disjoint from the convex hull of F_{K^c}.

(iii) Show that $\sum_{1\le i\le d+1} \alpha_i(x_i - x_0) = 0$ for some constants α_i that are not all zero. Define $\alpha_0 = -\sum_{1\le i\le d+1} \alpha_i$. Show that

$$\sum\nolimits_{0\le i\le d+1} \alpha_i x_i = 0 \qquad \text{AND} \qquad \sum\nolimits_{0\le i\le d+1} \alpha_i = 0.$$

Define $J = \{i \in [0 : d+1] : \alpha_i > 0\}$. Show that the convex hulls of the subsets $F_J = \{x_i : i \in J\}$ and F_J^c have a nonempty intersection. Thus $\mathcal{H}_d$ cannot pick out the subset F_J from F.

[3] Suppose $\mathcal{D}$, $\mathcal{D}_1, \ldots, \mathcal{D}_k$ are vc-classes of subsets of $\mathcal{X}$.

(i) Show that each of the following is also a VC-class: $\{D^c : D \in \mathcal{D}\}$ and $\mathcal{D}_1 \cup \mathcal{D}_2$ and $\{D_1 \cup D_2 : D_i \in \mathcal{D}_i\}$ and $\{D_1 \cap D_2 : D_i \in \mathcal{D}_i\}$. Hint: Use the polynomial bounds on the numbers of subsets that can be picked out.

(ii) (Dudley, 1978, Proposition 7.12) Write $\mathcal{A}(B_1, \ldots, B_k)$ for the algebra of subsets of $\mathcal{X}$ generated by subsets $B_1, \ldots, B_k$ of $\mathcal{X}$. For each finite k, show that $\cup\{\mathcal{A}(D_1, \ldots, D_k\} : D_1, \ldots, D_k \in \mathcal{D}\}$ is also a VC-class.

(iii) Suppose $\text{VCDIM}(\mathcal{D}_i) \le d$ for each i. Let $\mathcal{B}_k$ denote the class of all sets expressible by means of at most k *union, intersection, or complement* symbols. Find an increasing, integer-valued function $\beta(k)$ such that the VC-dimension of $\mathcal{B}_k$ is at most $\beta(k)d$.

Remark. As the Remarks following Definition <8> noted, the facts in the previous problem also apply to the shatter dimension of subsets of $\{0,1\}^n$, because shatter dimension is a special case of VC-dimension.

[4] Show that the function $h(r) := (en/r)^r$ is nondecreasing on the range $1 \le r \le n$. Hint: Show that $\log h(r)$ has derivative $\log(n/r)$.

[5] Lemma <13> stated an inversion inequality for the function G defined on $(0,\infty)$ by $G(r) = e^r/r$: if $G(r) \le w$ then $e \le w$ and $e^r \le c_0 w \log w$ where $c_0 := (1-e^{-1})^{-1} \approx 1.58$. Follow these steps to establish that inequality.

(i) Show that G is convex, with $\min_{r>0} G(r) = G(1) = e$.

(ii) If $w \ge G(r)$ then $w \ge e$. If $w = e$ then r must equal 1 and $e^r = e = w \log w$.

(iii) If $w > e$ then there is a unique R in $(1,\infty)$ for which $G(R) = w$ and $\{r > 0 : G(r) \le w\} \subset (0, R]$.

(iv) Suppose $G(r) \le w = G(R)$ with $R > 1$. Define $S := \log(R)$. Then

$$\log w = R - \log(R) = e^S - S = e^S\,(1 - 1/G(S)) \ge R\,(1 - 1/e)\,.$$

Consequently $e^r \le e^R = RG(R) \le c_0 w \log w$.

[6] Suppose we know that Theorem <23> holds when μ is uniform on $[\![n]\!]$ Follow these steps to extend to the case where μ is any probability measure $[\![n]\!]$.

(i) Suppose the columns of an $n \times N$ binary matrix M, with $\text{SDIM}(M) \le d$, all belong to $\mathcal{X}$ and they are ϵ-separated for $\mathfrak{H}_\mu$:

$$\min_{j\ne k} \sum\nolimits_{i\in[\![n]\!]} \mu_i\{M[i,j] \ne M[i,k]\} > \epsilon.$$

Let ν denote the probability measure with $\nu\{i\} = \ell_i/K$, where $\ell_i \in \mathbb{N}$ and $\sum_i \ell_i = K$. Show that

$$|\sum\nolimits_i \mu_i\{M[i,j] \ne M[i,k]\} - \sum\nolimits_i \nu_i\{M[i,j] \ne M[i,k]\}| \le \sum\nolimits_i |\mu_i - \nu_i|.$$

(ii) Let B denote the $(Kn) \times N$ matrix obtained by adding $\ell_i - 1$ copies of the row $M[i,\cdot]$, for each i. Show that $\text{SDIM}(B) = \text{SDIM}(M)$ and

$$(Kn)^{-1} \sum\nolimits_{i'\in[\![Kn]\!]} \{B[i,j] \ne B[i',k]\} = n^{-1} \sum\nolimits_{i\in[\![n]\!]} \{M[i,j] \ne M[i,k]\}$$

(iii) Appeal to Theorem <23> for B and the uniform distribution to complete the argument.

14.7 Notes

The bound stated in Theorem <4> appeared in the paper by Sauer (1972), who remarked that "a referee of this paper wrote that these results have also been established by S. Shelah", citing two works as "to appear". In an online blog (http://leon.bottou.org/news/vapnik-chervonenkis_sauer), Bottou made a case that the credit should go to Vapnik and Červonenkis, who first published a short summary (Vapnik and Červonenkis, 1968) of their result and then a longer version (Vapnik and Chervonenkis, 1971). Dudley (2014, page 211) (see also Dudley 1978, Section 7) noted that the 1971 bound was a little weaker than the Sauer result.

In his online blog `(https://gilkalai.wordpress.com/)` for 28 September 2008, Kalai presented another version of the VC Lemma then commented:

> It was mentioned (with an algebraic proof by Frankl and Pach) in Gowers' blog and also, in another context, in Kowalski's blog. Sauer proved it in response to a problem of Erdos. Shelah (with Perles) proved it as a useful lemma for Shelah's theory of stable models. (At some later time, Benjy Weiss asked Perles about such a result in the context of ergodic theory and Perles who forgot that he proved it once proved it again.) Vapnik and Chervonenkis proved it in the context of statistical learning theory.

Steele (1975, 1978) generalized the result described in Theorem <4> to matrices whose entries come from a finite alphabet. Haussler and Long (1995) generalized further.

I learned about the downshift method for proving Theorem <4> from Michel Talagrand. Ledoux and Talagrand (1991, pp. 411-412) used the set-theoretic version of the downshift method, attributing it (page 420) to Frankl (1983). Independently, both Alon (1983) and Frankl (1983) used the shifting method to prove a result that implies Theorem <4>. Frankl (1995, page 1298) noted that the shifting technique was introduced by Erdős, Ko, and Rado (1961).

The results in Sections 14.4 and 14.5 are based on Haussler (1995), who acknowledged (page 220) Linial for the downshifting method of proof for Theorem <19>, with the comment that the result had already been proved by Haussler, Littlestone, and Warmuth (1994). For an illuminating Bayesian interpretation of the probability method used to prove Theorem <23> see Haussler (1995, Section 3).

Truly a star-studded set of authors who played a role in the development and exploitation of a beautiful inequality.

References

Alon, N. (1983). On the density of sets of vectors. *Discrete Mathematics 46*, 199–202.

Alon, N. and J. H. Spencer (2000). *The Probabilistic Method* (second ed.). Wiley.

Alon, N. and M. Tarsi (1992). Colorings and orientations of graphs. *Combinatorica 12(2)*, 125–134.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability 6*, 899–929.

Dudley, R. M. (2014). *Uniform Central Limit Theorems* (2nd ed.), Volume 142 of *Cambridge studies in advanced mathematics*. Cambridge University Press. (First edition, 1999).

Erdős, P., C. Ko, and R. Rado (1961). Intersection theorems for systems of finite sets. *Quart. J. Math. Oxford 12 (2)*, 313–320.

Frankl, P. (1983). On the trace of finite sets. *Journal of Combinatorial Theory, Series A 34*, 41–45.

Frankl, P. (1987). *Surveys in combinatorics]*, Volume 123 of *London Math. Soc. Lecture Note Series*, pp. 81–110. Cambridge Univ. Press.

Frankl, P. (1995). Extremal set systems. In R. L. Graham, M. Grötschel, and L. Lovász (Eds.), *Handbook of Combinatorics*, Volume 2, Chapter 24, pp. 1293–1329. Elsevier.

Haussler, D. (1995). Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory 69*, 217–232.

Haussler, D., N. Littlestone, and M. K. Warmuth (1994). Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation 115*, 248–292.

Haussler, D. and P. M. Long (1995). A generalization of Sauer's lemma. *Journal of Combinatorial Theory 71*, 219–240.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. New York: Springer.

Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer.

Sauer, N. (1972). On the density of families of sets. *Journal of Combinatorial Theory 13*, 145–147.

Steele, J. M. (1975). *Combinatorial Entropy and Uniform Limit Laws*. Ph. D. thesis, Stanford University.

Steele, J. M. (1978). Existence of submatrices with all possible columns. *Journal of Combinatorial Theory, Series A 24*, 84–88.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Process: With Applications to Statistics*. Springer-Verlag.

Vapnik, V. N. and A. J. Červonenkis (1968). The uniform convergence of frequencies of the appearance of events to their probabilities. *Dokl. Akad. Nauk SSSR 181*, 781–783. In Russian. English translation: Soviet Math. Dokl. 9 (1968), 915–918. See Mathematical Review MR0231431 by R. M. Dudley.

Vapnik, V. N. and A. Ya. Chervonenkis (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications 16*, 264–280. Russian original received by the editors on May 7, 1969.

Vapnik, V. N. and A. Ya. Červonenkis (1981). Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and Its Applications 26*, 532–553. Russian original received by the editors on July 28, 1978.

Chapter 15
Fat shattering/surrounding

SECTION 15.1 *recalls (from Chapter 13) how symmetrization arguments generate "signed-sum" processes of the form* $\{\langle \mathfrak{s}, x\rangle : x \in \mathcal{X}\}$ *with* $\mathcal{X}$ *a subset of some euclidean space and* $\mathfrak{s}$ *a vector of independent sign variables.*

SECTION 15.2 *extends the combinatorial ideas from VC-classes of sets to subsets* $\mathcal{X}$ *of euclidean spaces, using an analog of the VC-dimension or shatter-dimension from Chapter 14.*

SECTION 15.3 *refines tha concept of shatter/surround dimension from Section 15.2. The strengthened idea (sometimes called "fat shattering") replaces a single shatter dimension by shatter function,* $\text{SDIM}(\delta, \mathcal{X})$*, depending on a sensitivity parameter* δ*.*

SECTION 15.4 *proves a key Lemma that relates* ℓ_μ^p *separation properties of finite subsets of euclidean spaces to counts of "lattice cubes" surrounded by lower-dimensional projections.*

SECTION 15.5 *derives a bound on* ℓ_μ^p *packing numbers for* $\mathcal{X} \subset [0,1]^n$*, to introduce some of the key ideas needed for more general* $\mathcal{X}$*.*

SECTION 15.7 *expands the method from the previous Section to prove the result for subsets of* ℓ_μ^{2p} *balls by means of facts about weak-*$\mathcal{L}^q$ *spaces.*

SECTION 15.8 *describes an iterative technique that enables integral conditions on square-roots of log packing numbers to be replaced by integral conditions on square-roots of the shatter function.*

15.1 Signed-sum processes

If $\mathcal{X}$ is a subset of some $\mathbb{R}^n$ and $\mathfrak{s} := (\mathfrak{s}_1, \ldots, \mathfrak{s}_n)$ is a vector of independent sign variables (that is $\mathbb{P}\{\mathfrak{s}_i = +1\} = \mathbb{P}\{\mathfrak{s}_i = -1\} = 1/2$) then the collection of all sums $\langle \mathfrak{s}, x\rangle = \sum_i \mathfrak{s}_i x_i$, for $x \in \mathcal{X}$, defines a very simple ***signed-sum*** stochastic process. Such processes often result from symmetrization arguments used to prove maximal inequalities, as in §13.4:

<1> **Theorem.** *Suppose* $S(t,\omega) := \sum_{i\in[\![n]\!]} (X_i(t,\omega) - \mathbb{P}X_i(t))$ *for a collection of independent stochastic processes* $\{X_i(t,\omega) : i \in [\![n]\!], t \in T\}$ *with* $X_i(t) \in \mathcal{L}^1(\mathbb{P})$ *for each* t*. Then (under some assumptions needed to ensure measurability)* $\mathbb{P}\sup_{t\in T}|S(t,\omega)| \le 2\mathbb{P}\mathbb{Q}^{\mathfrak{s}}\sup_{x\in\mathcal{X}_\omega}|\langle \mathfrak{s}, x\rangle|$*, where* $\mathbb{Q}$ *denotes the uniform distribution on* $\{-1,+1\}^n$ *and* $\mathcal{X}_\omega$ *denotes the set of all* x *in* $\mathbb{R}^n$ *of the form* $(X_1(t,\omega), \ldots, X_n(t,\omega))$ *for some* t *in* T*.* □

Signed-sum processes are eminently suited for control by chaining arguments because their increments have simple subgaussian distributions:

$$\mathbb{Q}e^{\lambda(\langle \mathfrak{s},x\rangle - \langle \mathfrak{s},y\rangle)} \le e^{\lambda^2 |x-y|_2^2/2} \qquad \text{for each } \lambda \text{ in } \mathbb{R}.$$

Here $|\cdot|_2$ denotes the usual ℓ^2 norm, $|w|_2 := \left(\sum_i w_i^2\right)^{1/2}$. As this Chapter will show, the chaining bounds become even more useful if we work with more general norms ℓ^p_μ, defined on $\mathbb{R}^n$ by $|x|_{p,\mu} := \left(\sum_{i\in[\![n]\!]} \mu_i |x_i|^p\right)^{1/p}$ for some probability measure μ on $[\![n]\!]$ and some $p \ge 1$.

Remark. Most of the results in this Chapter could be extended from subsets of $\mathbb{R}^n$ to collections of real-valued functions on some set $\mathbb{A}$. However, once we focus on a given vector $(a_1, \ldots, a_n)$ in $\mathbb{A}^n$ everything quicky reduces to subsets of $\mathbb{R}^n$.

