NOTES ON EMPIRICAL PROCESSES

R. M. Dudley

January 24, 2000

Preface

These are lecture notes for a course in Aarhus, August 1999. For a second printing of the notes, a few errors have been corrected. Chapter 1 is on the classical empirical process defined in terms of empirical distribution functions. A proof, expanding on one in a 1989 paper by Bretagnolle and Massart, is given for the Komlós-Major-Tusnády result on the speed of convergence of the empirical process to a Brownian bridge in the supremum norm. In this second printing Chapter 1 has been improved to prove the constants stated by Bretagnolle and Massart.

The rest of the notes are about the theory of empirical processes on general spaces, and are an abridgment of my book, *Uniform Central Limit Theorems*, or "UCLT," published by Cambridge University Press in 1999. In this abridgment, proofs are not given except for some very short ones and some not given in UCLT. Also omitted are several sections that summarize recent results without proofs, the Notes, and the problems except for those on Chapter 4. Because the Notes are left out, attributions of statements are not as thorough as they are in UCLT. Although some results of mine are included, the absence of an attribution to anyone else should not be interepreted as a claim of credit for myself. Due to various changes, numberings of statements differ between these notes and UCLT.

Some of the material appeared in an earlier form in my lecture notes for the Ecole d'été de probabilités de St.-Flour, 1982, published in *Lecture Notes in Math.* (Springer) vol. 1097 (1984), pp. 1-142.

I thank Uwe Einmahl, Evarist Giné, Stanisław Kwapien, David Mason, and Agata Smoktunowicz for information about some material in these notes which is not in UCLT.

For useful conversations on topics in UCLT I'm glad to thank Kenneth Alexander, Niels Trolle Andersen, Miguel Arcones, Patrice Assouad, Erich Berger, Lucien Birgé, Igor S. Borisov, Donald Cohn, Yves Derrienic, Uwe Einmahl, Joseph Fu, Evarist Giné, Sam Gutmann, David Haussler, Jørgen Hoffmann-Jørgensen, Yen-Chin Huang, Vladimir Koltchinskii, Lucien Le Cam, Pascal Massart, James Munkres, Rimas Norvaiša, Walter Philipp, Tom Salisbury, Rae Shortt, Michel Talagrand, He Sheng Wu, Joe Yukich, and Joel Zinn. I especially thank Peter Gaenssler and Franz Strobl, Evarist Giné, and Jinghua Qian, for providing lists of corrections and suggestions. I also thank Xavier Fernique and Evarist Giné very much for sending me copies of recent expositions, and Jørgen Hoffmann-Jørgensen for the invitation to give the lectures.

Throughout these notes, all references to "RAP" are to the author's book *Real Analysis and Probability*, Wadsworth and Brooks/Cole, Pacific Grove, Calif. 1989, reprinted by CRC, 1999, and with corrections by Chapman and Hall, New York, 1993; the latter is out of print at this writing.

Also, "$A := B$" means A is defined by B, whereas "$A =: B$" means B is defined by A.

Richard Dudley
Cambridge, Mass., January 24, 2000

Contents

Chapter 1

Empirical distribution functions: the KMT theorem

1.1 Introduction

Let $U[0,1]$ be the uniform distribution on $[0,1]$ and U its distribution function. Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with law U. Let $F_n(t)$ be the empirical distribution function based on $X_1, X_2, \ldots, X_n$,

$$F_n(t) \ := \ \frac{1}{n} \sum_{j=1}^{n} 1_{\{X_j \le t\}},$$

and $\alpha_n(t)$ the corresponding empirical process, i.e., $\alpha_n(t) = \sqrt{n}(F_n(t) - t), \ t \in [0,1]$. Here α_n may be called the *classical* empirical process. Recall that a *Brownian bridge* is a Gaussian stochastic process $B(t), \ 0 \le t \le 1$, with $EB(t) = 0$ and $EB(t)B(u) = t(1-u)$ for $0 \le t \le u \le 1$. Donsker (1952) proved (neglecting measurability problems) that $\alpha_n(t)$ converges in law to a Brownian bridge $B(t)$ with respect to the sup norm. Komlós, Major, and Tusnády (1975) stated a sharp rate of convergence, namely that on some probability space there exist X_i i.i.d. $U[0,1]$ and Brownian bridges B_n such that

$$P\left(\sup_{0 \le t \le 1} |\sqrt{n}(\alpha_n(t) - B_n(t))| > x + c \log n \right) \ < \ Ke^{-\lambda x} \tag{1.1}$$

for all n and x, where c, K, and λ are positive absolute constants. Komlós, Major and Tusnády (KMT) formulated a construction giving a joint distribution of α_n and B_n, and this construction has been accepted by later workers. But Komlós, Major and Tusnády gave hardly any proof for (1.1). Csörgő and Révész (1981) sketched a method of proof of (1.1) based on lemmas of G. Tusnády, Lemmas 1.2 and 1.4 below. The implication from Lemma 1.4 to 1.2 is not difficult, but Csörgő and Révész did not include a proof of Lemma 1.4. Bretagnolle and Massart (1989) gave a proof of the lemmas and of the inequality (1.1) with specific constants, Theorem 1.1 below. Bretagnolle and Massart's proof was rather compressed and some readers have had difficulty following it. Csörgő and Horváth (1993), pp. 116-139, expanded the proof while making it more elementary and gave a proof of Lemma 1.4 for $n \ge n_0$ where n_0 is at least 100. The purpose of the present chapter is to give a detailed and in some minor details corrected version of the original Bretagnolle and Massart proof of the lemmas for all n, overlapping in

part with the Csörgő and Horváth proof, then to prove (1.1) for some constants, as given by Bretagnolle and Massart and largely following their proof.

Mason and van Zwet (1987) gave another proof of the inequality (1.1) and an extended form of it for subintervals $0 \le t \le d/n$ with $1 \le d \le n$ and $\log n$ replaced by $\log d$, without Tusnády's inequalities and without specifying the constants c, K, λ. Some parts of the proof sketched by Mason and van Zwet are given in more detail by Mason (1998).

Acknowledgments. I am very grateful to Evarist Giné, David Mason, Jon Wellner, and Uwe Einmahl for conversations and correspondence on the topic.

1.2 Statements: the theorem and Tusnády's lemmas

The main result of the present chapter is:

Theorem 1.1. (Bretagnolle and Massart) *The approximation (1.1) of the empirical process by the Brownian bridge holds with $c = 12$, $K = 2$ and $\lambda = 1/6$ for $n \ge 2$.*

The rest of this chapter will give a proof of the theorem. In a preprint, Rio (1991, Theorem 5.1) states in place of (1.1)

$$P\left(\sup_{0\le t\le 1} |\sqrt{n}(\alpha_n(t) - B_n(t))| > ax + b\log n + \gamma \log 2\right) \quad < \quad Ke^{-x} \tag{1.2}$$

for $n \ge 8$ where $a = 3.26$, $b = 4.86$, $\gamma = 2.70$, and $K = 1$. This implies that for $n \ge 8$, (1.1) holds with $c = 5.76$, $K = 1$, and $\lambda = 1/3.26$, where all three constants are better than in Theorem 1.1.

Tusnády's lemmas are concerned with approximating symmetric binomial distributions by normal distributions. Let $\mathcal{B}(n, 1/2)$ denote the symmetric binomial distribution for n trials. Thus if B_n has this distribution, B_n is the number of successes in n independent trials with probability $1/2$ of success on each trial. For any distribution function F and $0 < t < 1$ let $F^{-1}(t) := \inf\{x : F(x) \ge t\}$. Here is one of Tusnády's lemmas (Lemma 4 of Bretagnolle and Massart (1989)).

Lemma 1.2. *Let Φ be the standard normal distribution function and Y a standard normal random variable. Let Φ_n be the distribution function of $\mathcal{B}(n, 1/2)$ and set $C_n := \Phi_n^{-1}(\Phi(Y)) - n/2$. Then*

$$|C_n| \ \le \ 1 + (\sqrt{n}/2)|Y|, \tag{1.3}$$

$$|C_n - (\sqrt{n}/2)Y| \ \le \ 1 + Y^2/8. \tag{1.4}$$

Recall the following well known and easily checked facts:

Theorem 1.3. *Let X be a real random variable with distribution function F.*
(a) If F is continuous then $F(X)$ has a $U[0, 1]$ distribution.
(b) For any F, if V has a $U[0, 1]$ distribution then $F^{-1}(V)$ has distribution function F.

Thus $\Phi(Y)$ has a $U[0, 1]$ distribution and $\Phi_n^{-1}(\Phi(Y))$ has distribution $\mathcal{B}(n, 1/2)$. Lemma 1.2 will be shown (by a relatively short proof) to follow from:

Lemma 1.4. *Let Y be a standard normal variable and let β_n be a binomial random variable with distribution $\mathcal{B}(n, 1/2)$. Then for any integer j such that $0 \le j \le n$ and $n + j$ is even, we have*

$$P(\beta_n \ge (n+j)/2) \;\ge\; P(\sqrt{n}Y/2 \ge n(1 - \sqrt{1 - j/n})), \tag{1.5}$$

$$P(\beta_n \ge (n+j)/2) \;\le\; P(\sqrt{n}Y/2 \ge (j-2)/2). \tag{1.6}$$

Remarks. The restriction that $n + j$ be even is not stated in the formulation of the lemma by Bretagnolle and Massart (1989), but $n + j$ is always even in their proof. If (1.5) holds for $n + j$ even it also holds directly for $n + j$ odd, but the same is not clear for (1.6). It turns out that only the case $n + j$ even is needed in the proof of Lemma 1.2, so I chose to restrict the statement to that case.

The following form of Stirling's formula with remainder is used in the proof of Lemma 1.4.

Lemma 1.5. *Let $n! = (n/e)^n \sqrt{2\pi n} A_n$ where $A_n = 1 + \beta_n/(12n)$, which defines A_n and β_n for $n = 1, 2, \cdots$. Then $\beta_n \downarrow 1$ as $n \to \infty$.*

1.3 Stirling's formula: Proof of Lemma 1.5

It can be checked directly that $\beta_1 > \beta_2 > \cdots > \beta_8 > 1$. So it suffices to prove the lemma for $n \ge 8$. We have $A_n = \exp((12n)^{-1} - \theta_n/(360n^3))$ where $0 < \theta_n < 1$, see Whittaker and Watson (1927), p. 252 or Nanjundiah (1959). Then by Taylor's theorem with remainder,

$$A_n \;=\; \left(1 + \frac{1}{12n} + \frac{1}{288n^2} + \frac{1}{6(12n)^3}\phi_n e^{1/12n}\right) \exp(-\theta_n/(360n^3))$$

where $0 < \phi_n < 1$. Next,

$$\beta_{n+1} \;\le\; 12(n+1)\left[\exp\left(\frac{1}{12(n+1)}\right) - 1\right]$$

$$\le\; 1 + \frac{1}{24(n+1)} + \frac{1}{6(12(n+1))^2} e^{1/(12(n+1))},$$

from which $\limsup_{n\to\infty} \beta_n \le 1$, and

$$\beta_n \;=\; 12n[A_n - 1] \;\ge\; 12n\left[\left(1 + \frac{1}{12n} + \frac{1}{288n^2}\right)\exp(-1/(360n^3)) - 1\right].$$

Using $e^{-x} \ge 1 - x$ gives

$$\beta_n \;\ge\; 12n\left[\frac{1}{12n} + \frac{1}{288n^2} - \frac{1}{360n^3}\left(1 + \frac{1}{12n} + \frac{1}{288n^2}\right)\right]$$

$$=\; 1 + \frac{1}{24n} - \frac{1}{30n^2}\left(1 + \frac{1}{12n} + \frac{1}{288n^2}\right).$$

Thus $\liminf_{n\to\infty} \beta_n \ge 1$ and $\beta_n \to 1$ as $n \to \infty$. To prove $\beta_n \ge \beta_{n+1}$ for $n \ge 8$ it will suffice to show that

$$1 + \frac{1}{24(n+1)} + \frac{e^{1/108}}{6 \cdot 144n^2} \;\le\; 1 + \frac{1}{24n} - \frac{1}{30n^2}\left[1 + \frac{1}{96} + \frac{1}{288 \cdot 8^2}\right]$$

or

$$\frac{e^{1/108}}{6\cdot 144n^2}+\frac{1}{30n^2}\left[1+\frac{1}{96}+\frac{1}{288\cdot 64}\right] \;\le\; \frac{1}{24n(n+1)}$$

or that $0.035/n^2 \le 1/[24n(n+1)]$ or $0.84 \le 1-1/(n+1)$, which holds for $n \ge 8$, proving that β_n decreases with n. Since its limit is 1, Lemma 1.5 is proved. □

1.4 Proof of Lemma 1.4

First, (1.5) will be proved. For any $i = 0, 1, \cdots, n$ such that $n+i$ is even, let $k := (n+i)/2$ so that k is an integer, $n/2 \le k \le n$, and $i = 2k-n$. Let $p_{ni} := P(\beta_n = (n+i)/2) = P(\beta_n = k) = \binom{n}{k}/2^n$ and $x_i := i/n$. Define $p_{ni} := 0$ for $n+i$ odd. The factorials in $\binom{n}{k}$ will be approximated via Stirling's formula with correction terms as in Lemma 1.5. To that end, let

$$CS(u,v,w,x,n) \;:=\; \frac{1+u/(12n)}{(1+v/[6n(1-x)])(1+w/[6n(1+x)])}.$$

By Lemma 1.5, we can write for $0 \le i < n$ and $n+i$ even

$$p_{ni} \;=\; CS(x_i,n)\sqrt{2/\pi n}\exp(-ng(x_i)/2 \;-\; (1/2)\log(1-x_i^2)) \tag{1.7}$$

where $g(x) := (1+x)\log(1+x)+(1-x)\log(1-x)$ and $CS(x_i,n) := CS(\beta_n,\beta_{n-k},\beta_k,x_i,n)$. By Lemma 1.5 and since $k \ge n/2$,

$$1^+ \;:=\; 1.013251 \ge 12(e(2\pi)^{-1/2}-1) = \beta_1 \ge \beta_{n-k} \ge \beta_k \ge \beta_n > 1.$$

Thus, for $x := x_i$, by clear or easily checked monotonicity properties,

$$CS(x,n) \;\le\; CS(\beta_n,\beta_k,\beta_k,x,n) \;=$$

$$\left(1+\frac{\beta_n}{12n}\right)\left[1+\frac{\beta_k}{3n(1-x^2)}+\frac{\beta_k^2}{36n^2(1-x^2)}\right]^{-1}$$

$$\le\; CS(\beta_n,\beta_k,\beta_k,0,n) \;\le\; CS(\beta_n,\beta_n,\beta_n,0,n)$$

$$\le\; CS(1,1,1,0,n) \;=\; \left(1+\frac{1}{12n}\right)\left[1+\frac{1}{3n}+\frac{1}{36n^2}\right]^{-1}.$$

It will be shown next that $\log(1+y)-2\log(1+2y) \le -3y+7y^2/2$ for $y \ge 0$. Both sides vanish for $y=0$. Differentiating and clearing fractions, we get a clearly true inequality. Setting $y := 1/(12n)$ then gives

$$\log CS(x_i,n) \;\le\; -1/(4n) \;+\; 7/(288n^2). \tag{1.8}$$

To get a lower bound for $CS(x,n)$ we have by an analogous string of inequalities

$$CS(x,n) \;\ge\; \left(1+\frac{1}{12n}\right)\left\{1+\frac{1^+}{3n(1-x^2)}+\frac{(1^+)^2}{36n^2(1-x^2)}\right\}^{-1}. \tag{1.9}$$

The inequality (1.5) to be proved can be written as

$$\sum_{i=j}^{n} p_{ni} \;\geq\; 1 - \Phi(2\sqrt{n}(1 - \sqrt{1 - j/n})). \tag{1.10}$$

When $j = 0$ the result is clear. When $n \leq 4$ and $j = n$ or $n - 2$ the result can be checked from tables of the normal distribution. Thus we can assume from here on

$$n \geq 5. \tag{1.11}$$

CASE I. Let $j^2 \geq 2n$, in other words $x_j \geq \sqrt{2/n}$. Recall that for $t > 0$ we have $P(Y > t) \leq (t\sqrt{2\pi})^{-1} \exp(-t^2/2)$, e.g. Dudley (1993), Lemma 12.1.6(a). Then (1.10) follows easily when $j = n$ and $n \geq 5$. To prove it for $j = n - 2$ it is enough to show

$$n(2 - \log 2) - 4\sqrt{2n} + \log(n + 1) + 4 + \log[2\sqrt{2\pi}(\sqrt{n} - \sqrt{2})] \;\geq\; 0, \quad n \geq 5.$$

The left side is increasing in n for $n \geq 5$ and is ≥ 0 at $n = 5$.

For $5 \leq n \leq 7$ we have $(n - 4)^2 < 2n$, so we can assume in the present case that $2n \leq j^2 \leq (n - 4)^2$ and $n \geq 8$. Let $y_i := 2\sqrt{n}(1 - \sqrt{1 - i/n})$. Then it will suffice to show

$$p_{ni} \;\geq\; \int_{y_i}^{y_{i+2}} \phi(u)du, \quad i = j, j + 2, \cdots, n - 4, \tag{1.12}$$

where ϕ is the standard normal density function. Let

$$f_n(x) \;:=\; \sqrt{n/2\pi(1 - x)} \exp(-2n(1 - \sqrt{1 - x})^2). \tag{1.13}$$

By the change of variables $u = 2\sqrt{n}(1 - \sqrt{1 - x})$, (1.12) becomes

$$p_{ni} \;\geq\; \int_{x_i}^{x_{i+2}} f_n(x)dx. \tag{1.14}$$

Clearly $f_n > 0$. To see that $f_n(x)$ is decreasing in x for $\sqrt{2/n} \leq x \leq 1 - 4/n$, note that

$$2(1 - x)f_n'/f_n \;=\; 1 - 4n[\sqrt{1 - x} - 1 + x],$$

so f_n is decreasing where $\sqrt{1 - x} - (1 - x) > 1/(4n)$. We have $\sqrt{y} - y \geq y$ for $y \leq 1/4$, so $\sqrt{y} - y > 1/(4n)$ for $1/(4n) < y \leq 1/4$. Let $y := 1 - x$. Also $\sqrt{1 - x} - (1 - x) > x/4$ for $x < 8/9$, so $\sqrt{1 - x} - (1 - x) > 1/(4n)$ for $1/n < x < 8/9$. Thus $\sqrt{1 - x} - (1 - x) > 1/(4n)$ for $1/n < x < 1 - 1/(4n)$, which includes the desired range. Thus to prove (1.14) it will be enough to show that

$$p_{ni} \;\geq\; (2/n)f_n(x_i), \quad i = j, j + 2, \cdots, n - 4. \tag{1.15}$$

So by (1.7) it will be enough to show that for $\sqrt{2/n} \leq x \leq 1 - 4/n$ and $n \geq 8$,

$$CS(x, n)(1 + x)^{-1/2} \exp[n\{4(1 - \sqrt{1 - x})^2 - g(x)\}/2] \;\geq\; 1. \tag{1.16}$$

Let

$$J(x) \;:=\; 4(1 - \sqrt{1 - x})^2 - g(x). \tag{1.17}$$

Then J is increasing for $0 < x < 1$, since its first and second derivatives are both 0 at 0, while its third derivative is easily checked to be positive on $(0,1)$. In light of (1.9), to prove (1.16) it suffices to show that

$$\left(1+\frac{1}{12n}\right)e^{nJ(x)/2} \;\geq\; \sqrt{1+x}\left(1+\frac{1^+}{3n(1-x^2)}+\frac{(1^+)^2}{36n^2(1-x^2)}\right). \tag{1.18}$$

When $x \leq 1-4/n$ and $n \geq 8$ the right side is less than 1.5, using first $\sqrt{1+x} \leq \sqrt{2}$, next $x \leq 1 - 4/n$, and lastly $n \geq 8$. For $x \geq 0.55$ and $n \geq 8$ the left side is larger than 1.57, so (1.18) is proved for $x \geq 0.55$. We will next need the inequality

$$J(x) \;\geq\; x^3/2 + 7x^4/48, \quad 0 \leq x \leq 0.55. \tag{1.19}$$

To check this one can calculate $J(0) = J'(0) = J''(0) = 0$, $J^{(3)}(0) = 3$, $J^{(4)}(0) = 7/2$, so that the right side of (1.19) is the Taylor series of J around 0 through fourth order. One then shows straightforwardly that $J^{(5)}(x) > 0$ for $0 \leq x < 1$.

It follows since $nx^2 \geq 2$ and $n \geq 8$ that $nJ(x)/2 \geq x/2 + 7/24n$. Let $K(x) := \exp(x/2)/\sqrt{1+x}$ and $\kappa(x) := (K(x)-1)/x^2$. We will next see that $\kappa(\cdot)$ is decreasing on $[0,1]$. To show $\kappa' \leq 0$ is equivalent to $e^{x/2}[4+4x-x^2] \geq 4(1+x)^{3/2}$, which is true at $x = 0$. Differentiating, we would like to show $e^{x/2}[6-x^2/2] \geq 6\sqrt{1+x}$, or squaring that and multiplying by 4, $e^x(144-24x^2+x^4) \geq 144(1+x)$. This is true at $x = 0$. Differentiating, we would like to prove $e^x(144-48x-24x^2+4x^3+x^4) \geq 144$. Using $e^x \geq 1+x$ and algebra gives this result for $0 \leq x \leq 1$.

It follows that $K(x) \geq 1 + 0.3799/n$ when $\sqrt{2/n} \leq x \leq 0.55$. It remains to show that for $x \leq 0.55$,

$$\left(1+\frac{1}{12n}\right)\left(1+\frac{0.3799}{n}\right)e^{7/(24n)} \;\geq\; 1+\frac{1^+}{3n(1-x^2)}+\frac{(1^+)^2}{36n^2(1-x^2)}.$$

At $x = 0.55$ the right side is less than $1 + 0.543/n$, so Case I is completed since $0.543 \leq 1/12 + 0.3799 + 7/24$.

CASE II. The remaining case is $j < \sqrt{2n}$. For any integer k, $P(\beta_n \geq k) = 1 - P(\beta_n \leq k-1)$. For $k = (n+j)/2$ we have $k - 1 = (n+j-2)/2$. If n is odd, then $P(\beta_n \geq n/2) = 1/2 = P(Y \geq 0)$. If n is even, then $P(\beta_n \geq n/2) - p_{n0}/2 = 1/2 = P(Y \geq 0)$. So, since $p_{n0} = 0$ for n odd, (1.5) is equivalent to

$$\frac{1}{2}p_{n0} + \sum_{0<i\leq j-2} p_{ni} \;\leq\; P(0 \leq Y \leq 2\sqrt{n}(1-\sqrt{1-j/n})). \tag{1.20}$$

Given $j < \sqrt{2n}$, a family $I_0, I_1, \cdots, I_K$ of adjacent intervals will be defined such that for n odd,

$$p_{ni} \;\leq\; P(\sqrt{n}Y/2 \in I_k) \;\text{ with }\; i = 2k+1,\; 0 \leq k \leq K \;:=\; (j-3)/2, \tag{1.21}$$

while for n even,

$$p_{ni} \;\leq\; P(\sqrt{n}Y/2 \in I_k) \;\text{ with }\; i = 2k,\; 1 \leq k \leq K \;:=\; (j-2)/2, \tag{1.22}$$

and

$$p_{n0}/2 \;\leq\; P(\sqrt{n}Y/2 \in I_0). \tag{1.23}$$

In either case,

$$I_0 \cup I_1 \cup \cdots \cup I_K \subset [0, n(1-\sqrt{1-j/n})]. \tag{1.24}$$

The intervals will be defined by

$$\delta_{k+1} \;:=\; (k+1)/n + k(k+1/2)(k+1)/n^{3/2}, \quad k \geq 0, \tag{1.25}$$

$$\Delta_{k+1} \;:=\; \delta_{k+1} + k + 1/2 = \delta_{k+1} + (i+1)/2, \quad i = 2k, \quad n \text{ even}, \tag{1.26}$$

$$\Delta_{k+1} \;:=\; \delta_{k+1} + k + 1 = \delta_{k+1} + (i+1)/2, \quad i = 2k+1, \quad n \text{ odd}, \tag{1.27}$$

$$I_k \;:=\; [\Delta_k, \Delta_{k+1}] \;\text{ with }\; \Delta_0 = 0. \tag{1.28}$$

It will be shown that $I_0, I_1, \cdots, I_K$ defined by (1.25) through (1.28) satisfy (1.21) through (1.24). Recall that $n \geq 5$ (1.11) and $x_i \;:=\; i/n$.

Proof of (1.24). It needs to be shown that $\Delta_{K+1} \leq n(1-\sqrt{1-x_j})$. Since $j < \sqrt{2n}$, we have $K \leq j/2 - 1 < \sqrt{n/2} - 1$ and

$$\delta_{K+1} \;\leq\; (K+1)/n \;+\; K(K+1/2)/(n\sqrt{2}) \;\leq\; x_j/2 + nx_j^2/(4\sqrt{2}).$$

We have $\Delta_{K+1} = nx_j/2 - 1/2 + \delta_{K+1}$. It will be shown next that

$$1 - \sqrt{1-x} \;\geq\; x/2 + x^2/8, \quad 0 \leq x \leq 1. \tag{1.29}$$

The functions and their first derivatives agree at 0 while the second derivative of the left side is clearly larger.

It then remains to prove that

$$1/2 + nx_j^2(1/8 \;-\; 1/4\sqrt{2}) - x_j/2 \;\geq\; 0.$$

This is true since $nx_j^2 \leq 2$ and $x_j \leq (2/8)^{1/2} = 1/2$, so (1.24) is proved.

Proof of (1.21)-(1.23). First it will be proved that

$$p_{ni} \;\leq\; \frac{\sqrt{2}}{\sqrt{\pi n}} \exp\left[-\frac{1}{4n} + \frac{7}{288n^2} - \frac{(n-1)i^2}{2n^2} + \frac{(i/n)^{2n}}{2n(1-i^2/n^2)}\right]. \tag{1.30}$$

In light of (1.7) and (1.8), it is enough to prove, for $x \;:=\; i/n$, that

$$-[ng(x) + \log(1-x^2) - (n-1)x^2]/2 \;\leq\; x^{2n}/2n(1-x^2). \tag{1.31}$$

It is easy to verify that for $0 \leq t < 1$,

$$g(t) \;=\; (1+t)\log(1+t) + (1-t)\log(1-t) \;=\; \sum_{r=1}^{\infty} t^{2r}/r(2r-1).$$

Thus the left side of (1.31) can be expanded as $\sum_{r\geq 2} x^{2r}(1 - n/(2r-1))/2r = A + B$ where $A = \sum_{r=2}^{n-1}$ and $B = \sum_{r\geq n}$. We have

$$d^2A/dx^2 = \sum_{2\leq r\leq (n+1)/2} (2r-n-1)(x^{2r-2} - x^{2n-2r})$$

which is ≤ 0 for $0 \le x \le 1$. Since $A = dA/dx = 0$ for $x = 0$ we have $A \le 0$ for $0 \le x \le 1$. Then, $2nB \le x^{2n}/(1-x^2)$, so (1.30) is proved.

We have for $n \ge 5$ and $x \le (\sqrt{2n}-2)/n$ that $x^{2n}/(1-x^2) < 10^{-3}$, since $n \mapsto (\sqrt{2n}-2)/n$ is decreasing in n for $n \ge 8$ and the statement can be checked for $n = 5, 6, 7, 8$. So (1.30) yields

$$p_{ni} \;\le\; \sqrt{2/\pi n}\exp[-0.249/n + 7/288n^2 - (n-1)i^2/2n^2]. \tag{1.32}$$

Next we will need:

Lemma 1.6. *For any $0 \le a < b$ and a standard normal variable Y,*

$$P(Y \in [a,b]) \;\ge\; \sqrt{1/2\pi}(b-a)\exp[-a^2/4 - b^2/4]\phi(a,b) \tag{1.33}$$

where $\phi(a,b) \;:=\; [4/(b^2-a^2)]\sinh[(b^2-a^2)/4] \;\ge\; 1.$

Proof. Since the Taylor series of sinh around 0 has all coefficients positive, and $(\sinh u)/u$ is an even function, clearly $\sinh u/u \ge 1$ for any real u. The conclusion of the lemma is equivalent to

$$\frac{a+b}{2}\int_a^b \exp(-u^2/2)du \;\ge\; \exp(-a^2/2) - \exp(-b^2/2). \tag{1.34}$$

Letting $x \;:=\; b-a$ and $v \;:=\; u-a$ we need to prove

$$\left(a + \frac{x}{2}\right)\int_0^x \exp(-av - v^2/2)dv \;\ge\; 1 - \exp(-ax - x^2/2).$$

This holds for $x = 0$. Taking derivatives of both sides and simplifying, we would like to show

$$\int_0^x \exp(-av - v^2/2)dv \;\ge\; x\exp(-ax - x^2/2).$$

This also holds for $x = 0$, and differentiating both sides leads to a clearly true inequality, so Lemma 1.6 is proved. □

For the intervals I_k, Lemma 1.6 yields

$$P(\sqrt{n}Y/2 \in I_k) \;\ge\; \sqrt{2/\pi n}\phi_k \exp[-(\Delta_{k+1}^2 + \Delta_k^2)/n + \log(\Delta_{k+1} - \Delta_k)] \tag{1.35}$$

where $\phi_k \;:=\; \phi(2\Delta_k/\sqrt{n}, 2\Delta_{k+1}/\sqrt{n})$. The aim is to show that the ratio of the bounds (1.35) over (1.32) is at least 1.

First consider the case $k = 0$. If n is even, this means we want to prove (1.23). Using (1.32) and (1.35) and $\phi_0 \ge 1$, it suffices to show that

$$0.249/n - 7/288n^2 - 1/4n - 1/n^2 - 1/n^3 + \log(1 + 2/n) \;\ge\; 0.$$

Since $\log(1+u) \ge u - u^2/2$ for $u \ge 0$ by taking a derivative, it will be enough to show that

$$(E)_n \;:=\; 1.999/n - 3/n^2 - 7/288n^2 - 1/n^3 \;\ge\; 0,$$

and it is easily checked that $n(E)_n > 0$ since $n \ge 5$.

If n is odd, then (1.32) applies for $i = 2k+1 = 1$ and we have $\Delta_0 = 0$, $\Delta_1 = \delta_1 + 1 = 1 + 1/n$ so (1.35) yields

$$P(\sqrt{n}Y/2 \in I_0) \ \geq \ \sqrt{2/\pi n}\exp[-(1+1/n)^2/n + \log(1+1/n)].$$

Using $\log(1+u) \geq u - u^2/2$ again, the desired inequality can be checked since $n \geq 5$. This completes the case $k = 0$.

Now suppose $k \geq 1$. In this case, $i < \sqrt{2n} - 2$ implies $n \geq 10$ for n even and $n \geq 13$ for n odd. Let $s_k \ := \ \delta_k + \delta_{k+1}$ and $d_k \ := \ \delta_{k+1} - \delta_k$. Then for i as in the definition of Δ_{k+1},

$$\Delta_{k+1} + \Delta_k \ = \ i + s_k, \tag{1.36}$$

$$\Delta_{k+1} - \Delta_k \ = \ 1 + d_k, \tag{1.37}$$

$$s_k \ = \ \frac{2k+1}{n} + \frac{2k^3 + k}{n^{3/2}}, \tag{1.38}$$

and

$$d_k \ = \ \frac{1}{n} + \frac{3k^2}{n^{3/2}}. \tag{1.39}$$

From the Taylor series of sinh around 0 one easily sees that $(\sinh u)/u \geq 1 + u^2/6$ for all u. Letting $u \ := \ (\Delta_{k+1}^2 - \Delta_k^2)/n \geq i/n$ gives

$$\log \phi_k \ \geq \ \log(1 + i^2/6n^2). \tag{1.40}$$

We have

$$d_k \leq 3/(2\sqrt{n}) \tag{1.41}$$

since $2k \leq \sqrt{2n} - 2$ and $n \geq 10$. Next we have another lemma:

Lemma 1.7. $\log(1+x) \geq \lambda x$ *for* $0 \leq x \leq \ \alpha$ *for each of the pairs* $(\alpha, \lambda) = (0.207, 0.9)$, $(0.195, 0.913)$, $(0.14, 0.93)$, $(0.04, 0.98)$.

Proof. Since $x \mapsto \log(1+x)$ is concave, or equivalently we are proving $1 + x \geq e^{\lambda x}$ where the latter function is convex, it suffices to check the inequalities at the endpoints, where they hold. □

Lemma 1.7 and (1.40) then give

$$\log \phi_k \ \geq \ 0.98 i^2/6n^2 \tag{1.42}$$

since $i^2/(6n^2) \leq 1/3n \leq 0.04$, $n \geq 10$. Next,

Lemma 1.8. *We have* $\log(\Delta_{k+1} - \Delta_k) \ \geq \ \lambda d_k$ *where* $\lambda = 0.9$ *when* n *is even and* $n \geq 20$, $\lambda = 0.93$ *when* n *is odd and* $n \geq 25$, *and* $\lambda = 0.913$ *when* $k = 1$ *and* $n \geq 10$. *Only these cases are possible (for* $k \geq 1$*).*

Proof. If n is even and $k \geq 2$, then $4 \leq i = 2k < \sqrt{2n} - 2$ implies $n \geq 20$. If n is odd and $k \geq 2$, then $5 \leq i = 2k + 1 < \sqrt{2n} - 2$ implies $n \geq 25$. So only the given cases are possible.

We have $k \leq k_n \ := \ \sqrt{n/2} - 1$ for n even or $k_n \ := \ \sqrt{n/2} - 3/2$ for n odd. Let $d(n) \ := \ 1/n \ + \ 3k_n^2/n^{3/2}$ and $t \ := \ 1/\sqrt{n}$. It will be shown that $d(n)$ is decreasing in n,

separately for n even and odd. For n even we would like to show that $3t/2 + (1 - 3\sqrt{2})t^2 + 3t^3$ is increasing for $0 \le t \le 1/\sqrt{20}$ and in fact its derivative is > 0.04. For n odd we would like to show that $3t/2 + (1 - 9/\sqrt{2})t^2 + 27t^3/4$ is increasing. We find that its derivative has no real roots and so is always positive as desired.

Since $d(\cdot)$ is decreasing for $n \ge 20$, its maximum for n even, $n \ge 20$ is at $n = 20$ and we find it is less than 0.207 so Lemma 1.7 applies to give $\lambda = 0.9$. Similarly for n odd and $n \ge 25$ we have the maximum $d(25) < 0.14$ and Lemma 1.7 applies to give $\lambda = 0.93$.

If $k = 1$ then $n \mapsto n^{-1} + 3/n^{3/2}$ is clearly decreasing. Its value at $n = 10$ is less than 0.195 and Lemma 1.7 applies with $\lambda = 0.913$. So Lemma 1.8 is proved. □

It will next be shown that for $n \ge 10$

$$s_k \;\le\; n^{-1} + k/\sqrt{n}. \tag{1.43}$$

By (1.38) this is equivalent to $2/\sqrt{n} + (2k^2 + 1)/n \le 1$. Since $k \le \sqrt{n/2} - 1$ one can check that (1.43) holds for $n \ge 14$. For $n = 10, 11, 12, 13$ note that k is an integer, in fact $k \le 1$, and (1.43) holds.

After some calculations, letting $s := s_k$ and $d := d_k$ and noting that

$$\Delta_k^2 + \Delta_{k+1}^2 \;=\; \frac{1}{2}[(\Delta_{k+1} - \Delta_k)^2 + (\Delta_k + \Delta_{k+1})^2],$$

to show that the ratio of (1.35) to (1.32) is at least 1 is equivalent to showing that

$$-\frac{is}{n} - \frac{d}{n} - \frac{s^2}{2n} - \frac{d^2}{2n} - \frac{1}{2n} - \frac{7}{288n^2} - \frac{i^2}{2n^2} + \frac{0.249}{n} + \log(1+d) + \log\phi_k \;\ge\; 0. \tag{1.44}$$

Proof of (1.44). First suppose that n is even and $n \ge 20$ or n is odd and $n \ge 25$. Apply the bound (1.41) for $d^2/2n$, (1.42) for $\log\phi_k$, (1.43) for s and Lemma 1.8 for $\log(1+d)$. Apply the exact value (1.39) of d in the d/n and λd terms. We assemble together terms with factors k^2, k and no factor of k, getting a lower bound A for (1.44) of the form

$$A \;:=\; \alpha[k^2/n^{3/2}] - 2\beta[k/n^{5/4}] + \gamma[1/n] \tag{1.45}$$

where, if n is even, so $i = 2k$ and $\lambda = 0.9$, we get

$$\alpha = 0.7 - [2.5 - 2(0.98)/3]/\sqrt{n} - 3/n,$$

$$\beta = n^{-3/4} + n^{-5/4}/2,$$

$$\gamma = 0.649 - [17/8 + 7/288]/n - 1/2n^2.$$

Note that for each fixed n, A is $1/n$ times a quadratic in $k/n^{1/4}$. Also, α and γ are increasing in n while β is decreasing. Thus for $n \ge 20$ the supremum of $\beta^2 - \alpha\gamma$ is attained at $n = 20$ where it is < -0.06. So the quadratic has no real roots and since $\alpha > 0$ it is always positive, thus (1.44) holds.

When n is odd, $i = 2k + 1$, $\lambda = 0.93$ and $n \ge 25$. We get a lower bound A for (1.44) of the same form (1.45) where now

$$\alpha = 0.79 - [2.5 - 2(0.98)/3]/\sqrt{n} - 3/n,$$

$$\beta = 1/2n^{1/4} + 2(1 - 0.98/6)/n^{3/4} + 1/2n^{5/4},$$
$$\gamma = 0.679 - (3.625 + 7/288 - 0.98/6)/n - 1/2n^2.$$

For the same reasons, the supremum of $\beta^2 - \alpha\gamma$ for $n \geq 25$ is now attained at $n = 25$ and is negative (less than -0.015), so the conclusion (1.44) again holds.

It remains to consider the case $k = 1$ where n is even and $n \geq 10$ or n is odd and $n \geq 13$. Here instead of bounds for s_k and d_k we use the exact values (1.38) and (1.39) for $k = 1$. We still use the bounds (1.42) for $\log \phi_k$ and Lemma 1.8 for $\log(1+d_k)$. When n is even, $i = 2k = 2$, and we obtain a lower bound A' for (1.44) of the form $a_1/n + a_2/n^{3/2} + \cdots$. All terms n^{-2} and beyond have negative coefficients. Applying the inequality $-n^{-(3/2)-\alpha} \geq -n^{-3/2} \cdot 10^{-\alpha}$ for $n \geq 10$ and $\alpha = 1/2, 1, \cdots$, I found a lower bound $A' \geq 0.662/n - 1.115/n^{3/2} > 0$ for $n \geq 10$. The same method for n odd gave $A' \geq 0.662/n - 1.998/n^{3/2} > 0$ for $n \geq 13$. The proof of (1.5) is complete.

Proof of (1.6). For n odd, (1.6) is clear when $j = 1$, so we can assume $j \geq 3$. For n even, (1.6) is clear when $j = 2$. We next consider the case $j = 0$. By symmetry we need to prove that $p_{n0} \leq P(\sqrt{n}|Y|/2 \leq 1)$. This can be checked from a normal table for $n = 2$. For $n \geq 4$ we have $p_{n0} \leq \sqrt{2/\pi n}$ by (1.32). The integral of the standard normal density from $-2/\sqrt{n}$ to $2/\sqrt{n}$ is clearly larger than the length of the interval times the density at the endpoints, namely $2\sqrt{2/\pi n}\exp(-2/n)$. Since $\exp(-2/n) \geq 1/2$ for $n \geq 4$ the proof for n even and $j = 0$ is done.

We are left with the cases $j \geq 3$. For $j = n$, we have $p_{nn} = 2^{-n}$ and can check the conclusion for $n = 3, 4$ from a normal table. Let ϕ be the standard normal density. We have the inequality, for $t > 0$,

$$P(Y \geq t) \;\geq\; \psi(t) \;:=\; \phi(t)[t^{-1} - t^{-3}], \tag{1.46}$$

Feller (1968), p. 175. Feller does not give a proof. For completeness, here is one:

$$\psi(t) \;=\; -\int_t^\infty \psi'(x)dx \;=\; \int_t^\infty \phi(x)(1 - 3x^{-4})dx \;\leq\; P(Y \geq t).$$

To prove (1.6) via (1.46) for $j = n \geq 5$ we need to prove

$$1/2^n \;\leq\; \phi(t_n)t_n^{-1}(1 - t_n^{-2})$$

where $t_n \;:=\; (n-2)/\sqrt{n}$. Clearly $n \mapsto t_n$ is increasing. For $n \geq 5$ we have $1 - t_n^{-2} \geq 4/9$ and $(2\pi)^{-1/2}e^{2-2/n} \cdot 4/9 \geq 0.878$. Thus it suffices to prove

$$n(\log 2 - 0.5) + 0.5 \log n - \log(n-2) + \log(0.878) \;\geq\; 0, \quad n \geq 5.$$

This can be checked for $n = 5, 6$ and the left side is increasing in n for $n \geq 6$, so (1.6) for $j = n \geq 5$ follows.

So it will suffice to prove $p_{ni} \leq P(\sqrt{n}Y/2 \in [(i-2)/2, i/2])$ for $j \leq i < n$. From (1.30) and Lemma 1.6, and the bound $\phi_k \geq 1$, it will suffice to prove, for $x \;:=\; i/n$,

$$-\frac{1}{4n} + \frac{7}{288n^2} - \frac{(n-1)x^2}{2} + \frac{x^{2n}}{2n(1-x^2)} \;\leq\; -\frac{n[(x - 2/n)^2 + x^2]}{4}$$

where $3/n \leq x \leq 1 - 2/n$. Note that $2n(1 - x^2) \geq 4$. Thus it is enough to prove that

$$x - x^2/2 - x^{2n}/4 \;\geq\; 3/4n \;+\; 7/288n^2$$

for $3/n \le x \le 1$ and $n \ge 5$, which holds since the function on the left is concave, and the inequality holds at the endpoints. Thus (1.6) and Lemma 1.4 are proved. □

1.5 Proof of Lemma 1.2

Let $G(x)$ be the distribution function of a normal random variable Z with mean $n/2$ and variance $n/4$ (the same mean and variance as for $\mathcal{B}(n,1/2)$). Let $B(k,n,1/2) := \sum_{0\le i\le k}\binom{n}{i}2^{-n}$. Lemma 1.4 directly implies

$$G(\sqrt{2kn} - n/2) \le B(k,n,1/2) \le G(k+1) \text{ for } k \le n/2. \tag{1.47}$$

Specifically, letting $k := (n-j)/2$, (1.6) implies

$$B(k,n,1/2) \le P(Z \ge n-k-1) = P(k+1 \ge n - Z) = G(k+1)$$

since $n - Z$ has the same distribution as Z. (1.5) implies

$$B(k,n,1/2) \ge P\left(\frac{n}{2} - \frac{\sqrt{n}}{2}Y \le -\frac{n}{2} + \sqrt{2kn}\right) = G(\sqrt{2kn} - n/2).$$

Let

$$\eta := \Phi_n^{-1}(G(Z)). \tag{1.48}$$

This definition of η from Z is called a quantile transformation. By Theorem 1.3, $G(Z)$ has a $U[0,1]$ distribution and η a $\mathcal{B}(n,1/2)$ distribution. It will be shown that

$$Z - 1 \le \eta \le Z + (Z-n/2)^2/2n + 1 \text{ if } Z \le n/2, \tag{1.49}$$

and

$$Z - (Z-n/2)^2/2n - 1 \le \eta \le Z + 1 \text{ if } Z \ge n/2. \tag{1.50}$$

Define a sequence of extended real numbers $-\infty = c_{-1} < c_0 < c_1 < \cdots < c_n = +\infty$ by $G(c_k) = B(k,n,1/2)$ Then one can check that $\eta = k$ on the event $A_k := \{\omega : c_{k-1} < Z(\omega) \le c_k\}$. By (1.47), $G(c_k) = B(k,n,1/2) \le G(k+1)$ for $k \le n/2$. So, on the set A_k for $k \le n/2$ we have $Z - 1 \le c_k - 1 \le k = \eta$. Note that for n even, $n/2 < c_{n/2}$ while for n odd, $n/2 = c_{(n-1)/2}$. So the left side of (1.49) is proved.

If Y is a standard normal random variable with distribution function Φ and density ϕ then $\Phi(x) \le \phi(x)/x$ for $x > 0$, e.g. Dudley (1993), Lemma 12.1.6(a). So we have

$$P(Z \le -n/2) = P\left(\frac{n}{2} + \frac{\sqrt{n}}{2}Y \le -\frac{n}{2}\right) =$$

$$P\left(\frac{\sqrt{n}}{2}Y \le -n\right) = \Phi(-2\sqrt{n}) \le \frac{e^{-2n}}{2\sqrt{2\pi n}} < \frac{1}{2^n}.$$

So $G(-n/2) < G(c_0) = 2^{-n}$ and $-n/2 < c_0$. Thus if $Z \le -n/2$ then $\eta = 0$. Next note that $Z + (Z-n/2)^2/2n = (Z+n/2)^2/2n \ge 0$ always. Thus the right side of (1.49) holds when $Z \le -n/2$ and whenever $\eta = 0$. Now assume that $Z \ge -n/2$. By (1.47), for $1 \le k \le n/2$

$$G((2(k-1)n)^{1/2} - n/2) \le B(k-1,n,1/2) = G(c_{k-1}),$$

from which it follows that $(2(k-1)n)^{1/2} - n/2 \leq c_{k-1}$ and

$$k - 1 \leq (c_{k-1} + n/2)^2/2n. \tag{1.51}$$

The function $x \mapsto (x + n/2)^2$ is clearly increasing for $x \geq -n/2$ and thus for $x \geq c_0$. Applying (1.51) we get on the set A_k for $1 \leq k \leq n/2$

$$\eta = k \leq (Z + n/2)^2/2n + 1 = Z + (Z - n/2)^2/2n + 1.$$

Since $P(Z \leq n/2) = 1/2 \leq P(\eta \leq n/2)$, and η is a non-decreasing function of Z, $Z \leq n/2$ implies $\eta \leq n/2$. So (1.49) is proved.

It will be shown next that (η, Z) has the same joint distribution as $(n - \eta, n - Z)$. It is clear that η and $n - \eta$ have the same distribution and that Z and $n - Z$ do. We have for each $k = 0, 1, \cdots, n$, $n - \eta = k$ if and only if $\eta = n - k$ if and only if $c_{n-k-1} < Z \leq c_{n-k}$. We need to show that this is equivalent to $c_{k-1} \leq n - Z < c_k$, in other words $n - c_k < Z \leq n - c_{k-1}$. Thus we want to show that $c_{n-k-1} = n - c_k$ for each k. It is easy to check that $G(n - c_k) = P(Z \geq c_k) = 1 - G(c_k)$ while $G(c_k) = B(k, n, 1/2)$ and $G(c_{n-k-1}) = B(n - k - 1, n, 1/2) = 1 - B(k, n, 1/2)$. The statement about joint distributions follows. (1.49) thus implies (1.50).

Some elementary algebra, (1.49) and (1.50) imply

$$|\eta - Z| \leq 1 + (Z - n/2)^2/2n \tag{1.52}$$

and since $Z < n/2$ implies $\eta \leq n/2$ and $Z > n/2$ implies $\eta \geq n/2$,

$$|\eta - n/2| \leq 1 + |Z - n/2|. \tag{1.53}$$

Letting $Z = (n + \sqrt{n}Y)/2$ and noting that then $G(Z) \equiv \Phi(Y)$, (1.48), (1.52), and (1.53) imply Lemma 1.2 with $C_n = \eta - n/2$. □

1.6 Inequalities for the separate processes

We will need facts providing a modulus of continuity for the Brownian bridge and something similar for the empirical process (although it is discontinuous). Let $h(t) := +\infty$ if $t \leq -1$ and

$$h(t) := (1 + t)\log(1 + t) - t, \quad t > -1. \tag{1.54}$$

Lemma 1.9. *Let ξ be a binomial random variable with parameters n and p. Then for any $x \geq 0$ and $m := np$ we have*

$$P(\xi - m \geq x) \leq \inf_{s>0} e^{-sx} E e^{s(\xi - m)} = \left(\frac{m}{m+x}\right)^{m+x} \left(\frac{n-m}{n-m-x}\right)^{n-m-x}. \tag{1.55}$$

If $p \leq 1/2$ then bounds for the right side of (1.55) give

$$P(\xi \geq m + x) \leq \exp\left(-\frac{m}{1-p} h\left(\frac{x}{m}\right)\right) \tag{1.56}$$

and

$$P(\xi \leq m - x) \leq \exp(-x^2/[2p(1-p)]). \tag{1.57}$$

Proof. The first inequality in (1.55) is clear. Let $E(k,n,p)$ denote the probability of at least k successes in n independent trials with probability p of success on each trial, and $B(k,n,p)$ the probability of at most k successes. According to Chernoff's inequalities (Chernoff, 1954), we have with $q := 1-p$

$$E(k,n,p) \le (np/k)^k (nq/(n-k))^{n-k} \quad \text{if} \quad k \ge np,$$

and symmetrically

$$B(k,n,p) \le (np/k)^k (nq/(n-k))^{n-k} \quad \text{if} \quad k \le np.$$

These inequalities hold for k not necessarily an integer; for this and the equality in (1.55) see also Hoeffding (1963). Then for $p \le 1/2$, (1.56) is a consequence proved by Bennett (1962), see also Shorack and Wellner (1986, p. 440, (3)), and (1.57) is a consequence proved by Okamoto (1958) and extended by Hoeffding (1963). □

Let F_n be an empirical distribution function for the uniform distribution on $[0,1]$ and $\alpha_n(t) := \sqrt{n}(F_n(t)-t)$, $0 \le t \le 1$, the corresponding empirical process. The previous lemma extends via martingales to a bound for the empirical process on intervals.

Lemma 1.10. *For any b with $0 < b \le 1/2$ and $x > 0$,*

$$P(\sup_{0 \le t \le b} |\alpha_n(t)| > x/\sqrt{n}) \le 2\exp\left(-\frac{nb}{1-b} h\left(\frac{x(1-b)}{nb}\right)\right)$$
$$\le 2\exp(-nb(1-b)h(x/(nb))). \tag{1.58}$$

Remark. The bound given by (1.58) is Lemma 2 of Bretagnolle and Massart (1989). Lemma 1.2 of Csörgő and Horváth (1993), p. 116, has instead the bound $2\exp(-nbh(x/(nb)))$. This does not follow from Lemma 1.10, while the converse implication holds by (1.83) below, but I could not follow Csörgő and Horváth's proof of their form.

Proof. From the binomial conditional distributions of multinomial variables we have for $0 \le s \le t < 1$

$$E(F_n(t)|F_n(u),\ u \le s) = E(F_n(t)|F_n(s))$$
$$= F_n(s) + \frac{t-s}{1-s}(1-F_n(s)) = \frac{t-s}{1-s} + \frac{1-t}{1-s}F_n(s),$$

from which it follows directly that

$$E\left(\frac{F_n(t)-t}{1-t}\Big| F_n(u),\ u \le s\right) = \frac{F_n(s)-s}{1-s},$$

in other words, the process $(F_n(t)-t)/(1-t)$, $0 \le t < 1$ is a martingale in t (here n is fixed). Thus, $\alpha_n(t)/(1-t)$, $0 \le t < 1$, is also a martingale, and for any real s the process $\exp(s\alpha_n(t)/(1-t))$ is a submartingale, e.g. Dudley (1993), 10.3.3(b). Then

$$P(\sup_{0 \le t \le b} \alpha_n(t) > x/\sqrt{n}) \le P(\sup_{0 \le t \le b} \alpha_n(t)/(1-t) > x/\sqrt{n})$$

which for any $s > 0$ equals

$$P\left(\sup_{0\le t\le b} \exp(s\alpha_n(t)/(1-t)) > \exp(sx/\sqrt{n})\right).$$

By Doob's inequality (e.g. Dudley (1993), 10.4.2, for a finite sequence increasing up to a dense set) the latter probability is

$$\le \inf_{s>0} \exp(-sx/\sqrt{n}) E \exp(s\alpha_n(b)/(1-b)) \ \le \ \exp\left(-\frac{nb}{1-b} h\left(\frac{x(1-b)}{nb}\right)\right)$$

by Lemma 1.9, (1.56). In the same way, by (1.57) we get

$$P(\sup_{0\le t\le b} (-\alpha_n(t)) > x/\sqrt{n}) \ \le \ \exp(-x^2(1-b)/(2nb))). \tag{1.59}$$

It is easy to check that $h(u) \le u^2/2$ for $u \ge 0$, so the first inequality in Lemma 1.10 follows. It is easily shown by derivatives that $h(qy) \ge q^2 h(y)$ for $y \ge 0$ and $0 \le q \le 1$. For $q = 1-b$, the bound in (1.58) then follows. □

We next have a corresponding inequality for the Brownian bridge.

Lemma 1.11. *Let $B(t)$, $0 \le t \le 1$, be a Brownian bridge, $0 < b < 1$ and $x > 0$. Let Φ be the standard normal distribution function. Then*

$$P(\sup_{0\le t\le b} B(t) > x) \ = \ 1 - \Phi(x/\sqrt{b(1-b)})$$

$$+ \exp(-2x^2)\left(1 - \Phi\left(\frac{(1-2b)x}{\sqrt{b(1-b)}}\right)\right). \tag{1.60}$$

If $0 < b \le 1/2$, then for all $x > 0$,

$$P(\sup_{0\le t\le b} B(t) > x) \ \le \ \exp(-x^2/(2b(1-b))). \tag{1.61}$$

Proof. Let $X(t)$, $0 \le t < \infty$ be a Wiener process. For some real α and value of $X(1)$ let $\beta := X(1) - \alpha$. It will be shown that for any real α and y

$$P\{\sup_{0\le t\le 1} X(t) - \alpha t > y | X(1)\} \ = \ 1_{\{\beta > y\}} + \exp(-2y(y-\beta)) 1_{\{\beta \le y\}}. \tag{1.62}$$

Clearly, if $\beta > y$ then $\sup_{0\le t\le 1} X(t) - \alpha t > y$ (let $t = 1$). Suppose $\beta \le y$. One can apply a reflection argument as in the proof of Dudley (1993), Proposition 12.3.3, where details are given on making such an argument rigorous. Let $X(t) = B(t) + tX(1)$ for $0 \le t \le 1$, where $B(\cdot)$ is a Brownian bridge. We want to find $P(\sup_{0\le t\le 1} B(t) + \beta t > y)$. But this is the same as $P(\sup_{0\le t\le 1} Y(t) > y | Y(1) = \beta)$ for a Wiener process Y. For $\beta \le y$, the probability that $\sup_{0\le t\le 1} Y(t) > y$ and $\beta \le Y(1) \le \beta + dy$ is the same by reflection as $P(2y - \beta \le Y(1) \le 2y - \beta + dy)$. Thus the desired conditional probability, for the standard normal density ϕ, is $\phi(2y-\beta)/\phi(\beta) = \exp(-2y(y-\beta))$ as stated. So (1.62) is proved.

We can write the Brownian bridge B as $W(t) - tW(1)$, $0 \le t \le 1$, for a Wiener process W. Let $W_1(t) := b^{-1/2}W(bt)$, $0 \le t < \infty$. Then W_1 is a Wiener process. Let $\eta := W(1) - W(b)$. Then η has a normal $N(0, 1-b)$ distribution and is independent of $W_1(t)$, $0 \le t \le 1$. Let $\gamma := ((1-b)W_1(1) - \sqrt{b}\eta)\sqrt{b}/x$. We have

$$P(\sup_{0 \le t \le b} B(t) > x | \eta, W_1(1)) = P\left(\sup_{0 \le t \le 1} (W_1(t) - (bW_1(1) + \sqrt{b}\eta)t) > x/\sqrt{b} | \eta, W_1(1)\right).$$

Now the process $W_1(t) - (bW_1(t) + \sqrt{b}\eta)t$, $0 \le t \le 1$, has the same distribution as a Wiener process $Y(t)$, $0 \le t \le 1$, given that $Y(1) = (1-b)W_1(1) - \sqrt{b}\eta$. Thus by (1.62) with $\alpha = 0$,

$$P(\sup_{0 \le t \le b} B(t) > x | \eta, W_1(1)) = 1_{\{\gamma > 1\}} + 1_{\{\gamma \le 1\}} \exp(-2x^2(1-\gamma)/b). \tag{1.63}$$

Thus, integrating gives

$$P(\sup_{0 \le t \le b} B(t) > x) = P(\gamma > 1) + \exp(-2x^2/b) E\left(\exp(2x^2\gamma/b) 1_{\{\gamma \le 1\}}\right).$$

From the definition of γ it has a $N(0, b(1-b)/x^2)$ distribution. Since x is constant, the latter integral with respect to γ can be evaluated by completing the square in the exponent and yields (1.60).

We next need the inequality, for $x \ge 0$,

$$1 - \Phi(x) \le \frac{1}{2} \exp(-x^2/2). \tag{1.64}$$

This is easy to check via the first derivative for $0 \le x \le \sqrt{2/\pi}$. On the other hand we have the inequality $1 - \Phi(x) \le \phi(x)/x$, $x > 0$, e.g. Dudley (1993), 12.1.6(a), which gives the conclusion for $x \ge \sqrt{2/\pi}$.

Applying (1.64) to both terms of (1.60) gives (1.61), so the Lemma is proved. □

1.7 Proof of Theorem 1.1

For the Brownian bridge $B(t)$, $0 \le t \le 1$, it is well known that for any $x > 0$

$$P(\sup_{0 \le t \le 1} |B(t)| \ge x) \le 2 \exp(-2x^2),$$

e.g. Dudley (1993), Proposition 12.3.3. It follows that

$$P(\sqrt{n} \sup_{0 \le t \le 1} |B(t)| \ge u) \le 2 \exp(-u/3)$$

for $u \ge n/6$. We also have $|\alpha_1(t)| \le 1$ for all t and

$$P(\sup_{0 \le t \le 1} |\alpha_n(t)| \ge x) \le D \exp(-2x^2), \tag{1.65}$$

which is the Dvoretzky-Kiefer-Wolfowitz inequality with a constant D. Massart (1990) proved (1.65) with the sharp constant $D = 2$. Earlier Hu (1985) proved it with $D = 4\sqrt{2}$. $D = 6$ suffices for present purposes. Given D, it follows that for $u \geq n/6$,

$$P(\sqrt{n} \sup_{0 \leq t \leq 1} |\alpha_n(t)| \geq u) \ \leq \ D \exp(-u/3).$$

For $x < 6 \log 2$, we have $2e^{-x/6} > 1$ so the conclusion of Theorem 1.1 holds. holds. For $x > n/3 - 12 \log n$, $u \ := \ (x + 12 \log n)/2 > n/6$ so the left side of (1.1) is bounded above by $(2 + D)n^{-2}e^{-x/6}$. We have $(2 + D)n^{-2} \leq 2$ for $n \geq 2$ and $D \leq 6$.

Thus it will be enough to prove Theorem 1.1 when

$$6 \log 2 \leq x \leq n/3 \ - \ 12 \log n. \tag{1.66}$$

The function $t \mapsto t/3 - 12 \log t$ is decreasing for $t < 36$, increasing for $t > 36$. Thus one can check that for (1.66) to be non-vacuous is equivalent to

$$n \ \geq \ 204. \tag{1.67}$$

Let N be the largest integer such that $2^N \leq n$, so that $\nu \ := \ 2^N \leq n < 2\nu$. Let Z be a ν-dimensional normal random variable with independent components, each having mean 0 and variance $\lambda \ := \ n/\nu$. For integers $0 \leq i < m$ let $A(i, m) \ := \ \{i + 1, \cdots, m\}$. For any two vectors $a \ := \ (a_1, \cdots, a_\nu)$ and $b \ := \ (b_1, \cdots, b_\nu)$ in $\mathbb{R}^\nu$, we have the usual inner product $(a, b) \ := \ \sum_{i=1}^{\nu} a_i b_i$. For any subset $D \subset A(0, \nu)$ let 1_D be its indicator function as a member of $\mathbb{R}^\nu$. For any integers $j = 0, 1, 2, \cdots$ and $k = 0, 1, \cdots$, let

$$I_{j,k} \ := \ A(2^j k, 2^j (k + 1)), \tag{1.68}$$

let $e_{j,k}$ be the indicator function of $I_{j,k}$ and for $j \geq 1$, let $e'_{j,k} \ := \ e_{j-1,2k} - e_{j,k}/2$. Then one can easily check that the family $\mathcal{E} \ := \ \{e'_{j,k} : \ 1 \leq j \leq N, \ 0 \leq k < 2^{N-j}\} \cup \{e_{N,0}\}$ is an orthogonal basis of $\mathbb{R}^\nu$ with $(e_{N,0}, e_{N,0}) = \nu$ and $(e'_{j,k}, e'_{j,k}) = 2^{j-2}$ for each of the given j, k. Let $W_{j,k} \ := \ (Z, e_{j,k})$ and $W'_{j,k} \ := \ (Z, e'_{j,k})$. Then since the elements of $\mathcal{E}$ are orthogonal it follows that the random variables $W'_{j,k}$ for $1 \leq j \leq N, \ 0 \leq k < 2^{N-j}$ and $W_{N,0}$ are independent normal with

$$EW'_{j,k} \ = \ EW_{N,0} \ = \ 0, \ \ \mathrm{Var}(W'_{j,k}) = \lambda 2^{j-2}, \ \mathrm{Var}(W_{N,0}) = \lambda \nu. \tag{1.69}$$

Recalling the notation of Lemma 1.2, let Φ_n be the distribution function of a binomial $\mathcal{B}(n, 1/2)$ random variable, with inverse Φ_n^{-1}. Now let $G_m(t) \ := \ \Phi_m^{-1}(\Phi(t))$.

We will begin defining the construction that will connect the empirical process with a Brownian bridge. Let

$$U_{N,0} \ := \ n \tag{1.70}$$

and then recursively as j decreases from $j = N$ to $j = 1$,

$$U_{j-1,2k} \ := \ G_{U_{j,k}}((2^{2-j}/\lambda)^{1/2} W'_{j,k}), \quad U_{j-1,2k+1} \ := \ U_{j,k} - U_{j-1,2k}, \tag{1.71}$$

$k = 0, 1, \cdots, 2^{N-j} - 1$. Note that by (1.69), $(2^{2-j}/\lambda)^{1/2} W'_{j,k}$ has a standard normal distribution, so Φ of it has a $U[0, 1]$ distribution. It is easy to verify successively for $j = N, N - 1, \cdots, 0$ that the random vector $\{U_{j,k}, \ 0 \leq k < 2^{N-j}\}$ has a multinomial distribution with parameters

$n, 2^{j-N}, \cdots, 2^{j-N}$. Let $X := (U_{0,0}, U_{0,1}, \cdots, U_{0,\nu-1})$. Then the random vector X has a multinomial distribution with parameters $n, 1/\nu, \cdots, 1/\nu$.

The random vector X is equal in distribution to

$$\{n(F_n((k+1)/\nu) - F_n(k/\nu)),\ 0 \le k \le \nu - 1\}, \tag{1.72}$$

while for a Wiener process W, Z is equal in distribution to

$$\{\sqrt{n}(W((k+1)/\nu) - W(k/\nu)),\ 0 \le k \le \nu - 1\}. \tag{1.73}$$

Without loss of generality, we can assume that the above equalities in distribution are actual equalities for some uniform empirical distribution functions F_n and Wiener process $W = W_n$. Specifically, consider a vector of i.i.d. uniform random variables $(x_1, \cdots, x_n) \in \mathbb{R}^n$ such that

$$F_n(t) := \frac{1}{n} \sum_{j=1}^{n} 1_{\{x_j \le t\}}$$

and note that W has sample paths in $C[0,1]$. Both $\mathbb{R}^n$ and $C[0,1]$ are separable Banach spaces. Thus one can let $(x_1, \cdots, x_n)$ and W be conditionally independent given the vectors in (1.72) and (1.73) which have the joint distribution of X and Z, by the Vorob'ev-Berkes-Philipp theorem, see Berkes and Philippp (1979), Lemma A1. Then we define a Brownian bridge by $B_n(t) := W_n(t) - tW_n(1)$ and the empirical process $\alpha_n(t) := \sqrt{n}(F_n(t) - t)$, $0 \le t \le 1$. By our choices, we then have

$$\{n(F_n(j/\nu) - j/\nu)\}_{j=0}^{\nu} = \left\{ \sum_{i=0}^{j-1} \left(X_i - \frac{n}{\nu} \right) \right\}_{j=0}^{\nu} \tag{1.74}$$

and

$$\{\sqrt{n} B_n(j/\nu)\}_{j=0}^{\nu} = \left\{ \left(\sum_{i=0}^{j-1} Z_i \right) - \frac{j}{\nu} \sum_{r=0}^{\nu-1} Z_r \right\}_{j=0}^{\nu}. \tag{1.75}$$

Theorem 1.1 will be proved for the given B_n and α_n. Specifically, we want to prove

$$P_0 := P\left(\sup_{0 \le t \le 1} |\alpha_n(t) - B_n(t)| > (x + 12 \log n)/\sqrt{n} \right) \le 2 \exp(-x/6). \tag{1.76}$$

It will be shown that $\alpha_n(j/\nu)$ and $B_n(j/\nu)$ are not too far apart for $j = 0, 1, \cdots, \nu$ while the increments of the processes over the intervals between the lattice points j/ν are also not too large.

Let $C := 0.29$. Let M be the least integer such that

$$C(x + 6 \log n) \le \lambda 2^{M+1}. \tag{1.77}$$

Since $n \ge 204$ (1.67) and $\lambda < 2$ this implies $M \ge 2$. We have by definition of M and (1.66)

$$2^M \le \lambda 2^M \le C(x + 6 \log n) \le Cn/3 < 0.1 \cdot 2^{N+1} < 2^{N-2}$$

so $M \le N - 3$.

For each $t \in [0,1]$, let $\pi_M(t)$ be the nearest point of the grid $\{i/2^{N-M},\ 0 \le i \le 2^{N-M}\}$, or if there are two nearest points, take the smaller one. Let $D := X - Z$ and $D(m) := \sum_{i=1}^{m} D_i$. Let $C' := 0.855$ and define

$$\Theta := \{U_{j,k} \le \lambda(1+C')2^j \ \text{ whenever } \ M+1 < j \le N,\ 0 \le k < 2^{N-j}\}$$

$$\cap\, \{U_{j,k} \ge \lambda(1-C')2^j \ \text{ whenever } \ M < j \le N,\ 0 \le k < 2^{N-j}\}.$$

Then

$$P_0 \ \le \ P_1 + P_2 + P_3 + P(\Theta^c)$$

where

$$P_1 := P\left(\sup_{0\le t\le 1} |\alpha_n(t) - \alpha_n(\pi_M(t))| > 0.28(x + 6\log n)/\sqrt{n}\right), \tag{1.78}$$

$$P_2 := P\left(\sup_{0\le t\le 1} |B_n(t) - B_n(\pi_M(t))| > 0.22(x + 6\log n)/\sqrt{n}\right), \tag{1.79}$$

and, recalling (1.74) and (1.75),

$$P_3 := 2^{N-M} \max_{m\in A(M)} P\left\{\left(|D(m) - \frac{m}{\nu}D(\nu)| > 0.5x + 9\log n\right) \cap \Theta\right\}, \tag{1.80}$$

where $A(M) := \{k2^M : \ k = 1, 2, \cdots\} \cap A(0,\nu)$.