This Chapter will focus on methods for bounding the packing numbers of sets like $\mathcal{X}_\omega$, under various metrics, with an eye towards chaining arguments. Better chaining bounds might be established using the methods described in Chapters 11 and 12. Talagrand (2021, Chap 6) would be a good place to start exploring this possibility.

15.2 Surround dimension

The shatter-dimension, $\text{SDIM}(\mathcal{X})$, for $\mathcal{X} \subset \{0,1\}^n$, was defined in §14.2 as the largest d for which there exists an $I \subset [\![n]\!]$ with $|\mathcal{X}[I]| = 2^d$. That is, for each pattern u in $\{0,1\}^I$ there exists an x in $\mathcal{X}$ for which $x[I] = u$, meaning that $x_i = u_i$ for every i in I. For $\text{SDIM}(\mathcal{X})$ much smaller than n the concept turned out to be very useful because the ϵ-packing numbers (with respect to some weighted HAMMING distance) could be controlled by a polynomial in $1/\epsilon$ of degree bounded by a constant multiple of $\text{SDIM}(\mathcal{X})$.

The SDIM idea can be extended to more general subsets $\mathcal{X}$ of $\mathbb{R}^n$ if we replace the equalities by sets of inequalities.

<2> **Definition.** *Suppose $\mathcal{X}$ is a non-empty subset of $\mathbb{R}^n$ and I is a non-empty subset of $[\![n]\!]$ and $r \in \mathbb{R}^I$ and $u \in \{0,1\}^I$.*

*(i) Say that an x in $\mathcal{X}$ picks out the **pattern** u around r if $x_i > r_i$ when $u_i = 1$ and $x_i < r_i$ when $u_i = 0$.*

*(ii) Say that $\mathcal{X}[I]$ **surrounds** r if each pattern $u \in \{0,1\}^I$ around r can be picked out by some member of $\mathcal{X}[I]$.*

(iii) The surround-dimension, $\text{SDIM}(\mathcal{X})$, equals the largest d for which there exists an I with $|I| = d$ and an $r \in \mathbb{R}^I$ such that $\mathcal{X}[I]$ surrounds r. □

The initial 's' in SDIM now stands for 'surround'. As you will soon see, some of the arguments from Chapter 14 carry over to this extended $\text{SDIM}(\cdot)$ concept, with useful consequences for signed-sum processes.

Remark. If $\mathcal{X} \subseteq \{0,1\}^n$ then the SDIM($\mathcal{X}$) as in Definition <2> agrees with the shatter-dimension defined in §14.2—it suffices to consider the vector $r = (1/2, \ldots, 1/2)$ for such an $\mathcal{X}$.

The set $\mathcal{X}[I]$ surrounds r iff it puts at least one $x[I]$ in each the $2^{|I|}$ orthants define by r:

$$\mathcal{Q}_{u,r} := \prod\nolimits_{i\in I} H_i \qquad \text{where } H_i := \begin{cases} (r_i, \infty) & \text{if } u_i = 1 \\ (-\infty, r_i) & \text{if } u_i = 0 \end{cases}.$$

To avoid trivialities we could require that $\mathcal{X}$ contains at least two points, which ensures that SDIM($\mathcal{X}$) ≥ 1.

<3> **Example.** Let $\mathcal{H}$ be a d-dimension subspace of $\mathbb{R}^n$, for some $d < n$. Here is an argument to show SDIM($\mathcal{H}$) $\leq d$.

Suppose $I \subset [\![n]\!]$ and $r \in \mathbb{R}^I$, with $|I| > d$. As $\mathcal{H}[I]$ is a subspace of $\mathbb{R}^I$ with dimension at most d there must exist some nonzero vector z in $\mathbb{R}^I$ such that $\langle x[I], z\rangle := \sum_{i\in I} x_i z_i = 0$ for all $x \in \mathcal{H}$. The vector z defines two patterns, u, v:

$$u_i = 1 \text{ iff } z_i > 0 \quad \text{AND} \quad v_i = 1 \text{ iff } z_i < 0.$$

That is, u is the indicator function of $\{i \in I : z_i > 0\}$ and v is the indicator function of $\{i \in I : z_i < 0\}$. The set $\mathcal{X}$ cannot pick out both of these patterns: if $x, y \in \mathcal{X}$ and

$$x_i > r_i > y_i \text{ if } z_i > 0 \quad \text{AND} \quad x_i < r_i < y_i \text{ if } z_i < 0$$

then $x_i - y_i$ has the same sign as z_i. However, orthogonality of the vector z to both $x[I]$ and $y[I]$ would give

$$-\sum\nolimits_{i\in I} z_i r_i = \sum\nolimits_{i\in I} z_i(x_i - r_i) = \sum\nolimits_{i\in I} z_i(y_i - r_i),$$

□ implying $\sum_{i\in I} z_i(x_i - y_i) = 0$.

The extended definition of SDIM still enjoys the polynomial bound that makes the concept of a VC-class of sets so useful.

<4> **Lemma.** *If $\mathcal{X}$ is a subset of $\mathbb{R}^n$ with* SDIM($\mathcal{X}$) $\leq d$ *and $R \in \mathbb{R}^n$ then $\mathcal{X}$ picks out at most $\beta(n,d) := \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d} \leq (en/d)^d$ distinct patterns*
□ *around each R in $\mathbb{R}^n$.*

Proof. Let M be the $n \times N$ binary matrix whose columns list all the distinct patterns around R. For each nonempty subset I of $[\![n]\!]$ the patterns picked out by $\mathcal{X}[I]$ around $R[I]$ appear (possibly repeatedly) as columns of $M[I, \cdot]$. If $|I| > d$ then at least one member of $\{0,1\}^I$ fails to appear as a column of $M[I, \cdot]$. By the Shatter Theorem from §14.3 we have im-
□ ply $N \leq \beta(n,d)$.

For subsets $\mathcal{X}$ of $\{0,1\}^n$ you saw in §14.2 how an upper bound on $\text{SDIM}(\mathcal{X})$ controls the packing numbers for $\mathcal{X}$ under various weighted HAMMING distances. Something very similar holds for subsets $\mathcal{X}$ of $[0,1]^n$ when weighted HAMMING distances are replaced by the weighted ℓ_μ^p distance, for μ a probability measure on $[\![n]\!]$ and $p \geq 1$:

$$|x-y|_{p,\mu} := \Big(\sum\nolimits_{i\in[\![n]\!]} \mu_i \big|x[i]-y[i]\big|^p\Big)^{1/p} \qquad \text{for } x,y \in \mathbb{R}^n.$$

Note that $|x-y|_{p,\mu}^p \leq |x-y|_{1,\mu}$ for $x,y \in [0,1]^n$, implying, for $\mathcal{X} \subset [0,1]^n$,

<5> $$\text{PACK}(\epsilon, \mathcal{X}, \ell_{p,\mu}) \leq \text{PACK}(\epsilon^p, \mathcal{X}, \ell_{1,\mu}) \qquad \text{for all } \epsilon > 0 \text{ and } \mu.$$

<6> **Theorem.** *Suppose $\mathcal{X}$ is a subset of $[0,1]^n$ with $\text{SDIM}(\mathcal{X}) \leq d$. Then there exists a universal constant $C \approx 105.2$ for which*

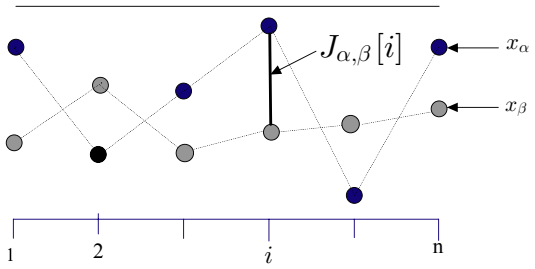
$$\text{PACK}(\epsilon, \mathcal{X}, \ell_\mu^p)^{1/d} \leq C(1/\epsilon^p)\log(1/\epsilon^p) \leq C(1/\epsilon)^{2p} \qquad \text{for } 0 < \epsilon \leq 1/2.$$

□ *for each probability measure μ on $[\![n]\!]$ and each $p \geq 1$.*

Proof. By inequality <5> it suffices to consider just the $p = 1$ case.

Suppose $x_1, \ldots, x_N$ are δ-separated members of $\mathcal{X}$ (with $\delta = \epsilon^p$), that is, $|x_\alpha - x_\beta|_{1,\mu} > \delta$ for $1 \leq \alpha < \beta \leq N$. Define

$$\begin{aligned} J_{\alpha,\beta}[i] &:= \{w \in \mathbb{R} : \min(x_\alpha[i], x_\beta[i]) < w < \max(x_\alpha[i], x_\beta[i])\} \\ &= \{w : x_\alpha[i] < w < x_\beta[i]\} \cup \{w : x_\beta[i] < w < x_\alpha[i]\}. \end{aligned}$$



For x_α and x_β to pick out different patterns around some $r \in \mathbb{R}^n$ there must be some i for which $r_i \in J_{\alpha,\beta}[i]$.

With $m := \lceil 2(\log N)/\delta \rceil$, let $\mathcal{I}_1, \ldots, \mathcal{I}_m, R_1, \ldots, R_m$ be independent random variables with $\mathcal{I}_j \sim \mu$ and $R_j \sim \text{UNIF}(0,1)$ for each j. For each pair of distinct indices α, β from $[\![N]\!]$ and $j \in [\![m]\!]$,

$$\begin{aligned} \mathbb{P}\{R_j \in J_{\alpha,\beta}[\mathcal{I}_j]\} &= \sum\nolimits_{i\in[\![n]\!]} \mathbb{P}\{\mathcal{I}_j = i\}\mathbb{P}\{R_j \in J_{\alpha,\beta}[i]\} \\ &= \sum\nolimits_i \mu_i |x_\alpha[i] - x_\beta[i]| > \delta. \end{aligned}$$

By independence, it follows that

$$\mathbb{P} \cap_{j\in[\![m]\!]} \{R_j \notin J_{\alpha,\beta}[\mathcal{I}_j]\} \le (1-\delta)^m \le \exp(-m\delta).$$

A union bound then gives

$$\mathbb{P} \cup_{\alpha<\beta} \cap_{j\in[\![m]\!]} \{R_j \notin J_{\alpha,\beta}[\mathcal{I}_j]\} \le \binom{N}{2} e^{-m\delta} \le \tfrac{1}{2}\exp(2\log N - m\delta) < 1.$$

The complementary event, $\cap_{\alpha<\beta}\cup_{j\in[\![m]\!]}\{R_j \in J_{\alpha,\beta}[\mathcal{I}_j]\}$, has nonzero probability. On that event, for each $\alpha<\beta$ there exists a j with R_j lying strictly between $x_\alpha[\mathcal{I}_j]$ and $x_\beta[\mathcal{I}_j]$. Thus there exists a realization of $\mathcal{I}, R$ such that the $x_\alpha[\mathcal{I}]$'s for $\alpha \in [\![N]\!]$ pick out N distinct patterns around $R[\mathcal{I}]$. By Lemma <4> we must have $N \le \beta(m,d) \le (em/d)^d$.

The rest of the argument (invert the inequality to bound N) is exactly the same as for $\{0,1\}^n$ in §14.2. □

The bound asserted by the Theorem tells us something useful only when $\theta := d/n$ is small. Indeed, if $\mathcal{X} \subset [0,1]^n$ we always have

$$\text{PACK}(\epsilon, \mathcal{X}, \ell_\mu^p) \le \text{PACK}(\epsilon, [0,1]^n, \ell_\mu^p) \le (2/\epsilon)^n \qquad \text{for } 0<\epsilon\le 1/2.$$

That is, $\log \text{PACK}(\epsilon, \mathcal{X}, \ell_\mu^p) \le n\log(2/\epsilon)$ whereas, the bound from the Theorem is $\log \text{PACK}(\epsilon, \mathcal{X}, \ell_\mu^p) \le n\theta\,(\log C_{<6>} + p\log(2/\epsilon))$.

Remark. To derive the bound on $\text{PACK}(\epsilon, [0,1]^n, \ell_\mu^p)$, partition the whole cube into at most $\lceil 1/\epsilon\rceil^n$ cubes of side $\le \epsilon$. Each smaller cube has ℓ_μ^p-diameter $\le \epsilon$. An ϵ-packing can put at most one point in each smaller cube.

The presence of an arbitrary probability measure μ in Theorem <6> brings a surprising bonus.

<7> **Corollary.** *Suppose $\mathcal{X} \subset \prod_i [a_i, b_i]$ for some vectors a, b in $\mathbb{R}^n$ with $a_i < b_i$ and $\text{SDIM}(\mathcal{X}) \le d$. Define $E := (b_i - a_i : i \in [\![n]\!])$. Then for each probability measure μ on $[\![n]\!]$ we have*

$$\text{PACK}(\epsilon|E|_{p,\mu}, \mathcal{X}, \ell_\mu^p)^{1/d} \le C_{<6>}\epsilon^{-p}\log(\epsilon^{-p}) \qquad \text{for } 0<\epsilon\le 1/2.$$

Proof. Suppose we have vectors $x_1, \ldots, x_N$ in $\mathcal{X}$ with

$$|x_\alpha - x_\beta|_{p,\mu} > \epsilon|E|_{p,\mu} \qquad \text{for } 1 \le \alpha < \beta \le N.$$

Define $\mathcal{Y} = \{y_\alpha : \alpha \in [\![N]\!]\}$, where $y_\alpha[i] := (x_\alpha[i] - a_i)/E_i \in [0,1]$ for each $i \in [\![n]\!]$. If $I \subset [\![n]\!]$ and $r \in \mathbb{R}^I$ and $\mathcal{Y}[I]$ surrounds $r[I]$ then $\mathcal{X}[I]$ surrounds $R[I]$, where $R := (a_i + E_i r_i : i \in [\![n]\!])$. Thus $\text{SDIM}(\mathcal{Y}) = \text{SDIM}(\mathcal{X}) \le d$.