First we bound $P(\Theta^c)$. Since by (1.71) $U_{j,k} = U_{j-1,2k} + U_{j-1,2k+1}$, we have

$$\Theta^c \subset \bigcup_{0\le k<2^{N-M-2}} \{U_{M+2,k} > (1+C')\lambda 2^{M+2}\} \cup \bigcup_{0\le k<2^{N-M-1}} \{U_{M+1,k} < (1-C')\lambda 2^{M+1}\}.$$

Since $U_{M+2,k}$ and $U_{M+1,k}$ are binomial random variables, Lemma 1.9 gives

$$P(\Theta^c) \ \le \ 2^{N-M-1}\left(\exp(-\lambda 2^{M+2}h(C')) + \exp(-\lambda 2^{M+1}h(-C'))\right).$$

Now $2h(C') \ge 0.5823 \ge h(-C') \ge 0.575$ (note that C' has been chosen to make $2h(C')$ and $h(-C')$ approximately equal). By definition of M (1.77), $\lambda 2^{M+1} \ge C(x + 6\log n)$, and $0.575C > 1/6$, so

$$P(\Theta^c) \ \le \ 2^{-M}\exp(-x/6). \tag{1.81}$$

Next, to bound P_1 and P_2. Let $b := 2^{M-N-1} \le 1/2$. Since $\alpha_n(t)$ has stationary increments, we can apply Lemma 1.10. Let $u := x + 6\log n$. We have by definition of M (1.77)

$$nb \ = \ n2^{M-N-1} \ < \ Cu/2. \tag{1.82}$$

By (1.66), $u < n/3$ so $b < C/6$. Recalling (1.54), note that $h'(t) \equiv \log(1+t)$. Thus h is increasing. For any given $v > 0$ it is easy to check that

$$y \mapsto yh(v/y) \ \text{ is decreasing for } \ y > 0. \tag{1.83}$$

Lemma 1.10 gives

$$P_1 \ \le \ 2^{N-M+2}\exp\left(-nb(1-b)h\left(\frac{0.28u}{nb}\right)\right)$$

$$< \ 2^{N-M+2} \exp\left(-\frac{C}{2}\left[1-\frac{C}{6}\right] uh\left(0.28\cdot\frac{2}{C}\right)\right)$$

by (1.83) and (1.82) and since $1-b > 1-C/6$, so one can calculate

$$P_1 \ \le \ 2^{N-M+2} e^{-u/6} \ \le \ 2^{2-M}\lambda^{-1}\exp(-x/6). \tag{1.84}$$

The Brownian bridge also has stationary increments, so Lemma 1.11, (1.61) and (1.82) give

$$P_2 \ \le \ 2^{N-M+2}\exp(-(0.22u)^2/(2nb))$$

$$\le \ 2^{N-M+2}\exp(-(0.22)^2 u/C) \ \le \ 2^{2-M}\lambda^{-1}e^{-x/6} \tag{1.85}$$

since $(0.22)^2/C > 1/6$.

It remains to bound P_3. Fix $m \in A(M)$. A bound is needed for

$$P_3(m) \ := \ P\left\{\left(|D(m) - \frac{m}{\nu}D(\nu)| > 0.5x + 9\log n\right) \cap \Theta\right\}. \tag{1.86}$$

For each $j = 1, \cdots, N$ take $k(j)$ such that $m \in I_{j,k(j)}$. By the definition (1.68) of $I_{j,k}$, $k(M) = m2^{-M} - 1$ and $k(j) = [k(j-1)/2]$ for $j = 1, \cdots, N$ where $[x]$ is the largest integer $\le x$. From here on each double subscript $j, k(j)$ will be abbreviated to the single subscript j, e.g. $e'_j \ := \ e'_{j,k(j)}$. The following orthogonal expansion holds in $\mathcal{E}$:

$$1_{A(0,m)} \ = \ \frac{m}{\nu}e_{N,0} + \sum_{M<j\le N} c_j e'_j, \tag{1.87}$$

where $0 \le c_j \le 1$ for $m < j \le N$. To see this, note that $1_{A(0,m)} \perp e'_{j,k}$ for $j \le M$ since 2^M is a divisor of m. Also, $1_{A(0,m)} \perp e'_{j,k}$ for $k \ne k(j)$ since $1_{A(0,m)}$ has all 0's or all 1's on the set where $e'_{j,k}$ has non-zero entries, half of which are $+1/2$ and the other half $-1/2$. In an orthogonal expansion $f = \sum_j c_j f_j$ we always have $c_j = (f, f_j)/\|f_j\|^2$ where $\|v\|^2 \ := \ (v, v)$. We have $\|e'_j\| = 2^{(j-2)/2}$. Now, $(1_{A(0,m)}, e'_j)$ is as large as possible when the components of e'_j equal $= 1/2$ only for indices $\le m$, and then the inner product equals 2^{j-2}, so $|c_j| \le 1$ as stated. The m/ν factor is clear.

We next have

$$e_j \ = \ 2^{j-N}e_{N,0} + \sum_{i>j}(-1)^{s(i,j,m)}2^{j+1-i}e'_i \tag{1.88}$$

where $s(i,j,m) = 0$ or 1 for each i, j, m so that the corresponding factors are ± 1, the signs being immaterial in what follows. Let $\Delta_j \ := \ (D, e'_j)$. Then from (1.87),

$$\left|D(m) - \frac{m}{\nu}D(\nu)\right| \ \le \ \sum_{M<j\le N}|\Delta_j|. \tag{1.89}$$

Recall that $W'_j = (Z, e'_j)$ (see between (1.68) and (1.69)) and $D = X - Z$. Let $\xi_j \ := \ (2^{2-j}/\lambda)^{1/2}W'_j$ for $M < j \le N$. Then by (1.69) and the preceding statement, $\xi_{M+1}, \cdots, \xi_N$ are i.i.d. standard normal random variables. We have $U_{j,k} = (X, e_{j,k})$ for all j and k from the definitions. Then $U_j = (X, e_j)$. Let $U'_j = (X, e'_j)$. By (1.71) and Lemma 1.2, (1.4),

$$|U'_j \ - \ \sqrt{U_j}\xi_j/2| \ \le \ 1 + \xi_j^2/8. \tag{1.90}$$

Let

$$L_j \;:=\; |W_j' - \sqrt{U_j}\xi_j/2| \;=\; |\xi_j||\sqrt{U_j} - \sqrt{\lambda 2^j}|/2$$

by definition of ξ_j. Thus

$$|\Delta_j| \;\le\; L_j + 1 + \xi_j^2/8. \tag{1.91}$$

Then we have on Θ

$$|\sqrt{U_j} - \sqrt{\lambda 2^j}| \;=\; |U_j - \lambda 2^j|/(\sqrt{\lambda 2^j} + \sqrt{U_j}) \;\le\; \frac{|U_j - \lambda 2^j|}{\sqrt{\lambda 2^j}} \cdot \frac{1}{1+\sqrt{1-C'}},$$

where as before $C' \;:=\; 0.855$. Then by (1.71), (1.88) and (1.3) of Lemma 1.2,

$$\begin{aligned} |U_j - \lambda 2^j| \;&\le\; 2^{j-N}|U_N - n| + 2\sum_{j<i\le N} 2^{j-i}|U_i'| \\ &\le\; 2 + (\lambda(1+C'))^{1/2} \sum_{j<i\le N} 2^{j-i/2}|\xi_i| \end{aligned}$$

on Θ, recalling that by (1.70), $U_N = U_{N,0} = n$. Let $C_2 \;:=\; 1/(1+\sqrt{1-C'})$. It follows that

$$L_j \;\le\; 2^{-j/2}C_2|\xi_j| + \frac{1}{2}C_2\sqrt{1+C'} \sum_{j<i\le N} 2^{(j-i)/2}|\xi_j||\xi_i|. \tag{1.92}$$

Applying the inequality $|\xi_i||\xi_j| \le (\xi_i^2 + \xi_j^2)/2$, we get the bound

$$\sum_{M<j\le N}\sum_{j<i\le N} 2^{(j-i)/2}|\xi_i\xi_j| \;\le\; \sum_{M<j\le N} A_j\xi_j^2 \tag{1.93}$$

where

$$A_j \;:=\; \frac{1}{2}\left(\sum_{M<r<j} 2^{(r-j)/2} + \sum_{j<i\le N} 2^{(j-i)/2}\right).$$

Then

$$\begin{aligned} A_j \;&\le\; \frac{1}{2}\left[\frac{2^{-1/2} - 2^{(M-j)/2}}{1-2^{-1/2}} + \frac{2^{-1/2}}{1-2^{-1/2}}\right] \\ &\le\; 1 + \sqrt{2} - 2^{(M-j-2)/2}/(1-2^{-1/2}). \end{aligned}$$

Let $C_3 \;:=\; C_2(1+\sqrt{2})\sqrt{1+C'}/2 \;\le\; 1.19067$. Then

$$\sum_{M<j\le N} L_j \;\le\; C_3 \sum_{M<j\le N}\xi_j^2 + \sum_{M<j\le N} 2^{-j/2}|\xi_j| C_2\left(1 - \frac{\sqrt{1+C'}}{2}2^{(M-2)/2}|\xi_j|/(1-2^{-1/2})\right). \tag{1.94}$$

Let

$$C_4 \;:=\; \frac{\sqrt{1+C'}}{4(1-2^{-1/2})} \;=\; \frac{\sqrt{2}\sqrt{1+C'}(\sqrt{2}+1)}{4},$$

and for each M let $c_M \;:=\; 1/(4C_4 2^{M/2})$. Then for any real number x, we have $x(1-C_4 2^{M/2}x) \le c_M$. It follows that

$$\sum_{M<j\le N} L_j \;\le\; \sum_{M<j\le N} C_3\xi_j^2 + c_M C_2 2^{-j/2}$$

$$\leq \; C_2 c_M 2^{-(M+1)/2}/(1-2^{-1/2}) \; + \sum_{M<j\leq N} C_3 \xi_j^2$$

$$\leq \; \frac{C_2 2^{-M}}{\sqrt{2}\sqrt{1+C'}} \; + \sum_{M<j\leq N} C_3 \xi_j^2.$$

Thus, combining (1.91) and (1.94) we get on Θ

$$\sum_{M<j\leq N} |\Delta_j| \; \leq \; N + \left(\frac{1}{8} + C_3\right) \sum_{M<j\leq N} \xi_j^2. \tag{1.95}$$

We have $E\exp(t\xi^2) = (1-2t)^{-1/2}$ for $t < 1/2$ and any standard normal variable ξ such as ξ_j for each j. Since $\xi_{M+1}, \cdots, \xi_N$ are independent we get

$$E\exp\left(\left(\frac{1}{3}\sum_{M<j\leq N} |\Delta_j|\right) 1_\Theta\right) \; \leq \; e^{N/3}\left(1 - \frac{2}{3}\left(C_3 + \frac{1}{8}\right)\right)^{(M-N)/2}$$

$$\leq \; e^{N/3} 2^{1.513(N-M)} \; \leq \; 2^{2N-1.5M}.$$

Markov's inequality and (1.89) then yield

$$P_3(m) \; \leq \; e^{-x/6} n^{-3} 2^{2N-1.5M}.$$

Thus

$$P_3 \; \leq \; e^{-x/6} n^{-3} 2^{3N-2.5M} \; \leq \; 2^{-2.5M} e^{-x/6}. \tag{1.96}$$

Collecting (1.81), (1.84), (1.85) and (1.96) we get that $P_0 \leq (2^{3-M}\lambda^{-1} + 2^{-M} + 2^{-2.5M})e^{-x/6}$. By (1.77) and (1.67) and since $x \geq 6\log 2$ (1.66) and $M \geq 2$, it follows that Theorem 1.1 holds. □

REFERENCES

Bennett, George W. (1962). Probability inequalities for the sum of bounded random variables. *J. Amer. Statist. Assoc.* **57**, 33–45.

Berkes, I., and Philipp, W. (1979). Approximation theorems for independent and weakly dependent random vectors. *Ann. Probab.* **7**, 29-54.

Bretagnolle, J., and Massart, P. (1989). Hungarian constructions from the nonasymptotic viewpoint. *Ann. Probab.* **17**, 239–256.

Chernoff, H. (1952). A measure of efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.* **23**, 493-507.

Csörgő, M., and Horváth, L. (1993). *Weighted Approximations in Probability and Statistics.* Wiley, Chichester.

Csörgő, M., and Révész, P. (1981). *Strong Approximations in Probability and Statistics.* Academic, New York.

Donsker, Monroe D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.* **23**, 277–281.

Dudley, Richard M. (1984). *A Course on Empirical Processes.* Ecole d'été de probabilités de St.-Flour, 1982. Lecture Notes in Math. **1097**, 1-142, Springer.

Dudley, R. M. (1993). *Real Analysis and Probability.* Second printing, corrected. Chapman and Hall, New York.

Feller, William (1968). *An Introduction to Probability Theory and Its Applications.* Vol. 1, 3d ed. Wiley, New York.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13-30.

Hu, Inchi (1985). A uniform bound for the tail probability of Kolmogorov-Smirnov statistics. *Ann. Statist.* **13**, 821-826.

Komlós, J., Major, P., and Tusnády, G. (1975). An approximation of partial sums of independent RV'-s and the sample DF. I. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **32**, 111–131.

Mason, D. M. (1998). Notes on the the KMT Brownian bridge approximation to the uniform empirical process. Preprint.

Mason, D. M., and van Zwet, W. (1987). A refinement of the KMT inequality for the uniform empirical process. *Ann. Probab.* **15**, 871-884.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* **18**, 1269-1283.

Nanjundiah, T. S. (1959). Note on Stirling's formula. *Amer. Math. Monthly* **66**, 701-703.

Okamoto, Masashi (1958). Some Inequalities Relating to the Partial Sum of Binomial Probabilities. *Ann. Inst. Statist. Math.* **10**, 29-35.

Rio, E. (1991). Local invariance principles and its application to density estimation. Prépubl Math. Univ. Paris-Sud 91-71.

Rio, E. (1994). Local invariance principles and their application to density estimation. *Probab. Theory Related Fields* **98**, 21-45.

Shorack, G., and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

Whittaker, E. T., and Watson, G. N. (1927). *Modern Analysis*, 4th ed., Cambridge Univ. Press, Repr. 1962.

Chapter 2

Gaussian Measures and Processes; Sample Continuity

Let $X_1, X_2, \ldots,$ be independent, identically distributed real-valued random variables with $EX_1 = 0$ and $EX_1^2 = \sigma^2 < \infty$. Let $S_n := X_1 + \cdots + X_n$. Then the one-dimensional central limit theorem says that the distribution of $S_n/n^{1/2}$ converges as $n \to \infty$ to the normal distribution $N(0, \sigma^2)$, which (if $\sigma > 0$) has a density $\sigma^{-1}(2\pi)^{-1/2} \exp(-x^2/(2\sigma^2))$ with respect to Lebesgue measure on $\mathbb{R}$ (RAP, Theorem 9.5.6). Also, if the X_i are i.i.d. with values in $\mathbb{R}^k$, $EX_1 = 0$ and $E|X_1|^2 < \infty$, then the distribution of $S_n/n^{1/2}$ converges to a normal distribution $N(0, C)$ where C is the covariance matrix of X_1 (*ibid.*).

These notes are mainly about extensions of the central limit theorem to infinite-dimensional situations. Here the limit distributions will be normal distributions on infinite-dimensional spaces. Since their behavior is not as simple as in the finite-dimensional case, this chapter is devoted to a study of normal or Gaussian measures.

2.1 Some definitions.

Let X be a real vector space. Recall that a *seminorm* is a function $\|\cdot\|$ from X into the nonnegative real numbers such that $\|x+y\| \leq \|x\| + \|y\|$ for all x and y in X and $\|cx\| = |c|\|x\|$ for all real c and $x \in X$. The seminorm $\|\cdot\|$ is called a *norm* if $\|x\| = 0$ only for $x = 0$ in X, and then $(X, \|\cdot\|)$ is called a *normed linear space*. A norm defines a metric by $d(x, y) := \|x - y\|$. A normed linear space complete for this metric is called a *Banach space*. As with any metric space, it is called *separable* if it has a countable dense subset. A probability distribution P defined on a separable Banach space will be assumed to be defined on the Borel σ-algebra generated by the open sets, unless another σ-algebra is specified. Then P will be called a *law*.

Let $(X, \|\cdot\|)$ be a separable Banach space. A law P on X will be called *Gaussian* or *normal* iff for every continuous linear form $f \in X'$, $P \circ f^{-1}$ is a normal law on $\mathbb{R}$. Recall that a law on a finite-dimensional real vector space is normal if and only if every real linear form is normally distributed (RAP, Theorem 9.5.13).

2.2 Gaussian vectors are probably not very large.

First, let's have some bounds for one-dimensional Gaussian variables. Let Φ be the standard normal distribution function and ϕ its density function, so $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ for all real x, and $\Phi(x) = \int_{-\infty}^{x} \phi(u)du$.

Proposition 2.1. *Let X be a real-valued random variable with a normal distribution $N(0, \sigma^2)$. Then*
(a) for any $M > 0$, $\Pr(|X| > M) \leq \exp\left(-M^2/\left(2\sigma^2\right)\right)$;
(b) if $M/\sigma \geq 1$, then

$$\frac{\sigma}{M}\phi\left(\frac{M}{\sigma}\right) \leq \Pr(|X| > M) \leq \frac{2\sigma}{M}\phi\left(\frac{M}{\sigma}\right).$$

This section will give an extension of inequality (a) to infinite-dimensional Gaussian variables such as those taking values in separable Banach spaces. It will be said that a law P on a separable Banach space $(X, \|\cdot\|)$ has mean 0 if $\int \|x\| dP(x) < \infty$ and $\int f(x)dP(x) = 0$ for each $f \in X'$. Here is one of the main results.

Theorem 2.2. (Landau-Shepp-Marcus-Fernique) *Let P be a normal law with mean 0 on a separable Banach space X. For $f \in X'$ let $\sigma^2(f) := \int f^2 dP$. Then $\tau^2 := \sup\{\sigma^2(f) : \ \|f\|' \leq 1\} < \infty$ and*

$$\int \exp(\alpha\|x\|^2)dP(x) \quad < \quad \infty \quad \textit{for any} \quad \alpha < 1/(2\tau^2).$$

By Proposition 2.1(a), the theorem holds in the one-dimensional case, and by the left side of part (b), the condition $\alpha < 1/(2\tau^2)$ is best possible. Some related facts will be given.

Definition. Let X be a real vector space and $\mathcal{B}$ a σ-algebra of subsets of X. Then $(X, \mathcal{B})$ is called a *measurable vector space* if both
(a) addition is jointly measurable from $X \times X$ to X, and
(b) scalar multiplication is jointly measurable from $\mathbb{R} \times X$ to X (for the usual Borel σ-algebra on $\mathbb{R}$).

Example. Let X be a topological vector space, namely a vector space with a topology for which (a) and (b) hold with "measurable" replaced by "continuous". Suppose the topology of X is metrizable and separable. For a Cartesian product of two separable metric spaces, since their topologies have countable bases, the Borel σ-algebra in the product equals the product σ-algebra of the Borel σ-algebras in the two spaces (RAP, Proposition 4.1.7). Thus X with its Borel σ-algebra is a measurable vector space.

The notion of normal law can't be defined for general measurable vector spaces by way of linear forms, as it was for Banach spaces in the last section, since there exist measurable vector spaces, such as spaces $L^p[0,1]$ for $0 < p < 1$, which have non-trivial normal measures but turn out to have no non-trivial measurable linear forms (Appendix F). Fernique (1970) proposed the following ingenious definition:

Definition. A probability measure P on a measurable vector space $(X, \mathcal{B})$ will be called *centered Gaussian* iff for variables U and V independent with law P (say, coordinates on the product $X \times X$ for the product law $P \times P$) and any θ with $0 < \theta < 2\pi$, $U\cos\theta + V\sin\theta$ and $-U\sin\theta + V\cos\theta$ are also independent with distribution P.

If $X = \mathbb{R}$, the transformation of $(U, V) \in \mathbb{R}^2$ in the last definition is a rotation through an angle θ. Normal laws with mean 0 on finite-dimensional real vector spaces are centered Gaussian in this sense, as can be seen from covariances. Conversely, a law on $X = \mathbb{R}$ satisfying the above definition of "centered Gaussian", even for one value of θ with $\sin(2\theta) \neq 0$, must be normal according to the "Darmois-Skitovič" theorem. We will not need the full strength of the latter theorem below, but we have:

Proposition 2.3. *A centered Gaussian law P on $\mathbb{R}$ is a law $N(0, \sigma^2)$ for some $\sigma^2 \geq 0$.*

Given a normal measure $P = N(m, C)$ on a finite-dimensional space X and a vector subspace Y of X, it follows from the structure of normal measures (RAP, Theorem 9.5.7) that $P(Y) = 0$ or 1. This fact extends to general measurable vector spaces:

Theorem 2.4. *(0-1 law) Let $(X, \mathcal{B})$ be a measurable vector space and Y a vector subspace with $Y \in \mathcal{B}$. Then for any centered Gaussian law P on X, $P(Y) = 0$ or 1.*

A measurable function $\|\cdot\|$ from a measurable vector space X into $[0, \infty]$ will be called a *pseudo-seminorm* iff $Y := \{x \in X\colon\ \|x\| < \infty\}$ is a vector subspace of X and $\|\cdot\|$ is a seminorm on Y, that is, $\|cx\| = |c|\|x\|$ for each real c and $x \in Y$, and so for all $x \in X$, with $0 \cdot \infty := 0$, and $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in Y$, and so for all $x, y \in X$.

By the 0–1 law (Theorem 2.4), for any pseudo-seminorm $\|\cdot\|$ and centered Gaussian P on X, $P(\|\cdot\| < \infty) = 0$ or 1. Likewise, $P(\|\cdot\| = 0) = 0$ or 1.

A real-valued *stochastic process* consists of a set T, a probability space $(\Omega, \mathcal{A}, P)$ and a map $(t, \omega) \mapsto X_t(\omega)$ from $T \times \Omega$ into $\mathbb{R}$, such that for each $t \in T$, $X_t(\cdot)$ is measurable from Ω into $\mathbb{R}$. A *sample function* of the process is a function $t \mapsto X_t(\omega)$ for any fixed ω. The process is called *Gaussian* iff for every finite subset F of T, the law $\mathcal{L}(\{X_t\}_{t\in F})$ is a normal distribution on $\mathbb{R}^F$.

If S is a countable set, then $\mathbb{R}^S$, the set of all real-valued functions on S, with product topology, is a separable metric topological linear space, hence a measurable vector space. If P is the law of a Gaussian stochastic process $\{x_t,\ t \in S\}$ on $\mathbb{R}^S$, with $Ex_t = 0$ for all $t \in S$, then P is centered Gaussian on $\mathbb{R}^S$. The supremum "norm" $\|\{y_t,\ t \in S\}\| := \sup_t |y_t|$ is clearly a pseudo-seminorm on $\mathbb{R}^S$.

Theorem 2.2 is a corollary of the following fact:

Theorem 2.5. *Let $(X, \mathcal{B})$ be a measurable vector space and P a centered Gaussian law on X. Let $\{y_n\}_{n\geq 1}$ be a sequence of measurable linear forms: $X \mapsto \mathbb{R}$. Let $\|x\| := \sup_n |y_n(x)|$. Suppose that $P(\|x\| < \infty) > 0$. Then $\tau := (\sup_n \int y_n^2 dP)^{1/2} < \infty$, and $E\exp(\alpha\|x\|^2) < \infty$ if and only if $\alpha < 1/(2\tau^2)$.*

2.3 Inequalities and comparisons for Gaussian distributions.

The main result of this section will say that if a set of Gaussian random variables is large enough in the sense of metric entropy (as defined in Appendix K), meaning that the number of variables more than ε apart grows rather fast as $\varepsilon \downarrow 0$, then it is almost surely unbounded. The proof is based on some inequalities, one due to Slepian and another to Sudakov and Chevet.

Theorem 2.6. *Slepian's inequality. Let $X_1, \cdots, X_n$ be real random variables with a normal joint distribution $N(0, r)$ on $\mathbb{R}^n$. Let $P_n(r) := \Pr\{X_j \geq 0$ for all $j = 1. \cdots, n\}$. Let q be another covariance matrix, with $r_{ii} = q_{ii} = 1$ for all $i = 1, \cdots, n$. If $r_{ij} \geq q_{ij}$ for all i and j, then $P_n(r) \geq P_n(q)$.*

Remarks. Since each X_j has distribution $N(0,1)$, clearly $P_n(r) \leq 1/2$, with $P_n(r) = 1/2$ when $r_{ij} = 1$ for all i and j. On the other hand, $P_n(r) = 0$ if $r_{ij} = -1$ for some $i \neq j$.

Recall that the *correlation (coefficient)* $r(X, Y)$ of two non-constant variables X and Y with finite second moments is defined by

$$r(X,Y) \;\; := \;\; E((X - EX)(Y - EY))/(\sigma_X \sigma_Y)$$

where $\sigma_X := \sigma(X)$ is the standard deviation $(E(X - EX)^2)^{1/2}$.

Corollary 2.7. *Let $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_n$ be two sets of jointly normally distributed variables with mean 0, $\sigma(X_i) > 0$ and $\sigma(Y_i) > 0$ for all i, and $r(X_i, X_j) \geq r(Y_i, Y_j)$ for all $i \neq j = 1, \cdots, n$. Then*

$$\Pr\{X_i \geq 0, \;\; i = 1, \cdots, n\} \;\; \geq \;\; \Pr\{Y_i \geq 0, \;\; i = 1, \cdots, n\}.$$

Proof. Replacing each X_i by $X_i/\sigma(X_i)$ and Y_i by $Y_i/\sigma(Y_i)$ does not change the events being considered or the correlations, and gives covariances to which Slepian's inequality applies. □

Let C be a set of Gaussian random variables with mean 0, and with the $\mathcal{L}^2$ metric $d(X, Y) := (E(X - Y)^2)^{1/2}$. Recall that for $\varepsilon > 0$, $D(\varepsilon, C) := D(\varepsilon, C, d) := \sup\{n :$ for some $X_1, \cdots, X_n \in C, \;\; d(X_i, X_j) > \varepsilon, \;\; 1 \leq i < j \leq n\}$ (Appendix K).

Theorem 2.8. (Sudakov-Chevet) *If* $\limsup_{\varepsilon \downarrow 0} \varepsilon^2 \log D(\varepsilon, C) = +\infty$, *then* $\sup\{|X| \colon X \in C\} = +\infty$ *almost surely.*

Example. Let G_n be i.i.d. $N(0,1)$ variables and $X_n := G_n/(\log n)^{1/2}$, $n \geq 2$. Then $d(X_j, X_k) > (\log j)^{-1/2}$ for $j < k$, so $D(\varepsilon, \{X_j\}_{2 \leq j \leq n}) \geq n - 1$ if $\varepsilon \leq (\log(n-1))^{-1/2}$. So for $0 < \varepsilon < 1$, $D(\varepsilon, \{X_j\}_{j \geq 2}) \geq [\exp(\varepsilon^{-2})] \geq \exp(\varepsilon^{-2})/2$ where $[x]$ denotes the greatest integer $\leq x$. So $\log D(\varepsilon, \{X_j\}_{j \geq 2}) \geq \varepsilon^{-2} - \log 2 \geq \varepsilon^{-2}/2$ for $0 < \varepsilon < 1/2$. On the other hand for each n, Proposition 2.1 gives $\Pr\{|X_n| \geq 2\} \leq \exp(-2 \log n) = 1/n^2$ so by the Borel-Cantelli Lemma, $\limsup_{n \to \infty} |X_n| \leq 2$ and $\sup_n |X_n| < +\infty$ a.s., so Theorem 2.8 is sharp.

Here is another comparison of Gaussian laws.

Theorem 2.9. *Let $N(0, C)$ and $N(0, D)$ be two normal measures with mean 0 on $\mathbb{R}^n$. Let $X = \{X_i\}_{i=1}^n$ have law $N(0, C)$ and let $Y = \{Y_i\}_{i=1}^n$ have law $N(0, D)$. Suppose that for all $i, j = 1, \cdots, n$, we have $E((Y_i - Y_j)^2) \leq E((X_i - X_j)^2)$, in other words for each i, j,*

$$D_{ii} + D_{jj} - 2D_{ij} \;\; \leq \;\; C_{ii} + C_{jj} - 2C_{ij}. \tag{2.1}$$

Then
(a) $E\{\max_{1 \leq i,j \leq n}(Y_i - Y_j)\} \;\; \leq \;\; E\{\max_{1 \leq i,j \leq n}(X_i - X_j)\}$ *and*
(b) $E \max_i Y_i \;\; \leq \;\; E \max_i X_i$.

Let $M_n := \max(Z_1, \cdots, Z_n)$ be the maximum of n i.i.d. standard normal variables Z_i. Then EM_n is bounded below as follows.

Lemma 2.10. *For all $n \geq 1$, $EM_n \geq (\log n)^{1/2}/12$.*

Remark. The constant $1/12$ can be improved to $(\pi \log 2)^{-1/2}$, by a less elementary proof (Fernique, 1997, (1.7.1) and references given there).

2.4 Gaussian measures and convexity.

There are several useful inequalities about normal measures and convex sets. Just one will be given as a sample in these notes. Convex sets were treated in RAP, sections 6.2 and 6.6.

A set C in a vector space is called *symmetric* if $-C := \{-x:\ x \in C\} = C$. A function f is called *even* if $f(-x) = f(x)$ for all x. Thus, the indicator function of a set is even if and only if the set is symmetric.

Theorem 2.11. *Let C be a convex, symmetric set in $\mathbb{R}^k$. Let f be a nonnegative, even function in $\mathcal{L}^1(\mathbb{R}^k, \mathcal{B}, V)$ where V is Lebesgue measure λ^k and $\mathcal{B}$ is the Borel σ-algebra. Suppose that for every $t > 0$, $K_t := \{x:\ f(x) > t\}$ is convex. Then for $0 \le \alpha \le 1$, any $y \in \mathbb{R}^k$, and $dx := dV(x)$,*

$$\int_C f(x + \alpha y)dx \quad \ge \quad \int_C f(x + y)dx.$$

2.5 The isonormal process: sample boundedness and continuity.

The isonormal Gaussian process L on a Hilbert space H is defined so that for any $x \in H$, $L(x)$ is a Gaussian random variable with distribution $N(0, \|x\|^2)$, and for any $x_1, \cdots, x_n \in H$, $(L(x_1), \cdots, L(x_n))$ have a jointly normal distribution with covariance given by the inner products (x_i, x_j). Then L exists by the nonnegative definiteness property of the inner product (e.g. RAP, Theorem 12.1.4). In these notes, usually H will be separable and infinite-dimensional. We have that $L(\cdot)$ is linear, in other words, for any $c \in \mathbb{R}$ and $x, y \in H$ we have $L(cx + y) = cL(x) + L(y)$ almost surely. (One can check that $L(cx + y) - cL(x) - L(y)$ has mean and variance 0.)

Historically, an isonormal process was first defined on the Hilbert space $L^2([0, \infty), \lambda)$, where λ is Lebesgue measure, as the "stochastic integral" $L(f) = \int f(t)dx_t$ where x_t is a Brownian motion process, a Gaussian process with mean 0 and covariance $Ex_s x_t = \min(s, t)$. Instead, the isonormal process is easy to define in general, and one can set $x_t = L(1_{[0,t]})$. Thus there is no need to define a stochastic integral for non-random integrands f with respect to x_t.

Two stochastic processes $\{X_t,\ t \in T\}$ and $\{Y_t,\ t \in T\}$ defined on the same index set, but possibly on different probability spaces, will be said to have *the same laws,* or one will be said to be a *version* of the other, if for each finite subset F of T, $\{X_t,\ t \in F\}$ and $\{Y_t,\ t \in F\}$ have the same law. If X_t and Y_t are defined on the same probability space then one is said to be a *modification* of the other if for each $t \in T$, we have $P(X_t = Y_t) = 1$. On the relationship of versions and modifications, especially for the isonormal process restricted to a set, see Appendix I.

Any Gaussian process with mean 0 can be factored through the isonormal process in the following sense. Let $X_t,\ t \in T$, be a Gaussian stochastic process where T is any set and $EX_t = 0$ for all $t \in T$. Then for some probability space $(\Omega, \mathcal{A}, P)$, $X_t(\cdot) \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ for each t. Let $t \mapsto h_t$ be a map from T into a Hilbert space H with inner product $(\cdot, \cdot)$ such that for all $s, t \in T$, $(h_s, h_t) = E(X_s X_t)$. Then $L(h_t)$ is a version of X_t. One can take $H = L^2(\Omega, \mathcal{A}, P)$ and $h_t \equiv X_t$.

It will be seen that sample boundedness and continuity of Gaussian processes can be largely treated by way of the isonormal process restricted to suitable subsets. For sample continuity this is done in section 2.8 below. For boundedness see Theorem 2.13.

For any subset $A \subset H$, $L(A)^*$ will be defined as $\operatorname{ess.sup}_{x\in A} L(x)$, the smallest random variable Y such that $Y \geq L(x)$ a.s. for all $x \in A$, where Y is determined up to a.s. equality. Here "ess.sup" stands for "essential supremum." Similarly let $|L(A)|^* := \operatorname{ess.sup}_{x\in A} |L(x)|$.

Lemma 2.12. *For any subset $A \subset H$, $L(A)^*$ and $|L(A)|^*$ are well-defined up to almost sure equality.*

Definitions. A set C in H is called a *GB-set* iff $|L(C)|^* < \infty$ a.s. Also, C will be called a *GC-set* iff it is totally bounded and the restriction of L to C can be chosen so that each of its sample functions $x \mapsto L(x)(\omega)$, $x \in C$, is uniformly continuous on C.

Since a uniformly continuous function on a totally bounded set must be bounded, every GC-set is a GB-set.

Theorem 2.13. *Let X_t, $t \in T$, be a Gaussian process with mean 0 on a probability space $(\Omega, \mathcal{A}, P)$. Then the process has a version with bounded sample functions if and only if $C := \{X_t(\cdot) : t \in T\}$ is a GB-set in $L^2(\Omega, \mathcal{A}, P)$.*

Proof. If: take a version of L with bounded sample functions on C, then $t \mapsto L(X_t)$ is a version of X_t and has bounded sample functions on T.

Only if: C must be totally bounded by the Sudakov-Chevet theorem 2.8 and so has a countable dense set D. There is some $S \subset T$ such that $t \mapsto X_t(\cdot)$ maps S 1-1 onto D. Since $L(X_t)$ is a version of X_t on S, it follows that D is a GB-set. Since $|L(C)|^* = |L(D)|^*$, C is also a GB-set. □

Definition. A function f from a set C in a vector space V into $\mathbb{R}$ will be called *prelinear* iff for any $c_1, \cdots, c_n \in C$ and $a_1, \cdots, a_n \in \mathbb{R}$ such that $a_1c_1 + \cdots + a_nc_n = 0$, we have $a_1f(c_1) + \cdots + a_nf(c_n) = 0$.

Lemma 2.14. *For any prelinear function f on a set C in a real vector space V into $\mathbb{R}$, let*

$$g(x_1c_1 + \cdots + x_nc_n) := x_1f(c_1) + \cdots + x_nf(c_n)$$

for any $x_1, \cdots, x_n \in \mathbb{R}$ and $c_1, \cdots, c_n \in C$. Then g is a well-defined linear function from the linear span of C into $\mathbb{R}$ which extends f. Such an extension exists if and only if f is prelinear.

Now, a *finite-dimensional projection* (fdp) will be an orthogonal projection (RAP, end of §5.3) of H onto a finite-dimensional subspace of H. For a sequence $\{\pi_n\}$ of such projections, $\pi_n \uparrow I$ will mean that the range of π_n is included in that of π_{n+1} for all n and that the union of all the ranges is dense in H. Since $\pi_n x$ is the nearest point to x in the range of π_n (RAP, Theorems 5.3.6 and 5.3.8), it follows that $\|\pi_n x - x\| \to 0$ as $n \to \infty$ for all $x \in H$. For any orthogonal projection π, let $\pi^\perp := I - \pi$, the orthogonal projection onto the orthogonal complement of the range of π (RAP, 5.3.8).

Lemma 2.15. *Whenever fdp's $\pi_n \uparrow I$, there is an orthonormal basis of H which includes an orthonormal basis of the range of π_n for each n.*

If C is a totally bounded set in a metric space, then the set V_1 of all uniformly continuous real-valued functions on C is a vector space. For any real function f on C let $\|f\|_C := \sup\{|f(x)|: \ x \in C\}$. Then V_1 with norm $\|\cdot\|_C$ is naturally isometric to the space $C(K)$ of all continuous functions on the completion K of C, where K is compact, so $C(K)$ is separable for the supremum norm (RAP, Corollary 11.2.5).

Let $C \subset H$ and let V_2 be the set of prelinear elements of V_1. Each element h of H defines a function on C by $x \mapsto (x,h), \ x \in C$. Let H_C be the completion of H for $\|\cdot\|_C$. Note that each element of H_C naturally defines a uniformly continuous, prelinear function on C as a uniform limit of uniformly continuous, prelinear functions. Let V_3 be the set of functions on C so defined. Then $V_3 \subset V_2$. (Often, $V_3 = V_2$, but whether $V_3 = V_2$ in all cases will not be settled here.)

Let V be a set of functions on C. Say that L on C can be *realized* on V if there is a probability measure μ on V such that the process $(v,x) \mapsto v(x), \ v \in V, \ x \in C$, has the joint distributions of L restricted to C: for any $x_1, \cdots, x_n$ in C, $v \mapsto v(x_i)$ are jointly Gaussian with mean 0 and covariances $(x_i, x_j), \ i,j = 1, \cdots, n$.

From the definition, μ would be defined on the smallest σ-algebra $\mathcal{B}_C$ making all evaluations $v \mapsto v(x)$ measurable for $x \in C$. If D is a countable dense set in C, V is a set of continuous functions on C and $v, w \in V$, then

$$\|v - w\|_C \ := \ \sup\{|(v-w)(y): \ y \in C\} \ = \ \sup\{|(v-w)(y): \ y \in D\},$$

so $v \mapsto \|v - w\|_C$ is $\mathcal{B}_C$ measurable for w fixed. If also V is a set of bounded functions on C, separable for $\|\cdot\|_C$, as V_1, V_2 and V_3 are, then all open sets for the $\|\cdot\|_C$ topology are in $\mathcal{B}_C$ (RAP, Proposition 2.1.4 and its proof), so $\mathcal{B}_C$ equals the Borel σ-algebra.

Given a set A in a vector space, the *symmetric convex hull* of A is the smallest convex set including A and $-A = \{-x: \ x \in A\}$, and is the set of all finite convex combinations $\sum_{i=1}^n \lambda_i a_i, \ a_i \in A \cup -A$, with $\lambda_i \geq 0$ and $\sum_{i=1}^n \lambda_i = 1$, for all positive integers n. The *closed symmetric convex hull* of A for some topology (in this case the Hilbert norm) is the closure of the symmetric convex hull. Here is a set of characterizations of GC-sets:

Theorem 2.16. *The following are equivalent for a totally bounded set C in H :*

(a) *C is a GC-set;*
(a′) *The closed, symmetric convex hull of C is a GC-set;*
(b) *For any $\varepsilon > 0$, $\Pr(|L(C)|^* < \varepsilon) \ > \ 0$;*
(c) *There exist fdp's $\pi_n \uparrow I$ such that $\liminf_{n\to\infty} |L(\pi_n^\perp C)|^* = 0$ a.s.;*
(d) *For some sequence $\pi_n \uparrow I$ of fdp's, $|L(\pi_n^\perp C)|^* \to 0$ in probability;*
(d′) *For some sequence $\pi_n \uparrow I$ of fdp's, $|L(\pi_n^\perp C)|^* \to 0$ almost surely;*
(e) *For every sequence $\pi_n \uparrow I$ of fdp's, $|L(\pi_n^\perp C)|^* \to 0$ in probability;*
(e′) *For every sequence $\pi_n \uparrow I$ of fdp's, $|L(\pi_n^\perp C)|^* \to 0$ almost surely;*
(f) *L can be realized on the completion V_3 of H for $\|\cdot\|_C$;*
(g) *L on C can be realized on the space V_2 of uniformly continuous, prelinear functions;*
(h) *L on C can be realized on the space V_1 of uniformly continuous functions.*

Corollary 2.17. *For any two GC-sets C, D, their union $C \cup D$ is also a GC-set.*

Proof. Condition (e) or (e′) in Theorem 2.16 holds on C and D and so on $C \cup D$. □

Recall that a Borel probability measure on a separable Banach space B is called *Gaussian* if every continuous linear form in B' has a Gaussian distribution. It follows that the norm

$\|\cdot\|$ on B satisfies some inequalities on the upper tail of its distribution for μ (Landau-Shepp-Marcus-Fernique bounds, Theorem 2.2). In particular, $\int \|x\|^2 d\mu(x) < \infty$.

Theorem 2.18. *Let $(B, \|\cdot\|)$ be a separable Banach space. Let μ be a Gaussian probability measure with mean 0 on the Borel sets of B. Then the unit ball $B_1' := \{f:\ \|f\|' \le 1\}$ in the dual Banach space B' is a compact GC-set in $L^2(B, \mu)$.*

Next, here is a lower bound based on packing numbers (Appendix K).

Theorem 2.19. *For any countable subset S of a Hilbert space H with its usual metric d, and any $\varepsilon > 0$,*

$$E \sup_{x \in S} L(x) \ \ge \ \tfrac{1}{17} \varepsilon (\log D(\varepsilon, S, d))^{1/2}.$$

Remark The constant $1/17$ can be improved to $(2\pi \log 2)^{-1/2}$ (Fernique, 1997, Theorem 4.1.4).

The rest of this section is not in UCLT:

Proposition 2.20. *If C is any GB-set then $|L(C)|^* = \sup_n |X_n|$ a.s. for some sequence $\{X_n\}$ of jointly Gaussian random variables with mean 0, and the result of Theorem 2.5 applies to $\|\cdot\| = |L(C)|^*$.*

Proof. C is totally bounded by the Sudakov-Chevet theorem 2.8. Thus C is separable. Let $\{x_n\}_{n \ge 1}$ be a countable dense set in C. Let $X_n := L(x_n)$ for each n. Lemma 2.12 and its proof show that $|L(C)|^* = \sup_n |X_n|$ a.s. The product $\mathbb{R}^\infty$ of a sequence of copies of $\mathbb{R}$ with product σ-algebra (the smallest for which the coordinates are measurable) is clearly a measurable vector space, so Theorem 2.5 applies to $\{X_n\} \in \mathbb{R}^\infty$ where y_n are the coordinate functions, $y_n(\{X_j\}_{j \ge 1}) := X_n$. □

Proposition 2.21. *If C is a GC-set and the isonormal process L is chosen to be uniformly continuous on C a.s., then*

$$\lim_{\delta \downarrow 0} E\{\sup |L(x) - L(y)| :\ \|x - y\| \le \delta,\ x, y \in C\} \ = \ 0.$$

Proof. The quantity in braces decreases to 0 a.s. as $\delta \downarrow 0$. Its expectation is finite for any fixed $\delta > 0$ by Proposition 2.20. Thus the result follows from dominated convergence or monotone convergence. □

The following theorem gives a necessary condition for the GC property, corresponding to Theorem 2.8 for the GB property. Fernique (1997) attributes both to Sudakov.

Theorem 2.22. *If C is a GC-set then $\varepsilon^2 \log D(\varepsilon, C) \to 0$ as $\varepsilon \downarrow 0$.*

Proof. Given $h > 0$, take $\delta > 0$ by Proposition 2.21 such that for $\overline{B}(x, \delta) := \{y :\ \|x - y\| \le \delta\}$, we have for all $x \in C$,

$$E \sup\{L(y) :\ y \in \overline{B}(x, \delta) \cap C\} \ \le \ h/34.$$

Let $N := D(\delta, C)$. Then C is covered by some closed balls $\overline{B}(x_i, \delta)$, $i = 1, \cdots, N$. We have $D(u, C) \le N \max_i D(u, \overline{B}(x_i, \delta) \cap C)$. Using $\sqrt{a + b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$ and Theorem 2.19 applied to each $S = \overline{B}(x_i, \delta) \cap C$ we get $u\sqrt{\log D(u, C)} \ \le \ u\sqrt{\log N} + h/2$ for $0 < u < \infty$. For $u < h/(2\sqrt{\log N})$ we get $u\sqrt{\log D(u, C)} \le h$. □ .

2.6 A metric entropy sufficient condition for sample continuity.

Recall that a stochastic process $X_t(\omega)$, $t \in T$, is said to be *sample-bounded* on T if $\sup_{t\in T} X_t$ is finite for almost all ω. If T is a topological space, then the process is said to be *sample-continuous* if for almost all ω, $t \mapsto X_t(\omega)$ is continuous. The isonormal process is not sample-continuous on the Hilbert space H : let $\{e_n\}$ be an orthonormal sequence. Then $L(e_n)$ are i.i.d. $N(0,1)$ variables. Thus if $a_n \to 0$ slowly enough, specifically if $a_n(\log n)^{1/2} \to \infty$ as $n \to \infty$, $L(a_n e_n)$ are almost surely unbounded (by Theorem 2.8). So not all bounded sets or even compact sets are GB-sets or GC-sets. This section will state a sufficient condition based on metric entropy (defined in Appendix K), while §2.7 will give characterizations based on what are called majorizing measures.

A metric entropy sufficient condition for sample continuity of L will actually give a quantitative bound for the continuity. Let (T,d) be a metric space. A function f will be called a *sample modulus* for a real stochastic process $\{X_t,\ t \in T\}$ iff there is a process Y_t with the same laws as X_t and such that for almost all ω there is an $M(\omega) < \infty$ such that for all $s,\ t \in T,\ |Y_s - Y_t|(\omega) \le M(\omega) f(d(s,t))$.

Whenever f is a sample modulus for L on $C \subset H$, and $\{X_t,\ t \in T\}$ is a Gaussian process with mean 0 and $\{X_t(\cdot):\ t \in T\} = C$, then f is also a sample modulus for the process X_t, with the intrinsic pseudo-metric $d(s,t) := (E(X_s - X_t)^2)^{1/2}$ on T.

See in Appendix K the definitions of $N(\varepsilon, C)$ and $H(\varepsilon, C)$, for the usual metric $d(x,y) := \|x - y\|$ on H. Now the main theorem of this section can be stated:

Theorem 2.23. *For any $C \subset H$, if $\int_0^\infty (\log N(t,C))^{1/2} dt < \infty$ then C is a GC-set, and if*

$$f(x) \ := \ \int_0^x (\log N(t,C))^{1/2} dt, \qquad x > 0,$$

then f is a sample modulus for L on C.

Note. If C is bounded, then $N(t,C) = 1$ and $\log N(t,C) = 0$ for t large enough, and $N(\cdot, C)$ is a nonincreasing function, so integrability of $(\log N(t,C))^{1/2}$ is only an issue near $t = 0$. If $f(x) = +\infty$ for some $x > 0$, then $f(x) = +\infty$ for all $x > 0$, so it still provides a sample modulus but only a trivial one. By Theorem K.1 of Appendix K, $N(t,C)$ could be replaced equivalently by $D(t,C)$.

It follows from Theorem 2.23 that C is a GC-set if as $\varepsilon \downarrow 0$, $N(\varepsilon, C) = O(\exp(\varepsilon^{-p}))$ for some $p < 2$, or if $N(\varepsilon, C) = O(\exp(\varepsilon^{-2} |\log \varepsilon|^{-r}))$ for some $r > 2$. On the other hand Theorem 2.8 implies that C is not a GB-set if as $\varepsilon \downarrow 0$, eventually $N(\varepsilon, C) \ge \exp(\varepsilon^{-p})$ for some $p > 2$ or $N(\varepsilon, C) \ge \exp(\varepsilon^{-2} |\log \varepsilon|^{s})$ for some $s > 0$. It turns out that the gap cannot be closed further: if $N(\varepsilon, C)$ is of the order of $\exp(\varepsilon^{-2} |\log \varepsilon|^{-r})$ for $0 \le r \le 2$, there are examples showing that C may or may not be a GB-set. So a characterization of the GB-property can't be given in terms of metric entropy, although it comes rather close. For a characterization in other terms, see the next section.

Remark. If C is a GC-set, then $L(\cdot)(\cdot)$ can be chosen such that for all ω, $x \mapsto L(x)(\omega)$ is continuous for $x \in C$. Then for any countable dense subset A of C, $L(C)^* = \sup_{x\in A} L(x)$ a.s.

Next, the same integral as in Theorem 2.23 yields a bound for expectations of certain suprema.

Theorem 2.24. *Let $C \subset H$ be non-empty and let $D := \operatorname{diam} C = \sup_{x,y\in C} \|x-y\|$. Let $B := \{x-y : x, y \in C\}$. Then for f as in Theorem 2.23,*
(a) $E|L(B)|^* \leq 81 f(D/4)$ and
(b) $EL(C)^* \leq 81 f(D/4)$.

Remarks. All three quantities in (a), (b) are invariant under translation, replacing C by $\{c+u : c \in C\}$ for any fixed u. But $E|L(C)|^*$ does not have such invariance, and becomes unbounded as $\|u\| \to \infty$, so for it we cannot have an upper bound $Kf(D)$, $K < \infty$.

If the constant $81/4$ is replaced by a larger one, one can have, instead of the quantities on the left in (a) and (b), Young-Orlicz norms (Appendix H) $\|\cdot\|_g$ where $g(x) := \exp(x^2) - 1$, see Theorem 2.25 next.

Let g be a convex, increasing function from $[0,\infty)$ onto itself. If Y is a random variable such that $Eg(\delta Y) < \infty$ for some $\delta > 0$, let $\|Y\|_g := \inf\{c > 0 : Eg(|Y|/c) \leq 1\}$. Then $\|\cdot\|_g$ is a seminorm on such random variables (Appendix H). If there is no such $\delta > 0$, let $\|Y\|_g := +\infty$.

Theorem 2.25. *There is an absolute constant $M < \infty$ such that for any subset C of a Hilbert space H, and $g(x) := \exp(x^2) - 1$,*

$$\|L(C)^*\|_g \leq \| |L(C)|^* \|_g \leq ME(|L(C)|^*) \leq 2ME(L(C)^*).$$

2.7 Majorizing measures.

This section will state characterizations of GB-sets and GC-sets in terms of majorizing measures, to be defined next. The characterizations are due to X. Fernique and M. Talagrand. For a metric space (T,d), $r > 0$, and $x \in T$, the open ball of center x and radius r is $B(x,r) := \{y : d(x,y) < r\}$.

Definition. Let (T,d) be a metric space and $\mathcal{P}(T)$ the set of all laws (Borel probability measures) on T. For $m \in \mathcal{P}(T)$ let

$$\gamma_m(T) := \sup_{x\in T} \int_0^\infty \left(\log\left(\frac{1}{m(B(x,r))}\right)\right)^{1/2} dr.$$

If $\gamma_m(T) < \infty$, then m is called a *majorizing measure* for (T,d). Let $\gamma(T) := \gamma(T,d) := \inf\{\gamma_m(T) : m \in \mathcal{P}(T)\}$.

Then $\gamma(T,d) < \infty$ if and only if there exists a majorizing measure on T. If m is a majorizing measure on T, then for all $x \in T$, $m(B(x,r)) > 0$ for all $r > 0$ and does not approach 0 too fast as $r \downarrow 0$. For example, if T is finite, then $\gamma_m(T) < \infty$ if and only if $m(\{x\}) > 0$ for all $x \in T$. On $[0,1]$ with usual metric, Lebesgue measure λ is a majorizing measure, as is any law having an absolutely continuous component with a density $h \geq c$ for some $c > 0$.

Two theorems, which together characterize GB-sets as sets in Hilbert space having majorizing measures, will be stated. Here $T = C$ will be a subset of a Hilbert space with the usual Hilbert metric. Recall $L(C)^* := \operatorname{ess.sup}_{x\in C} L(x)$ and $|L(C)|^*$ as defined by Lemma 2.12.

Theorem 2.26. (Fernique, 1975) *If C is a subset of a Hilbert space H and $\gamma(C) < \infty$ then C is a GB-set. For some absolute constant K, $EL(C)^* \leq K\gamma(C)$.*

Notes. Translation of C, replacing it by $\{c + h : \ c \in C\}$ for some fixed h, preserves all of $N(\varepsilon, C)$, $\gamma(C)$, and $EL(C)^*$ but not $E|L(C)|^*$. The word "majorizing" apparently refers to the inequality in Theorem 2.26.

Theorem 2.27. (Talagrand, 1987) *If C is a GB-set then $\gamma(C) < \infty$. For some absolute constant K' and all $C \subset H$, $\gamma(C) \leq K'EL(C)^*$.*

The original proof has been considerably shortened: Talagrand (1992), Fernique (1997).

Theorem 2.28. *Let (T, d) be a totally bounded metric space. Suppose that m is a law (Borel probability measure) on T such that for some $M < \infty$ and all $r > 0$, $\sup_{x \in T} m(B(x, r)) < M \inf_{y \in T} m(B(y, r))$. Then the following are equivalent:*

(i) *m is a majorizing measure for T;*

(ii) *$\gamma(T) < \infty$;*

(iii) *$\int_0^1 (\log D(\varepsilon, T))^{1/2} d\varepsilon < \infty$.*

Corollary 2.29. *Let (T, d) be a metric space such that there is a group G of 1–1 transformations g of T onto itself for which*

(a) *d is G-invariant: for all $s, \ t \in T$ and $g \in G$,*

$$d(g(s), g(t)) \quad = \quad d(s, t);$$

(b) *There is a law (Borel probability measure) m on T which is G-invariant, i.e. $m \circ g^{-1} = m$ for all $g \in G$;*

(c) *G acts transitively on T : for all $s, \ t \in T$ there is a $g \in G$ with $g(s) = t$.*

Then the hypotheses and thus the conclusion of Theorem 2.28 hold.

Proof. The hypotheses imply that for each $r > 0$, $m(B(x, r))$ is the same for all $x \in T$. Thus Theorem 2.28 applies for any $M \geq 1$. □

Notes. Fernique (1975) proved that for Gaussian processes satisfying a homogeneity condition like that in Corollary 2.29, the metric entropy integral condition (Theorem 2.28(iii)) is necessary and sufficient for sample continuity. Theorems 2.26 and 2.27 above, with Theorem 2.28, show that the metric entropy integral condition is equivalent to sample continuity for the isonormal process on T for a subset T of a Hilbert space satisfying the conditions of Theorem 2.28.

For an example of the situation in Corollary 2.29 let T be the unit circle $x^2 + y^2 = 1$ in $\mathbb{R}^2$, let G be the group of rotations, let d be the usual metric on $\mathbb{R}^2$, and let $dm(\theta) = d\theta/(2\pi)$. Likewise T could be a sphere of any dimension, with the orthogonal group G.

To see how Theorem 2.28 applies beyond Corollary 2.29, suppose one wants to prove sample-continuity of a Gaussian process on a locally compact but not compact metric space, such as a Euclidean space or a non-compact manifold. Then it suffices to prove sample continuity on each of a family of compact sets whose interiors form a base for the topology, such as balls in Euclidean spaces. Then one can often define a measure, such as Lebesgue measure in a Euclidean space, restrict it to a compact set C and normalize it to have mass 1 to get a law m. Then $m(B(x, r))$ may not depend on x while $B(x, r)$ is included in the interior of C, but

become smaller as x approaches the boundary of C, yet the hypothesis of Theorem 2.28 still holds.

There is also a criterion for the GC property where the majorizing measure condition is strengthened, as follows. Fernique proved "if" and Talagrand (1987) "only if."

Theorem 2.30. (Fernique-Talagrand) *Let H be a Hilbert space and $C \subset H$. Then C is a GC-set if and only if there exists a probability measure μ on C such that*

$$\limsup_{\varepsilon \downarrow 0} \sup_{x \in C} \int_0^\varepsilon \left[\log \left(\frac{1}{\mu(B(x,r))} \right) \right]^{1/2} dr \;=\; 0.$$

Proof. A proof for "only if" will be given here since it is not given in UCLT. Let C be a GC-set and let L be chosen to be uniformly continuous on C a.s. Let R be the diameter of C. Recall that $\overline{B}(x,r) := \{y : \|x-y\| \le r\}$. In integrals for majorizing measures, $B(x,r)$ can be replaced equivalently by $\overline{B}(x,r)$ since $\mu(B(x,r))$ differs from $\mu(\overline{B}(x,r))$ only for r in a countable set: $r \mapsto \mu(B(x,r))$ and $r \mapsto \mu(\overline{B}(x,r))$ are the left-continuous and right-continuous versions of the same non-decreasing function of r.

For $n = 1, 2, \cdots$, let $M_n := M(n) := D(R/2^n, C)$. Let $S_n := \{x_{ni}\}_{i=1}^{M(n)}$ be a maximal set with $\|x_{ni} - x_{nj}\| > R/2^n$ for $i \ne j$. Let $T_{ni} := \overline{B}(x_{ni}, R/2^n) \cap C$ for each n and i. Then $C = \bigcup_{i=1}^{M(n)} T_{ni}$ for each n. By Theorem 2.27, for any constant $A > K'$, choose for each n and $i = 1, \cdots, M_n$ a majorizing measure μ_{ni} on T_{ni} such that

$$\int_0^{R/2^{n-1}} \left[\log \frac{1}{\mu_{ni}(T_{ni} \cap \overline{B}(t,u))} \right]^{1/2} du \;\le\; AE \sup\{L(x) : \; x \in T_{ni}\}$$

for all $t \in T_{ni}$. Let

$$\mu \;:=\; \sum_{n=0}^{\infty} 2^{-n-1} \sum_{i=1}^{M(n)} \mu_{ni}/M_n,$$

a probability measure on C. For each $t \in C$ and each n, we have $t \in T_{ni}$ for some $i \le M_n$. Thus for $0 < u < R/2^n$, $\mu(\overline{B}(t,u)) \;\ge\; 2^{-n-1}\mu_{ni}(\overline{B}(t,u))/M_n$. We have $\sqrt{a+b+c} \le \sqrt{a} + \sqrt{b} + \sqrt{c}$ for $a, b, c \ge 0$. So

$$\int_0^{R/2^n} \left[\log(1/\mu(\overline{B}(t,u))) \right]^{1/2} du$$

$$\le\; [\sqrt{(n+1)\log 2} + \sqrt{\log M_n}] R/2^n + AE \sup\{L(x) : \; x \in T_{ni}\}. \tag{2.2}$$

For $n = 0$, $M_0 = 1$ so $\log M_0 = 0$. I claim that $R \le \sqrt{2\pi} E \sup_{t \in C} L(t)$. To prove this let $x, y \in H$, $\|x - y\| = r$. Then

$$E \max\{L(x), L(y)\} \;=\; E[L(x) + \max(0, L(y) - L(x))] \;=\; E \max(0, rZ)$$

where Z has a $N(0,1)$ distribution. Then $E \max(0, Z) = 1/\sqrt{2\pi}$. Letting $r \uparrow R$ the claim follows.

We then have for all $t \in C$

$$\int_0^R \left[\log(1/\mu(\overline{B}(t,u))) \right]^{1/2} du \;\le\; (A + \sqrt{2\pi}) E \sup_{x \in C} L(x).$$

Since C is a GC-set, it is a GB-set and $EL(C)^* < \infty$ by Proposition 2.20. Thus μ is a majorizing measure on C. For other values of n, (2.2) gives that

$$\sup_{t\in C} \int_0^{R/2^n} \left[\log(1/\mu(\overline{B}(t,u)))\right]^{1/2} du$$

is bounded by a sum of three terms. As $n \to \infty$, the first goes to 0 clearly, the second by Theorem 2.22, and the third by Proposition 2.21. Thus "only if" in Theorem 2.30 is proved. □

2.8 Sample continuity and compactness.

In this section it will be seen that for a Gaussian process X_t indexed by a compact metric space, or other suitable parameter space such as an open or closed set in a Euclidean space, sample continuity reduces to that of the isonormal process on some subsets, and continuity of the non-random function $t \mapsto EX_t$.

Let $(T, \mathcal{T})$ and $(W,\mathcal{U})$ be two topological spaces. Let $\{X_t,\ t \in T\}$ be a stochastic process defined over a probability space $(\Omega, \mathcal{B}, P)$ with values in W, meaning that for each $t \in T$ and Borel set $B \subset W$, $X_t^{-1}(B) \in \mathcal{B}$. (Recall that the σ-algebra of Borel sets is generated by the open sets and that it's equivalent to assume $X_t^{-1}(U) \in \mathcal{B}$ for each $U \in \mathcal{U}$.) A process $\{X_t\}_{t\in T}$ will be called *version-continuous* iff there is a process Y with the same laws, possibly defined over a different probability space $(\Omega', \mathcal{B}', P')$, such that for all $\omega' \in \Omega'$, $t \mapsto Y_t(\omega')$ is continuous from T into W. (Equivalently, continuity need only hold for almost all ω'.)

Theorem 2.31. *A Gaussian process $\{X_t\}$ indexed by a metric space T, defined on a probability space (Ω, P), is version-continuous if and only if both*
(a) *the non-random function $t \mapsto EX_t$ is continuous, and*
(b) *the process $\{X_t - EX_t\}$ is version-continuous.*
Then, $t \mapsto X_t(\cdot)$ is continuous into $L^2(P)$.

So in studying sample continuity or version-continuity of Gaussian processes we may as well restrict ourselves to processes with mean 0. Let X_t be such a process, $t \in T$. Each $X_t(\cdot)$ is an element of a Hilbert space H, namely $L^2(P)$. Consider the isonormal process L on this H. Then since L is Gaussian, has mean 0 and preserves covariances, we see that $L(X_t)$ has the same laws as X_t.

If $h(\cdot)$ is a continuous function from T into a Hilbert space H, with range $C := \{h(t):\ t \in T\}$, and if L restricted to C is version-continuous, then the process $L \circ h$ is clearly version-continuous. Conversely, if (T, e) is compact and h is 1–1, then h is a a homeomorphism (RAP, Theorem 2.2.11). Then, version continuity of L on C and $L \circ h$ on T are equivalent. So, for (T, e) compact and $t \mapsto X_t(\cdot)$ one-to-one, version continuity of the Gaussian process X_t reduces to that of L on a subset C of H. (Theorem 2.32 and Corollary 2.33 below will show that the 1–1 assumption is not actually necessary.) If T is locally compact, for example an open or closed subset of some $\mathbb{R}^k$, then continuity is equivalent to continuity on each compact subset.

The next fact holds for general, not necessarily Gaussian processes.

Theorem 2.32. *If (T, e) is a compact metric space, h is a continuous function from T onto a metric space K and $Y(x, \omega)$, $x \in K$, $\omega \in \Omega$, is a stochastic process on K with values in a*

complete separable metric space S, then $Y \circ h$ is version-continuous on T if and only if Y is on K.

Remark. If $Y(x, \omega) \equiv Y(x)$, a non-random function, then the result is a known fact in general topology (RAP, Theorem 2.2.11). The difficulty in the proof is that if $Y \circ h$ is version-continuous, it is not clear that the corresponding sample-continuous process X can be written as $Y' \circ h$ for a process Y' on K.

Corollary 2.33. *A Gaussian process $\{X_t,\ t \in T\}$ with mean 0 on a compact metric space (T, e) is version-continuous if and only if both $t \mapsto X_t(\cdot) \in H := L^2(P)$ is continuous and its range K is a GC-set.*

Example. If X_t is a Gaussian process defined for $t \in \mathbb{R}$, suppose X_t is periodic of period 2π, $X_t \equiv X_{t+2\pi}$ for all t. Suppose that $E((X_t - X_s)^2) > 0$ for $|s - t| < 2\pi$. Then we can write the process as $X_t = Y(e^{it})$ where Y is a process indexed by the unit circle $T^1 := \{z\colon\ |z| = 1\}$ in the complex plane, which is compact. Version continuity for X and Y are equivalent, and $z \mapsto Y(z)(\cdot)$ is 1–1 from T^1 into $H := L^2(P)$, so version continuity is equivalent to that of L on the range of Y in H (without needing Theorem 2.32 and Corollary 2.33). On the other hand any process indexed by $\mathbb{R}$ is version continuous if and only if it is so on each compact interval $[-N, N]$, where in this example for $N \geq \pi$, the process is not 1–1 into H.

Recall that a sample function of a stochastic process X_t is a function $t \mapsto X_t(\omega)$ for a fixed ω. The usual metric on Hilbert space is the natural one for an isonormal process, but the GC-property holds for other metrics in the following sense:

Theorem 2.34. *Let C be a subset of Hilbert space H. Then the following are equivalent:*

(I) *C is a GC-set;*

(II) *L on C has a version with bounded, uniformly continuous sample functions;*

(III) *there exists a metric ρ on C such that (C, ρ) is totally bounded, and the sample functions of the isonormal process L on C can be chosen to be ρ-uniformly continuous a.s.*

Proof. (I) implies (II) since by definition a GC-set is totally bounded, so a uniformly continuous function on it must be bounded. (II) implies (III) directly where ρ is the usual metric.

Suppose (III) holds. Take a version of L such that on a set of probability one, the sample functions of L are ρ-uniformly continuous on C. Then L extends to a Gaussian process $t \mapsto X_t$ on the compact completion M of C for ρ. Here X_t is version-continuous and so by Corollary 2.33, C is included in, and thus is, a GC-set. □

REFERENCES

Andersen, Niels Trolle, and Dobrić, Vladimir (1988). The central limit theorem for stochastic processes II. *J. Theoret. Probab.* **1**, 287–303.

Anderson, Theodore W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proc. Amer. Math. Soc.* **6**, 170–176.

Bonnesen, T., and Fenchel, W. (1934). *Theorie der konvexen Körper.* Berlin, Springer; repr. Chelsea, New York, 1948.

Borell, Christer (1974). Convex measures on locally convex spaces. *Ark. Mat.* **12**, 239–252.

Borell, C. (1975a). Convex set functions in d-space. *Period. Math. Hungar.* **6**, 111–136.

Borell, C. (1975b). The Brunn-Minkowski inequality in Gauss space. *Invent. Math.* **30**, 207–216.