Define a new probability measure ν on $[\![n]\!]$ by $\nu_i := \mu_i E_i^p / |E|_{p,\mu}^p$. Then, for $1 \le \alpha < \beta \le N$,

$$|y_\alpha - y_\beta|_{p,\nu}^p = \sum\nolimits_i \nu_i |x_\alpha[i] - x_\beta[i]|^p / E_i^p = |x_\alpha - x_\beta|_{p,\mu}^p / |E|_{p,\mu}^p > \epsilon^p.$$

Thus $N \le \text{PACK}(\epsilon, \mathcal{Y}, \ell_\nu^p) \le C_{<6>}\epsilon^{-p}\log(\epsilon^{-p})$. □

Even though the SDIM concept has proven quite useful for asymptotic analysis in various statistical estimation problems, it does suffer from a regrettable lack of "robustness" against small changes in the vectors.

<8> **Example.** Consider the rather trivial subset $\mathcal{X}$ of $\mathbb{R}^n$ consisting of the vectors $x_t := t\mathbb{1}$ for $0 \le t \le 1$. That is, $x_t[i] = t$ for every i. Note that $|x_s - x_t|_{p,\mu} = |s-t|$ for each $p \ge 1$ and each probability measure μ. The set $\mathcal{X}$ is just a thinly disguised version of the unit interval $[0,1]$.

It is easy to show that $\text{SDIM}(\mathcal{X}) = 1$ and

$$1/(2\epsilon) \le \text{COVER}(\epsilon, \mathcal{X}, \ell^p_\mu) \le 2 + 1/(2\epsilon) \le 3/(2\epsilon) \qquad \text{for } 0 < \epsilon \le 1/2,$$

the upper bound coming from the set $\{x_t : t = 1/2 \pm 2k\epsilon,\ k \in [\![k_\epsilon]\!]\}$, where k_ϵ is the smallest integer for which $1/2 + 2k\epsilon \ge 1 - \epsilon$.

Arbitrarily small perturbations can increase the surround dimension to n with insignificant effects on the packing numbers for ϵ. Here is an example. Use $\{\mathfrak{d}_k : k \in [\![2^n]\!]\} = \{-1,+1\}^n$ to make small perturbations: with $0 < \theta < (1/3)^n$ define $T_n := \{t(k) : k \in [\![2^n]\!]\}$ where $t(k) := 1/2 + k\theta^2$. Replace $\mathcal{X}$ by $\widetilde{\mathcal{X}} = \{\widetilde{x}_t : 0 \le t \le 1\}$ where

$$\widetilde{x}_t := \begin{cases} x_t + \theta\mathfrak{d}_k & \text{if } t \in T_n \\ x_t & \text{otherwise} \end{cases} .$$

Then $\text{SDIM}(\widetilde{\mathcal{X}}) = n$ because the set $\widetilde{\mathcal{X}}$ surrounds $\frac{1}{2}\mathbb{1}$.

Let S_ϵ be an ϵ-covering set for $[0,1]$, under the usual euclidean distance, with $|S_\epsilon| \le 2/\epsilon$ If $\epsilon \ge \theta$ then $\{x_t : t \in S_\epsilon\}$ is a 2ϵ covering of $\widetilde{\mathcal{X}}$. And
□ if $\epsilon < \theta$ then $\{\widetilde{x}_t : t \in S_\epsilon \cup T_n\}$ is ϵ-covering of $\widetilde{\mathcal{X}}$ of size less than $3/\epsilon$.

15.3 Fattening up the concept of surround dimension

We need to strengthen the concept of 'surrounding' if we wish to avoid the sort of instability illustrated by Example <8>. The following Definition has the desired effect. It also comes closer to capturing the essential property for existence of good maximal inequalities. (See Notes.)

<9> **Definition.** *Let $\mathcal{X}$ be a subset of $\mathbb{R}^n$ and δ be a strictly positive number.*

(i) Suppose $I \subseteq [\![n]\!]$ and $r \in \mathbb{R}^I$. Say that an x in $\mathcal{X}$ picks out a ***pattern*** *$u \in \{0,1\}^I$ around r at level δ if $x_i \ge r_i + \delta/2$ for $u_i = 1$ and $x_i \le r_i - \delta/2$ for $u_i = 0$.*

(ii) If $r \in \mathbb{R}^I$, say that $\mathcal{X}$ ***surrounds*** *r* ***at level*** *δ if each pattern around r can be picked out (at level δ) by some member of $\mathcal{X}$.*

(iii) Define $\text{SDIM}(\delta, \mathcal{X})$ to be the largest d for which there exists an $I \subset [\![n]\!]$
□ *with $|I| = d$ and an r in $\mathbb{R}^I$ such that $\mathcal{X}$ surrounds r (at level δ).*

Analogs of $\text{SDIM}(\cdot, \mathcal{X})$ for general collections of function seem to have first appeared in the machine learning literature in the 1990's, sometimes under the name ***fat shattering***. A similar idea had already appeared in a completely different context (Talagrand, 1984, §9) as a condition for establishing pointwise compactness of a set of real-valued functions. See the Notes.

The most elegant results concerning the connections between $\text{SDIM}(\cdot, \mathcal{X})$ and packing numbers were worked out in three papers: Mendelson and Vershynin (2003), Vershynin (2005) and R-V = Rudelson and Vershynin (2006). The following Theorem, based on Corollary 5.4 of R-V, gives the flavor of the Mendelson-Rudelson-Vershynin (M-R-V) results. The papers contain many other interesting results.

<10> **Theorem.** *Suppose $p \geq 1$ and $\mathcal{X}$ is a subset of $\mathbb{R}^n$ contained within an ℓ_μ^{2p} ball $\{y \in \mathbb{R}^n : |y - y_0|_{2p,\mu} \leq R\}$, for some y_0 in $\mathbb{R}^n$, some radius R, and some probability measure μ on $[[n]]$. Then*

$$\log \text{PACK}(t, \mathcal{X}, \ell_\mu^p) \leq C_1 p^2 \text{SDIM}(C_2 t, \mathcal{X}) \log(C_3 R/t) \qquad \text{for } 0 < t \leq 2R,$$

□ *where C_1, C_2, C_3 are universal constants.*

Remark. As $|x - y|_{p,\mu} \leq 2R$ for $x, y \in \mathcal{X}$, there is no need to consider values of t that are $> 2R$.

A general y_0 will needed in Section 15.8. As neither the packing numbers nor $\text{SDIM}(\cdot, \mathcal{X})$ are affected by translation of $\mathcal{X}$, for the Proof of Theorem <10>, in Section 15.7, there will be no loss of generality in assuming that $y_0 = 0$ and $|x|_{2p,\mu} \leq R$ for each x in $\mathcal{X}$.

The Proof will involve some subtle arguments, including the "M-R-V trick" (Section 15.4) of relating packing numbers to the number of ***lattice cubes***—unit cubes centered at points with integer coordinates—surrounded by the packing set. This idea will play a role similar to Lemma <4> in Section 15.2. In order to preview the method in a much simpler setting, Section 15.5 will apply the lattice-cube argument with the ℓ_μ^{2p} ball replaced by $[0, 1]^n$.

The Proof of the more general result will involve ***weak*** ℓ^q spaces. I found this part in the R-V argument the most challenging, probably because I knew very little about such spaces when I first looked at the paper. Fortunately, it turned out that the necessary details (which are summarized in Section 15.6) were actually not so complicated.

15.4 Lattice cubes surrounded by well separated vectors

The M-R-V idea is elegantly simple. If a subset $\mathcal{X}$ of $\mathbb{R}^n$ surrounds a point r in $\mathbb{R}^I$ at level δ then for each pattern $u \in \{0, 1\}^I$ there exists an x in $\mathcal{X}$

with $x_i/\delta \geq y_i + 1/2$ if $u_i = 1$ and $x_i/\delta \leq y_i - 1/2$ if $u_i = 0$, where $y := r/\delta$. That is, the rescaled set $\mathcal{X}/\delta := \{x/\delta : x \in \mathcal{X}\}$ surrounds the set

$$\text{cube}_I(y) := \prod_{i \in I} [y_i - 1/2, y_i + 1/2] \subset \mathbb{R}^I,$$

in the sense that $\mathcal{X}[I]$ puts at least one point in each of the $2^{|I|}$ orthants that share a single vertex with the cube. The cube lies inside the convex hull of those $2^{|I|}$ points.

For the purposes of a counting argument it helps if we restrict attention to ***lattice cubes***, $\text{cube}_I(\zeta)$ with $\zeta \in \mathbb{Z}^I$. (V and R-V used the name "integer cell" and required vertices in $\mathbb{Z}^I$.)

<11> **Definition.** *For $\mathcal{W} \subset \mathbb{R}^n$, define $\mathbb{L}(\mathcal{W})$ to be the set of all pairs (I, ζ), with $\emptyset \neq I \subseteq [\![n]\!]$ and $\zeta \in \mathbb{Z}^I$, for which $\mathcal{W}$ surrounds $\text{cube}_I(\zeta)$, that is, for each pattern u in $\{0,1\}^I$ there is a w in $\mathcal{W}$ such that $w_i \geq \zeta_i + 1/2$ if $u_i = 1$ and $w_i \leq \zeta_i - 1/2$ if $u_i = 0$.* □

Remark. Following R-V, we could also allow $I = \emptyset$, defining its contribution to $\mathbb{L}(\mathcal{W})$ as 1 for every nonempty $\mathcal{W}$. That choice makes some of the formulas more elegant but I always worry about nasty surprises that might lie hidden behind such conventions.

Now suppose $\mathcal{K}$ is a bounded, convex subset of $\mathbb{R}^n$ with $\mathcal{W} \subset \mathcal{K}$. If $\emptyset \neq I \subset [\![n]\!]$ and $\mathcal{W}$ surrounds $\text{cube}_I(\zeta)$ then that lattice cube must lie inside the convex set $\mathcal{K}[I]$, the projection of $\mathcal{K}$ onto the coordinate space $\mathbb{R}^I$. Each such cube has $|I|$-dimensional LEBESGUE measure, $\mathfrak{m}_I \text{cube}_I(\zeta)$, equal to 1; and two distinct lattice cubes can at worst intersect in a set of zero $\mathfrak{m}_I$ measure. It follows that

<12> $$\mathfrak{m}_I(\mathcal{K}[I]) \geq \#\{\zeta \in \mathbb{Z}^I : \mathcal{W} \text{ surrounds } \text{cube}_I(\zeta)\}.$$

If $|I| > \text{SDIM}(1, \mathcal{W})$ then $\mathcal{W}$ cannot surround any $\text{cube}_I(y)$ with $y \in \mathbb{R}^I$, let alone any lattice cubes. Only subsets I with $1 \leq |I| \leq \text{SDIM}(1, \mathcal{W})$ contribute to the union that defines $\mathbb{L}(\mathcal{W})$. Thus

<13> $$\sum_{\emptyset \neq I \subseteq [\![n]\!],\ |I| \leq \text{SDIM}(1,\mathcal{W})} \mathfrak{m}_I(\mathcal{K}[I]) \geq \#\mathbb{L}(\mathcal{W}).$$

This inequality together with the following Lemma will provide the main technical tool needed in the next two Sections.

<14> **Lemma.** *If, for some $p \geq 1$ and some probability measure μ on $[\![n]\!]$, a subset $\mathcal{W}$ of $\mathbb{R}^n$ contains vectors $w_1, \ldots, w_N$ that are very well separated, in the sense that $\min_{\alpha \neq \beta} |w_\alpha - w_\beta|_{p,\mu} > 10$, then $1 + \#\mathbb{L}(\mathcal{W}) \geq N^{1/(2p)}$.*

Proof. Think of the w_j's as the columns of an $n \times N$ matrix W, that is, $w_j = W[\cdot, j]$. The projection of w_j onto $\mathbb{R}^I$ then becomes $W[I, j]$ for each subset I of $[\![n]\!]$, and the separation assumption becomes

$$\sum_{i \le n} \mu_i |W[i,\alpha] - W[i,\beta]|^p > 10^p \qquad \text{for all } 1 \le \alpha < \beta \le N. \tag{15}$$

The asserted inequality for $\#\mathbb{L}(\mathcal{W})$ is trivial when $N = 1$. For the purposes of an inductive proof, suppose that $N \ge 2$ and that the analogous assertion is true for $1, 2, \ldots, N-1$. The inductive step will depend on a cunning way of finding two disjoint subsets, $J_\oplus$ and $J_\ominus$ of $[\![N]\!]$ for which

$$|J_\ominus|^\theta + |J_\oplus|^\theta > N^\theta \qquad \text{where } \theta := (2p)^{-1}. \tag{16}$$

The strict inequality ensures $\max(|J_\ominus|, |J_\oplus|) < N$, which will allow us to invoke the inductive hypothesis for the submatrices $W[\cdot, J_\ominus]$ and $W[\cdot, J_\oplus]$. The argument will consist of three steps:

1. Show that there exists $i_0 \in [\![n]\!]$ and an integer ℓ such that

$$\mathbb{P}\{X \le \ell - 1/2\}^\theta + \mathbb{P}\{X \ge \ell + 1/2\}^\theta > 1, \tag{17}$$

where $X := W[i_0, \mathcal{J}]$ with $\mathcal{J}$ is uniformly distributed on $[\![N]\!]$.

2. Use i_0 and ℓ to define the two candidates, $W^\ominus$ and $W^\oplus$, for attention from the inductive hypothesis:

$$\begin{aligned} J_\ominus &:= \{j \in [\![N]\!] : W[i_0, j] \le \ell - 1/2\} \quad \text{with} \quad W^\ominus := W[\cdot, J_\ominus], \\ J_\oplus &:= \{j \in [\![N]\!] : W[i_0, j] \ge \ell + 1/2\} \quad \text{with} \quad W^\oplus := W[\cdot, J_\oplus]. \end{aligned} \tag{18}$$

3. Use <16> to show that

$$1 + \#\mathbb{L}(W) \ge 2 + \#\mathbb{L}(W^\ominus) + \#\mathbb{L}(W^\oplus), \tag{19}$$

which will complete the induction.

Step 1. Once again, the existence of suitable subsets can be proved by the probabilistic method. Let $\mathcal{J}$ and $\widetilde{\mathcal{J}}$ be random variables corresponding to a sample of size 2 without replacement from the uniform distribution on $[\![N]\!]$. (Equivalently, the pair $(\mathcal{J}, \widetilde{\mathcal{J}})$ is uniformly distributed over the set $[\![N]\!]^2$ minus its diagonal.) By construction $\mathbb{P}\{\mathcal{J} \ne \widetilde{\mathcal{J}}\} = 1$ so that inequality <15> implies $\mathbb{P} \sum_{i \le n} \mu_i |W[i, \mathcal{J}] - W[i, \widetilde{\mathcal{J}}]|^p > 10^p$. Take $\mathbb{P}$ inside the sum to see that there must exist at least one i_0 for which

$$\mathbb{P}|W[i_0, \mathcal{J}] - W[i_0, \widetilde{\mathcal{J}}]|^p > 10^p.$$

That is $\mathbb{P}|X - \widetilde{X}|^p > 10^p$ for $X := W[i_0, \mathcal{J}]$ and $\widetilde{X} := W[i_0, \widetilde{\mathcal{J}}]$. The probability measure μ plays no further role in the proof.