Chevet, Simone (1970). Mesures de Radon sur $\mathbb{R}^n$ et mesures cylindriques. *Ann. Fac. Sci. Univ. Clermont* #43 (math., 6° fasc.) 91–158.

Cohn, Donald L. (1980). *Measure Theory.* Birkhäuser, Boston.

Darmois, Georges (1951). Sur une propriété caractéristique de la loi de probabilité de Laplace. *Comptes Rendus Acad. Sci. Paris* **232**, 1999–2000.

Dudley, R. M. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Funct. Anal.* **1**, 290–330.

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

Dunford, N., and Schwartz, J. T. (1958). *Linear Operators, Part I: General Theory,* Interscience, New York; repr. Wiley, New York (1988).

Feldman, Jacob (1972). Sets of boundedness and continuity for the canonical normal process. *Proc. Sixth Berkeley Symposium Math. Statist. Prob.* **2**, pp. 357–367. Univ. of Calif. Press, Berkeley and Los Angeles.

Fernique, Xavier (1964). Continuité des processus gaussiens. *Comptes Rendus Acad. Sci. Paris* **258**, 6058–6060.

Fernique, X. (1970). Intégrabilité des vecteurs gaussiens. *Comptes Rendus Acad. Sci. Paris Sér.* A **270**, 1698–1699.

Fernique, X. (1971). Régularité de processus gaussiens. *Invent. Math.* **12**, 304–320.

Fernique, X. (1975). Régularité des trajectoires des fonctions aléatoires gaussiennes. Ecole d'été de probabilités de St.-Flour, 1974. *Lecture Notes in Math.* (Springer) **480**, 1–96.

Fernique, X. (1985). Sur la convergence étroite des mesures gaussiennes. *Z. Wahrschleinlichkeitsth. verw. Geb.* **68**, 331–336.

Fernique, X. (1997). Fonctions aléatoires gaussiennes, vecteurs aléatoires gaussiens. Publications du Centre de Recherches Mathématiques, Montréal.

Giné, Evarist, and Zinn, Joel (1986). Lectures on the central limit theorem for empirical processes. In *Probability and Banach Spaces*, Proc. Conf. Zaragoza, 1985, *Lecture Notes in Math.* (Springer) **1221**, 50–113.

Gordon, Yehoram (1985). Some inequalities for Gaussian processes and applications. *Israel J. Math.* **50**, 265–289.

Gross, Leonard (1962). Measurable functions on Hilbert space. *Trans. Amer. Math. Soc.* **105**, 372–390.

Hadwiger, H. (1957). *Vorlesungen über Inhalt, Oberfläche und Isoperimetrie.* Springer, Berlin.

Itô, Kiyosi, and McKean, H. P. Jr. (1974). *Diffusion processes and their sample paths.* Springer, Berlin, New York.

Kahane, Jean-Pierre (1986). Une inégalité du type de Slepian et Gordon sur les processus gaussiens. *Israel J. Math.* **55**, 109–110.

*Komatsu, Y. (1955). Elementary inequalities for Mills' ratio. *Rep. Statist. Appl. Res. Un. Japan. Sci. Engrs.* **4**, 69-70

Landau, H. J., and Shepp, Lawrence A. (1971). On the supremum of a Gaussian process. *Sankhyā* Ser. A **32**, 369–378.

Ledoux, M., and Talagrand, M. (1991). *Probability in Banach Spaces.* Springer, Berlin.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses.* 2d ed.; Wiley, New York; repr. 1997.

Leichtweiss, K. (1980). *Konvexe Mengen.* Springer, Berlin.

Marcus, M. B. (1974). The ε-entropy of some compact subsets of ℓ^p. *J. Approximation Th.* **10**, 304-312.

Marcus, Michael B., and Shepp, L. A. (1972). Sample behavior of Gaussian processes. *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* (1970) **2**, 423–441. Univ. of Calif. Press, Berkeley and Los Angeles.

McMullen, P. (1993). Valuations and dissections. In *Handbook of Convex Geometry*, eds. P. M. Gruber, J. M. Wills, North-Holland, Amsterdam, vol. B.

Milman, V. D., and Pisier, G. (1987). Gaussian processes and mixed volumes. *Ann. Probab.* **15**, 292-304.

Mityagin, B. S. (1961). Approximate dimension and bases in nuclear spaces. Russian Math. Surveys **16**, no. 4, 59-127.

Pisier, G. (1989). *The volume of convex bodies and Banach space geometry.* Cambridge University Press.

Pollard, D. (1990) *Empirical Processes: Theory and Applications.* NSF-CBMS Regional Conference Series in Probab. and Statist. **2**. Inst. Math. Statist. and Amer. Statist. Assoc.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications.* 2d ed. Wiley, New York.

Skitovič, V. P. (1954). Linear combinations of independent random variables and the normal distribution law. *Izv. Akad. Nauk SSSR Ser. Mat.* **18**, 185–200 (in Russian).

Slepian, David (1962). The one-sided barrier problem for Gaussian noise. *Bell Syst. Tech. J.* **41**, 463–501.

Sudakov, V. N. (1969). Gaussian measures, Cauchy measures and ε-entropy. *Soviet Math. Dokl.* **10**, 310–313.

Sudakov, V. N. (1971). Gaussian random processes and measures of solid angles in Hilbert space. *Soviet Math. Dokl.* **12**, 412–415.

Sudakov, V. N. (1973). A remark on the criterion of continuity of Gaussian sample function. In *Proc. Second Japan-USSR Symp. Probab. Theory, Kyoto, 1972, ed. G. Maruyama, Yu. V. Prokhorov; Lecture Notes in Math.* (Springer) **330**, 444-454.

Talagrand, M. (1987). Regularity of Gaussian processes. *Acta Math.* (Sweden) **159**, 99–149.

Talagrand, M. (1992). A simple proof of the majorizing measure theorem. *Geom. Funct. Anal.* **2**, 118–125.

*I found this reference from a secondary source and have not seen it in the original.

Chapter 3

Foundations of uniform central limit theorems; Donsker classes

3.1 Definitions: convergence in law

Let $(S, \mathcal{B}, P)$ be a probability space, to be called the *sample space*. Examples to have in mind for S are Euclidean spaces such as the plane $\mathbb{R}^2$. To form empirical measures, we'd like to take variables $X_1, X_2, \cdots$, i.i.d. with law P. To do this, take a countable product S^∞ of copies of $(S, \mathcal{B}, P)$ (RAP, Theorem 8.2.2) and let X_i be the coordinates on the product. A product may be taken with another probability space. *Throughout the rest of these notes, X_i will be defined as such coordinates* unless something is said to the contrary. This way of defining X_i will be called the *standard model*. An example showing that use of the standard model makes a difference will be given at the end of Section 5.3.

Then, we can form the *empirical measures* $P_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, and the *empirical process* $\nu_n := n^{1/2}(P_n - P)$. So P_n is a probability measure on S, defined on $\mathcal{B}$ and actually on all subsets of S, for any values of $X_1, \cdots, X_n$. Each ν_n is a finite signed measure of total charge 0.

As $n \to \infty$, for any set $A \in \mathcal{B}$, $\nu_n(A)$ converges in law to a normal law with mean 0 and variance $P(A)(1-P(A))$. Next, a general analogue of the Brownian bridge will be defined. Let G_P be a Gaussian process indexed by the set $\mathcal{L}^2(P)$ of measurable, square-integrable functions for P, having mean 0 and covariance $EG_P(f)G_P(g) = \int fgdP - \int fdP \int gdP$. To see that such a Gaussian process exists, we need to check that the given covariance is nonnegative definite, which holds since it is just the covariance for P, in other words

$$EG_P(f)G_P(g) \;=\; (f,g)_{0,P} \;:=\; \mathrm{Cov}_P(f,g) \;:=\; \int fgdP - \int fdP \int gdP,$$

and such a covariance is always nonnegative definite.

Let $\mathcal{L}_0^2(P)$ be the set of all functions $f \in \mathcal{L}^2(P)$ such that $\int fdP = 0$. Let $L_0^2(P)$ be the set of all equivalence classes of elements of $\mathcal{L}_0^2(P)$ for equality a.s. (P). On $\mathcal{L}_0^2(P)$, the covariance for G_P reduces to the ordinary $\mathcal{L}^2$ semi-inner product. In other words, restricted to the subspace $\mathcal{L}_0^2(P)$, G_P is an isonormal process. Let C be the one-dimensional space of constant functions c as a subspace of $\mathcal{L}^2(P)$. Then G_P is 0 on C, while the spaces C and $\mathcal{L}_0^2(P)$ are orthogonal complements of each other (RAP, Theorem 5.3.8) in $\mathcal{L}^2(P)$. For any f and g in $\mathcal{L}^2(P)$ and $c \in \mathbb{R}$, $G_P(cf+g) = cG_P(f) + G_P(g)$ a.s., since as is easily checked, $G_P(cf+g) - cG_P(f) - G_P(g)$ has mean and variance both 0.

Letting $\rho_P(f,g) := (E(G_P(f) - G_P(g))^2)^{1/2}$ defines a pseudo-metric on $\mathcal{L}^2(P)$. Then ρ_P equals the usual $\mathcal{L}^2(P)$ pseudometric on $\mathcal{L}_0^2(P)$. Moreover, with respect to the semi-inner product $(\cdot,\cdot)_{0,P}$ on $\mathcal{L}^2(P)$, G_P is the isonormal process and ρ_P the usual pseudo-metric. Thus, *the results of Chapter 2 apply to sample continuity of G_P with respect to ρ_P.*

The Brownian bridge process y_t is a special case of the G_P process where P is Lebesgue measure on $[0,1]$ and $y_t = G_P(1_{[0,t]})$; it is easily checked that this has the right covariance.

The Brownian bridge process can be taken to have continuous sample paths, in other words to be continuous as a function of t for each ω (RAP, Theorem 12.1.5). But G_P is not sample-continuous on the whole space $\mathcal{L}^2(P)$. In fact, unless P is concentrated in finitely many atoms, the spaces $\mathcal{L}^2(P)$ and $\mathcal{L}_0^2(P)$ are infinite-dimensional, in the sense that they contain infinite orthonormal sets, and we saw in Section 2.6 that an isonormal process on an infinite-dimensional Hilbert space H is not sample-continuous. We will be concerned then with suitable subsets of $\mathcal{L}^2(P)$. A class $\mathcal{F} \subset \mathcal{L}^2$ will be called *pregaussian* if a G_P process $(f,\omega) \mapsto G_P(f)(\omega)$ can be defined on some probability space such that for each ω, $f \mapsto G_P(f)(\omega)$ is bounded and uniformly continuous for ρ_P from $\mathcal{F}$ into $\mathbb{R}$.

Given $f \in \mathcal{L}^2(P)$, let $\pi_0(f) := f - \int f dP$. Then $\pi_0(f) \in \mathcal{L}_0^2(P)$. Given $\mathcal{F} \subset \mathcal{L}^2(P)$, let $\pi_0(\mathcal{F})$ be the set of all functions $\pi_0(f)$, $f \in \mathcal{F}$. For any $f \in \mathcal{L}^2(P)$, $\rho_P(f, \pi_0(f)) = 0$, and $G_P(f) = G_P(\pi_0(f))$ a.s. I claim that $\mathcal{F}$ is pregaussian if and only if $\pi_0(\mathcal{F})$ is: if $\pi_0(\mathcal{F})$ is pregaussian, then $f \mapsto G_P(\pi_0(f))$ has the desired properties on $\mathcal{F}$, while if $\mathcal{F}$ is pregaussian, then for any $g \in \pi_0(\mathcal{F})$, select arbitrarily (by the axiom of choice!) an $f_g \in \mathcal{F}$ with $\pi_0(f_g) = g$. Then $g \mapsto G_P(f_g)$ has the desired properties.

Now recall the definitions of GB-set and GC-set from Section 2.5 above. Note that if a set C in a Hilbert space is a GB-set, then it must be totally bounded by Theorem 2.8. If C is also a GC-set, then uniformly continuous sample functions of the isonormal process on C must be bounded. Thus since G_P is isonormal on $L_0^2(P)$, any set $\mathcal{G} \subset \mathcal{L}_0^2(P)$ is pregaussian if and only if the corresponding set in $L_0^2(P)$ is a GC-set.

Recall the notion of prelinear function (Lemma 2.14). A G_P process Y on a class $\mathcal{F} \subset \mathcal{L}^2(S,\mathcal{B},P)$ for a probability space $(S,\mathcal{B},P)$ will be called *coherent* if for each $\omega \in S$, the function $f \mapsto Y(f)(\omega)$ on $\mathcal{F}$ is bounded, ρ_P-uniformly continuous and prelinear.

Theorem 3.1. *Let $(S,\mathcal{B},P)$ be any probability space. Then for any class $\mathcal{F} \subset \mathcal{L}^2(S,\mathcal{B},P)$, $\mathcal{F}$ is pregaussian if and only if there exists a coherent G_P process on $\mathcal{F}$.*

Proof. "If" follows from the definition of pregaussian. For the converse, note that π_0 above is linear and G_P is isonormal on $\mathcal{L}_0^2(S,\mathcal{B},P)$, and apply Theorem 2.16. □

Now, if a class $\mathcal{F} \subset \mathcal{L}^2(P)$ is pregaussian we can ask whether ν_n converges in distribution, or in law, to G_P with respect to uniform convergence over $\mathcal{F}$. Recall that for random variables Y_n with values in a separable metric space S, convergence in law of Y_n to Y_0 is defined to mean that $Eg(Y_n) \to Eg(Y_0)$ as $n \to \infty$ for every bounded continuous function g on S. But empirical processes, even empirical distribution functions as treated in Chapter 1 above, take values in nonseparable metric spaces, and for an arbitrary bounded continuous g, $g(\nu_n)$ may not be measurable. A new definition of convergence in law is needed to take care of this non-measurability. The definition will involve the notion of upper integral. Let g be a real-valued, not necessarily measurable function defined on a space X where $(X,\mathcal{S},\mu)$ is a measure space. Let $\overline{\mathbb{R}}$ be the set $[-\infty,\infty]$ of extended real numbers. Then the *upper integral* is defined by

$$\int^* g d\mu \ := \ \inf\{\int h d\mu : \ h \geq g, \ h \text{ measurable and } \overline{\mathbb{R}}\text{-valued}\},$$

which will be undefined if there exists a measurable $h \geq g$ with $\int h d\mu = \infty - \infty$ undefined, unless there is also a measurable $\psi \geq g$ with $\int \psi d\mu = -\infty$, in which case $\int^* g d\mu$ will also be defined as $-\infty$. There always exists at least one measurable $h \geq g$, namely $h \equiv +\infty$.

We will be dealing often with compositions of functions. If f is a function whose domain includes the range of g then either $f(g)$ or $f \circ g$ will denote the function such that $(f \circ g)(x) \equiv f(g(x))$.

Now here is a definition of convergence *in* law, where only the limit variable necessarily *has* a law:

Definition. (J. Hoffmann-Jørgensen) Let (S, d) be any metric space. Let $(\Omega_n, \mathcal{A}_n, Q_n)$ be probability spaces for $n = 0, 1, 2, \cdots$, and $Y_n,\ n \geq 0$, functions from Ω_n into S. Suppose that Y_0 takes values in some separable subset of S and is measurable for the Borel sets on its range. Then Y_n will be said to converge to Y_0 *in law* as $n \to \infty$, in symbols $Y_n \Rightarrow Y_0$, if for every bounded continuous real-valued function g on S,

$$\int^* g(Y_n) dQ_n \ \to \ \int g(Y_0) dQ_0 \quad \text{as} \quad n \to \infty.$$

For g bounded, $\int^* g(Y_n) dP$ is always defined and finite. Then, here is a general definition of when the central limit theorem for empirical measures holds with respect to uniform convergence over a class $\mathcal{F}$ of functions. The metric space S will be the space $\ell^\infty(\mathcal{F})$ of all bounded real-valued functions on $\mathcal{F}$, with metric given by the supremum norm $\|H\|_\mathcal{F} := \sup\{|H(f)| : f \in \mathcal{F}\}$.

Definition. Let $(\Omega, \mathcal{A}, P)$ be a probability space and $\mathcal{F} \subset \mathcal{L}^2(P)$. Then $\mathcal{F}$ will be called a *Donsker class* for P, or *P-Donsker class*, or be said to satisfy the central limit theorem (for empirical measures) for P, if $\mathcal{F}$ is pregaussian for P and $\nu_n \Rightarrow G_P$ in $\ell^\infty(\mathcal{F})$.

Later on, a number of rather large classes $\mathcal{F}$ of functions will be seen to be Donsker classes for various laws P. The next few sections develop some of the needed theory.

3.2 Measurable cover functions

In the last section, convergence in law was defined in terms of upper integrals. The notion of upper integral is related to that of measurable cover. Let $(\Omega, \mathcal{A}, P)$ be a probability space. Then for a possibly non-measurable set $A \subset \Omega$, a set B is called a *measurable cover* of A if $A \subset B,\ B \in \mathcal{A}$, and $P(B) = \inf\{P(C) :\ A \subset C,\ C \text{ measurable}\}$. If B and C are measurable covers of the same set A, then clearly so is $B \cap C$. It follows that $B = C$ up to a set of measure 0, in other words $P(B \Delta C) = 0$ where Δ denotes the symmetric difference, or equivalently $P(1_B = 1_C) = 1$.

For any set $A \subset \Omega$ let $P^*(A) := \inf\{P(B) :\ B \text{ measurable}, A \subset B\}$. Then for any measurable cover B of A, clearly $P^*(A) = P(B)$.

Let $\mathcal{L}^0 := \mathcal{L}^0(\Omega, \mathcal{A}, P, \overline{\mathbb{R}})$ denote the set of all measurable functions from Ω into $\overline{\mathbb{R}}$. Then $\mathcal{L}^0$ is a lattice: for any $f, g \in \mathcal{L}^0$, $f \vee g := \max(f, g)$ and $f \wedge g := \min(f, g)$ are in $\mathcal{L}^0$. But this $\mathcal{L}^0$ is not a vector space since we could have for example $f = +\infty$ and $g = -\infty$, so $f + g$ would be undefined.

The map $y \mapsto \tan^{-1} y$ is one-to-one from $\overline{\mathbb{R}}$ onto $[-\pi/2, \pi/2]$. Then a metric on $\overline{\mathbb{R}}$ is defined from the usual metric on $[-\pi/2, \pi/2]$ by $\bar{d}(x, y) := |\tan^{-1} x - \tan^{-1} y|$. On $\mathcal{L}^0$ we have the Ky

Fan metric (RAP, Theorem 9.2.2) defined by

$$d(f,g) \ := \ \inf\{\varepsilon > 0 : \ P(\bar{d}(f(x),g(x)) > \varepsilon) \ \leq \ \varepsilon\}.$$

Then $d(f,g) = 0$ if and only if $P(f = g) = 1$.

For any set $\mathcal{J} \subset \mathcal{L}^0(\Omega, \mathcal{A}, P, \overline{\mathbb{R}})$, a function $f \in \mathcal{L}^0$ is called an *essential infimum* of $\mathcal{J}$, or $f := \text{ess.inf}\,\mathcal{J}$, iff for all $j \in \mathcal{J}$, $f \leq j$ a.s. and for any $g \in \mathcal{L}^0$ such that $g \leq j$ a.s. for all $j \in \mathcal{J}$, we have $g \leq f$ a.s. If f and g are two essential infima of the same set $\mathcal{J}$, then clearly $f = g$ a.s. A set $\mathcal{J}$ of functions will be called a *lower semilattice* if for any $f, g \in \mathcal{J}$, we have $\min(f,g) \in \mathcal{J}$.

Theorem 3.2. *For any probability space $(\Omega, \mathcal{A}, P)$ and class $\mathcal{J} \subset \mathcal{L}^0(\Omega, \mathcal{A}, P, \overline{\mathbb{R}})$, an essential infimum of $\mathcal{J}$ exists. If for some function $f : \ \Omega \mapsto \mathbb{R}$ we have $\mathcal{J} = \{j \in \mathcal{L}^0 : j \geq f$ everywhere$\}$, then $f^* := ess.\inf \mathcal{J}$ can be chosen so that $f^* \geq f$ everywhere. Also, $\int f^* dP$ and $E^* f \ := \ \int^* f dP$ are both defined and equal if either of them is well-defined (possibly infinite), for example if f^* is bounded below.*

Here f^* is called the *measurable cover function* of f. Recall that in Section 2.5, $L(A)^*$ was the essential supremum of $L(x)$ for $x \in A$, and so, the essential infimum of random variables Y such that for each $x \in A$, $Y \geq L(x)$ a.s. - a different, although related, notion. If f is real valued and bounded above by some finite valued measurable function then f^* is a measurable real-valued function. But whenever there exist non-measurable sets $A_n \downarrow \emptyset$ with $P^*(A_n) \ \equiv \ 1$, as for Lebesgue measure (e.g. RAP, Section 3.4, problem 2), let $f := n$ on $A_n \setminus A_{n+1}$. Then f is real valued but $f^* = +\infty$ a.s.

The next two lemmas on measurable cover functions are basic.

Lemma 3.3. *For any two functions $f, g : \ \Omega \mapsto (-\infty, \infty]$, we have*
(a) $(f+g)^* \leq f^* + g^*$ a.s., and
(b) $(f-g)^* \geq f^* - g^*$ whenever both sides are defined a.s.

Lemma 3.4. *Let S be a vector space with a seminorm $\|\cdot\|$. Then for any two functions X, Y from Ω into S, $\|X+Y\|^* \leq (\|X\| + \|Y\|)^* \leq \|X\|^* + \|Y\|^*$ a.s. and $\|cX\|^* = |c|\|X\|^*$ a.s. for any real c.*

Proof. The first inequality is clear, the second follows from Lemma 3.3 and the equation is clear (for $c = 0$ and $c \neq 0$). □

Next, in some cases of independence, the upper-star operation can be distributed over products or sums.

Lemma 3.5. *Let $(\Omega_j, \mathcal{A}_j, P_j)$, $j = 1, \cdots, n$, be any n probability spaces. Let f_j be functions from Ω_j into $\overline{\mathbb{R}}$. Suppose either*
(a) $f_j \geq 0$, $j = 1, \cdots, n$, or
(b) $f_1 \ \equiv \ 1$ and $n = 2$.

Then on the Cartesian product $\Pi_{j=1}^n (\Omega_j, \mathcal{A}_j, P_j)$, if $x := (x_1, \cdots, x_n)$ and we have $f(x) := \Pi_{j=1}^n f_j(x_j)$, then $f^*(x) = \Pi_{j=1}^n f_j^*(x_j)$ a.s., where $0 \cdot \infty$ is set equal to 0.
(c) Or, if $f_j(x_j) > -\infty$ for all x_j, $j = 1, \cdots, n$, and $g(x_1, \cdots, x_n) := f_1(x_1) + \cdots + f_n(x_n)$, then $g^*(x_1, \cdots, x_n) = f_1^*(x_1) + \cdots + f_n^*(x_n)$ a.s.

For the next fact here is some notation: given two functions f, g and a σ-algebra $\mathcal{S}$ on the range of f, let $(f,g)(x) := (f(x), g(x))$ and $f^{-1}(\mathcal{S}) := \{f^{-1}(A) : A \in \mathcal{S}\}$.

Lemma 3.6. *Let $(\Omega, \mathcal{A}, P) = \Pi_{i=1}^{3}(\Omega_i, \mathcal{S}_i, P_i)$ with coordinate projections Π_i: $\Pi_i(x_1, x_2, x_3) := x_i$, $i = 1, 2, 3$. Let $\mathcal{S}_1 \otimes \mathcal{S}_2$ denote the product σ-algebra on $\Omega_1 \times \Omega_2$. Then for any bounded real function f on $\Omega_1 \times \Omega_3$ and $g(x_1, x_2, x_3) := f(x_1, x_3)$, conditional expectations of g^* satisfy*

$$E(g^* | (\Pi_1, \Pi_2)^{-1}(\mathcal{S}_1 \otimes \mathcal{S}_2)) = E(g^* | \Pi^{-1}(\mathcal{S}_1)) \quad \text{a.s. for } P.$$

Lemma 3.7. *Let X be a real-valued function on a probability space $(\Omega, \mathcal{A}, P)$. Then for any $t \in \mathbb{R}$,*
(a) $P^*(X > t) = P(X^* > t)$.
(b) For any $\varepsilon > 0$, $P^*(X \geq t) \leq P(X^* \geq t) \leq P^*(X \geq t - \varepsilon)$.

Let $(\Omega, \mathcal{A}, P)$ be a probability space. For a function f from Ω into $[-\infty, \infty]$ let $\int_* f dP := \sup\{\int g dP : g \text{ measurable}, g \leq f\}$. Let f_* be the essential supremum of all measurable functions $g \leq f$. Then just as for f^*, f_* is well-defined up to a.s equality and $\int_* f dP = \int f_* dP$ whenever either side is defined, as in Theorem 3.2.

It's easy to check that $f_* = -((-f)^*)$ and that $\int_* f dP = -(\int^* -f dP)$. So the convergence in law $Y_n \Rightarrow Y_0$ as defined in Section 3.1 implies that $\int_* g(Y_n) dQ_n \to \int g(Y_0) dQ_0$.

Next, here is a one-sided Tonelli-Fubini theorem for starred functions:

Theorem 3.8. *Let $(X, \mathcal{A}, P) \times (Y, \mathcal{B}, Q)$ be a product of two probability spaces. For a real-valued function $f \geq 0$ on $X \times Y$, define f^* with respect to $P \times Q$. For each $x \in X$ let $(E_2^* f)(x) := \int^* f(x, y) dQ(y)$. Then*

$$E_1^* E_2^* f(x, y) \leq \int f^*(x, y) d(P \times Q)(x, y), \tag{3.1}$$

where $E_1^* = E^*$ with respect to P. Also, if Q is purely atomic, so that $\sum_j Q(\{y_j\}) = 1$ for some $y_j \in Y$, and $E_2(\cdot) := \int \cdot \, dQ$, then

$$E_1^* E_2 f(X, Y) \leq E^* f(X, Y) = E_2 E_1^* f(X, Y). \tag{3.2}$$

We also have a one-sided monotone convergence theorem with stars:

Theorem 3.9. *Let $(\Omega, \mathcal{A}, P)$ be a probability space and let f_j be real-valued functions on Ω such that $f_j \uparrow f$, i.e. $f_j(x) \uparrow f(x)$ for all $x \in \Omega$. If $E^* f_1 > -\infty$ then $E^* f_j \uparrow E^* f$ as $j \to \infty$.*

Note that there exist subsets $A_j := A(j)$ of $[0, 1]$ with outer measure $\lambda^*(A_j) = 1$ for all j and $A_j \downarrow \emptyset$ (RAP, Problem 3.4.2). Letting $f_j := 1_{A(j)}$, we have $f_j \downarrow 0$, and $E^* f_j = 1$ for all j, so that the monotone convergence theorem fails for E^* for decreasing sequences. Next is a Fatou lemma with stars.

Theorem 3.10. *Let $(\Omega, \mathcal{A}, P)$ be a probability space and f_j any nonnegative real-valued functions on Ω. Then*

$$E^* \liminf_{j \to \infty} f_j \leq \liminf_{j \to \infty} E^* f_j.$$

3.3 Almost uniform convergence and convergence in outer probability

In Section 3.1, the definition of convergence of laws was adapted to define convergence in law for random elements which may not have laws defined. In this section the same will be done for convergence in probability and almost sure convergence.

Let $(\Omega, \mathcal{A}, P)$ be a probability space, (S, d) a metric space, and f_n functions from Ω into S. Then f_n will be said to converge to f_0 in *outer probability* if $d(f_n, f_0)^* \to 0$ in probability as $n \to \infty$, or equivalently, by Lemma 3.7, for every $\varepsilon > 0$, $P^*\{d(f_n, f_0) > \varepsilon\} \to 0$ as $n \to \infty$. Also, f_n is said to converge to f_0 *almost uniformly* if as $n \to \infty$, $d(f_n, f_0)^* \to 0$ almost surely.

The following is immediate:

Proposition 3.11. *Almost uniform convergence always implies convergence in outer probability.*

If f_n are all measurable functions, then clearly convergence in outer probability is equivalent to the usual convergence in probability, and almost uniform convergence to almost sure convergence. Now some definitions will be given for Glivenko-Cantelli properties, which are laws of large numbers for empirical measures.

Definition. If $(X, \mathcal{A}, P)$ is a probability space and $\mathcal{F}$ is a class of integrable real-valued functions, $\mathcal{F} \subset \mathcal{L}^1(X, A, P)$, then $\mathcal{F}$ will be called a *strong* (resp. *weak) Glivenko-Cantelli class for* P iff as $n \to \infty$, $\|P_n - P\|_{\mathcal{F}} \to 0$ almost uniformly (resp. in outer probability).

In the following Proposition, part (C) is usually called "Egorov's theorem" for almost surely convergent sequences of measurable functions (RAP, Theorem 7.5.1).

Proposition 3.12. *Let $(\Omega, \mathcal{A}, P)$ be a probability space, (S, d) a metric space, and f_n any functions from Ω into S for $n = 0, 1, \cdots$. Then the following are equivalent:*
(A) $f_n \to f_0$ almost uniformly;
(B) For any $\varepsilon > 0$, $P^*\{\sup_{n \geq m} d(f_n, f_0) > \varepsilon\} \downarrow 0$ as $m \to \infty$.
(C) For any $\delta > 0$ there is some $B \in \mathcal{A}$ with $P(B) > 1 - \delta$ such that $f_n \to f_0$ uniformly on B.
(D) There exist measurable $h_n \geq d(f_n, f_0)$ with $h_n \to 0$ a.s.

Example. In $[0, 1]$ with Lebesgue measure P let $A_1 \supset A_2 \supset \cdots$ be sets with $P^*(A_n) = 1$ and $\bigcap_{n=1}^{\infty} A_n = \emptyset$ (e.g. RAP, Section 3.4, problem 2; Cohn, 1980, p. 35). Then $1_{A_n} \to 0$ everywhere and, in that sense, almost surely, but not almost uniformly. Note also that 1_{A_n} doesn't converge to 0 in law as defined in Section 3.1. To avoid such pathology, almost uniform convergence is helpful.

Proposition 3.13. *Let (S, d) and (Y, e) be two metric spaces and $(\Omega, \mathcal{A}, P)$ a probability space. Let f_n be functions from Ω into S for $n = 1, 2, \cdots$, such that $f_n \to f_0$ in outer probability as $n \to \infty$. Assume that f_0 has separable range and is measurable (for the Borel σ-algebra on S). Let g be a continuous function from S into Y. Then $g(f_n) \to g(f_0)$ in outer probability.*

Note. If $\int G(f_0) dP$ is defined for all bounded continuous real G (as it must be if $f_n \Rightarrow f_0$, by definition) then the image measure $P \circ f_0^{-1}$ is defined on all Borel subsets of S (RAP, Theorem 7.1.1). Such a law does have a separable support except perhaps in some set-theoretically pathological cases (Appendix C).

Lemma 3.14. *Let $(\Omega, \mathcal{A}, P)$ be a probability space and $\{g_n\}_{n=0}^{\infty}$ a uniformly bounded sequence of real-valued functions on Ω such that g_0 is measurable. If $g_n \to g_0$ in outer probability then* $\limsup_{n\to\infty} \int^* g_n dP \leq \int g_0 dP$.

On any metric space, the σ-algebra will be the Borel σ-algebra unless something is said to the contrary.

Corollary 3.15. *If f_n are functions from a probability space into a metric space, $f_n \to f_0$ in outer probability and f_0 is measurable with separable range, then $f_n \Rightarrow f_0$.*

Proof. Apply Proposition 3.13 to $g = G$ for any bounded continuous G and Lemma 3.14. □

3.4 Perfect functions

For a function g defined on a set A let $g[A] := \{g(x) : x \in A\}$. It will be useful that under some conditions on a measurable function g and general real-valued f, $(f \circ g)^* = f^* \circ g$. Here are some equivalent conditions (cf. Andersen, 1985b, Sec. II.2):

Theorem 3.16. *Let $(X, \mathcal{A}, P)$ be a probability space, $(Y, \mathcal{B})$ any measurable space, and g a measurable function from X to Y. Let Q be the restriction of $P \circ g^{-1}$ to $\mathcal{B}$. For any real-valued function f on Y, define f^* for Q. Then the following are equivalent:*
(a) For any $A \in \mathcal{A}$ there is a $B \in \mathcal{B}$ with $B \subset g[A]$ and $Q(B) \geq P(A)$;
(b) For any $A \in \mathcal{A}$ with $P(A) > 0$ there is a $B \in \mathcal{B}$ with $B \subset g[A]$ and $Q(B) > 0$;
(c) For every real function f on Y, $(f \circ g)^* = f^* \circ g$ a.s.;
(d) For any $D \subset Y$, $(1_D \circ g)^* = 1_D^* \circ g$ a.s.

Note. In (a) or (b), if the direct image $g[A] \in \mathcal{B}$, we could just set $B := g[A]$. But, for any uncountable complete separable metric space Y, there exists a complete separable metric space S (for example, a countable product $\mathbb{N}^{\infty}$ of copies of $\mathbb{N}$) and a continuous function f from S into Y such that $f[B]$ is not a Borel set in Y (RAP, Theorem 13.2.1, Proposition 13.2.5). If f is only required to be Borel measurable, then S can also be any uncountable complete metric space (RAP, Theorem 13.1.1).

A function g satisfying any of the four conditions in Theorem 3.16 will be called *perfect* or *P-perfect*. Coordinate projections on a product space are, as one would hope, perfect:

Proposition 3.17. *Suppose $A = X \times Y$, P is a product probability $v \times m$, and g is the natural projection of A onto Y. Then g is P-perfect.*

Theorem 3.18. *Let $(\Omega, \mathcal{A}, P)$ be a probability space and (S, d) a metric space. Suppose that for $n = 0, 1, \cdots$, $(Y_n, \mathcal{B}_n)$ is a measurable space, g_n a perfect measurable function from Ω into Y_n, and f_n a function from Y_n into S, where f_0 has separable range and is measurable. Let $Q_n := P \circ g_n^{-1}$ on $\mathcal{B}_n$ and suppose $f_n \circ g_n \to f_0 \circ g_0$ in outer probability as $n \to \infty$. Then $f_n \Rightarrow f_0$ as $n \to \infty$ for f_n on $(Y_n, \mathcal{B}_n, Q_n)$.*

Here is a related example:

Proposition 3.19. *Theorem 3.18 can fail without the hypothesis that g_n be perfect.*

Proof. Let $C \subset I := [0,1]$ satisfy $0 = \lambda_*(C) < \lambda^*(C) = 1$ for Lebesgue measure λ (RAP, Theorem 3.4.4). Let $P = \lambda^*$, giving a probability measure on the Borel sets of C (RAP, Theorem 3.3.6). Let $\Omega = C$, $f_0 \equiv 0$, $Y_n = I$, $f_n := 1_{I \setminus C}$, and let g_n be the identity from C into Y_n for all n. Then $f_n \circ g_n \equiv 0$ for all n, so $f_n \circ g_n \to f_0 \circ g_0$ in outer probability (and in any other sense). Let $\mathcal{B}_n$ be the Borel σ-algebra on $Y_n = I$ for each n. Let G be the identity from I into $\mathbb{R}$. Then $\int^* G(f_n) dQ_n = \int^* f_n d\lambda = 1$ for $n \geq 1$, while $\int G(f_0) dQ_0 = 0$, so f_n does not converge to f_0 in law. □

From Theorem 3.18, it follows that the g_n in the last proof are not perfect, as can also be seen directly, from condition (c) or (d) in Theorem 3.16.)

In Proposition 3.17, $X \times Y$ could be an arbitrary product probability space, but projection is a rather special function. The following fact will say that all measurable functions on reasonable domain spaces are perfect.

Recall that a metric space (S, d) is called *universally measurable (u.m.)* if for every law P on the completion of S, S is measurable for the completion of P (RAP, Section 11.5). So any complete metric space, or any Borel set in its completion, is u.m.

Theorem 3.20. *Let (S, d) be a u.m. separable metric space. Let P be a probability measure on the Borel σ-algebra of S. Then any Borel measurable function g from S into a separable metric space Y is perfect for P.*

Note. In view of Appendix C below the hypothesis that S be separable is not very restrictive.

3.5 Almost surely convergent realizations

First let's recall a theorem of Skorohod (RAP, Theorem 11.7.2): if (S, d) is a complete separable metric space, and P_n are laws on S converging to a law P_0, then on some probability space there exist S-valued measurable functions X_n such that $\mathcal{L}(X_n) = P_n$ for all n and $X_n \to X_0$ almost surely. This section will give an extension of Skorohod's theorem to our current setup.

Having almost uniformly convergent realizations shows that the definition of convergence in law for random elements is reasonable and is useful in some proofs on convergence in law.

Suppose $f_n \Rightarrow f_0$ where f_n are random elements, in other words functions not necessarily measurable except for $n = 0$, defined on some probability spaces (Ω_n, Q_n) into a possibly nonseparable metric space S. We want to find random elements Y_n "having the same laws" as f_n for each n such that $Y_n \to Y_0$ almost surely or better, almost uniformly. At first look it isn't clear what "having the same laws" should mean for random elements f_n, $n \geq 1$, not having laws defined on any non-trivial σ-algebra . A way that turns out to work is to define $Y_n = f_n \circ g_n$ where g_n are functions from some other probability space Ω with probability measure Q into Ω_n such that each g_n is measurable and $Q \circ g_n^{-1} = Q_n$ for each n. Thus the argument of f_n will have the same law Q_n as before. It turns out moreover that the g_n should be not only measurable but perfect.

Before stating the theorem, here is an example to show that there may really be no way to define a σ-algebra on S on which laws could be defined and yield an equivalence as in the next theorem, even if S is a finite set.

Example. Let $(X_n, \mathcal{A}_n, Q_n) = ([0,1], \mathcal{B}, \lambda)$ for all n (λ = Lebesgue measure, $\mathcal{B}$ = Borel σ-algebra). Take sets $C(n) \subset [0,1]$ with $0 = \lambda_*(C(n)) < \lambda^*(C(n)) = 1/n^2$ (RAP, Theorem

3.4.4). Let S be the two-point space $\{0,1\}$ with usual metric. Then $f_n := 1_{C(n)} \to 0$ in law and almost uniformly, but each "law" $\beta_n := Q_n \circ f_n^{-1}$ is only defined on the trivial σ-algebra $\{\emptyset, S\}$. The only larger σ-algebra on S is 2^S, but no β_n for $n \geq 1$ is defined on 2^S.

Theorem 3.21. *Let (S,d) be any metric space, $(X_n, \mathcal{A}_n, Q_n)$ any probability spaces, and f_n a function from X_n into S for each $n = 0, 1, \cdots$. Suppose f_0 has separable range S_0 and is measurable (for the Borel σ-algebra on S_0). Then $f_n \Rightarrow f_0$ if and only if there exists a probability space $(\Omega, \mathcal{S}, Q)$ and perfect measurable functions g_n from $(\Omega, \mathcal{S})$ to $(X_n, \mathcal{A}_n)$ for each $n = 0, 1, \cdots$, such that $Q \circ g_n^{-1} = Q_n$ on $\mathcal{A}_n$ for each n and $f_n \circ g_n \to f_0 \circ g_0$ almost uniformly as $n \to \infty$.*

Notes. Proposition 3.19 and the "if and only if" in Theorem 3.21 show that the hypothesis that g_n be perfect can't just be dropped from the Theorem. "If" follows from Proposition 3.11 and Theorem 3.18. "Only if" can be proved very much as in RAP (Theorem 11.7.2).

3.6 Conditions equivalent to convergence in law

Conditions equivalent to convergence of laws on separable metric spaces are given in the portmanteau theorem (RAP, Theorem 11.1.1) and metrization theorem (RAP, Theorem 11.3.3). Here, the conditions will be extended to general random elements for the theory being presented in this chapter.

For any probability space $(\Omega, \mathcal{A}, P)$ and real-valued function f on Ω let $E^* f := \int^* f dP$ and $E_* f := \int_* f dP$. If (S,d) is a metric space and f is a real-valued function on S, recall (RAP, Section 11.2) that the Lipschitz seminorm of f is defined by

$$\|f\|_L := \sup\{|f(x) - f(y)|/d(x,y) : x \neq y\}$$

and f is called a *Lipschitz* function if $\|f\|_L < \infty$. The bounded Lipschitz norm is defined by $\|f\|_{BL} := \|f\|_L + \|f\|_\infty$ where $\|f\|_\infty := \sup_x |f(x)|$. Then f is called a bounded Lipschitz function if $\|f\|_{BL} < \infty$, and $\|\cdot\|_{BL}$ is a norm on the space of all such functions.

The extended portmanteau theorem about to be stated is an adaptation of RAP, Theorem 11.1.1 and some further facts based on the last section (Theorem 3.21), cf. Andersen and Dobrič (1987), Remark 2.13.

Theorem 3.22. *Let (S,d) be any metric space. For $n = 0, 1, 2, \cdots$, let $(X_n, \mathcal{A}_n, Q_n)$ be a probability space and f_n a function from X_n into S. Suppose f_0 has separable range S_0 and is measurable. Let $P := Q_0 \circ f_0^{-1}$ on S. Then the following are equivalent:*
(a) $f_n \Rightarrow f_0$;
(a') $\limsup_{n\to\infty} E^* G(f_n) \leq E G(f_0)$ for all bounded continuous real-valued G on S;
(b) $E^* G(f_n) \to E G(f_0)$ as $n \to \infty$ for every bounded Lipschitz function G on S;
(b') (a') holds for all bounded Lipschitz G on S;
(c) $\sup\{|E^* G(f_n) - E G(f_0)| : \|G\|_{BL} \leq 1\} \to 0$ as $n \to \infty$;
(d) For any closed $F \subset S$, $P(F) \geq \limsup_{n\to\infty} Q_n^*(f_n \in F)$;
(e) For any open $U \subset S$, $P(U) \leq \liminf_{n\to\infty} (Q_n)_*(f_n \in U)$;
(f) For any continuity set A of P in S, $Q_n^*(f_n \in A) \to P(A)$ and $(Q_n)_*(f_n \in A) \to P(A)$ as $n \to \infty$;

(g) There exist a probability space $(\Omega, \mathcal{S}, Q)$ and measurable functions g_n from Ω into X_n and h_n from Ω into S such that the g_n are perfect, $Q \circ g_n^{-1} = Q_n$ and $Q \circ h_n^{-1} = P$ for all n, and $d(f_n \circ g_n, h_n) \to 0$ almost uniformly.

Moreover, (g) remains equivalent if any of the following changes are made in it: "almost uniformly" can be replaced by "in outer probability"; we can take take $h_n = f_0 \circ \gamma_n$ for some measurable functions γ_n from Ω into X_0, which can be taken to be perfect; and we can take γ_n to be all the same, $\gamma_n \equiv \gamma_1$ for all n.

It will be seen that convergence in law is also equivalent to convergence in some analogues of the Prohorov and dual-bounded-Lipschitz metrics which metrize convergence of laws on separable metric spaces as shown in RAP, Theorem 11.3.3. We will have analogues of metrics, rather than actual metrics, because of the non-symmetry between the non-measurable random elements f_n and limiting measurable random variable f_0.

Definitions. Let $(X_m, \mathcal{A}_m, Q_m)$ be probability spaces, $m = 0, 1$, and (S, d) a metric space. Let f_m be functions from X_m into S, $m = 0, 1$, such that f_0 is measurable and has separable range. Let $P := Q_0 \circ f_0^{-1}$. Then let

$$\beta(f_1, f_0) := \sup\{|E^* G(f_1) - EG(f_0)| : \|G\|_{BL} \le 1\}, \quad \text{and}$$

$$\rho(f_1, f_0) := \inf\{\varepsilon > 0 : P(F) \le (Q_1)_*(f_1 \in F^\varepsilon) + \varepsilon \text{ for every non-empty closed set } F \subset S\}.$$

Theorem 3.23. *For any metric space (S, d), probability spaces $(X_m, \mathcal{A}_m, Q_m)$ and functions f_m from X_m into S, where f_0 has separable range and is measurable, the following are equivalent:*
(i) $f_m \Rightarrow f_0$;
(ii) $\beta(f_m, f_0) \to 0$ as $m \to \infty$;
(iii) $\rho(f_m, f_0) \to 0$ as $m \to \infty$.

Next is a version of Theorem 3.23 where we have two indices.

Theorem 3.24. *Let (S, d) be a metric space and $(\Omega, \mathcal{S}, Q)$ a probability space. Suppose that for each $m, n = 1, 2, \cdots$, f_{mn} is a function from Ω into S, and f_0 is a measurable function from Ω into S with separable range. Then the following are equivalent:*
(i) $f_{m,n} \Rightarrow f_0$, i.e. for every bounded continuous real function G on S,

$$\int^* G(f_{m,n}) dQ \to \int G(f_0) dQ \quad \text{as} \quad m, n \to \infty;$$

(ii) $\beta(f_{mn}, f_0) \to 0$ as $m, n \to \infty$;
(iii) $\rho(f_{mn}, f_0) \to 0$ as $m, n \to \infty$.

We have the following;

Theorem 3.25. *Continuous mapping theorem. Under the conditions of Theorem 3.23, if (T, e) is another metric space, G is a continuous function from S into T, and $f_m \Rightarrow f_0$, then $G(f_m) \Rightarrow G(f_0)$.*

Proof. This follows directly from the definition of convergence in law $\Rightarrow$. □

We also have

Theorem 3.26. *Suppose f_m, $m \geq 0$, are measurable random variables taking values in a separable metric space S, so that laws $\mathcal{L}(f_m)$ exist on the Borel σ-algebra of S. Then convergence $f_m \Rightarrow f_0$ is equivalent to convergence of the laws $\mathcal{L}(f_m) \to \mathcal{L}(f_0)$ in the usual sense.*

Proof. For any bounded continuous real-valued function G on S, the functions $G(f_m)$ are measurable, so upper integrals reduce to integrals, and the result follows from the definitions. □

3.7 Asymptotic equicontinuity and Donsker classes

Recall from Section 3.1 the definitions of the empirical measures P_n, empirical process ν_n, pseudo-metric ρ_P and the Gaussian process G_P. Recall that a set $\mathcal{F} \subset \mathcal{L}^2(P)$ is called *pregaussian* if a G_P process restricted to $\mathcal{F}$ exists whose sample functions $f \mapsto G_P(f)(\omega)$ are (almost) all bounded and uniformly continuous for ρ_P on $\mathcal{F}$. Recall that such a G_P process is called coherent if in addition, its sample functions are prelinear on $\mathcal{F}$ as in Lemma 2.14.

Proposition 3.27. *For any probability measure P and pregaussian set $\mathcal{F} \subset \mathcal{L}^2(P)$, the symmetric convex hull $sco(\mathcal{F})$ of $\mathcal{F}$ is also pregaussian and there exists a G_P process defined on the linear span of $\mathcal{F}$ and constant functions which is 0 on the constant functions and coherent on $sco(\mathcal{F})$.*

Proof. A G_P process is isonormal on the space $\mathcal{L}^2(P)$ for the semi-inner product $(f,g)_{0,P} := P(fg) - PfPg$, and G_P is 0 on constant functions. So Theorem 2.16 on isonormal processes applies and gives the result. □

For a signed measure ν and measurable function f such that $\int f d\nu$ is defined, let $\nu(f) := \int f d\nu$.

A class $\mathcal{F}$ of functions will be said to satisfy the *asymptotic equicontinuity condition* for P and a pseudometric τ on $\mathcal{F}$, or $\mathcal{F} \in AEC(P,\tau)$ for short, if for every $\varepsilon > 0$ there is a $\delta > 0$ and an n_0 large enough such that for $n \geq n_0$,

$$\Pr^*\{\sup\{|\nu_n(f-g)| : \ f,g \in \mathcal{F}, \ \tau(f,g) < \delta\} > \varepsilon\} \ < \ \varepsilon.$$

Then $\mathcal{F} \in AEC(P)$ will mean $\mathcal{F} \in AEC(P,\rho_P)$.

Theorem 3.28. *Let $\mathcal{F} \subset \mathcal{L}^2(X,\mathcal{A},P)$. Then the following are equivalent:*
(I) $\mathcal{F}$ is a Donsker class for P, in other words $\mathcal{F}$ is P-pregaussian and $\nu_n \Rightarrow G_P$ in $\ell^\infty(\mathcal{F})$;
(II) (a) $\mathcal{F}$ is totally bounded for ρ_P and (b) $\mathcal{F}$ satisfies the asymptotic equicontinuity condition for P, $\mathcal{F} \in AEC(P)$;
(III) There is a pseudometric τ on $\mathcal{F}$ such that $\mathcal{F}$ is totally bounded for τ and $\mathcal{F} \in AEC(P,\tau)$.

3.8 Unions of Donsker classes

The union of any two Donsker classes $\mathcal{F}$ and $\mathcal{G}$ is a Donsker class. This is not surprising: one might think it was enough, given the asymptotic equicontinuity conditions for the separate

classes, for a given $\varepsilon > 0$, to take the larger of the two n_0's and the smaller of the two δ's. But it is not so easy as that. For example, $\mathcal{F}$ and $\mathcal{G}$ could both be finite sets, with distinct elements of $\mathcal{F}$ at distance, say, more than 0.2 apart for ρ_P, and likewise for $\mathcal{G}$, but there may be some element of $\mathcal{F}$ very close to an element of $\mathcal{G}$. So the equicontinuity condition on the union won't just follow from the conditions on the separate families.

Theorem 3.29. (K. Alexander) *Let $(\Omega, \mathcal{A}, P)$ be a probability space and let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two Donsker classes for P. Then $\mathcal{F} := \mathcal{F}_1 \cup \mathcal{F}_2$ is also a Donsker class for P.*

3.9 Sequences of sets and functions

Theorem 3.30. *Let $(X, \mathcal{A}, P)$ be a probability space and $\{C_m\}_{m \geq 1}$ a sequence of measurable sets. If*

$$\sum_{m=1}^{\infty} (P(C_m)(1 - P(C_m)))^r < \infty \quad \text{for some} \quad r < \infty, \tag{3.3}$$

then the sequence $\{C_m\}_{m \geq 1}$ is a Donsker class for P. Conversely, if the sets C_m are independent for P, then the sequence is a Donsker class only if (3.3) holds.

Next, let's consider sequences of functions. For a probability space $(A, \mathcal{A}, P)$ and $f \in \mathcal{L}^2(A, \mathcal{A}, P)$ let $\sigma_P^2(f) := \int f^2 dP - (\int f dP)^2$ (the variance of f). Here is a sufficient condition for the Donsker property of a sequence $\{f_m\}$ which is easy to prove, yet turns out to be optimal of its kind:

Theorem 3.31. *If $\{f_m\}_{m \geq 1} \subset \mathcal{L}^2(P)$ and $\sum_{m=1}^{\infty} \sigma_P^2(f_m) < \infty$, then $\{f_m\}_{m \geq 1}$ is a Donsker class for P.*

The following shows that Theorem 3.31, although it does not imply the first half of Theorem 3.30, is sharp in one sense:

Proposition 3.32. *Let $A := [0,1]$ and $P := U[0,1] :=$ Lebesgue measure on A. Let $a_m > 0$ satisfy $\sum_{m=1}^{\infty} a_m = +\infty$. Then there is a sequence $\{f_m\} \subset \mathcal{L}^2(A, \mathcal{A}, P)$ with $\sigma_P^2(f_m) \leq a_m$ for all m where $\{f_m\}$ is not a Donsker class.*

REFERENCES FOR CHAPTER 3

Alexander, K. S. (1987). The central limit theorem for empirical processes on Vapnik-Červonenkis classes. *Ann. Probab.* **15**, 178-203.

Andersen, Niels Trolle (1985a). The central limit theorem for non-separable valued functions. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **70**, 445-455.

Andersen, N. T. (1985b). The calculus of non-measurable functions and sets. Various Publ. Ser. no. 36, Matematisk Institut, Aarhus Universitet.

Andersen, N. T., and Dobrić, V. (1987). The central limit theorem for stochastic processes. *Ann. Probab.* **15**, 164-177.

Andersen, N. T., and Dobrić, V. (1988). The central limit theorem for stochastic processes II. *J. Theoret. Probab.* **1**, 287-303.

Bauer, H. (1981). *Probability Theory and Elements of Measure Theory*, 2d. ed. Academic Press, London.

Blumberg, Henry (1935). The measurable boundaries of an arbitary function. *Acta Math.* (Uppsala) **65**, 263-282.

Cohn, D. L. (1980). *Measure Theory.* Birkhäuser, Boston.

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. (1968). Distances of probability measures and random variables. *Ann. Math. Statist.* **39**, 1563-1572.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6**, 899-929; Correction **7** (1979) 909-911.

Dudley, R. M. (1981). Donsker classes of functions. In *Statistics and Related Topics* (Proc. Symp. Ottawa, 1980), North-Holland, New York, 341-352.

Dudley, R. M. (1984). A course on empirical processes. *École d'été de probabilités de St.-Flour*, 1982. *Lecture Notes in Math.* (Springer) **1097**, 1-142.

Dudley, R. M. (1985). An extended Wichura theorem, definitions of Donsker class, and weighted empirical distributions. In *Probability in Banach Spaces V* (Proc. Conf. Medford, 1984), *Lecture Notes in Math.* (Springer) **1153**, 141-178.

Dudley, R. M. (1990). Nonlinear functionals of empirical measures and the bootstrap. In *Probability in Banach Spaces 7*, Proc. Conf. Oberwolfach, 1988, *Progress in Probability* **21**, Birkhäuser, Boston, 63-82.

Dudley, R. M. (1994). Metric marginal problems for set-valued or non-measurable variables. *Probability Theory and Related Fields* **100**, 175-189.

Dudley, R. M., and Philipp, Walter (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Z. Wahrschinlichkeitstheorie verw. Gebiete* **62**, 509-552.

Eames, W., and May, L. E. (1967). Measurable cover functions. *Canad. Math. Bull.* **10**, 519-523.

Eršov, M. P. (1975). The Choquet theorem and stochastic equations. *Analysis Math.* **1**, 259-271.

Giné, E., and J. Zinn (1986). Lectures on the central limit theorem for empirical processes. In *Probability and Banach Spaces*, Proc. Conf. Zaragoza, 1985, *Lecture Notes in Math.* (Springer) **1221**, 50-113.

Gnedenko, B. V., and Kolmogorov, A. N. (1949). *Limit Distributions for Sums of Independent Random Variables.* Moscow. Transl. and ed. by K. L. Chung, Addison-Wesley, Reading, Mass., 1954, rev. ed. 1968.

Goffman, C., and Zink, R. E. (1960). Concerning the measurable boundaries of a real function. *Fund. Math.* **48**, 105-111.

Hoffmann-Jørgensen, Jørgen (1984). *Stochastic processes on Polish spaces.* Published (1991): Various Publication Series no. 39, Matematisk Institut, Aarhus Universitet (278 pp.).

Hoffmann-Jørgensen, Jørgen (1985). The law of large numbers for non-measurable and non-separable random elements. *Astérisque* **131**, 299-356.

Luxemburg, W. A. J., and Zaanen, A. C. (1983). *Riesz Spaces,* vol. 2, North-Holland, Amsterdam.

Pachl, Jan K. (1979). Two classes of measures. *Colloq. Math.* **42**, 331-340.

Ryll-Nardzewski, C. (1953). On quasi-compact measures. *Fund.*

Math. **40**, 125-130.

Sazonov, V. V. (1962). On perfect measures (in Russian). *Izv. Akad. Nauk SSSR* **26**, 391-414.

Skorohod, Anatolii Vladimirovich (1956). Limit theorems for stochastic processes. *Theor. Probab. Appl.* **1**, 261-290.

Skorohod, A. V. (1976). On a representation of random variables. *Theor. Probab. Appl.* **21**, 628-632 (English), 645-648 (Russian).

Topsϕe, Flemming (1970). *Topology and Measure. Lecture Notes in Math.* (Springer) **133**.

van der Vaart, Aad (1996). New Donsker classes. *Ann. Probab.* **24**, 2128-2140.

Vulikh, B. Z. (1961). *Introduction to the Theory of Partially Ordered Spaces* (transl. by L. F. Boron, 1967). Wolters-Noordhoff, Groningen.

Wichura, Michael J. (1970). On the construction of almost uniformly convergent random variables with given weakly convergent image laws. *Ann. Math. Statist.* **41**, 284-291.

Chapter 4

Vapnik-Červonenkis combinatorics

This chapter will treat some classes of sets satisfying a combinatorial condition. In Chapter 6, it will be seen that under a mild measurability condition to be treated in Chapter 5, these classes have the Donsker property, for all probability measures P on the sample space, and satisfy a law of large numbers (Glivenko-Cantelli property) uniformly in P. Moreover, for either of these limit-theorem properties of a class of sets (without assuming any measurability), the Vapnik-Červonenkis property is necessary (Section 6.4).

The present chapter will be self-contained, not depending on anything earlier in these notes, except in some examples.

4.1 Vapnik-Červonenkis classes

Let X be any set and $\mathcal{C}$ a collection of subsets of X. For $A \subset X$ let $\mathcal{C}_A := \mathcal{C} \sqcap A := A \sqcap \mathcal{C} := \{C \cap A : \ C \in \mathcal{C}\}$. Let $\mathrm{card}(A) := |A|$ denote the cardinality (number of elements) of A and $2^A := \{B : \ B \subset A\}$. Let $\Delta^{\mathcal{C}}(A) := |\mathcal{C}_A|$. If $A \sqcap \mathcal{C} = 2^A$, then $\mathcal{C}$ is said to *shatter* A. If A is finite, then $\mathcal{C}$ shatters A if and only if $\Delta^{\mathcal{C}}(A) = 2^{|A|}$.

Let $m^{\mathcal{C}}(n) := \max\{\Delta^{\mathcal{C}}(F) : \ F \subset X, \ |F| = n\}$ for $n = 0, 1, ...,$ or if $|X| < n$ let $m^{\mathcal{C}}(n) := m^{\mathcal{C}}(|X|)$. Then $m^{\mathcal{C}}(n) \leq 2^n$ for all n. Let

$$V(\mathcal{C}) \ := \inf\{n : m^{\mathcal{C}}(n) < 2^n\}, \text{ if this is finite,}$$

$$+\infty, \text{ if } \ m^{\mathcal{C}}(n) = 2^n \text{ for all } n,$$

$$S(\mathcal{C}) \ := \ \sup\{n : \ m^{\mathcal{C}}(n) = 2^n\},$$

$$-1, \text{ if } \mathcal{C} \text{ is empty.}$$

Then $S(\mathcal{C}) \equiv V(\mathcal{C}) - 1$, and $S(\mathcal{C})$ is the largest cardinality of a set shattered by $\mathcal{C}$, or $+\infty$ if arbitarily large finite sets are shattered. So, $V(\mathcal{C})$ is the smallest n such that no set of cardinality n is shattered by $\mathcal{C}$. If $V(\mathcal{C}) < \infty$, or equivalently if $S(\mathcal{C}) < \infty$, $\mathcal{C}$ will be called a *Vapnik-Červonenkis class* or *VC class.*

If X is finite, with n elements, then clearly 2^X is a VC class, with $S(2^X) = n$. Let ${}_N C_{\leq k} := \sum_{j=0}^{k} \binom{N}{j}$, where

$$\binom{N}{j} := N!/(j!(N-j)!), \quad j = 0, 1, ..., N,$$

$$0, \quad j > N.$$

Then ${}_NC_{\le k}$ is the number of combinations of N things, at most k at a time. (In an older notation ${}_NC_k := \binom{N}{k}$.) "Pascal's triangle" of identities for binomial coefficients extends to the ${}_NC_{\le k}$:

Proposition 4.1. *${}_NC_{\le k} = {}_{N-1}C_{\le k} + {}_{N-1}C_{\le k-1}$ for $k = 1, 2, \cdots$, and $N = 1, 2, \cdots$.*

For a non-VC class $\mathcal{C}$ we have $m^{\mathcal{C}}(n) = 2^n$ for all n. For a VC class, the next fact, which is fundamental in the Vapnik-Červonenkis theory, will imply that $m^{\mathcal{C}}(n)$ only grows as a polynomial rather than exponentially in n.

Theorem 4.2. *Sauer's Lemma. If $m^{\mathcal{C}}(n) > {}_nC_{\le k-1}$, where $k \ge 1$, then $m^{\mathcal{C}}(k) = 2^k$. Hence if $S(\mathcal{C}) < \infty$, then $m^{\mathcal{C}}(n) \le {}_nC_{\le S(\mathcal{C})}$ for all n.*

For fixed k, ${}_nC_{\le k}$ is easily seen to be a polynomial in n of degree k, with leading term $n^k/k!$. Thus, the next fact is not far from optimal:

Proposition 4.3. *For any nonnegative integers n and k with $n \ge k+2$, ${}_nC_{\le k} \le 1.5n^k/k!$.*

Theorem 4.2 and Proposition 4.3 give $m^{\mathcal{C}}(n) \le 1.5n^k/k!$ for $n \ge k+2$ where $k := S(\mathcal{C})$.

To see that Theorem 4.2 is sharp, let X be an infinite set and $\mathcal{C}$ the collection of all subsets of X with cardinality k. Then $S(\mathcal{C}) = k$ and the inequality in the second sentence of the theorem becomes an equality for all n.

Let dens($\mathcal{C}$) be defined, following P. Assouad (1983), as

$$\inf\{r > 0 : \text{ for some } K < \infty,\ m^{\mathcal{C}}(n) \le Kn^r \text{ for all } n \ge 1\}.$$

Then we have

Corollary 4.4. *For any set X and $\mathcal{C} \subset 2^X$, dens($\mathcal{C}$) $\le S(\mathcal{C})$. Conversely if dens($\mathcal{C}$) $< \infty$ then $S(\mathcal{C}) < \infty$.*

Note that $S(\mathcal{C})$ can be determined by one large shattered set while dens($\mathcal{C}$) has to do with the behavior of $\mathcal{C}$ on arbitrarily large finite sets. For example if X is a set with card$(X) = n$ and $\mathcal{C} = 2^X$ then $S(\mathcal{C}) = n$ while dens($\mathcal{C}$) $= 0$.

For any set X, it is immediate that if $\mathcal{C} \subset \mathcal{D} \subset 2^X$ then $S(\mathcal{C}) \le S(\mathcal{D})$ and dens($\mathcal{C}$) $\le$ dens($\mathcal{D}$).

The following is straightforward since for any set X, the map $A \mapsto X \setminus A$ is one-to-one from 2^X onto itself, and for any $A, B, C \subset X$, $A \cap B \ne C \cap B$ if and only if $(X \setminus A) \cap B \ne (X \setminus C) \cap B$:

Proposition 4.5. *If X is any set, $\mathcal{C} \subset 2^X$ and $\mathcal{D} := \{X \setminus A : A \in \mathcal{C}\}$ then for all $B \subset X$, $\Delta^{\mathcal{C}}(B) = \Delta^{\mathcal{D}}(B)$, so $m^{\mathcal{C}}(n) = m^{\mathcal{D}}(n)$ for all n, $S(\mathcal{D}) = S(\mathcal{C})$ and dens($\mathcal{D}$) =dens($\mathcal{C}$).*

4.2 Generating Vapnik-Červonenkis classes

Let's begin with some examples of *non*-VC classes for which some limit theorems for empirical measures will fail.

First, let $X = [0, 1]$ and let $\mathcal{C}$ be the class of all finite subsets of X. Let P be the uniform (Lebesgue) law on $[0, 1]$. Clearly, $S(\mathcal{C}) = +\infty$, and $\mathcal{C}$ is not a VC class. Also, for any possible

value of P_n, we will have $P_n(A) = 1$ for some $A = \{X_1, \cdots, X_n\} \in \mathcal{C}$ while $P(A) = 0$. Thus $\sup_{A \in \mathcal{C}} (P_n - P)(A) = 1$ for all n, so $\mathcal{C}$ is not a Glivenko-Cantelli class for P, in other words

$$\|P_n - P\|_{\mathcal{C}} := \sup_{A \in C} |(P_n - P)(A)|$$

doesn't approach 0 as $n \to \infty$ in any sense, e.g. in outer probability, since it is identically 1. It follows that $\mathcal{C}$ is also not a Donsker class for P.

Note that all functions 1_A for $A \in \mathcal{C}$ equal 0 almost surely for P. Thus, the whole class $\mathcal{F} := \{1_A : A \in \mathcal{C}\}$ reduces to the one point 0 in the space $L^2(P)$ of equivalence classes for equality almost everywhere of functions in $\mathcal{L}^2(P)$, that is, measurable, square-integrable functions. Thus for purposes of empirical processes, functions equal a.s. P are not the same and we need to deal with classes $\mathcal{F} \subset \mathcal{L}^2(P)$ of actual real-valued functions, not equivalence classes. Then, the integral $\int f d(P_n - P)$ will be well-defined for any $f \in \mathcal{L}^2(P)$. This integral is linear in f and thus prelinear for $f \in \mathcal{F}$ for any set $\mathcal{F} \subset \mathcal{L}^2(P)$. For the empirical process $\nu_n = n^{1/2}(P_n - P)$ we will not be taking versions or modifications as was done for Gaussian processes (Appendix I).

Next, let $\mathcal{C}_2$ be the collection of all closed, convex subsets of $\mathbb{R}^2$. Let S^1 be the unit circle $\{(x, y) : x^2 + y^2 = 1\}$. For any finite subset F of S^1, the convex polygon with vertices in F (a singleton if $|F| = 1$, or a line segment if $|F| = 2$) is in $\mathcal{C}_2$ and its intersection with S^1 is F. Thus $S(\mathcal{C}) = +\infty$ and $\mathcal{C}$ is not a VC class. Let P be the uniform law $dP(\theta) = 2\theta/(2\pi)$ on S^1. Then the Glivenko-Cantelli and Donsker properties fail for P just as in the previous example.

Classes with $S(\mathcal{C})$ finite, in other words Vapnik-Červonenkis classes, can be formed in various ways. Here is one. Let G be a collection of real-valued functions on a set X. Let

$$\mathrm{pos}(g) := \{x : g(x) > 0\}, \ \mathrm{nn}(g) := \{x : g(x) \geq 0\}, \ g \in G,$$

$$\mathrm{pos}(G) := \{\mathrm{pos}(g) : g \in G\}, \ \mathrm{nn}(G) := \{\mathrm{nn}(g) : g \in G\},$$

$$U(G) := \mathrm{pos}(G) \cup \mathrm{nn}(G).$$

Theorem 4.6. *Let H be an m-dimensional real vector space of functions on a set X, f any real function on X, and $H_1 := \{f + h : h \in H\}$. Then $S(pos(H_1)) = S(nn(H_1)) = m$. If H contains the constants then also $S(U(H_1)) = m$.*

Examples. (I) Let $H := \mathcal{P}_{d,k}$ be the space of all polynomials of degree at most k on $\mathbb{R}^d$. Then for each d and k, H is a finite-dimensional vector space of functions, so $\mathrm{pos}(H)$ is a Vapnik-Červonenkis class. For $k = 2$, it follows specifically that the set of all ellipsoids in $\mathbb{R}^d$ is included in a Vapnik-Červonenkis class and thus is one.
(II) Let $X = \mathbb{R}$. Let H be the 1-dimensional space of linear functions $f(x) = cx$, $x \in \mathbb{R}$, $c \in \mathbb{R}$. Then $S(\mathrm{pos}(H)) = S(\mathrm{nn}(H)) = 1$ by Theorem 4.6, but $U(H)$ shatters $\{0, 1\}$. Since sets in $U(H)$ are convex (half-lines), it follows that $S(U(H)) = 2$. So the condition that H contains the constants can't just be dropped from Theorem 4.6 for $U(H)$.