To prove existence of the desired ℓ we can use a "separation lemma" (Rudelson and Vershynin, 2006, Lemma 3.3), whose proof (Problem [11]) requires only elementary probability calculations: If Y is a random variable with zero median and $\mathbb{P}|Y|^p > 2.5$ for some $p \geq 1$ then

<20> $$\mathbb{P}\{Y \leq \lambda - 1\}^\theta + \mathbb{P}\{Y \geq \lambda\}^\theta > 1 \qquad \text{where } \theta := 1/(2p).$$

for at least one real number λ. Here you should read $\mathbb{P}\{\dots\}^\theta$ as a probability raised to the power θ.

Remark. R-V worked with inequalities involving LORENTZ spaces of functions. As I was not able to understand the advantages of such generality I have restricted their proofs to the more familiar setting involving ℓ^p-style norms on $\mathbb{R}^n$. (As they acknowledged, on p. 627, their main interest was the case of L^p spaces of functions.)

The random variables X and $\widetilde{X}$ are not independent but they do have the same marginal distribution. Define $Y := (X - M)/2$ and $\widetilde{Y} := (\widetilde{X} - M)/2$, with M a median for X. Then we have

$$4\,\|Y\|_p = 2\,\|Y\|_p + 2\left\|\widetilde{Y}\right\|_p \geq \left\|2Y - 2\widetilde{Y}\right\|_p = \left\|X - \widetilde{X}\right\|_p > 10.$$

The random variable Y has a zero median and $\mathbb{P}|Y|^p > (10/4)^p \geq 2.5$. Thus <20> holds for some λ. Equivalently,

$$\mathbb{P}\{X \leq R - 1/2\}^\theta + \mathbb{P}\{X \geq R + 3/2\}^\theta > 1 \qquad \text{for } R := M + 2(\lambda - 1) + 1/2.$$

The gap of width 2 gives us enough wiggle room to squeeze in the integer $\ell := \lceil R \rceil$, giving <17>

Step 2. As $\mathcal{J}$ is uniformly distributed on $[\![N]\!]$,

$$\mathbb{P}\{X \leq \ell - 1/2\} = \mathbb{P}\{\mathcal{J} \in J_\ominus\} = |J_\ominus|/N,$$
$$\mathbb{P}\{X \geq \ell + 1/2\} = \mathbb{P}\{\mathcal{J} \in J_\oplus\} = |J_\oplus|/N.$$

Thus <17> implies inequality <16>.

Step 3. An appeal to the inductive hypothesis for both $W^\ominus$ and $W^\oplus$ followed by an appeal to <16> gives

$$\big(1 + \#\mathbb{L}(W^\ominus)\big) + \big(1 + \#\mathbb{L}(W^\ominus)\big) \geq |J_\ominus|^\theta + |J_\ominus|^\theta > N^\theta.$$

which means that to complete the Proof it now suffices to prove <19>.

Clearly $\mathbb{L}(W) \supseteq \mathbb{L}(W^\ominus) \cup \mathbb{L}(W^\oplus)$. If we just add up the number of lattice cubes (I, ζ) in $\mathbb{L}(W^\ominus)$ and $\mathbb{L}(W^\oplus)$ we double count the cubes in $\mathbb{B} := \mathbb{L}(W^\ominus) \cap \mathbb{L}(W^\oplus)$. To offset this double count we can construct a one-to-one map $(I, \zeta) \mapsto (I^*, \zeta^*)$ from $\mathbb{B}$ into $\mathbb{E} := \mathbb{L}(W) \setminus (\mathbb{L}(W^\ominus) \cup \mathbb{L}(W^\oplus))$. And to pick up the extra 1 we can show that $(\{i_0\}, \ell)$ belongs to $\mathbb{E}$.

By construction, $W[i_0, j] \le \ell - 1/2$ if $j \in J_\ominus$ and $W[i_0, j] \ge \ell + 1/2$ if $j \in J_\oplus$. These inequalities ensure that $i_0 \notin I$ if $(I, \zeta) \in \mathbb{B}$: if $\zeta_{i_0} \ge \ell$ then no j in $J_\ominus$ can pick out a pattern u with $u[i_0] = +1$, and similarly with $J_\oplus$ if $\zeta_{i_0} \le \ell$. In particular, $(\{i_0\}, \ell) \notin \mathbb{B}$, but it does belong to $\mathbb{L}(W)$ because neither $J_\oplus$ nor $J_\ominus$ is empty. Now define $I^* := I \cup \{i_0\}$ and $\zeta^*[i] := \zeta[i]\{i \in I\} + \ell\{i = i_0\}$. (That is, ζ^* extends the domain of ζ to I^* by defining $\zeta^*[i_0] := \ell$.) To show that $(I^*, \zeta^*) \in \mathbb{L}(W)$, consider any pattern u^* from $\{-1, +1\}^{I^*}$. As the subpatterm $u = u^*[I]$ belongs to $\{-1, +1\}^I$ and $(I, \zeta) \in \mathbb{B}$, there exist columns $j_u^\ominus \in J_\ominus$ and $j_u^\oplus \in J_\oplus$ for which $W[I, j_u^\ominus]$ and $W[I, j_u^\oplus]$ both pick out pattern u (at level 1). And by definition of $J_\ominus$ and $J_\oplus$,

$$W[i_0, j_u^\ominus] + 1/2 \le \ell \le W[i_0, j_u^\oplus] - 1/2.$$

□ It follows that $(I^*, \zeta^*) \in \mathbb{L}(W)$.

15.5 Packing for subsets of $[0,1]^n$

This Section is intended only as a warm-up for the more subtle calculations needed to prove Theorem <10> in Section 15.7, by deriving a result similar to Theorem 1 of Mendelson and Vershynin (2003).

<21> **Theorem.** *If $\mathcal{X}$ is a subset of $[0,1]^n$ with $\text{SDIM}(\epsilon, \mathcal{X}) \le d$ for some $\epsilon > 0$ then, for each $p \ge 1$ and each probability measure μ on $[\![n]\!]$,*

□ $$\text{PACK}(C\epsilon, \mathcal{X}, \ell_\mu^p) \le \left(cp(1/\epsilon)^{2p+1}\right)^{4pd}$$

where C and c are universal constants.

Proof. Let me start by sketching how the bound from Section 15.4 gets into the argument, even though the resulting inequalty will need some fine-tuning to match the result asserted by the Theorem.

Suppose $x_1, \ldots, x_N$ are $C\epsilon$-separated in $\mathcal{X}$, that is,

<22> $$|x_\alpha - x_\beta|_{p,\mu} > C\epsilon \qquad \text{for } 1 \le \alpha < \beta \le N.$$

Define $w_\alpha := 10x_\alpha/(C\epsilon)$ for $1 \le \alpha \le N$ and $\mathcal{K} := [0, 10/(C\epsilon)]^n$. Then $\mathcal{W} := \{w_1, \ldots, w_N\}$ satisfies the assumptions of Lemma <14>. Note that $\mathfrak{m}_I \mathcal{K}[I] = (10/(C\epsilon))^{|I|}$ and there are $\binom{n}{k}$ different I's with $|I| = k$. Thus inequality <13> and the Lemma <14> together give

$$\begin{aligned} 1 + (10/(C\epsilon))^d \beta(n,d) &\ge 1 + \sum\nolimits_{1k \le d} \binom{n}{k} (10/(C\epsilon))^k \\ &\ge 1 + \#\mathbb{L}(\mathcal{W}) \ge N^{1/(2p)}. \end{aligned} \tag{23}$$

Inversion of inequality <23> would provide an upper bound on N that depended on n. Not very useful. We need to replace n by a quantity that

depends only on N and ϵ. A sampling argument, using the HOEFFDING inequality (§7.3) will take care of the problem. (Compare with sampling argument used to prove Theorem <6>.)

For an m still to be determined, let $\mathcal{I} := (\mathcal{I}_1, \ldots, \mathcal{I}_m)$ be vector of independent random variables taking values in $[\![n]\!]$, each having distribution μ. For each pair α, β with $1 \le \alpha < \beta \le N$ define $g_{\alpha,\beta}[i] := |x_\alpha[i] - x_\beta[i]|^p$. For each k in $[\![m]\!]$ we have

$$\mathbb{P} g_{\alpha,\beta}[\mathcal{I}_k] = \mu g_{\alpha,\beta} = |x_\alpha - x_\beta|^p_{p,\mu} > 2\delta := (C\epsilon)^p \qquad \text{by <22>.}$$

As $0 \le g_{\alpha,\beta}[i] \le 1$ for each i HOEFFDING gives

$$\mathbb{P}\{m^{-1} \sum\nolimits_k g_{\alpha,\beta}(\mathcal{I}_k) \le \mu g_{\alpha,\beta} - \delta\} \le \exp\left(-2m^2\delta^2/m\right).$$

Take a union bound over all $\binom{N}{2}$ pairs α, β to deduce that

$$\mathbb{P}\{\exists \alpha < \beta : m^{-1} \sum\nolimits_k g_{\alpha,\beta}(\mathcal{I}_k) \le \mu g_{\alpha,\beta}/2\} \le \tfrac{1}{2} \exp\left(2\log N - 2m\delta^2\right),$$

which is < 1 if we choose $m := \lceil \delta^{-2} \log N \rceil$. Thus there exists a realization of $\mathcal{I} := (\mathcal{I}_1, \ldots, \mathcal{I}_m)$ for which

$$m^{-1} \sum\nolimits_j |x_\alpha[\mathcal{I}_j] - x_\beta[\mathcal{I}_j]|^p > (C\epsilon)^p/2 \qquad \text{for } 1 \le \alpha < \beta \le N.$$

We are now almost back at <22>. The only changes are: $\mathcal{X}$ has been replaced by $\mathcal{Y} := \mathcal{X}[\mathcal{I}]$ for some realization of $\mathcal{I}$, a subset of $[\![0,1]\!]^m$; and μ has been replaced by μ_m, the uniform distribution on $[\![m]\!]$ (which can also be thought of as the empirical measure, defined by $\mathcal{I}_1, \ldots, \mathcal{I}_m$.) Most importantly $\text{SDIM}(\epsilon, \mathcal{Y}) \le \text{SDIM}(\epsilon, \mathcal{X})$. Slight modifications of the argument following <22> would now give an inequality very similar to <23> except that n is replaced by m. From there, another inversion argument like the one in §14.2 should bound $N^{1/d}$ by a quantity like the one asserted by Theorem <21>.

The details are not important at the moment, because this Section is merely a warm-up for the Proof of Theorem <10> in Section 15.7. You should just take away two ideas: Theorem <14> brings $\text{SDIM}(\,\cdot\,, \mathcal{X})$ into the upper bound for N; and a preliminary sampling from μ can replace n by a smaller m. □

15.6 Weak ℓ^q spaces

The argument for Theorem <10> will parallel the argument sketched in Section 15.5 for subsets of $[0,1]^n$. The main difference will appear during the step where a sampling trick replaces n by a much smaller m, because the HOEFFDING inequality would no longer provide a useful bound.

Remark. The ball $\{x \in \mathbb{R}^n : \sum_i \mu_i |x_i|^p \le R^p\}$ lies inside $[-R/\kappa, R/\kappa]^n$, where $\kappa := \min_i \mu_i \le 1/n$. but a factor of at least nR would render the HOEFFDING bound rather useless.

The trick is to work with "weak-$\mathcal{L}^q$ norms" for $q = 3p/2$. For a general probability space $(\mathbb{A}, \mathcal{A}, \mu)$ and an $\mathcal{A}$-measurable real-valued function f on $\mathbb{A}$ define

<24> $$\|f\|_{q,\infty,\mu} := \left(\sup_{t>0} t^q \mu\{|f| > t\}\right)^{1/q}.$$

The set $\mathcal{L}_{q,\infty,\mu} := \{f : \|f\|_{q,\infty,\mu} < \infty\}$ is a vector space. Despite the use of the familar symbol for a norm, $\|\cdot\|_{q,\infty,\mu}$ is ***not a norm*** on $\mathcal{L}_{q,\infty,\mu}$. See Problem [13]. However (Garling, 2007, §10.2) for $q > 1$ there does exist a true norm ρ (or semi-norm, if we don't work with μ-equivalence classes) on $\mathcal{L}_{q,\infty,\mu}$ for which $(1 - 1/q)\rho(f) \le \|f\|_{q,\infty,\mu} \le \rho(f)$. Maybe $\|\cdot\|_{q,\infty,\mu}$ should be called a ***quasi-norm*** to remind us not to invoke a triangle inequality.

Quasi-norms provide a convenient way to control $\mathcal{L}^p(\mu)$ and $\mathcal{L}^p(\mu_m)$ norms, by means of arguments involving tail probabilities. As before, μ_m denotes the empirical measure for a sample $\mathcal{I} = (\mathcal{I}_1, \ldots, \mathcal{I}_m)$ from μ;

$$\mu_m x := m^{-1} \sum\nolimits_{k \in [\![m]\!]} x[\mathcal{I}_k] \qquad \text{for } x \in \mathbb{R}^n.$$

The following non-trivial facts about general quasi-norms—where x is replaced by a general measurable function f and the $\mathcal{I}_k$'s are replaced by independent observations on a general μ—can be proved by means of standard probabilistic tricks. They will be invoked with $p \ge 1$ and $q = 3p/2$ for the Proof of Theorem <10>.

- $\|\lambda f\|_{q,\infty,\mu} = \lambda \|f\|_{q,\infty,\mu}$ for each constant $\lambda > 0$ and

<25> $$b_{p,q} \|f\|_{p,\mu} \le \|f\|_{q,\infty,\mu} \le \|f\|_{q,\mu} \le \|f\|_{2p,\mu} \qquad \text{if } 1 \le p < q < 2p,$$

  where $b_{p,q} := (1 - p/q)^{1/p}$. Analogous inequalities (with the same $b_{p,q}$) hold with μ replaced by μ_m. See Problem [14].