Let X be a real vector space of dimension m. Let H be the space of all real *affine* functions on X, in other words functions of the form $h + c$ where h is real linear and c is any real constant. Then H has dimension $m+1$ and $\mathrm{pos}(H)$ is the set of all open half-spaces of X. Letting $f = 0$ in Theorem 4.6 for this H gives a special case known as Radon's Theorem. On the other hand, Theorem 4.6 for $f = 0$ with general X and H follows from Radon's theorem via the following stability fact.

Theorem 4.7. *If X and Y are sets, F is a function from X into Y, $\mathcal{C} \subset 2^Y$, and $F^{-1}(\mathcal{C}) := \{F^{-1}(A) : A \in \mathcal{C}\}$, then $S(F^{-1}(\mathcal{C})) \le S(\mathcal{C})$. If F is onto Y then $S(F^{-1}(\mathcal{C})) = S(\mathcal{C})$.*

Now let X be any set and G a finite-dimensional real vector space of real functions on X. Then there is a natural map $F : x \mapsto \delta_x$ from X into the space of linear functions on G. Then by Theorem 4.7 one could deduce Theorem 4.6 from its special case where X is an m- or $(m+1)$-dimensional real vector space and f and all functions in H are affine, so that sets in $\text{pos}(H_1)$ are open half-spaces.

Next it will be seen that a bounded number of Boolean operations preserves the Vapnik-Červonenkis property.

Theorem 4.8. *Let X be a set, $\mathcal{C} \subset 2^X$, and for $k = 1, 2, ...$, let $\mathcal{C}^{(k)}$ be the union of all (Boolean) algebras generated by k or fewer elements of $\mathcal{C}$. Then $dens(\mathcal{C}^{(k)}) \le k \cdot dens(\mathcal{C})$, so if $S(\mathcal{C}) < \infty$ then $S(\mathcal{C}^{(k)}) < \infty$.*

Note that a Boolean algebra generated by k sets can have as many as 2^{2^k} elements, and 2^{2^k} is very large if k is at all large.

Vapnik-Červonenkis classes can be generated by combining Theorems 4.6 and 4.8. For example, half-spaces in $\mathbb{R}^d$ form a VC class. Intersections of at most k half-spaces give convex polytopes with at most k faces, so these form a VC class.

Remarks. Let X be an infinite set, $r = 1, 2, ...$, and $\mathcal{C}_r$ the collection of all subsets of X with at most r elements. Then clearly $\text{dens}(\mathcal{C}_r) = S(\mathcal{C}_r) = r$. It's easy to check that $\mathcal{D} := \mathcal{C}_r^{(k)}$ consists of all sets B such that either B or $X \setminus B$ has at most kr elements. Thus $m^{\mathcal{D}}(n) \le 2({}_nC_{\le kr})$, with $m^{\mathcal{D}}(n) = 2({}_nC_{\le kr})$ for $n \ge 2kr + 1$. So $\text{dens}(\mathcal{D}) = kr$ since ${}_nC_{\le j}$ is a polynomial in n of degree j. Thus the inequality $\text{dens}(\mathcal{C}^{(k)}) \le k \cdot \text{dens}(\mathcal{C})$ is sharp. But it does not always hold for $S(\cdot)$ in place of $\text{dens}(\cdot)$: if $\mathcal{C}$ is the collection of open half-spaces in $\mathbb{R}^d$, $d \ge 1$, then $S(\mathcal{C}) = d + 1$ by Radon's theorem, while, taking for example the d half-spaces $\{x_j > 0\}$ for $j = 1, \cdots, d$, we see that $\mathcal{C}^{(d)}$ shatters a set of 2^d points, one in each coordinate orthant, so $S(\mathcal{C}^{(d)}) \ge 2^d > d(d+1)$ for $d \ge 5$.

Classes with $V(\mathcal{C}) = 0$ or 1 are easily characterized:

Proposition 4.9. *A class $\mathcal{C}$ of subsets of a set X has $V(\mathcal{C}) = 0$, or equivalently $S(\mathcal{C}) = -1$, if and only if $\mathcal{C}$ is empty. Also, $V(\mathcal{C}) = 1$, or equivalently $S(\mathcal{C}) = 0$, if and only if $\mathcal{C}$ contains exactly one set. Thus $S(\mathcal{C}) \ge 1$ if and only if $\mathcal{C}$ contains at least two sets.*

Here are two sufficient conditions for $S(\mathcal{C}) = 1$:

Theorem 4.10. *If $\mathcal{C}$ is a collection of at least two subsets of a set X, then $S(\mathcal{C}) = 1$ if either*
(a) $\mathcal{C}$ is linearly ordered by inclusion, or
(b) Any two sets in $\mathcal{C}$ are disjoint.

Section 4.4 will go more into detail about classes of index 1.

4.3 Maximal classes

Let $\mathcal{C} \subset \mathcal{A}$ be classes of subsets of a set X. Then $\mathcal{C}$ will be called $(\mathcal{A}, n)$-*maximal* if $S(\mathcal{C}) = n$ and if $\mathcal{C} \subset \mathcal{D}$ strictly and $\mathcal{D} \subset \mathcal{A}$, then $S(\mathcal{D}) > n$. If $\mathcal{A}$ is the class 2^X of all subsets of X, then

$\mathcal{C}$ will be called *n-maximal*. If $\mathcal{C}$ is n-maximal, then clearly $\mathcal{C}$ is $(\mathcal{A}, n)$-maximal for any $\mathcal{A}$ such that $\mathcal{C} \subset \mathcal{A} \subset 2^X$.

In view of Proposition 4.9, classes $\mathcal{C}$ with $S(\mathcal{C}) = i,\ i = -1$ or 0, are empty or contain just one set respectively, and so are always i-maximal. Thus n-maximality only becomes interesting for $n \geq 1$.

Examples. 1. For any set X, let $\mathcal{C}$ consist of $\emptyset$ (the empty set) and all singletons $\{x\}$ for $x \in X$. Then $\mathcal{C}$ is clearly 1-maximal.
2. Let $X = \mathbb{R}$. Let $\mathcal{LH}$ consist of $\emptyset$, $\mathbb{R}$, and all left half-lines, closed $(-\infty, x]$ or open $(-\infty, x)$, for $x \in \mathbb{R}$. In other words, $\mathcal{LH}$ is the collection of all subsets $A \subset \mathbb{R}$ such that whenever $x < y \in A$ then also $x \in A$. Then clearly $S(\mathcal{LH}) = 1$ since for $x < y$ and $A \in \mathcal{LH}$, $A \cap \{x, y\} \neq \{y\}$. On the other hand if any subset of $\mathbb{R}$ not in $\mathcal{LH}$ is adjoined, then some 2-element set is shattered, so $\mathcal{LH}$ is 1-maximal.
3. Let $X = \mathbb{R}$ and let Co consist of all subintervals of $\mathbb{R}$, namely $\emptyset$, $\mathbb{R}$, any closed or open, left or right half-line, and any bounded interval, open or closed at either end. In other words Co is the class of all convex subsets of $\mathbb{R}$. Then $S(Co) = 2$, in fact Co shatters every 2-element subset of $\mathbb{R}$, while if $x < y < z$ and $A \in Co$, then $A \cap \{x, y, z\} \neq \{x, z\}$. On the other hand if any set not in Co is adjoined to it, its index becomes 3, so Co is 2-maximal.

Here is an existence theorem for maximal classes, easily provable with Zorn's lemma.

Theorem 4.11. *Let X be a set and $\mathcal{D} \subset 2^X$. Suppose that $\mathcal{C} \subset \mathcal{D}$ and $S(\mathcal{C}) = n$. Then there exists a $(\mathcal{D}, n)$-maximal class $\mathcal{B}$ with $\mathcal{C} \subset \mathcal{B}$.*

The following fact is straightforward:

Proposition 4.12. *For any set X, $Y \subset X$, $\mathcal{C} \subset 2^X$, and $\mathcal{C}_Y := \mathcal{C} \sqcap Y$, we have $S(\mathcal{C}_Y) \leq S(\mathcal{C})$.*

Recall that $\mathbb{Z}_2 := \{0, 1\}$ with addition mod 2, in other words the usual addition except that $1 + 1 = 0$. For any set X, the group $\mathbb{Z}_2^X$ of all functions from X into $\mathbb{Z}_2$, with the natural addition $(f + g)(x) := f(x) + g(x)$ in $\mathbb{Z}_2$, provides a group structure for the collection 2^X of all subsets of X. Addition of indicator functions mod 2 corresponds to the symmetric difference $A \Delta B := (A \setminus B) \cup (B \setminus A)$, so that $1_A + 1_B = 1_{A \Delta B}$ mod 2. For any fixed set $A \subset X$, the translation $1_B \mapsto 1_A + 1_B$ takes $\mathbb{Z}_2^X$ one-to-one and onto itself. If the functions are restricted to a subset $Y \subset X$, translation still takes $\mathbb{Z}_2^Y$ one-to-one and onto itself. For any $A \subset X$ and $\mathcal{C} \subset 2^X$, let $A \Delta \Delta \mathcal{C} := \{A \Delta C :\ C \in \mathcal{C}\}$. Then for any finite $F \subset X$, $\mathcal{C}$ shatters F if and only if $A \Delta \Delta \mathcal{C}$ does. It follows that:

Proposition 4.13. *For any fixed set $A \subset X$ and class $\mathcal{C} \subset 2^X$, $S(\mathcal{C}) = S(A \Delta \Delta \mathcal{C})$. Also, $\mathcal{C}$ is n-maximal if and only if $A \Delta \Delta \mathcal{C}$ is.*

Next, we have:

Proposition 4.14. *If $\mathcal{C}$ is an n-maximal class of subsets of a set X, and $n \geq 1$, then $\bigcup_{A \in \mathcal{C}} A = X$ and $\bigcap_{A \in \mathcal{C}} A = \emptyset$.*

On $\mathbb{Z}_2^X = 2^X$ there is a product topology coming from the discrete topology on $\mathbb{Z}_2$. The product topology is compact by Tychonoff's theorem (RAP, Theorem 2.2.8).

Proposition 4.15. *For any set X, any n-maximal class $\mathcal{C} \subset 2^X$ is closed and so compact in 2^X.*

A class $\mathcal{C}$ of subsets of a set X will be called *complemented* if $X \setminus A \in \mathcal{C}$ for every $A \in \mathcal{C}$.

Theorem 4.16. *If $S(\mathcal{C}) = n$, $\mathcal{C} \subset \mathcal{A}$ strictly, and $\mathcal{C}$ is complemented, then $\mathcal{C}$ is not $(\mathcal{A}, n)$-maximal.*

If $\mathcal{F}$ is a k-dimensional real vector space of real-valued functions on a set X containing the constants and $\mathcal{C}$ is the collection $U(\mathcal{F})$ of all sets $\{x : f(x) > 0\}$ or $\{x : f(x) \geq 0\}$ for all $f \in \mathcal{F}$ and real t, then $S(\mathcal{C}) = k$ by Theorem 4.6. Since $\mathcal{C}$ is complemented, it is never k-maximal.

Let X be any set and $\mathcal{C} := \mathcal{C}_k$ the collection of all subsets of X with at most k elements. Then clearly $S(\mathcal{C}) = k$. Also, $\mathcal{C}$ is k-maximal since if $A \notin \mathcal{C}$, $A \subset X$, then $|A| > k$ and if B is any subset of A with $|B| = k + 1$ then B is shattered by $\mathcal{C} \cup \{A\}$. For $\mathcal{C} = \mathcal{C}_k$ we have $m^{\mathcal{C}}(n) \equiv {}_nC_{\leq k}$, which is the maximum possible value of $m^{\mathcal{C}}(n)$ by Sauer's Lemma (Theorem 4.2). The following example shows that not all k-maximal classes have these values of $m^{\mathcal{C}}(n)$:

Example. Let $X = \{1, 2, 3, 4\}$, $\mathcal{G} = \{\{4\}, \{1, 3\}, \{2, 3\}, \{3, 4\}, \{1, 2, 3\}, \{1, 2, 3, 4\}\}$. Let $\mathcal{C}$ be the complement of $\mathcal{G}$ in 2^X. Then it can be checked that $\mathcal{C}$ is 2-maximal but $|\mathcal{C}| = 10 < {}_4C_{\leq 2} = 11$.

4.4 Classes of index 1

In this section the structure of classes $\mathcal{C}$ with $S(\mathcal{C}) = 1$ will be treated. Recall for classes of two or more sets, that disjoint classes and classes linearly ordered by inclusion have $S(\mathcal{C}) = 1$ (Theorem 4.10). A common extension of these two kinds of classes is given by treelike partial orderings, defined as follows.

A binary relation $\leq$ on a set X will be called a *quasi-order* if it is transitive: $x \leq y$ and $y \leq z$ imply $x \leq z$, and reflexive: $x \leq x$ for all $x \in X$. The quasi-order is called a *partial order* if also $x \leq y$ and $y \leq x$ imply $x = y$. For any set S, inclusion ($\subset$) is a partial order on 2^S or any subset of 2^S.

Let $\leq$ be a quasi-order on a set X. Then two elements x and y of X are called *comparable* if at least one of $x \leq y$ and $y \leq x$ holds, or *incomparable* if neither holds. A quasi-order $\leq$ on X will be called *fully comparable* if any two elements of X are comparable. A fully comparable partial order is called *linear*. A quasi-order $\leq$ will be called *sub-fully comparable* if for any $y \in X$ and $L_y := \{x : x \leq y\}$, the restriction of $\leq$ to L_y is fully comparable. A sub-fully comparable partial order will be called *treelike*.

Theorem 4.17. *Let $\mathcal{C} \subset 2^X$ contain at least two sets and satisfy, for any $x \neq y$ in X,*

$$A \cap \{x, y\} = \emptyset \text{ for some } A \in \mathcal{C}. \tag{4.1}$$

Then the following are equivalent:
(a) $S(\mathcal{C}) = 1$;
(b) For every $Y \subset X$, the inclusion partial ordering of $\mathcal{C}_Y := Y \sqcap \mathcal{C}$ is treelike;
(c) For every $Y \subset X$ with $|Y| = 2$, the partial ordering of $\mathcal{C}_Y := Y \sqcap \mathcal{C}$ by inclusion is treelike.

Proposition 4.18. *Let X be a set and $\mathcal{A} \subset 2^X$ where $\emptyset \in \mathcal{A}$ and for any B and C in $\mathcal{A}$, $B \cap C \in \mathcal{A}$. If $\mathcal{C}$ is $(\mathcal{A}, 1)$-maximal and satisfies (4.1) for any $x \neq y$ in X, then $\emptyset \in \mathcal{C}$ and $B \cap C \in \mathcal{C}$ for any B and C in $\mathcal{C}$.*

Proposition 4.19. *Let X be a set and $\mathcal{C}$ a finite class of subsets of X with $S(\mathcal{C}) = 1$ such that for any $x \neq y$ in X, (4.1) holds. Let $\mathcal{D} := \mathcal{D}(\mathcal{C})$ consist of $\emptyset$ and all intersections of nonempty subclasses of $\mathcal{C}$. Then $S(\mathcal{D}) = 1$. For each non-empty set $D \in \mathcal{D}$ there is a $C := C(D) \in \mathcal{D}$ such that $C \subset D$ strictly ($C \neq D$) and if B is any set in $\mathcal{D}$ with $B \subset D$ strictly, then $B \subset C$.*

Proposition 4.20. *Under the hypotheses of Proposition 4.19, the sets $D \setminus C(D)$ for distinct non-empty $D \in \mathcal{D}$ are all disjoint and are nonempty.*

A *graph* is a nonempty set S together with a set E of unordered pairs $\{x, y\}$ for some $x \neq y$ in S. Then S will be called the set of *nodes* and E the set of *edges* of the graph. The graph (S, E) is called a *tree* if
(a) it is connected, in other words, for any x and y in S there is a finite n and $x_i \in S,\ i = 0, 1, ..., n$, such that $x_0 = x,\ x_n = y$, and $\{x_{k-1}, x_k\} \in E$ for $k = 1, ..., n$.
(b) the graph is *acyclic*, which means that there is no cycle, where a *cycle* is a set of distinct $x_1, ..., x_n \in S$ such that $n \geq 3$, and letting $x_0 := x_n$, $\{x_{k-1}, x_k\} \in E$ for $k = 1, ..., n$.

Theorem 4.21. *(a) For m nodes, for any positive integer m, there exist connected graphs with $m - 1$ edges.*
(b) A connected graph with m nodes cannot have fewer than $m - 1$ edges.
(c) A connected graph with m nodes has exactly $m - 1$ edges if and only if it is a tree.

Let the class $\mathcal{D}$ in Propositions 4.19 and 4.20 form the nodes of a graph G whose edges are the pairs $\{C(D), D\}$ for $D \in \mathcal{D},\ D \neq \emptyset$.

Proposition 4.22. *The graph G is a tree.*

Proposition 4.23. *Let X be a finite set. Let $\mathcal{C}$ be 1-maximal in X and suppose (4.1) holds for all $x \neq y$ in X. Then $\mathcal{C} = \mathcal{D}(\mathcal{C})$ as defined in Proposition 4.19. The sets $D \setminus C(D)$ for non-empty $D \in \mathcal{C}$ are all the singletons $\{x\},\ x \in X$. If $|X| = m$ then $|\mathcal{C}| = m + 1$.*

Suppose in this paragraph that $\mathcal{C}$ is a class of two or more sets such that (4.1) holds with $\emptyset$ replaced by $\{x, y\}$. Then the class of complements, $\mathcal{N} := \{X \setminus C :\ C \in \mathcal{C}\}$ satisfies the original hypotheses of Theorem 4.17. If $\mathcal{C}$ is 1-maximal, so is $\mathcal{N}$ by Proposition 4.13. So Theorem 4.17 and Propositions 4.18 through 4.22 apply to $\mathcal{N}$, and so does Proposition 4.23 if X is finite. Then, $\mathcal{C}$ itself has a "cotreelike" ordering, where for each $C \in \mathcal{C}$, $\{D \in \mathcal{C} :\ C \subset D\}$ is linearly ordered by inclusion. Propositions 4.18 and 4.19 apply to $\mathcal{C}$ if $\emptyset$ is replaced by X and intersections by unions; in Proposition 4.19, we will have an immediate successor $D(C) \supset C$ instead of a predecessor; and sets $D(C) \setminus C$ instead of $D \setminus C(D)$ in Propositions 4.20 and 4.23. The resulting tree (Proposition 4.22) then branches out as sets become smaller rather than larger.

Next will be several facts in the general case, i.e. without the hypothesis (4.1).

Theorem 4.24. *Let X be any set and $\mathcal{C}$ any collection of subsets with $S(\mathcal{C}) = 1$. Then for any $C \in \mathcal{C}$, the collection $\mathcal{C}_{X \setminus C} := \{B \setminus C :\ B \in \mathcal{C}\}$ satisfies (4.1) for any $x \neq y$ as a collection of subsets of $X \setminus C$. Likewise, $\mathcal{C}_{C \setminus} := \{C \setminus B :\ B \in \mathcal{C}\}$ satisfies (4.1) for any $x \neq y$ as a collection of subsets of C, $S(\mathcal{C}_{C \setminus}) \leq 1$ and $S(\mathcal{C}_{X \setminus C}) \leq 1$.*

So, for an arbitrary class $\mathcal{C}$ with $S(\mathcal{C}) = 1$, we have by Theorem 4.17 a treelike inclusion partial ordering in one part $X \setminus C$ of X and a cotreelike ordering in the complementary part C, for any $C \in \mathcal{C}$. If also $X \setminus C$ happens to be in $\mathcal{C}$, both orderings are linear. To see how the two orderings fit together in general, Proposition 4.13 gives:

Corollary 4.25. *Let $\mathcal{C}$ be any class of sets with $S(\mathcal{C}) = 1$ and $A \in \mathcal{C}$. Let $\mathcal{D} := A\Delta\Delta\mathcal{C}$. Then $S(\mathcal{D}) = 1$ and $\emptyset \in \mathcal{D}$. If $\mathcal{C}$ is 1-maximal, so is $\mathcal{D}$. Then Theorem 4.17, Proposition 4.18, and if $\mathcal{C}$ is finite, Propositions 4.19, 4.20, 4.22, and if X is finite, 4.23, apply to $\mathcal{D}$.*

The last sentence in Proposition 4.23 has a converse and extension:

Proposition 4.26. *Let X be finite with m elements and $\mathcal{C} \subset 2^X$ with $S(\mathcal{C}) = 1$. Then $\mathcal{C}$ is 1-maximal if and only if $|\mathcal{C}| = m + 1$.*

Now, $m + 1 = {}_mC_{\leq 1}$, which is the maximum value of $m^{\mathcal{C}}(m)$ for $S(\mathcal{C}) = 1$ by Sauer's Lemma (Theorem 4.2). The example at the end of Section 4.3 shows that Proposition 4.26 in the form $|\mathcal{C}| = {}_mC_{\leq k}$, $k = 1$, does not extend to k-maximality for $k > 1$.

Next it will be seen that 1-maximality can be relativized to subsets. For a set X, a subset $Y \subset X$, and a class $\mathcal{C} \subset 2^X$, recall that $\mathcal{C}_Y := \mathcal{C} \sqcap Y := \{A \cap Y : A \in \mathcal{C}\}$.

Theorem 4.27. *If $\mathcal{C}$ is 1-maximal and $\emptyset \neq Y \subset X$, then $\mathcal{C}_Y$ is a 1-maximal class of subsets of Y.*

For any set X and $\mathcal{C} \subset 2^X$, let $x \leq_{\mathcal{C}} y$ iff $x = y$ or $y \in \bigcup_{B \in \mathcal{C}} B$ and for all $A \in \mathcal{C}$, $y \in A$ implies $x \in A$. Then $\leq_{\mathcal{C}}$ is a quasi-order (as defined early in this section) but in general not a partial order. The treelike partial orderings as in Theorem 4.17 were on collections of sets. Now orderings will be defined on X.

Theorem 4.28. *If $S(\mathcal{C}) = 1$, $\bigcup_{B \in \mathcal{C}} B = X$, and $\mathcal{C}$ satisfies (4.1) for all $x \neq y$, then $\leq_{\mathcal{C}}$ is a treelike quasi-order on X. If $\mathcal{C}$ is 1-maximal, then $\leq_{\mathcal{C}}$ is a partial order. Conversely, for any quasi-order $\leq$ on a set X, let $\mathcal{C} := \mathcal{C}_{\leq} := \{A \subset X : A$ is linearly quasi-ordered by $\leq$ and $x \in A$ whenever $x \leq y \in A\}$. Then $S(\mathcal{C}) \leq 1$. If $\leq$ is a treelike partial order and $X \neq \emptyset$, then $\mathcal{C}$ is 1-maximal.*

Example. Let $X = \{1, 2, 3, 4, 5\}$ and

$$\mathcal{G} = \{\emptyset, \{1\}, \{5\}, \{2, 5\}, \{1, 2, 4\}, \{1, 2, 3, 4\}, \{1, 2, 3, 5\}, \{1, 2, 3, 4, 5\}\}.$$

Let $\mathcal{C}$ be the complement of $\mathcal{G}$ in 2^X. Then it can be checked that $\mathcal{C}$ is 3-maximal but for $Y = \{1, 2, 3, 4\}$, $\mathcal{C}_Y$ is not 3-maximal in Y. So in Theorem 4.27, "1-maximal" and $Y \neq \emptyset$ cannot be replaced by "3-maximal" and "$|Y| \geq 3$" respectively.

The next theorem is not in UCLT. Recall that a linearly ordered subset of a partially ordered set is called a *chain*.

Theorem 4.29. *Let $\mathcal{C}$ be a 1-maximal class of subsets of a set X containing $\emptyset$.*
(I) Then $B \in \mathcal{C}$ if and only if both (a) B is a chain for $\leq_{\mathcal{C}}$ and (b) if $x \leq_{\mathcal{C}} y \in B$ then $x \in B$.
(II) If X is finite, $B \in \mathcal{C}$ if and only if $B = \emptyset$ or for some $z \in X$, $B = \{x : x \leq_{\mathcal{C}} z\}$.

Proof. To prove "only if" in (I), (b) holds by definition of $\leq_{\mathcal{C}}$. To prove (a), supppose $x, y \in B$ are not comparable for $\leq_{\mathcal{C}}$. By Theorem 4.27 applied to singletons Y, we have $X = \cup_{C \in \mathcal{C}} C$. Thus for some $D \in \mathcal{C}$, $y \in D$ and $x \notin D$, and for some $E \in \mathcal{C}$, $x \in E$ and $y \notin E$. So $\mathcal{C} \supset \{\emptyset, D, E, B\}$ shatters $\{x, y\}$, a contradiction. Thus (a) holds.

Conversely, suppose (a) and (b) hold. Suppose $\mathcal{C} \cup \{B\}$ shatters some $\{x, y\}$. If $x \leq_{\mathcal{C}} y$ then $C \cap \{x, y\} \neq \{y\}$ for $C \in \mathcal{C}$ or $C = B$, a contradiction. So x and y are not comparable

for $\leq_{\mathcal{C}}$. Then $\mathcal{C} \sqcap \{x, y\}$ contains $\emptyset$, $\{x\}$ and $\{y\}$, and so not $\{x, y\}$. Also $B \cap \{x, y\} \neq \{x, y\}$, giving another contradiction. So $S(C \cup \{B\}) = 1$ and since $\mathcal{C}$ is 1-maximal, $B \in \mathcal{C}$, proving "if."

For (II), a B of the given form satisfies (b) clearly, and (a) holds because $\leq_{\mathcal{C}}$ is treelike by Theorem 4.28, so $B \in \mathcal{C}$ by part (I). Conversely, if $B \in \mathcal{C}$ it is a chain for $\leq_{\mathcal{C}}$ by (I) so if it is non-empty it has a largest element z and then $B = \{x : \ x \leq_{\mathcal{C}} z\}$ by (a) and (b). □

4.5 Combining VC classes

Recalling the density as in Corollary 4.4, the following is clear:

Theorem 4.30. *For any set X, if $\mathcal{A} \subset 2^X$ and $\mathcal{C} \subset 2^X$ then*
$$dens(\mathcal{A} \cup \mathcal{C}) = \max(dens(\mathcal{A}), dens(\mathcal{C})).$$

For the Vapnik-Červonenkis index we have instead:

Proposition 4.31. *For any set X, $\mathcal{A} \subset 2^X$ and $\mathcal{C} \subset 2^X$, $S(\mathcal{A} \cup \mathcal{C}) \leq S(\mathcal{A}) + S(\mathcal{C}) + 1$. This bound is best possible: for any nonnegative integers k and m there exist X, $\mathcal{A}$ and $\mathcal{C} \subset 2^X$ with $S(\mathcal{A}) = k$, $S(\mathcal{C}) = m$ and $S(\mathcal{A} \cup \mathcal{C}) = k + m + 1$.*

Let X be a set and $\mathcal{C}, \mathcal{D}$ any two collections of subsets of X. Let

$$\mathcal{C} \sqcap \mathcal{D} \ := \ \{C \cap D : \ C \in \mathcal{C}, \ D \in \mathcal{D}\}, \quad \mathcal{C} \sqcup \mathcal{D} \ := \ \{C \cup D : \ C \in \mathcal{C}, \ D \in \mathcal{D}\}.$$

If $\mathcal{A}$ is a class of subsets of another set Y let

$$\mathcal{C} \boxtimes \mathcal{A} \ := \ \{C \times A : \ C \in \mathcal{C}, \ A \in \mathcal{A}\}.$$

Theorem 4.32. *For any $\mathcal{C} \subset 2^X$ and $\mathcal{D} \subset 2^X$ or 2^Y let $k := dens(\mathcal{C})$ and $m :=dens(\mathcal{D})$. Then we have: $dens(\mathcal{C} \boxdot \mathcal{D}) \leq k + m$ for $\boxdot = \sqcap, \ \sqcup$ or $\boxtimes$.*

For the Vapnik-Červonenkis index the behavior of the $\boxdot$ operations is not so simple. For $k, m = 0, 1, 2, ...,$ and $\boxdot = \sqcup, \sqcap$ or $\boxtimes$ let $\boxdot(k, m) := \max\{S(\mathcal{C} \boxdot \mathcal{D}) : \ S(\mathcal{C}) = k, \ S(\mathcal{D}) = m\}$. Here the maximum is taken where X and Y may be infinite sets. Then we have:

Theorem 4.33. *For any $k = 0, 1, 2, \cdots$ and $m = 0, 1, 2, \cdots$, and $\boxdot = \sqcup, \ \sqcap$ or $\boxtimes$, we have $\boxdot(k, m) < \infty$.*

Theorem 4.34. *For any $k, m = 0, 1, 2, ...$, $\sqcap(k, m) = \sqcup(k, m) = \boxtimes(k, m)$.*

Let $S(k, m)$ be the common value of the quantities in Theorem 4.34. Theorem 4.33 can be improved as follows. For any nonnegative integers j, k let $\theta(j, k) := \sup\{r \in \mathbb{N} : ({}_rC_{\leq j})({}_rC_{\leq k}) \geq 2^r\}$. Then $\theta(j, k) < \infty$ for each j, k by Proposition 4.3 and we have:

Proposition 4.35. *$S(j, k) \leq \theta(j, k)$ for any $j, k \in \mathbb{N}$.*

Can the values $S(k, m)$ be computed? The next two theorems and proposition (not in UCLT) will give some information.

Theorem 4.36. *Let X be a set, $\mathcal{C}, \mathcal{D} \subset 2^X$, and $\mathcal{C} \sqcup \mathcal{D} = 2^X$. Let $A \subset X$ and suppose for all $B \in \mathcal{C}$, either $B \subset A$ or $B \subset A^c$. Then $\mathcal{D}$ shatters either A or A^c.*

Proof. Supppose $\mathcal{D}$ doesn't shatter A. Then take $H \subset A$ such that $D \cap A \neq H$ for all $D \in \mathcal{D}$. Take any $E \subset A^c$. Then $E \cup H = C \cup D$ for some $C \in \mathcal{C}$ and $D \in \mathcal{D}$. If $C \subset A^c$ then $D \cap A = H$, a contradiction. So $C \subset A$ and $D \cap A^c = E$. Thus $\mathcal{D}$ shatters A^c. □

For any set Y recall that $|Y|$ denotes the number of elements of Y. Here is an upper bound for $S(1,k)$ that will turn out to be exact for $k = 1, 2$.

Theorem 4.37. *For any $k = 1, 2, \cdots$, $S(1,k) \leq 2k+1$.*

Proof. Suppose $|X| = 2k+2$, $\mathcal{C} \sqcup \mathcal{D} = 2^X$, $S(\mathcal{C}) = 1$, and $S(\mathcal{D}) = k$. We can assume by Theorem 4.11 that $\mathcal{C}$ is 1-maximal. Then $\cup_{B \in \mathcal{C}} B = X$ by Theorem 4.27 applied to singletons Y. We have $\emptyset \in \mathcal{C} \cap \mathcal{D}$. Thus by Theorem 4.17, $\mathcal{C}$ has a treelike partial ordering by inclusion, which induces such a treelike partial ordering on X by Theorem 4.28. Let Y be the set of elements of X having at least one predecessor for this ordering. Each $y \in Y$ has a smallest predecessor $f(y) \notin Y$. For each $B \subset Y$, we have $B = C \cup D$, $C \in \mathcal{C}$, $D \in \mathcal{D}$, where $C = \emptyset$, $B = D$ since if $y \in C \cap Y$, $f(y) \in C \setminus Y$. So $\mathcal{D}$ shatters Y and Y has at most k elements.

Let r be the number of values of f, say $t_1, \cdots, t_r$. Then Y is decomposed into disjoint subsets $Y_1, \cdots, Y_r$ such that $f = t_j$ on Y_j for each j. Let $C := (Y \cup \mathrm{ran} f)^c$. Then $|C| \geq 2$. Let $n_j := |Y_j|$, $j = 1, \cdots, r$. Then

$$2k+2 = |C| + \textstyle\sum_{j=1}^{r} (n_j + 1). \tag{4.2}$$

It will be shown that there exist subsets $E \subset C$ and $I \subset \{1, \cdots, r\}$ such that

$$|E| + \textstyle\sum_{j \in I} (n_j + 1) = k+1. \tag{4.3}$$

Let K be the largest possible value $\leq k+1$ of the left side of (4.3). Suppose $K \leq k$. Then

$$K = |C| + \textstyle\sum_{j \in J} (n_j + 1) \tag{4.4}$$

for some $J \subset \{1, \cdots, r\}$ since elements of C could be put into E one at a time. We then have by (4.2)

$$\textstyle\sum_{j \notin J} (n_j + 1) = 2k+2-K \geq k+2. \tag{4.5}$$

Let n_0 be the smallest value of n_j for $j \notin J$. Then $n_0 \geq |C| + 1$, or another j could be put in I for a suitable E on the left side of (4.3), giving a larger K. Since each $n_j \leq k$, there must be at least two $j \notin J$ by (4.5). Thus $r - 2 + 2n_0 \leq |Y| \leq k$, $r \leq k - 2|C|$, and by (4.2),

$$2k+2-|C| = \textstyle\sum_{j=1}^{r} (n_j + 1) \leq 2k - 2|C|$$

and $|C| \leq -2$, a contradiction, so $K = k+1$ and (4.3) is proved.