- There exists a universal constants c_1, c_2 for which: If g is a measurable function on $\mathbb{A}$ with $1 \ge \|g\|_{2p,\mu} > 0$ for some $p \ge 1$ and $q := 3p/2$, then

<26> $$\mathbb{P}\{\|g\|_{q,\mu_m} \le \|g\|_{p,\mu}/8\} \le e^{-c_1 m (\|g\|_{p,\mu}/4)^{6p}} \qquad \text{see Problem [15]}$$

<27> $$\mathbb{P}\{\|g\|_{q,\mu_m} \le c_2\} > 3/4 \qquad \text{see Problem [16].}$$

When μ lives on $[\![n]\!]$ then $\|\cdot\|_{p,\mu}$ means the same as $|\cdot|_{p,\mu}$, and so on.

15.7 Packing for subsets of an ℓ^q ball

Recall that the Theorem asserts that

<28> $$\log \text{PACK}(t, \mathcal{X}, \ell^p_\mu) \le C_1 p^2 \text{SDIM}(C_2 t, \mathcal{X}) \log(C_3 R/t) \qquad \text{for } t > 0.$$

The Proof will be similar to the one in Section 15.5 except that we no longer have a useful upper bound on $\max_i |x[i]|$ for $x \in \mathcal{X}$. For a given t, we start from a t-separated set $\{x_\alpha : \alpha \in [\![N]\!]\}$ of $\mathcal{X}$. That is,

$$|x_\alpha - x_\beta|_{p,\mu} > t \quad \text{AND} \quad \max_\alpha |x_\alpha|_{2p,\mu} \le R \qquad \text{for } 1 \le \alpha < \beta \le N.$$

We need to find universal constants for which $\log N \le C_1 p^2 d \log(C_3 R/t)$, with $d := \text{SDIM}(C_2 t, \mathcal{X})$, a result that holds trivially if $\log N \le d$. Thus we can assume $\log N > d$.

The plan of attack involves three steps:

1. For some m (to be chosen) let μ_m denote the empirical measure defined by a vector $\mathcal{I} = (\mathcal{I}_1, \ldots, \mathcal{I}_m)$ of independent observations from μ. For a given $t > 0$ we hope to find a realization of $\mathcal{I}$ for which there exists a subset J of $[\![N]\!]$, with $|J| \ge N/2$ such that

   $$\{x_\alpha[\mathcal{I}] : \alpha \in J\} \text{ is } t/8\text{-separated in } \mathcal{L}^q_{\mu_m}.$$

   When suitably rescaled this set will give a 10-separated set $\mathcal{W}$ in $\mathbb{R}^m$ for which $\text{SDIM}(\cdot, \mathcal{W})$ can be bounded using $\text{SDIM}(\cdot, \mathcal{X})$. Lemma <14> will then give $1 + \#\mathbb{L}(\mathcal{W}) \ge |N/2|^{1/2q}$.

2. Find an upper bound for $\#\mathbb{L}(\mathcal{W})$ by using upper bounds for the LEBESGUE measure of ℓ^q balls in various dimensions, as in <13>.

3. Combine the bounds from the first two steps into an inequality that can be inverted, as for Lemma <6>, to get the upper bound for $\log N$.

Step 1: reduction to $\mathbb{R}^m$.
With $g = (x_\alpha - x_\beta)/(2R)$ for some $\alpha < \beta$, inequalities <26> and $|g|_{p,\mu} > t/(2R)$ give

$$\mathbb{P}\{|x_\alpha - x_\beta| q, \mu_m \le t/8\} \le \exp\left(-c_1 m (t/8R)^{6p}\right).$$

A union bound over $\binom{N}{2}$ pairs then gives

$$\mathbb{P}\{\exists \alpha < \beta : |x_\alpha - x_\beta|_{q,\mu_m} \le t/8\} < \tfrac{1}{2} \exp(2 \log N - c_1 m (t/8R)^{6p}).$$

The right-hand side is $< 1/2$ if

<29> $$m := \lceil (2 \log N)(8R/t)^{6p}/c_1 \rceil.$$

Put another way, $\mathbb{P}G > 1/2$ for $G := \cap_{\alpha<\beta}\{|x_\alpha - x_\beta|_{q,\mu_m} > t/8\}$.

In order to show that a good number of x_α's lie in some $\ell^q_{\mu_m}$ ball define

$$J := J(\mathfrak{I}) := \{\alpha \in [\![N]\!] : |x_\alpha|_{q,\mu_m} \le c_2 R\},$$
$$Z(\mathfrak{I}) := N^{-1}\sum\nolimits_{\alpha\in[\![N]\!]}\{|x_\alpha|_{q,\mu_m} \le c_2 R\}.$$

Note that $0 \le Z \le |J|/N \le 1$ and $\mathbb{P}Z > 3/4$, by multiple appeals to <27> with $g = x_\alpha/R$. The elementary inequality

$$3/4 < \mathbb{P}Z \le \mathbb{P}\{Z \ge 1/2\} + (1/2)\mathbb{P}\{Z < 1/2\}$$

implies $\mathbb{P}\{Z \ge 1/2\} > 1/2$, so that $\mathbb{P}G \cap \{Z \ge 1/2\} > 0$. (Compare with: if $\min(\mathbb{P}A, \mathbb{P}B) < 1/2$ then $\mathbb{P}A \cup B < 1$.) The intersection $G \cap \{Z \ge 1/2\}$ is non-empty. Thus there exists some realization of $\mathfrak{I}$ for which $J(\mathfrak{I}) \ge N/2$ and

<30> $|x_\alpha|_{q,\mu_m} \le c_2 R$ for $\alpha \in J(\mathfrak{I})$,

<31> $|x_\alpha - x_\beta|_{q,\mu_m} > t/8$ for all $\alpha < \beta$ with $\alpha, \beta \in J(\mathfrak{I})$.

From now on $\mathfrak{I}$ and J stay fixed.

Step 2: counting lattice cubes.
For each α in J define $w_\alpha := (80/t)x_\alpha[\mathfrak{I}] \in \mathbb{R}^{\mathfrak{I}}$. That is, $w_\alpha[k] := (80/t)x_\alpha[\mathfrak{I}_k]$ for each k in $[\![m]\!]$. Define $\mathcal{W} := \{w_\alpha : \alpha \in J\}$.

Let ν denote the uniform distribution on $[\![m]\!]$. Then for distinct α, β in J,

$$|w_\alpha - w_\beta|^q_{q,\nu}/(80/t)^q = m^{-1}\sum\nolimits_k |x_\alpha[\mathfrak{I}_k] - x_\beta[\mathfrak{I}_k]|^q = |x_\alpha - x_\beta|^q_{q,\mu_m} > (t/8)^q.$$

Thus $|w_\alpha - w_\beta|_{q,\nu} > 10$. Similarly

$$|w_\alpha|_{q,\nu} = (80/t)|x_\alpha|_{q,\mu_m} \le \lambda := c_2 80R/t.$$

That is, $\mathcal{W} \subset B := \{y \in \mathbb{R}^m : |y|_{q,\nu} \le \lambda\} = \{y \in \mathbb{R}^m : \sum_k |y_k|^q \le m\lambda^q\}$, the ℓ^q ball in $\mathbb{R}^m$ of radius $m^{1/q}\lambda$.

Consider a nonempty subset K of $[\![m]\!]$, a pattern $u \in \{0,1\}^K$, and an $r \in \mathbb{R}^K$. If w_α picks out the pattern u around r at level 1 then x_α picks out pattern u around $(t/80)r$ at level $t/80$. Thus $\text{SDIM}(1, \mathcal{W}) \le d := \text{SDIM}(t/80, \mathcal{X})$.

For $\emptyset \ne K \subseteq [\![m]\!]$ the projection of B onto the $\mathbb{R}^K$ coordinate subspace is also an ℓ^q ball of radius λ, which has $|K|$-dimensional LEBESGUE measure equal to $\lambda^{|K|}V_{|K|}$ where, by Problem [9], $V_k \le 2^k c_0 (qe/k)^{k/q}$. It follows by <13> and Lemma <14> that

$$1 + \sum\nolimits_{1\le k\le d}\binom{m}{k}\lambda^k 2^k c_0 (qe/k)^{k/q} \ge 1 + \#\mathbb{L}(\mathcal{W}) \ge (N/2)^{1/2q}$$

$$
\begin{aligned}
(N/2)^{1/2q} &\le 1 + \#\mathbb{L}(\mathcal{W}) \\
&\le 1 + \sum\nolimits_{k=1}^{d} \binom{m}{k} \lambda^k V_k \\
&\le \lambda^d \beta(m,d) V_d \qquad \text{with } \beta(m,d) \le (em/d)^d \text{ as in <4>} \\
&\le \Big(80 c_2 R/t \times (em/d) \times 2 c_0^{1/d} (qe/d)^{1/q} \Big)^d .
\end{aligned}
$$

From <29> we have $m/d \le Cq(\log N^{1/2qd})(8R/t)^{6p}$, for some universal constant C. Consequently, for some other universal constants $\kappa_1, \kappa_2, \ldots$,

$$N^{1/2qd} \le \kappa_1 (R/t)(m/d) \le \kappa_2 (\log N^{1/2qd})(8R/t)^{7p+1}$$

Equivalently, if $A := \log N^{1/2qd}$ then $e^A/A \le \kappa_3 (8R/t)^{7p+1}$. As asserted in §14.2 (and proved in the Problems for Chapter 14), such an inequality implies $e^A \le \kappa_4 (8R/t)^{16p}$ and

$$\log N = 2qdA \le \log \kappa_4 + 32pqd \log(8R/t) \le C_1 p^2 d \log(8R/t),$$

as needed for <28>.

*15.8 Chaining bounds via surround dimension

As you saw in §10.6, the integral $\int_0^\delta \mathbb{K}(t)\, dt$ with

$$\mathbb{K}^2(t) = \log \textsc{pack}(t, T, d) \qquad \text{or} \qquad \mathbb{K}^2(t) = \log \textsc{cover}(t, T, d)$$

plays a central role in establishing maximal inequalities for stochastic processes $\{X_t : t \in T\}$ with subgaussian increments controlled by a metric d on T. In particular, those methods apply to signed-sum processes indexed by bounded subsets $\mathcal{X}$ of $\mathbb{R}^n$, with ℓ_μ^2 providing the metric.

Theorem <10>, for general p, provides a way of bounding ℓ_μ^p analogs of $\mathbb{K}(t)$. This assertion might seem implausible because, typically, applications involving $\mathcal{L}_\mu^2$ do not come with a some suitably small $\mathcal{L}_\mu^4$ ball containing $\mathcal{X}$.

A neat counterintuitive trick (Rudelson and Vershynin, 2006, §6) overcomes that deficiency. Suppose we wish to control

$$\mathbb{K}(t,p) := \big(\log \textsc{cover}(t, \mathcal{X}, \ell_\mu^p) \big)^{1/2} \qquad \text{for some } p \ge 1 \text{ and } t > 0$$

by means of the function $\mathbb{S}(t) := \textsc{sdim}(C_2 t, \mathcal{X})^{1/2}$. The idea is:

(i) Cover $\mathcal{X}$ by $\cup_{j \le N} B_j$ where B_j is an ℓ_μ^{2p} ball of radius $3t$ and $N := \textsc{cover}(3t, \mathcal{X}, \ell_\mu^{2p})$.

(ii) Then cover each $\mathcal{X} \cap B_j$ by a union of $N_j := \textsc{cover}(t, \mathcal{X} \cap B_j, \ell_\mu^p)$ balls of ℓ_μ^p radius t.

Remark. For the moment, do not worry about the use of the unknown $\text{COVER}(3t, \mathcal{X}, \ell_\mu^{2p})$. It will soon be replaced by another unknown quantity.

By Theorem <10>,

$$\log N_j \le C_1 p^2 \text{SDIM}(C_2 t, \mathcal{X} \cap B_j) \log(3C_3) \le C^2 p^2 \mathbb{S}(t)^2,$$

for some new universal constant C. Together (i) and (ii) produce a covering of $\mathcal{X}$ by at most NM balls having ℓ_μ^p radius t, where $M := \max_j N_j$.

We now have a very neat recursive bound:

$$\mathbb{K}(t,p) \le \sqrt{\log N + \log M} \le \mathbb{K}(3t, 2p) + Cp\mathbb{S}(t) \qquad \text{for } t > 0. \tag{32}$$

When iterated, this inequality can provide a bound on $\sum_i \delta_i \mathbb{K}(\delta_i, p)$, the sort of sum that was important in §10.6.

For a given $\delta_0 > 0$ and $\theta = 1/3$ and $p \ge 1$ define $\delta_i := \delta_0 \theta^i$ and $p_i := 2^i p$. Then, for $i \ge 1$,

$$\begin{aligned}
\mathbb{K}(\delta_i, p) &\le \mathbb{K}(\delta_{i-1}, p_1) + Cp\mathbb{S}(\delta_i) \\
\mathbb{K}(\delta_{i-1}, p_1) &\le \mathbb{K}(\delta_{i-2}, p_2) + Cp_1\mathbb{S}(\delta_{i-1}), \\
\mathbb{K}(\delta_{i-2}, p_2) &\le \mathbb{K}(\delta_{i-3}, p_3) + Cp_2\mathbb{S}(\delta_{i-2}),
\end{aligned}$$

and so on. Repeated substitutions then give

$$\mathbb{K}(\delta_i, p) \le \mathbb{K}(\delta_0, p_i) + C \sum\nolimits_{j=0}^{i-1} p_j \mathbb{S}(\delta_{i-j}) \tag{33}$$

and, for each m in $\mathbb{N}$,

$$\begin{aligned}
\sum_{i=0}^m \delta_i \mathbb{K}(\delta_i, p_i) &\le \sum_{i=0}^m \delta_i \mathbb{K}(\delta_0, p_i) + \sum_{i=0}^m \sum_{j=0}^m \{j < i\} \delta_i p_j \mathbb{S}(\delta_{i-j}) \\
&\le \delta_0 \sum_{i=0}^m \theta^i \mathbb{K}(\delta_0, p_i) + \sum_{j=0}^m p_j \delta_j \sum\nolimits_{i=j+1}^m \delta_{i-j} \mathbb{S}(\delta_{i-j}). \qquad \text{<34>}
\end{aligned}$$

Inequality <34> can be simplified in various ways.

- The fact that $|x|_{p,\mu}$ is an increasing function of p for each fixed x shows that

$$\sum\nolimits_{i=0}^m \delta_i \mathbb{K}(\delta_i, p_i) \ge \int_{\delta_m}^{\delta_0/3} \sqrt{\log \text{COVER}(t, \mathcal{X}, \ell_\mu^p)}\, dt.$$

The final integral is the kind of quantity that appears in chaining bounds for the signed-sum process obtainable by the methods of §10.6.

- The final sum in <34> equals

$$\delta_0 p \sum_{j=0}^{m} (2/3)^j \Big(\delta_1 \mathbb{S}(\delta_1) + \cdots + \delta_{m-j} \mathbb{S}(\delta_{m-j}) \Big).$$

  which is less than

$$\tfrac{3}{2}\delta_0 p \Big(\delta_1 \mathbb{S}(\delta_1) + \cdots + \delta_m \mathbb{S}(\delta_m) \Big) \le \tfrac{3}{2}\delta_0 p \int_{\delta_m}^{\delta_0} \mathbb{S}(t)\, dt.$$

- If $\mathcal{X} \subset [-R_0, R_0]^n$ and $\delta_0 = 2R_0$ then $|x - y|_{p,\mu} \le 2R_0$ for all $x, y \in \mathcal{X}$, which would imply $\mathbb{K}(\delta_0, p_i) = 0$.

- Note that $\delta_0 \sum_{i=0}^{m} \theta^i \mathbb{K}(\delta_0, p_i) \le \mathbb{K}(\delta_0, p) \le \frac{3}{2}\delta_0 \mathbb{K}(\delta_0, p)$.

Clearly, integrability of $\text{SDIM}(\,\cdot\,, \mathcal{X})$ could tell us useful about signed-sum processes indexed by $\mathcal{X}$.

15.9 Problems

[1] Suppose $\mathcal{X} \subset \mathbb{R}^n$ and $(r, u) \in \mathbb{R}^n \times \{-1, +1\}^n$ and $I := (i_1, \ldots, i_k) \in [\![n]\!]^k$ with $0 < k \le n$. For $J := (i_1, \ldots, i_m)$, with $0 < m < k$, suppose $\mathcal{X}[J]$ cannot pick out pattern $u[J]$ around $r[J]$. Show that $\mathcal{X}[I]$ cannot pick out pattern $u[I]$ around $r[I]$.

[2] The ***subgraph*** of a real-valued function f defined on some set $\mathbb{A}$ is defined as $\text{subgraph}(f) := \{(a, r) \in \mathbb{A} \times \mathbb{R} : f(a) \ge r\}$. A collection $\mathbb{F} := \{f_t : t \in T\}$ of such functions is called a ***VC-subgraph class*** if $\mathcal{D} := \{\text{subgraph}(f_t) : t \in T\}$ has finite VC-dimension.

Suppose $\text{VCDIM}(\mathcal{D}) = d < \infty$ and $A := (a_1, \ldots, a_n) \in \mathbb{A}^{[\![n]\!]}$ with $n \ge d$. Define $\mathcal{X}_a := \{(f_t[a_1], \ldots, f_t[a_n]) : t \in T\}$. Show that $\text{SDIM}(\mathcal{X}_a) \le d$. Hint: If $I \subset [\![n]\!]$ and $r \in \mathbb{R}^I$ then $\mathcal{X}_a[I]$ surrounds r iff for each u in $\{-1, +1\}^I$ there exists a t in T with $\min_{\alpha \in I} u[\alpha]\,(f_t[a_\alpha] - r[\alpha]) > 0$.

[3] For each pair of real numbers a, b define a real-valued function $f_{a,b}(y) := \min\big(|y - a|^2, |y - b|^2\big)$ for $y \in \mathbb{R}$. Then $\text{subgraph}(f_{a,b}) = D_a \cap D_b$ where, for each real c,

$$D_c := \{(y, r) \in \mathbb{R}^2 : |y - c|^2 \ge r\} = \{(y, r) \in \mathbb{R}^2 : |y|^2 - 2cy + c^2 - r \ge 0\}.$$

If you hunt through §14.2 and the Problems for Chapter 14 you will learn how to show that $\{D_c : c \in \mathbb{R}\}$ is a VC-class of sets and $\{f_{a,b} : a, b \in \mathbb{R}\}$ is a VC-subgraph class of functions. Similar results hold for the minimum of k quadratic functions on higher-dimensional euclidean spaces.

[4] Suppose $z_i = (x_i, r_i)$ for $i = 1, \ldots, n$ are points in $\mathcal{X} \times \mathbb{R}$ for which $x_1 = x_2$. Show that no set of functions can surround z. Hint: Without loss of

generality, $r_2 \geq r_1$. Can there exist a function f for which $r_1 > f(x_1)$ and $r_2 < f(x_2)$?

[5] Show that $\text{SDIM}(\mathcal{F}) = \sup_{\epsilon>0} \text{SDIM}(\epsilon, \mathcal{F})$.

[6] The traditional example where $\text{SDIM}(\,\cdot\,, \mathcal{X})$ has something to offer is the set of all $[0,1]$-valued non-decreasing functions on $\mathbb{R}$. When translated into the setting of Section 15.3 the example becomes: let $\mathcal{X}$ denote the set of all $x = (x_i : i \in [\![n]\!])$ in $[0,1]^n$ for which $0 \leq x_1 \leq x_2 \leq \cdots \leq x_n \leq 1$.

(i) Suppose $I = (i_1, \ldots, i_d)$ with $1 \leq i_1 < \cdots < i_d \leq n$ and $\mathcal{X}[I]$ surrounds $r = (r[i] : i \in I)$ at level δ. Show that there exists an x in $\mathcal{X}$ for which

$$r[i_\alpha] + \delta/2 \leq x[i_\alpha] \leq x[i_{\alpha+1}] \leq r[i_{\alpha+1}] - \delta_2 \qquad \text{for most } \alpha.$$

Deduce that $0 \leq r[i_\alpha] + \delta \leq r[i_{\alpha+1}]$ and hence $d \leq 1 + 1/\delta$. Thus $\text{SDIM}(\delta, \mathcal{X}) \leq 2/\delta$ for $0 < \delta < 1$.

(ii) Show that $\text{SDIM}(\mathcal{X}) = n$. Hint: Suppose $r_1 < r_2 < \cdots < r_n \leq 1$.

[7] Suppose $\mathcal{H}$ is a k-dimensional subspace of $\mathbb{R}^n$ and $\mathcal{X} \subset \mathbb{R}^n$ with the property: for each x in $\mathcal{X}$ there is an h in $\mathcal{H}$ for which $\max_{\alpha\in[\![n]\!]} |x_\alpha - h_\alpha| < \delta/2$. Argue as follows to show that $\text{SDIM}(\delta, \mathcal{X}) \leq k$.

(i) Suppose $\mathcal{X}[I]$ surrounds some r in $\mathbb{R}^I$ at level δ, for some I in $[\![n]\!]^m$. Then for each u in $\{-1, +1\}^I$ there exists an x in $\mathcal{X}$ for which $\min_\alpha u_\alpha(x_\alpha - r_\alpha) \geq \delta/2$. Find an h in $\mathcal{H}$ for which $\epsilon := \max_{\alpha\in[\![n]\!]} |x_\alpha - h_\alpha| < \delta/2$. Then $\min_\alpha(h_\alpha - r_\alpha) \geq \delta/2 - \epsilon > 0$.

(ii) Thus $\mathcal{H}[I]$ surrounds r in the sense of Definition <2>. It follows from Example <3> that $m = |I| \leq k$.

[8] There in an important difference between the ℓ^p and ℓ^p_μ norms on $\mathbb{R}^n$, for $p \geq 1$ and μ a probability measure on $[\![n]\!]$.

(i) Recall that $|x|_p := (\sum_i |x_i|^p)^{1/p}$. Show that $|x|_p \geq |x|_q$ for $q > p$. Hint: If $|x|_p = R$ then $\sum_i |x_i/R|^p \leq 1$, which implies $\sum_i |x_i/R|^q \leq 1$.

(ii) Recall that $|x|_{p,\mu} := (\sum_i \mu_i |x_i|^p)^{1/p}$. Show that $|x|_{p,\mu} \leq |x|_{q,\mu}$ for $q > p$. Hint: JENSEN.

(iii) Compare with: If $\mu\{i\} = 1/n$ for each i in $[\![n]\!]$ then $|x|_{p,\mu} = n^{-1/p}|x|_p$.

[9] (Based on Pisier, 1989, page 11.) Define a norm on $\mathbb{R}^n$ by $\|x\| := (\sum_i \sigma_i |x_i|^p)^{1/p}$ where the σ_i's are strictly positive constants and $p \geq 1$. Denote the corresponding unit ball, $\{x \in \mathbb{R}^n : \|x\| \leq 1\}$, by B. Remember that $\mathfrak{m}_n$ denotes LEBESGUE measure on $\mathbb{R}^n$.

(i) Show that $e^{-z} = \int_0^\infty e^{-r}\{z \leq r\}\, dr$ for $z \geq 0$.

(ii) Show that $\mathfrak{m}_n e^{-\|x\|^p} = \int_0^\infty e^{-r}\mathfrak{m}_n\{\|x\| \leq r^{1/p}\}\, dr = (\mathfrak{m}_n B)\Gamma(1 + n/p)$.

(iii) Show that

$$\mathfrak{m}_n e^{-\|x\|^p} = \prod_i 2\int_0^\infty e^{-\sigma_i|x_i|^p}dx_i = 2^n\prod_i \sigma_i^{-1/p}\Gamma(1+1/p).$$

(iv) Deduce that $\mathfrak{m}_n B = (\sigma_1\dots\sigma_n)^{-1/p}(2\Gamma(1+1/p))^n/\Gamma(1+n/p)$.

(v) Show that $\Gamma(x+1) \geq (x/e)^x/c_0$, where $c_0 = e/\sqrt{2\pi} \approx 1.085$, for all $x \geq 0$. Hint: By the continuous analog of the STIRLING formula (§2.5),

$$\log\Gamma(x+1) \geq f(x) := (x+1/2)\log(1+x) - (1+x) + \log\sqrt{2\pi}.$$

Define $g(x) := f(x) - x\log(x/e) + \log c_0 = (x+1/2)\log(1+x) - x\log x$. Show that $g(0) = 0$ and $\dot{g}(x) > 0$ for $x > 0$. (The c_0 is just an artifact of using a lower bound for $\Gamma(x+1)$. Numerical evidence shows that it could be replaced by 1 or even by a constant slightly smaller than 1.)

(vi) Use (v) and the fact that $\Gamma(x) \leq 1$ for $1 \leq x \leq 2$ to deduce that

$$(\sigma_1\dots\sigma_n)^{1/p}\mathfrak{m}_n B \leq 2^n/\Gamma(1+n/p) \leq 2^n c_0(pe/n)^{n/p}.$$

*[10] Prove the following assertions then explain how they are related.

(i) $\mathfrak{m}_n\{(x_1,\dots,x_n) \in \mathbb{R}^n : \sum_i |x_i| \leq 1\} = 2^2/n!$. Hint: previous Problem.

(ii) For independent random variables $U_1,\dots,U_n$, each distributed UNIF$(0,1)$ we have $\mathbb{P}\{U_1 < U_2 < \cdots < U_n\} = 1/n!$.

(iii) The $\mathfrak{m}_n$ measure of a pyramid in $\mathbb{R}^n$ having height h and base B is equal to $n^{-1}h\mathfrak{m}_{n-1}B$.

(iv) The unit cube $[0,1]^n$ has $2n$ faces.

[11] Lemma <20> is equivalent to the following result. Let Y be a real random variable with median 0. If $\theta := 1/(2p)$ for some constant $p \geq 1$ and

$$(\star) \qquad \mathbb{P}\{Y \leq \lambda - 1\}^\theta + \mathbb{P}\{Y \geq \lambda\}^\theta \leq 1 \qquad \text{for each } \lambda \text{ in } \mathbb{R}.$$

then $\mathbb{P}|Y|^p \leq 2.5$. Argue as follows.

(i) For each k in $\mathbb{N}$, use $(\star)$ to show that

$$\mathbb{P}\{Y \geq k\}^\theta \leq 1 - \mathbb{P}\{Y \leq k-1\}^\theta \leq \mathbb{P}\{Y > k-1\} \leq \mathbb{P}\{Y \geq k-1\}.$$

(ii) Deduce that

$$\mathbb{P}\{Y \geq k\} \leq \mathbb{P}\{Y \geq k-1\}^{2p} \leq \mathbb{P}\{Y \geq k-2\}^{(2p)^2} \leq \cdots \leq \mathbb{P}\{Y > 0\}^{(2p)^k}.$$

(iii) Note that, by definition of a median, $\min\left(\mathbb{P}\{Y \geq 0\}, \mathbb{P}\{Y \leq 0\}\right) \geq 1/2$. Deduce that $\max\left(\mathbb{P}\{Y < 0\}, \mathbb{P}\{Y > 0\}\right) \leq 1/2$.

(iv) Argue similarly with Y replaced by $-Y$ then add the two tail bounds to conclude that $\mathbb{P}\{|Y| \geq k\} \leq 2(1/2)^{(2p)^k}$ for $k \in \mathbb{N}$.

(v) Show that

$$\begin{aligned}\mathbb{P}|Y|^p &\le \mathbb{P}\{|Y| \le 1\} + \sum\nolimits_{k\in\mathbb{N}} (k+1)^p \mathbb{P}\{k < |Y| \le k+1\} \\ &\le 1 + 2\sum\nolimits_{k\in\mathbb{N}} \exp\left(p\log(k+1) - (2p)^k \log(2)\right).\end{aligned}$$

(vi) Show that the kth term in the last sum is a decreasing function of p, for $p \ge 1$. Deduce that

$$\mathbb{P}|Y|^p \le 1 + 2\sum\nolimits_{k\in\mathbb{N}} \exp\left(\log(k+1) - 2^k \log(2)\right) \le 2.5.$$

[12] Suppose S has a $\text{BIN}(m,\delta)$ distribution. Use the the BENNETT inequality (or the tail bounds from §3.7) to show that

$$\max\Big(\mathbb{P}\{S - m\delta \ge mt\}\,,\, \mathbb{P}\{S - m\delta \le -mt\}\Big) \le e^{-Cmt} \qquad \text{if } 2t \ge \delta(1-\delta),$$

for $C := \psi_{\text{benn}}(1/2)/4 \approx 0.216$.

> For the final four Problems, $(\mathbb{A}, \mathcal{A}, \mu)$ is a probability space and μ_m denotes the empirical measure defined by independent $\xi_1, \ldots, \xi_m$ observations on μ. That is, $\mu_m f := m^{-1}\sum_i f(\xi_i)$. When specialized to the case $\mathbb{A} = [\![n]\!]$ these Problems cover some of the probability details omitted from Section 15.7.

[13] Suppose $\mathbb{A} = \{0,1\}$ and $\mu\{0\} = \theta = 1 - \mu\{1\}$ for some θ in $(0,1)$. If $f(x) := (1+c)\{x=0\} + \{x=1\}$ and $g(x) := c\{x=1\}$ for some $c > 0$ then

$$\|f\|_{q,\infty,\mu} = \max\left(1, (1+c)\theta^{1/q}\right) \quad \text{AND} \quad \|g\|_{q,\infty,\mu} = c(1-\theta)^{1/q},$$

so that

$$\|f+g\|\, q,\infty,\mu = 1 + c > \|f\|_{q,\infty,\mu} + \|g\|_{q,\infty,\mu} \qquad \text{if } \frac{1+c}{c} > \frac{(1-\theta)^{1/q}}{1-\theta^{1/q}},$$

which happens for θ close enough to 0. Thus $\|\cdot\|\, q,\infty,\mu$ is not a norm.

[14] For a measurable f on $\mathbb{A}$ and $q \ge 1$ the $\mathcal{L}^q(\mu)$ quasi-norm is defined as $\|f\|\, q,\infty,\mu := (\sup_{t>0} t^q \mu\{|f| > t\})^{1/q}$.

(i) Show $\|\lambda f\|_{q,\infty,\mu} = \lambda \|f\|_{q,\infty,\mu}$ for each positive constant λ.

(ii) For $1 \le p < q < 2p$ show that

$$b_{p,q}\|f\|_{p,\mu} \le \|f\|_{q,\infty,\mu} \le \|f\|_{q,\mu} \le \|f\|_{2p,\mu} \qquad \text{where } b_{p,q} := (1 - p/q)^{1/p}.$$

Hint: By (i), there is no loss of generality in assuming $\|f\|_{q,\infty,\mu} = 1$, in which case $\mu|f|^p = \int_0^\infty pt^{p-1}\mu\{|f| > t\}\,dt \le 1 + \int_1^\infty pt^{p-1}t^{-q}dt$.

[15] Suppose g is a measurable function on $\mathbb{A}$ for which $1 \geq \|g\|_{2p,\mu} > 0$ for some $p \geq 1$. Show that there exists a universal constant C for which

$$\mathbb{P}\{\|g\|_{q,\mu_m} \leq \|g\|_{p,\mu}/2\} \leq e^{-Cm(\|g\|_{p,\mu}/4)^{6p}} \qquad \text{for } q := 3p/2$$

by justifying the following assertions.

(i) To simplify notation define $\theta := \|g\|_{p,\mu}/4$. By Problem [14],

$$(\sup_{t>0} t^q \mu\{|g| > t\})^{1/q} \geq b_{p,q} 4\theta > \theta \qquad \text{because } b_{p,q} = (1/3)^{1/p} > 1/4.$$

Thus there exists some $t_0 > 0$ for which $\delta := \mu\{|g| > t_0\} > (\theta/t_0)^q$.

(ii) The inequality $1 \geq \mu|g|^{2p} \geq t_0^{2p}\mu\{|g| > t_0\}$ gives $t_0^{-2p} \geq \delta$. Combine with the lower bound from (ii) to get $1/t_0 > \theta^3$ and $\delta > \theta^{4q} = \theta^{6p}$.

(iii) If $\theta/2 \geq \|g\|_{q,\mu_m}$ then $(\theta/2)^q \geq \mu_m|g|^q \geq t_0^q \mu_m\{|g| > t_0\}$ and, by (ii),

$$\delta/2 > \delta/2^q > (\theta/2t_0)^q \geq \mu_m\{|g| > t_0\}.$$

Consequently, $\mathbb{P}\{\|g\|_{q,\mu_m} \leq \theta/2\} \leq \mathbb{P}\{m\mu_m\{|g| > t_0\} \leq m\delta/2\}$. As the sum $m\mu_m\{|g| > t_0\}$ has a $\text{Bin}(m,\delta)$ distribution, it follows by Problem [12] (with $t = \delta/2$) that

$$\begin{aligned}\mathbb{P}\{\|g\|_{q,\mu_m} \leq \theta/2\} &\leq \mathbb{P}\{m\mu_m\{|g| > t_0\} \leq m\delta - m\delta/2\} \\ &\leq \exp(-Cm\delta/2) \leq \exp(-Cm\theta^{6p}/2),\end{aligned}$$

as hoped for.

[16] Again suppose g is a measurable function on $\mathbb{A}$ with $1 \geq \|g\|_{2p,\mu} > 0$ for some $p \geq 1$ and $q = 3p/2$. Show that there is a universal constant C for which $\mathbb{P}\{\|g\|_{q,\mu_m} \leq C\} > 3/4$ by justifying the following assertions.

(i) For each j in $\mathbb{N}$ define

$$t_j := 2^j, \quad G_j := \{|g| > t_j\}, \quad \delta_j := \mu G_j, \quad \theta_j := t_j^{-2p} \leq (1/4)^j.$$

The inequality $1 \geq \mu|g|^{2p}$ implies $\delta_j \leq \theta_j$.

(ii) Define $J_m := \max\{j \in \mathbb{N} : \theta_j \geq (8m)^{-1}\}$ and $A_m := \{\mu_m G_{J_m+1} = 0\}$ for $m \in \mathbb{N}$. Then $\delta_{J_m+1} \leq \theta_{J_m+1} < (8m)^{-1}$, which implies

$$\mathbb{P}A_m = (1 - \delta_{J_m+1})^m \geq \left(1 - (8m)^{-1}\right)^m \geq 7/8 \qquad \text{for all } m.$$

(iii) If $D \geq 1/2$, Problem [12] gives

$$\begin{aligned}&\sum_{j=1}^J \mathbb{P}\{\mu_m G_j \geq (1+D)\theta_j\} \\ &\quad\leq \sum_{j=1}^J \exp\left(-C_{[12]} mD\theta_j\right) \leq \sum_{j=1}^J \exp\left(-C_{[12]} D 4^{p(J-j)} m\theta_J\right) \\ &\quad\leq \sum_{i\in\mathbb{N}_0} \exp\left(-4^i C_{[12]} D/8\right) \qquad \text{because } m\theta_J \geq 1/8 \text{ and } p \geq 1.\end{aligned}$$

If D is large enough, the final sum is $< 1/8$. (**R** says $D = 77$ suffices.)

(iv) Define $B_m := A_m \cap \bigcap_{j=1}^{J_m} \{\mu_m G_j \le (1+D)\theta_j\}$. The results from (ii) and (iii) show that

$$\mathbb{P}B_m^c \le \mathbb{P}A_m^c + \sum_{j=1}^{J_m} \mathbb{P}\{\mu_m G_j \ge (1+D)\theta_j\} < 1/4.$$

Thus $\mathbb{P}B_m > 3/4$. On the event B_m we have

$$\begin{aligned}
\mu_m g^q &= \int_0^\infty q t^{q-1} \mu_m\{g \ge t\}\, dt \\
&\le t_1^q + \sum_{j=1}^{J_m} \int_{t_j}^{t_{j+1}} q t^{q-1} \mu_m\{g > t_j\}\, dt = t_1^q + \sum_{j=1}^{J_m} t_{j+1}^q \mu_m G_j \\
&\le 2^q + \sum_{j=1}^{J_m} 2^{(j+1)q}(1+D)2^{-2pj} \\
&\le C^q := 2^q \left(1 + (1+D)\sum_{j\in\mathbb{N}} 2^{-pj/2}\right).
\end{aligned}$$

R says $C < 66$ if $D = 77$.

15.10 Notes

Precursors for many of the ideas explained in this Chapter appeared in a pathbreaking paper by Vapnik and Chervonenkis (1981). These include: maximal inequalities via symmetrization bounds; approximation using ℓ^∞ covering numbers; bounding of packing numbers by volumes of convex sets; and sampling to reduce dimension.

More specifically, much of the Chapter could be thought of as an extension of Lemma 7.13 of Dudley (1978) for VC-classes of sets $\mathcal{D}$, which showed that the $\text{COVER}(\epsilon, \mathcal{D}, \mathcal{L}^1(P)) \le C\left(\epsilon^{-1}\log\epsilon^{-1}\right)^d$ for $0 < \epsilon \le 1/2$, where $\text{VCDIM}(\mathcal{D}) \le d$. That Lemma was the direct inspiration for Pollard (1982a, Theorem 9), where I derived bounds on $\mathcal{L}^2(P)$ covering numbers for collections of functions of the form $F(x)\{x \in D\}$ with $F \in \mathcal{L}^2(P)$ and D ranging over a VC-class of sets $\mathcal{D}$. I had studied such function classes precisely because I needed them to establish a CLT for k-means clustering on multi-dimensional euclidean spaces (Pollard, 1982b). (See Problems [2] and [3].) The extension to collections of functions $\mathcal{F}$ with $\sup_{f\in\mathcal{F}}|f| \le F \in \mathcal{L}^2(P)$ and $\{\text{subgraph}(f) : f \in \mathcal{F}\}$ a VC-class of sets in Pollard (1984, Lemma II.25) was based on a clever extension of Dudley's sampling argument due to Le Cam (1983, §3). (See Problem [2] for the definition of a VC-subgraph class of sets and the connection with SDIM.) By the time I came to give my Iowa lectures it had become clear that: it was not necessary to assume iid observations; and everything came down to control of ℓ^2 packing numbers for subsets $\mathcal{X}$ of $\mathbb{R}^n$. The published version of the lectures (Pollard, 1990, §4) derived bounds on those packing numbers by means of assumptions about the "pseudo-dimension" of $\mathcal{X}$, a

thinly disguised version of the concept of VC-dimension for a collection of subgraphs. The more descriptive name "surround dimension" (SDIM) now replaces the mysterious "pseudo-dimension". The last four Sections of the Iowa lecture notes presented detailed accounts of how the ideas from Section 15.2 can be applied to challenging statistical problems.

Nolan and Pollard (1987, Section 4), Pakes and Pollard (1989, Section 2), and Pollard (1989) abstracted the covering number properties (derived from pseudo-dimension assumptions) into the concept of a ***euclidean class of functions***, essentially defined by the property described in Corollary <7>. These classes share many of the properties associated in classical statistical theory with collections of functions indexed smoothly by a finite-dimensional parameter. They have proved useful for applications to modern statistical asymptotic theory. See Kim and Pollard (1990) for example.

I defer to Anthony and Bartlett (2002) regarding the history of "fat shattering" in the machine learning literature, with which I am not very familiar. They gave (p. 163) credit to Kearns and Schapire (1994) for the fat shattering idea (although those authors did not actually use that name) and (p. 267) to Alon, Ben-David, Cesa-Bianchi, and Haussler (1997) for "the first uniform convergence result in terms of the fat-shattering dimension".

As noted in Section 15.3, the SDIM idea was also introduced by Talagrand (1984, §9) with his discovery that absence of a particular form of "critical set" was the defining property of a condition that he called "stability". He later (Talagrand, 1996, Def 2.1) called existence of such a set a "witness of irregularity". Talagrand (1987, page 842) described some of the history from his perspective:

> "The notion of stable sets was developed by D. H. Fremlin and the author in the course of their study of pointwise compact sets of measurable functions, as being the most natural criterion for pointwise compactness. The author then undertook in [9] (where more historical comments can be found) a systematic study of this notion and its application to various questions of measure theory, mostly Pettis integration. One natural question about stable sets is as follows: Is the convex hull of a uniformly bounded stable set still stable? This is far from clear from the definition. To prove it, the author established Theorem (11-1-1) of [9]. In the language of the present paper, this result means that a uniformly bounded set of functions Z is stable if and only if D_n goes to zero in probability. Due to his ignorance of probability theory, the author became aware that he had essentially solved the Glivenko-Cantelli problem only after he discussed the material of [5] with Professor Hoffmann-Jørgensen."

Here [5] = Hoffmann-Jørgensen (1985) and [9] = Talagrand (1984). In probabilistic terminology, he had shown that the uniform (over a collection of functions $\mathcal{F}$ with envelope in $\mathcal{L}^1(P)$) SLLN for sampling from P fails if and only if there exists some ***witness of irregularity***: a nonatomic set A with strictly positive P measure and some constants $\alpha < \beta$ such that $\mathcal{F}$ picks out all possible patterns from (almost all) finite subsets of A. That is, if $x_1, \ldots, x_k \in A$ and $K \subseteq [\![k]\!]$ then there is some f in $\mathcal{F}$ for which

$$f(x_i) \qquad \begin{cases} \geq \beta & \text{for } i \in K \\ \leq \alpha & \text{for } i \in K^c \end{cases}$$

He later (Talagrand, 1996, page 375) pointed out, in the context of uniform SLLNs, that witnesses of irregularity

> "represent a strong pathology, and thereby are very easy to detect. This strong pathology is concentrated on the set A in the sense that $[\mathcal{F}]$ behaves wildly over all but exceptional finite subsets of A. This is in contrast with Proposition ..., that implies that when $[\mathcal{F}]$ is not a GC class, some rather diffuse pathology occurs."

Talagrand (2003, page 1566) also gave credit to Alon, Ben-David, Cesa-Bianchi, and Haussler (1997) for one of the equivalences established by Talagrand (1996).

References

Alon, N., S. Ben-David, N. Cesa-Bianchi, and D. Haussler (1997). Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM, 44 (4)*, 615–631. First announced in the 1993 Proceedings of the 34th Annual Symposium on Foundations of Computer Science.

Anthony, M. and P. Bartlett (2002). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability 6*, 899–929.

Garling, D. J. H. (2007). *Inequalities: A Journey into Linear Analysis*. Cambridge University Press.

Hoffmann-Jørgensen, J. (1985). The law of large numbers for non measurable and non separable elements. *Astérisque 131*, 299–356.

Kearns, M. J. and R. E. Schapire (1994). Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences 48*, 464–497.

Kim, J. and D. Pollard (1990). Cube root asymptotics. *Annals of Statistics 18*, 191–219.

Le Cam, L. (1983). A remark on empirical measures. In Bickel, Doksum, and Hodges (Eds.), *A Festschrift for Erich Lehmann in honor of his sixty-fifth birthday*, pp. 305–327. Wadsworth. Available online from https://www.stat.berkeley.edu/users/rice/LeCam/papers/.

Mendelson, S. and R. Vershynin (2003). Entropy and the combinatorial dimension. *Inventiones Mathematicae 152*, 37–55.

Nolan, D. and D. Pollard (1987). U-processes: rates of convergence. *Annals of Statistics 15*, 780–799.

Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica 57*, 1027–1058.

Pisier, G. (1989). *The volume of convex bodies and Banach space geometry*, Volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press.

Pollard, D. (1982a). A central limit theorem for empirical processes. *J. Austral. Math. Soc. Ser. A 33*, 235–248.

Pollard, D. (1982b). A central limit theorem for k-means clustering. *Annals of Probability 10*, 919–926.

Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer.

Pollard, D. (1989). Asymptotics via empirical processes (with discussion). *Statistical Science 4*, 341–366.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*, Volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Hayward, CA: Institute of Mathematical Statistics.

Rudelson, M. and R. Vershynin (2006). Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics 164*, 603–648.

Talagrand, M. (1984). *Pettis Integral and Measure Theory*, Volume 307 of *Memoirs AMS*. American Mathematical Society.

Talagrand, M. (1987). The Glivenko-Cantelli problem. *Annals of Probability 15*(3), 837–870.

Talagrand, M. (1996). The Glivenko-Cantelli problem, ten years later. *Journal of Theoretical Probability 9*(2), 371–384.

Talagrand, M. (2003). Vapnik-Chervonenkis type conditions and uniform Donsker classes of functions. *Annals of Probability 31*, 1565–1582.

Talagrand, M. (2021). *Upper and Lower Bounds for Stochastic Processes: Decomposition Theorems* (Second ed.), Volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag.

Vapnik, V. N. and A. Ya. Chervonenkis (1981). Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and Its Applications 26*, 532–553. Russian original received by the editors on July 28, 1978.

Vershynin, R. (2005). Integer cells in convex sets. *Advances in Mathematics 197*, 248–273.

Index

Symbols and acronyms

:= equals by definition
$\sim$ means 'has distribution'
$\mathbb{1}_n = (1, \ldots, 1) \in \mathbb{R}^n$
$|A|$ or $\#A$ = the size (aka cardinality) of a finite set A
$\mathbb{N}$ = the natural numbers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$
$\mathbb{Z}$ = the integers
$\mathbb{R}$ = real line and $\mathbb{C}$ = complex numbers
$[\![n]\!] = \{k \in \mathbb{N} : 1 \le k \le n\}$ for $n \in \mathbb{N}$
$[\![a{:}b]\!] = \{k \in \mathbb{Z} : a \le k \le b\}$ for $a, b \in \mathbb{Z}$
$\mathbb{R}^n := \mathbb{R}^{[\![n]\!]}$ n-dimensional euclidean space
$\mathbb{R}^T$ the set of all real-valued functions on T; §9.3

$A_k(\mathbf{p}) = \mathbb{P}\{S = k\}$ for $S \sim \text{PBIN}(\mathbf{p})$ and $\mathbf{p} = (p_1, \ldots, p_n)$; §4.1
$\mathcal{B}(\mathcal{X})$ = the BOREL σ-field on a topological space $\mathcal{X}$
BER(p) = probability measure P with $P\{1\} = 1 - P\{0\} = p$
$\beta(n, d) := \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}$; §14.2
BIN(n, p) = the Binomial distribution
BL$(\mathcal{X}, d)$ the set of bounded, LIPSCHITZ-continuous functions on $\mathcal{X}$; §9.2
$\|\cdot\|_{\text{BL}}, \|\cdot\|_{\text{Lip}}, \|\cdot\|_\infty$ norms on BL$(\mathcal{X}, d)$; §9.2
$\mathbb{B}(T)$ = set of all bounded real-valued functions on T; §9.7
CLT = central limit theorem
$C[0, 1]$ = the set of all continuous functions on $[0, 1]$; §9.3
$D[0, 1]$ = functions right-continuous with left limits; §9.1
d_u = uniform metric on sets of bounded functions; §9.1
$\mathbb{\Delta}(x) := 2(e^x - 1 - x)/x^2$; §8.1
$E_k(\mathbf{p}) = \mathbb{P}\{S \ge k\}$ for $S \sim \text{PBIN}(\mathbf{p})$ and $\mathbf{p} = (p_1, \ldots, p_n)$; §4.1
$\mathbb{f}(x) = e^x - 1 - x$; §2.2 and §8.1
$\mathfrak{F}(A) := \mathbb{P} \sup_{t \in A} X_t$ for $A \subset T$ and $\{X_t : t \in T\}$ centered gaussian; §12.1
$\Gamma(\cdot)$ = the Gamma function; §2.5
$\mathbb{h}(x) = (1 + x)\log(1 + x) - x$, the convex conjugate of $\mathbb{f}$
$\mathfrak{H}$, $\mathfrak{H}_\mu$ = HAMMING distances; §14.2
HEAD, TAIL = functions to provide an orientation for the edges of a graph; §14.4
HYPER(n,R,B) = HYPERGEOMETRIC distribution; sample w/o replacement
$\mathbb{L}(\mathcal{W})$ = a set of lattices cubes; 15.4
ℓ^p_μ, $|\cdot|_{p,\mu}$ = space of functions (norm) on $\mathbb{R}^n$; §15.2
$\mathcal{L}_{q,\infty,\mu}$ = weak $\mathcal{L}^q$ space; §15.6
$\|\cdot\|_{q,\infty,\mu}, |\cdot|_{q,\infty,\mu}$ = weak $\mathcal{L}^q$ space quasi-norm; §15.6

$\mathcal{L}^\Psi$ = space of functions f with $\|f\|_\Psi < \infty$; §5.1
MM = majorizing measure; §11.1
MGF = moment generating function
MVN = multivariate normal distribution, $N(\mu, V)$
$\|f\|_\Psi$ = ORLICZ norm of a function f; §5.1
OSC$(\cdot, x, T)$ = oscillation of a function x on a set T; §9.7
PBIN$(p_1, \ldots, p_n)$ = Poisson-Binomial distribution; §3.7 and §4.1
PGF = probability generating function
$\psi_{\text{benn}}(x) = 2\hbar(x)/x^2$, named after George Bennett; §2.2 and §8.1
R = a language for statistical computing, https://www.r-project.org
SDIM = shatter (or surround) dimension; §14.2 and §15.2
SLLN = strong law of large numbers
SUBGAMMA = distributions with MGF $\le \exp\left(\alpha\lambda^2/(1-\beta\lambda)\right)$; §3.6
SUBG(τ^2) = subgaussian distribution; MGF $\le \exp(\lambda^2\tau^2/2)$; §7.1
SVD = singular value decomposition (of a matrix); $A = \sum_i \sigma_i u_i v_i'$ UC(T, ρ) = all bounded, uniformly continuous real functions on (T, ρ); §9.7
UCLT = uniform central limit theorem; §9.7
UGMTP = *A User's Guide to Measure Theoretic Probability* (Pollard, 2001)
USLLN = uniform strong law of large numbers; §9.7
VCDIM = Vapnik-Červonenkis dimension; §14.2
$\mathcal{Y}$, $\mathcal{Y}_\infty$ = ORLICZ functions: convex, increasing functions with $\Psi(0) = 0$; §5.1
$\mathcal{Y}_{\exp}$ = ORLICZ functions with exponential rates of growth; §5.3

Concepts

admissible partitions, 247, 285

binary matrix, 330, 348
  hereditary, 333
bounded difference condition, 164

chaining
  admissible partitions, 247, 261
  comparison of methods, 268
  Fernique's approach, 255
  framework, 225
  via nested partitions, 263
  weighted, 255
concentration
  gaussian LIPSHITZ functionals, 127
concepts named after humans
  FENCHEL-LEGENDRE conjugate, 19
  BOHR-MOLLERUP theorem, 30
  HOEFFDING inequality
    independent summands, 159
    martingales, 164
  STIRLING formula
    for $\Gamma(x)$, 30
    for factorials, 21
  BENNETT inequality
    for BIN(n, p), 174
    independent summands, 173
    martingales, 178
BERNSTEIN, 54
BERNSTEIN inequality, 172
BROWNIAN
  bridge, 197
  motion, 202, 223
CHEN-STEIN method, 168, 182
DONSKER
  class, 213
  theorem, 212
DOOB
  separable, 206
DVORETZKY cross-sections, 142
ERDŐS-RÉNYI random graph, 182
FERNIQUE comparison inequality, 124
  Chatterjee's method, 144
gaussian
  isoperimetric inequality, 135
  stochastic calculus method, 131
gaussian process
  concentration of max, 279
  concentration of supremum, 123
  continuous paths, 284

extended SUDAKOV inequality, 279
functional $\mathfrak{F}(\cdot)$, 278
weighted partitions, 280
gaussian processes
path method, 125
GORDON inequality, 137, 141
HAMMING distance, 165, 329
HOEFFDING inequality
for sampling, 59
ISSERLIS/WICK formula, 147
ITÔ formula, 132
KIM-VU method, 180
KULLBACK-LEIBLER distance, 24
LIPSCHITZ
continuous, 199, 206
LUXEMBURG norm, 98
MCDIARMID inequality, 165
Mills ratio, 42, 69
ORLICZ
norm, 98
space, 97
space, linear functionals on, 112
ORLICZ function, 97
of exponential growth, 103, 247
PISIER-MAUREY method, 128
Sauer/Shelah/VC Lemma, 326
SLEPIAN inequality, 136
SMIRNOV statistic, 298
STEIN
i-b-p trick, 126
SUDAKOV minoration, 124
VC class of sets, 301
VC-dimension, VC-class, 325
WIENER measure, 203
YOUNG inequality, 109
conditional variance, 178
conjugate convex functions, 19, 38, 107
convergence in distribution
traditional, 198
with stars, 208
coupling
of Poisson-Binomials, 298
of random vectors, 338
covering number, 228
comparison with packing, 228
δ-net, 228
cumulants, 38

decoupling, 175
discrete hypercube, 327
orientation of edges, 336
subgraph, edges, 334
Doob's heuristic, 197

empirical measure, 357
empirical process, 304
classical, 197, 210, 298
envelopes, 305
general sense, 305
oscillation control, 310
essential supremum/infimum, 203
exponential tilting, 18

falling factorial function, 76
fat shattering, 352

hereditary, *see* binary matrix

increments
controlling by ORLICZ norm, 247
integration-by-parts trick, 126
inversion of $e^r/r \le w$, 329, 361
isonormal gaussian process, 149

Kolmogorov's extension theorem, 201

lattice cube, 353

majorizing measure, 247, 279
as smoother, 252
as weighted chaining, 257
Fernique's conjecture, 250
Marriage Lemma, 335
martingale
differences, 177
martingale differences, 162
maximal inequality, 221
via $\sqrt{\text{SDIM}}$, 361
via ORLICZ norms, 106
via packing numbers, 232
measurable cover, 204
median
POISSON-BINOMIAL, 91
method of paired comparisons, 296
metric entropy, 229
metric space
totally bounded, 228
MGF
convexity of log MGF, 17
mode(s)
BINOMIAL, 81

POISSON-BINOMIAL, 82

Newton's log concavity, 78
normal
  hazard rate, 42, 45
  Laplace bounds, 42
  maxima of independent, 43
  Stein's approximation, 47
  subgaussian tail, 40

orthant, 348
oscillation, 210
  of a sample path, 221
  continuity of sample paths, 237
  stochastic equicontinuity, 222

packing number, 229
  comparison with covering, 229
  δ-packing, 229
  δ-separated, 229
  from shatter dimension, 329
  Haussler's improved bound, 337
  packing framework, 230
  subsets of ℓ_μ^{2p} ball, 352
  subsets of unit cube, 349
  unit balls in $\mathbb{R}^n$, 229
  versus majorizing measure, 284
  via SDIM, 347, 352, 359
  via conditional variance, 337
patterns, 347
polynomials with real roots, 75
predictable stopping time, 178
probabilistic method of Erdős, 324, 354
project onto random subspace, 330, 349, 359

quasi-norm
  for $\mathcal{L}_{q,\infty,\mu}$, 358

R matrix notation, 328

semimetric, 210
separation lemma, 355
shatter, *see* surround
Shatter Theorem, 328, 348
shatter, shatter-dimension, 326, 328, 347
  to orient edges, 336
shifting, downshifting, 328
  control of edge set, 334
  proof of Shatter Theorem, 331
  to orient edges of a graph, 334
sigma-field
  cylinder, 201
sign variables, 156, 307, 318, 346
signed-sum process, 346
singular value decomposition, 175, 176
soft maximum, 144
special functions
  measurable cover, 204
square root trick, 312
standard distributions
  BINOMIAL, 56
  Gamma and chi-squared, 53
  hypergeometric
    as POISSON-BINOMIAL, 79
  HYPERGEOMETRIC
    BINOMIAL tail bound, 59
  POISSON, 51
  Poisson-Binomial, 297
  POISSON-BINOMIAL
    BINOMIAL tail bound, 58
  SUBGAMMA, 53
  SUBGAUSSIAN
    characterization, 157
    via sign variables, 156
  sum of bounded rv's
    BINOMIAL tail bound, 58
stochastic process
  DOOB-separable, 206
  asymptotic equicontinuity, 212
  control of increments, 222
  fidis, 201
  independent copy, 302
  oscillation, 210
  sample path, 200
subgaussian, 41, 155
  signed sums, 347
subsets picked out, 322
surround, *see* shatter
surround, surround-dimension, 347
symmetrization
  gaussian, 128

tail bounds
  from MGF, 36
tightness
  asymptotic, 212
  probability measure, 210
tricks
  Taylor representation, 15
  decoupling, 175

M-R-V, *see* fat shattering
restoring symmetry, 61
Rio's X_i^+, 55
upper/lower integral, 204
volume of unit ℓ^p ball in $\mathbb{R}^n$, 364
weak convergence, 199
weak $\mathcal{L}^q$ space, 352, 358