Thus there is a set $A \subset X$ with $|A| = k+1$, $A := E \cup \bigcup_{j \in I} Y_j \cup \{t_j\}$, with E and I from (4.3). Let $B \in \mathcal{C}$. Then by Theorem 4.29(I)(a), either $B \subset Y_j \cup \{t_j\}$ for some j or B is a singleton. Thus either $B \subset A$ or $B \subset A^c$. So Theorem 4.36 applies and $S(\mathcal{D}) \geq k+1$, a contradiction. □

For $k = 1, 2, 3$ we have, where the lower bound for $k = 2$ is due to L. Birgé,

Proposition 4.38. *$S(1,k) = 2k+1$ for $k = 1,2,3$.*

Proof. By Theorem 4.37 we need to show $S(1,k) \geq 2k+1$ for $k = 1,2,3$. For $k = 1$ let $X := \{1,2,3\}$, $\mathcal{C} := \{\emptyset, \{1\}, \{2\}, \{3\}\}$, and $\mathcal{D} := \{\emptyset, \{1\}, \{2\}, \{2,3\}\}$. Then clearly $S(\mathcal{C}) = 1$, $S(\mathcal{D}) = 1$, and $S(\mathcal{C} \sqcup \mathcal{D}) = 3$, so $S(1,1) \geq 3$.

For $k = 2$ let $X := \{1,2,3,4,5\}$, $\mathcal{C} := \{\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{4,5\}\}$, $\mathcal{D} := \{\emptyset, \{1\}, \{2\}, \{3\}, \{5\}, \{1,2\}, \{1,3\}, \{1,5\}, \{2,3\}, \{2,5\}, \{2,3,4\}, \{2,3,5\}, \{2,3,4,5\}\}$. Then one can check that $S(\mathcal{C}) = 1$, $S(\mathcal{D}) = 2$, and $\mathcal{C} \sqcup \mathcal{D} = 2^X$. So $S(1,2) \geq 5$.

To show that $S(1,3) = 7$, take the set $X := \{0,1,2,3,4,5,6\}$. We will find classes $\mathcal{C} \subset 2^X$ and $\mathcal{E} \subset 2^X$ with $S(\mathcal{C}) = 1$, $S(\mathcal{E}) = 3$, and $\mathcal{C} \sqcup \mathcal{E} = 2^X$. Sets $\{a, b, \ldots, d\}$ will be denoted $ab\cdots d$, e.g. $1246 := \{1,2,4,6\}$.

Let $\mathcal{C} := \{\emptyset, 0, 1, 12, 3, 34, 5, 56\}$. Then $\mathcal{C}$ has a treelike partial ordering by inclusion and $S(\mathcal{C}) = 1$.

A set with k elements is called a k-set. $\mathcal{E}$ will contain the following subsets of X: the 0-set $\emptyset$; all 1-sets; all 2-sets except 12 and 34; all 3-sets not including 12 or 34; all 4-sets included in 01234; and the 5-sets 01234 and 12346. Then $\mathcal{E}$ shatters some 3-sets, e.g. 246. To show $S(\mathcal{E}) = 3$ we need to show $\mathcal{E}$ shatters no 4-set. $\mathcal{E}$ shatters no 4-set containing 5 since there is no set A in $\mathcal{E}$ with cardinality $|A| \geq 4$ containing 5.

A 4-set $B \subset 01234$ includes at least one of the pairs 12 or 34. By symmetry, suppose $12 \subset B$. Each set C in $\mathcal{E}$ including 12 contains at least two of 0, 3 and 4, so $|C \cap B| \geq 3$ and $C \cap B \neq 12$. Thus $\mathcal{E}$ does not shatter B. It remains to consider 4-sets containing 6 and not 5. There is no $A \in \mathcal{E}$ including 06 with $|A| \geq 4$. Thus $\mathcal{E}$ does not shatter any 4-set including 06. The sets 1236 and 1246 are not shattered by $\mathcal{E}$ because the subset 126 is not cut from them. Likewise the sets 1346 and 2346 are not shattered because 346 is not cut from them. Thus $S(\mathcal{E}) = 3$.

To show $\mathcal{C} \sqcup \mathcal{E} = 2^X$, clearly $\mathcal{C} \sqcup \mathcal{E}$ contains all 0- and 1-sets and it is easy to check that it contains all 2-sets and all 3-sets A not including 12 or 34. If $A \supset 12$ then $A = 12 \cup c$ where $12 \in \mathcal{C}$ and $c \in \mathcal{E}$, and likewise for 34.

$\mathcal{C} \sqcup \mathcal{E}$ contains $X = 56 \cup 01234$, $012345 = 5 \cup 01234$, and $012346 = 0 \cup 12346$. Each other 6-set is the union of $56 \in \mathcal{C}$ and a 4-set in $\mathcal{E}$ included in 01234.

A 5-set containing 5 and not 6 is the union of $5 \in \mathcal{C}$ and a 4-set $\subset 01234$. A 5-set F containing 6 and not 5 includes at least one of the pairs $P_1 = 12$ or $P_2 = 34$. If it includes both pairs it is in $\mathcal{E}$. If it includes just one pair P_j we have

$$(*) \qquad F = A \cup (F \setminus A),\ A \in \mathcal{C},\ F \setminus A \in \mathcal{E},$$

for $A = P_j$.

If a 5-set $F \supset 56$ includes a pair P_j then (*) follows likewise. Otherwise it holds for $A = 56$. The remaining 5-set 01234 is in $\mathcal{E}$.

A 4-set $\subset 01234$ is in $\mathcal{E}$. A 4-set F containing 5 or 6 or both includes at most one pair P_j. If it includes P_j, (*) holds for $A = P_j$. So suppose F includes neither pair P_j. If $56 \subset F$ then (*) holds for $A = 56$. If $5 \in F$ and $6 \notin F$ then (*) holds for $A = 5$. The remaining case is $6 \in F$ and $5 \notin F$. At least one of $a = 0, 1$ or 3 is in F and (*) holds for $A = a$. The proof of the case $k = 3$ of Proposition 4.38 is complete. □

For classes satisfying stronger conditions, more is true:

Theorem 4.39. *Let X and Y be sets, $\mathcal{C} \subset 2^X$ and $\mathcal{D} \subset 2^Y$. If $\mathcal{C}$ is linearly ordered by inclusion, then $S(\mathcal{C} \boxtimes \mathcal{D}) \leq S(\mathcal{D}) + 1$.*

Theorem 4.40. *For any set X and $\mathcal{C}, \mathcal{D} \subset 2^X$, if $\mathcal{C}$ is linearly ordered by inclusion, then $S(\mathcal{C} \boxdot \mathcal{D}) \leq S(\mathcal{D}) + 1$ for $\boxdot = \sqcap$ or $\sqcup$.*

Then by Theorem 4.10 and induction we have:

Corollary 4.41. *Let $\mathcal{C}_i$ be classes of subsets of a set X and $\mathcal{C} := \{\bigcap_{i=1}^n C_i : \ C_i \in \mathcal{C}_i, \ i = 1, ..., n\}$, where each $\mathcal{C}_i$ is linearly ordered by inclusion. Then $S(\mathcal{C}) \leq n$.*

Definition. For any set X and Vapnik-Červonenkis class $\mathcal{C} \subset 2^X$, $\mathcal{C}$ will be called *bordered* if for some $F \subset X$, with $|F| = S(\mathcal{C})$, and $x \in X \setminus F$, F is shattered by sets in $\mathcal{C}$ all containing x.

Theorem 4.42. *Let $\mathcal{C}_i \subset 2^{X(i)}$ be bordered Vapnik-Červonenkis classes for $i = 1, 2$. Then $S(\mathcal{C}_1 \boxtimes \mathcal{C}_2) \geq S(\mathcal{C}_1) + S(\mathcal{C}_2)$.*

Theorem 4.42 extends by induction to any number of factors. One consequence is:

Corollary 4.43. *In $\mathbb{R}$ let $\mathcal{J}$ be the set of all intervals, which may be open or closed, bounded or unbounded on each side. In other words $\mathcal{J}$ is the set of all convex subsets of $\mathbb{R}$. In $\mathbb{R}^m$ let $\mathcal{C}$ be the collection of all rectangles parallel to the axes, $\mathcal{C} := \{\Pi_{i=1}^m J_i : \ J_i \in \mathcal{J}, \ i = 1, ..., m\}$. Then $S(\mathcal{C}) = 2m$. Let $\mathcal{D}$ be the set of all left half-lines $(-\infty, x]$ or $(-\infty, x)$ for $x \in \mathbb{R}$. Let $\mathcal{T} := \{\Pi_{i=1}^m H_i : \ H_i \in \mathcal{D}, \ i = 1, \cdots, m\}$, so $\mathcal{T}$ is the class of lower orthants parallel to the given axes. Then $S(\mathcal{C}) = m$.*

Proposition 4.44. *Let $\mathcal{I}$ be the set of all intervals in $\mathbb{R}$. Let Y be any set and $\mathcal{C} \subset 2^Y$ with $Y \in \mathcal{C}$. Then in $\mathbb{R} \times Y$, $S(\mathcal{J} \boxtimes \mathcal{C}) \leq 2 + S(\mathcal{C})$.*

Next is a necessary condition for a class $\mathcal{C}$ to be of index 1, not in UCLT. A *chain* of sets will be a class of sets linearly ordered by inclusion. For any class $\mathcal{D}$ of sets let $\mathcal{D}' := \{A^c : \ A \in \mathcal{D}\}$.

Theorem 4.45. (Smoktunowicz) *In a set X, let $\mathcal{C} \subset 2^X$ and $S(\mathcal{C}) = 1$.*
(i) If $\emptyset \in \mathcal{C}$ then for some chains $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{C} \subset \mathcal{A} \sqcap \mathcal{B}$.
(ii) In general, for some chains $\mathcal{A}_i$, $i = 1, 2, 3, 4$,

$$\mathcal{C} \subset (\mathcal{A}_1 \sqcap \mathcal{A}_2) \sqcup (\mathcal{A}_3 \sqcap \mathcal{A}_4)'.$$

Proof. For (ii), for any $A \in \mathcal{C}$, $\mathcal{C} \sqcap A^c$ and $\mathcal{C}' \sqcap A$ are VC classes of index 1, containing $\emptyset$, and

$$\mathcal{C} \subset (\mathcal{C} \sqcap A^c) \sqcup ((\mathcal{C}' \sqcap A)' \sqcap A).$$

Assuming (i), $\mathcal{C}' \sqcap A \subset \mathcal{B}_3 \sqcap \mathcal{B}_4$ for some chains $\mathcal{B}_3, \mathcal{B}_4$ of subsets of A. Letting $\mathcal{A}_j := \mathcal{B}_j \sqcup A^c$, $j = 3, 4$, we have $(\mathcal{C}' \sqcap A)' \sqcap A \subset (\mathcal{A}_3 \sqcap \mathcal{A}_4)'$. So (i) implies (ii).

To prove (i), by Theorem 4.11 we can assume $\mathcal{C}$ is 1-maximal. Thus since $\emptyset \in \mathcal{C}$, $\mathcal{C}$ has a treelike partial ordering by inclusion by Theorem 4.17. First suppose X is finite. By Theorem 4.28, take the treelike partial ordering of X induced by $\mathcal{C}$.

Any chain is included in a maximal chain (for inclusion), and in a finite set of n elements, a maximal chain is of the form

$$\{\emptyset, \{a_1\}, \{a_1, a_2\}, \cdots, \{a_1, \cdots, a_n\}\}$$

and thus is equivalent to defining a linear ordering of the set, $a_1 < a_2 < \cdots < a_n$. To define our two chains $\mathcal{A}$, $\mathcal{B}$ we will thus define two linear orderings $<_{\mathcal{A}}, <_{\mathcal{B}}$ of X. This will be done recursively as follows. Take the elements of X having no predecessors (there must be at least one) and call them $a_1, \cdots, a_k$ for some choice of indices. Let $a_1 <_{\mathcal{A}} a_2 <_{\mathcal{A}} \cdots <_{\mathcal{A}} a_k$, $a_k <_{\mathcal{B}} a_{k-1} <_{\mathcal{B}} \cdots <_{\mathcal{B}} a_1$.

Next, suppose a_j has immediate successors $a_{j1}, \cdots, a_{jr}$. Let $a_j <_{\mathcal{A}} a_{j1} <_{\mathcal{A}} \cdots <_{\mathcal{A}} a_{jr} <_{\mathcal{A}} a_{j+1}$ where "$<_{\mathcal{A}} a_{j+1}$" is omitted if $j = k$. Also let $a_j <_{\mathcal{B}} a_{jr} <_{\mathcal{B}} a_{j,r-1} <_{\mathcal{B}} \cdots <_{\mathcal{B}} a_{j1} <_{\mathcal{B}} a_{j-1}$ where "$<_{\mathcal{B}} a_{j-1}$" is omitted if $j = 1$. Iterating such definitions we get two linear orderings of X, each defining a chain of sets as above, so we get chains $\mathcal{A}, \mathcal{B}$.

By Theorem 4.29, every element of $\mathcal{C}$ is a set of the form $C := \{a_{j_1}, a_{j_1 j_2}, \cdots, a^{(m)} := a_{j_1 j_2 \cdots j_m}\}$ where $a_{j_1} \leq_{\mathcal{C}} a_{j_1 j_2} \leq_{\mathcal{C}} \cdots \leq_{\mathcal{C}} a^{(m)}$ and "$\leq_{\mathcal{C}}$" can be replaced by either $<_{\mathcal{A}}$ or $<_{\mathcal{B}}$. It can be checked easily that

$$C = \{x \in X : x \leq_{\mathcal{A}} a^{(m)}\} \cap \{x \in X : x \leq_{\mathcal{B}} a^{(m)}\}.$$

Thus C is the intersection of a set in $\mathcal{A}$ and a set in $\mathcal{B}$ and the finite case is done.

Now suppose X is infinite, $\emptyset \in \mathcal{C} \subset 2^X$ and $S(\mathcal{C}) = 1$. Let $\mathcal{H} := \{J \subset X : J \neq \emptyset, |J| < \infty\}$. For each $J \in \mathcal{H}$ take chains $\mathcal{P}_{Ji}$, $i = 1, 2$, such that $\mathcal{C} \sqcap J \subset \mathcal{P}_{J1} \sqcap \mathcal{P}_{J2}$. Let h be a non-point ultrafilter of subsets of $\mathcal{H}$, in other words:

(1) h is a non-empty collection of non-empty subsets of $\mathcal{H}$.

(2) If $\mathcal{A}, \mathcal{B} \in h$ then $\mathcal{A} \cap \mathcal{B} \in h$.

(3) If $\mathcal{A} \in h$ and $\mathcal{A} \subset \mathcal{B} \subset \mathcal{H}$ then $\mathcal{B} \in h$.

(4) For all $\mathcal{A} \subset \mathcal{H}$, either $\mathcal{A} \in h$ or $\mathcal{A}^c \in h$.

(5) For each $J \in \mathcal{H}$, $\{J\} \notin h$.

The first three conditions make h a filter, the fourth an ultrafilter, and the fifth a non-point ultrafilter. Non-point ultrafilters exist by the axiom of choice (RAP, Theorem 2.2.4 and the statement after its proof).

Given any indexed family of non-empty sets $\{A_J\}_{J \in \mathcal{H}}$, $\Pi_{J \in \mathcal{H}} A_J$ is the set of all $\{a_J\}_{J \in \mathcal{H}}$ such that $a_J \in A_J$ for all $J \in \mathcal{H}$. For $\{a_J\}_{J \in \mathcal{H}}, \{b_J\}_{J \in \mathcal{H}}$ in $\Pi_{J \in \mathcal{H}} A_J$, let $\{a_J\}_{J \in \mathcal{H}} \equiv_h \{b_J\}_{J \in \mathcal{H}}$ if and only if for some $\mathcal{A} \in h$, $a_J = b_J$ for all $J \in \mathcal{A}$. Let $\lim_h A_J$ be the set of equivalence classes of members of $\Pi_{J \in \mathcal{H}} A_J$ for the relation $\equiv_h$. Let $\{a_J\}^{(\equiv)}_{J \in \mathcal{H}}$ be the equivalence class to which $\{a_J\}_{J \in \mathcal{H}}$ belongs.

If $\mathcal{B}_J$ is a class of subsets of A_J for each $J \in \mathcal{H}$, then for each element $\mathcal{Z}$ of $\lim_h \mathcal{B}_J$, where $\mathcal{Z} = \{B_j\}^{(\equiv)}_{J \in \mathcal{H}}$ for some $B_J \in \mathcal{B}_J$ for each J, define a set $\mathcal{E}_{\mathcal{Z}} \subset \lim_h A_J$ by $\{a_J\}^{(\equiv)}_{J \in \mathcal{H}} \in \mathcal{E}_{\mathcal{Z}}$ if and only if for some $\mathcal{A} \in h$, $a_J \in B_J$ for all $J \in \mathcal{A}$. Let $(\lim)_h \mathcal{B}_J := \{\mathcal{E}_{\mathcal{Z}} : \mathcal{Z} \in \lim_h \mathcal{B}_J\}$.

Let $\overline{X} := \lim_h J$, $\overline{\mathcal{C}} := (\lim)_h \mathcal{C} \sqcap J$ and for $i = 1, 2$ let $\overline{\mathcal{P}}_i := (\lim)_h \mathcal{P}_{Ji}$. To see that each $\overline{\mathcal{P}}_i$ is a chain of subsets of $\overline{X}$, we can take $i = 1$. Let $\mathcal{W}, \mathcal{Z} \in \lim_h \mathcal{P}_{J1}$, $\{B_J\}_{J \in \mathcal{H}} \in \mathcal{W}$, $\{C_J\}_{J \in \mathcal{H}} \in \mathcal{Z}$. Let $\mathcal{J} := \{J \in \mathcal{H} : B_J \subset C_J\}$. Then either $\mathcal{J} \in h$ or $\mathcal{J}^c \in h$. If $\mathcal{J} \in h$ then clearly $\mathcal{E}_{\mathcal{W}} \subset \mathcal{E}_{\mathcal{Z}}$. Otherwise, $\mathcal{J}^c \in h$ and since $\mathcal{P}_{J1}$ is a chain for each J, $C_J \subset B_J$ for all $J \in \mathcal{J}^c$ and $\mathcal{E}_{\mathcal{Z}} \subset \mathcal{E}_{\mathcal{W}}$.

It is easy to check that $\overline{\mathcal{C}} \subset \overline{\mathcal{P}}_1 \sqcap \overline{\mathcal{P}}_2$. There is a natural 1-1 map i of X into $\overline{X}$ by $\{a_J\}^{(\equiv)}_{J \in \mathcal{H}} \in i(x)$ if and only if for some $\mathcal{A} \in h$, $a_J = x$ for all $J \in \mathcal{A}$. So we can view X as a

subset of $\overline{X}$. Each $\overline{\mathcal{P}}_j \sqcap X$ is a chain of subsets of X, and

$$\mathcal{C} \subset \overline{\mathcal{C}} \sqcap X \subset (\overline{P}_1 \sqcap X) \sqcap (\overline{\mathcal{P}}_2 \sqcap X),$$

completing the proof. □

Section 4.4 describes the structure of classes $\mathcal{C}$ with $S(\mathcal{C}) = 1$, but the structure of VC classes with $S(\mathcal{C}) = k$ for $k > 1$ apparently is not known in general. Smoktunowicz (1997) showed that the class $\mathcal{L}$ of lines in the plane, a VC class with $S(\mathcal{L}) = 2$, cannot be obtained from finitely many VC classes of index 1 and finitely many applications of the operations $\sqcap$, $\sqcup$ and '. Theorem 4.45 reduces the proof from "VC classes of index 1" to "chains."

4.6 Probability laws and independence

Let $(X, \mathcal{A}, P)$ be a probability space. Recall the pseudo-metric $d_P(A, B) := P(A \Delta B)$ on $\mathcal{A}$, where $A \Delta B := (A \setminus B) \cup (B \setminus A)$. Recall also that for any (pseudo-) metric space (S, d) and $\varepsilon > 0$, $D(\varepsilon, S, d)$ denotes the maximum number of points more than ε apart (Appendix K).

Definition. For a measurable space $(X, \mathcal{A})$ and $\mathcal{C} \subset \mathcal{A}$ let $s(\mathcal{C}) := \inf\{w :$ there is a $K = K(w, \mathcal{C}) < \infty$ such that for every law P on $\mathcal{A}$ and $0 < \varepsilon \leq 1$, $D(\varepsilon, \mathcal{C}, d_P) \leq K\varepsilon^{-w}\}$.

This index $s(\mathcal{C})$ turns out to equal the density:

Theorem 4.46. *For any measurable space $(X, \mathcal{A})$ and $\mathcal{C} \subset \mathcal{A}$, dens$(\mathcal{C}) = s(\mathcal{C})$.*

In fact in this case the infimum in the definition of $s(\mathcal{C})$ is attained:

Theorem 4.47. (Haussler). *For each $m = 1, 2, \cdots$, there is a $K_m < \infty$ such that for any class $\mathcal{C}$ with $S(\mathcal{C}) = m$ and any law P defined on a σ-algebra including $\mathcal{C}$, $D(\varepsilon, \mathcal{C}, d_P) \leq K_m \varepsilon^{-m}$ for $0 < \varepsilon < 1$.*

For a proof see van der Vaart and Wellner (1996), pp. 137-140.

There is a notion of independence for sets without probability. To define it, for any set X and subset $A \subset X$ let $A^1 := A$ and $A^{-1} := X \setminus A$. Sets $A_1, ..., A_m$ are called *independent*, or *independent as sets*, if for every function $s(\cdot)$ from $\{1, ..., m\}$ into $\{-1, +1\}$, $\bigcap_{j=1}^m A_j^{s(j)} \neq \emptyset$. Such intersections, when they are nonempty, are called *atoms* of the Boolean algebra generated by $A_1, ..., A_m$. Thus for $A_1, ..., A_m$ to be independent as sets means that the Boolean algebra they generate has the maximum possible number, 2^m, of atoms.

If $A_1, ..., A_m$ are independent as sets, then one can define a probability law on the algebra they generate for which they are jointly independent in the usual probability sense and for which $P(A_i) = 1/2$, $i = 1, ..., m$. For example, choose a point in each atom and put mass $1/2^m$ at each point chosen. Or, if desired, given any q_i, $0 \leq q_i \leq 1$, one can define a probability measure Q for which the A_i are jointly independent and have $Q(A_i) = q_i$, $i = 1, ..., n$.

For a set X and $\mathcal{C} \subset 2^X$ let

$$I(\mathcal{C}) := \sup\{m : A_1, ..., A_m \text{ are independent as sets for some } A_i \in \mathcal{C},$$

$$i = 1, ..., m\}.$$

Theorem 4.48. *For any set X, $\mathcal{C} \subset 2^X$, and $n = 1, 2, ...,$ if $S(\mathcal{C}) \geq 2^n$ then $I(\mathcal{C}) \geq n$. Conversely if $I(\mathcal{C}) \geq 2^n$ then $S(\mathcal{C}) \geq n$. So $I(\mathcal{C}) < \infty$ if and only if $S(\mathcal{C}) < \infty$. In both cases, 2^n cannot be replaced by $2^n - 1$.*

For any set X, $\mathcal{C} \subset 2^X$ and $Y \subset X$, recall that $\mathcal{C}_Y := Y \sqcap \mathcal{C} := \{Y \cap C : \ C \in \mathcal{C}\}$. Let $\mathrm{At}(\mathcal{C}|Y)$ be the set of atoms of the algebra of subsets of Y generated by $\mathcal{C}_Y$, where in the cases to be considered, $\mathcal{C}_Y$ will be finite because $\mathcal{C}$ or Y is. Let $\Delta_{\mathcal{C}}(Y) := |\mathrm{At}(\mathcal{C}|Y)|$ be the number of such atoms. Let $m_{\mathcal{C}}^Y(n) := \sup\{\Delta_{\mathcal{A}}(Y) : \ \mathcal{A} \subset \mathcal{C}, \ |\mathcal{A}| \leq n\} \leq 2^n$. Let

$$\mathrm{dens}^*(\mathcal{C}) \ := \ \inf\{s \geq 0 : \quad \text{for some } C < \infty, \quad m_{\mathcal{C}}^X(n) \leq Cn^s \text{ for all } n\}.$$

For any $x \in X$ let $\mathcal{C}_x := \{A \in \mathcal{C} : \ x \in A\}$. Let $\mathcal{C}'_Y := \{\mathcal{C}_y : \ y \in Y\}$.

Theorem 4.49. (Assouad) *For any set X and $\mathcal{A} \subset \mathcal{C} \subset 2^X$, with $\mathcal{A}$ finite,*
(a) $\Delta_{\mathcal{A}}(X) = \Delta^{\mathcal{C}'_X}(\mathcal{A})$.
(b) For $n = 1, 2, ...,$ $m_{\mathcal{C}}^X(n) = m^{\mathcal{C}'_X}(n)$.
(c) $S(\mathcal{C}'_X) = I(\mathcal{C})$.
(d) $dens^*(\mathcal{C}) \ = dens(\mathcal{C}'_X) \leq I(\mathcal{C})$.

4.7 Vapnik-Červonenkis properties of classes of functions

The notion of VC class of sets has several extensions to classes of functions.

Definitions. Let X be a set and $\mathcal{F}$ a class of real-valued functions on X. Let $\mathcal{C} \subset 2^X$. If f is any real-valued function, each set $\{f > t\}$ for $t \in \mathbb{R}$ will be called a *major set* of f. The class $\mathcal{F}$ will be called a *major class* for $\mathcal{C}$ if all the major sets of each $f \in \mathcal{F}$ are in $\mathcal{C}$. If $\mathcal{C}$ is a Vapnik-Červonenkis class, then $\mathcal{F}$ will be called a *VC major* class (for $\mathcal{C}$).

The *subgraph* of a real-valued function f will be the set $\{(x, t) \in X \times \mathbb{R} : \ 0 \leq t \leq f(x)$ or $f(x) \leq t \leq 0\}$. If $\mathcal{D}$ is a class of subsets of $X \times \mathbb{R}$, and for each $f \in \mathcal{F}$, the subgraph of f is in $\mathcal{D}$, then $\mathcal{F}$ will be called a *subgraph class* for $\mathcal{D}$. If $\mathcal{D}$ is a VC class in $X \times \mathbb{R}$ then $\mathcal{F}$ will be called a *VC subgraph class.*

The *symmetric convex hull* of $\mathcal{F}$ is the set of all functions $\sum_{i=1}^m c_i f_i$ for $f_i \in \mathcal{F}$, $c_i \in \mathbb{R}$, any finite m, and $\sum_{i=1}^m |c_i| \leq 1$. If $0 < M < \infty$, let $H(\mathcal{F}, M)$ denote M times the symmetric convex hull of $\mathcal{F}$.

Let $\bar{H}_s(\mathcal{F}, M)$ be the smallest class $\mathcal{G}$ of functions including $H(\mathcal{F}, M)$ such that whenever $g_n \in \mathcal{G}$ for all n and $g_n(x) \to g(x)$ as $n \to \infty$ for all x, we have $g \in \mathcal{G}$.

A class $\mathcal{F}$ of functions such that $\mathcal{F} \subset \bar{H}_s(\mathcal{G}, M)$ for some $M < \infty$ and a given $\mathcal{G}$ will be called a *VC subgraph hull class* if $\mathcal{G}$ is a VC subgraph class, and a *VC hull class* if $\mathcal{G} = \{1_C : \ C \in \mathcal{C}\}$ where $\mathcal{C}$ is a VC class of sets.

So there are at least four possible ways to extend the notion of VC class to classes of functions. Some implications hold between these different conditions, but no two of them are equivalent. The next theorem deals with some of the easier cases of implication or non-implication.

Theorem 4.50. *Let $\mathcal{F}$ be a uniformly bounded class of nonnegative real-valued functions on a set X. Then*
(a) If $\mathcal{F}$ is the set of indicators of members of a VC class of sets, then $\mathcal{F}$ is also a VC major class, a VC subgraph class, and a VC hull class.

(b) If $\mathcal{F}$ is a VC major class then it is a VC hull class.
(c) There exist VC hull classes $\mathcal{F}$ which are not VC major.
(d) There exist VC subgraph classes $\mathcal{F}$ which are not VC major.

4.8 Classes of functions and dual density

For a metric space (S, d) and $\varepsilon > 0$ $D(\varepsilon, S, d)$, is defined as the maximum number of points more than ε apart (Appendix K). For a probability measure Q and $1 \le p < \infty$ we have the L^p metric $d_{p,Q}(f, g) := (\int |f - g|^p dQ)^{1/p}$. For a class $\mathcal{F} \subset \mathcal{L}^p(Q)$ let $D^{(p)}(\varepsilon, \mathcal{F}, Q) := D(\varepsilon, \mathcal{F}, d_{p,Q})$. Let $D^{(p)}(\varepsilon, \mathcal{F})$ be the supremum of $D(\varepsilon, \mathcal{F}, d_{p,Q})$ over all laws Q concentrated in finite sets.

If $\mathcal{F}$ is a class of measurable real-valued functions on a measurable space $(X, \mathcal{A})$, let $F_{\mathcal{F}}(x) := \sup_{f \in \mathcal{F}} |f(x)|$. Then a measurable function F will be called an *envelope function* for $\mathcal{F}$ if and only if $F_{\mathcal{F}} \le F$. If $F_{\mathcal{F}}$ is measurable it will be called *the* envelope function of $\mathcal{F}$. For any law P on $(X, \mathcal{A})$, $F^*_{\mathcal{F}}$ is an envelope function for $\mathcal{F}$, which in general depends on P.

Given $\mathcal{F}$, an envelope function F for it, $\varepsilon > 0$ and $1 \le p < \infty$, let $D^{(p)}_F(\varepsilon, \mathcal{F}, Q)$ be the supremum of m such that there exist $f_1, ..., f_m \in \mathcal{F}$ for which $\int |f_i - f_j|^p dQ > \varepsilon^p \int F^p dQ$ for all $i \ne j$.

The next fact extends Theorem 4.47 to families of functions. See van der Vaart and Wellner (1996), Theorem 2.6.7.

Theorem 4.51. *Let $1 \le p < \infty$. Let $(X, \mathcal{A}, Q)$ be a probability space and $\mathcal{F}$ be a VC subgraph class of measurable real-valued functions on X. Let $\mathcal{F}$ have an envelope $F \in \mathcal{L}^p(X, \mathcal{A}, Q)$ with $0 < \int F dQ$. Let $\mathcal{C}$ be the collection of subgraphs in $X \times \mathbb{R}$ of functions in $\mathcal{F}$. Then there is an $A < \infty$ depending only on $S(\mathcal{C})$ such that*

$$D^{(p)}_F(\varepsilon, \mathcal{F}, Q) \ \le \ A/\varepsilon^{pS(\mathcal{C})} \quad \text{for} \quad 0 < \varepsilon \le 1. \tag{4.6}$$

The following fact is a continuation of Theorem 4.50.

Theorem 4.52. *Let $\mathcal{F}$ be a uniformly bounded class of functions on a set X.*
(a) If $\mathcal{F}$ is a VC subgraph class then

$$\text{For some } r < \infty \text{ and } M < \infty, \ D^{(2)}(\varepsilon, \mathcal{F}) \le M\varepsilon^{-r} \quad \text{for} \quad 0 < \varepsilon < 1. \tag{4.7}$$

(b) There exist classes $\mathcal{F}$ satisfying (4.7) which are not VC hull.
(c) There exist VC subgraph classes which are not VC hull.

It will be seen in Proposition 10.2 below that there are VC major (thus VC hull) classes which do not satisfy (4.7) and so are not VC subgraph classes.

Problems on Chapter 4

1. Let $\mathcal{C}$ be the class of all unions of two intervals in $\mathbb{R}$. Evaluate $S(\mathcal{C})$. Hint: try it first directly; if you like, look at the more general Problem 11.

2. If $S(\mathcal{C}) = 3$ find the upper bounds for $m^{\mathcal{C}}(n)$ given by Theorem 4.2 and by Proposition 4.3.

3. Show that for $\mathrm{dens}(\mathcal{C}) = 0$, $S(\mathcal{C})$, which is finite by Corollary 4.4, can be arbitrarily large. Hint: let $\mathcal{C}$ be finite.

4. Find the smallest n such that there is a set X with $|X| = n$ and $\mathcal{C} \subset 2^X$ with $S(\mathcal{C}) = 1$ where neither (a) nor (b) in Theorem 4.10 holds.

Hints on Problems 5-7: If $\mathcal{C}$ is a collection of convex sets in $\mathbb{R}^d$ and shatters a set F, then no point in F is in the convex hull of the other points. Then, the convex hull of F is a polyhedron of which each point of F is a vertex. In the plane, it's a polygon. To get a lower bound $S(\mathcal{C}) \geq k$ it's enough to find one set of k elements that is shattered. Try the vertices of a regular k-gon. To get upper bounds, use facts such as Theorem 4.6 and Proposition 4.35.

5. Let $\mathcal{C}$ be the set of all interiors of ellipses in $\mathbb{R}^2$, with arbitrary centers and semiaxes in any two perpendicular directions. Give upper and lower bounds for $S(\mathcal{C})$.

6. A half-plane in $\mathbb{R}^2$ is a set of the form $\{(x, y) : \ ax + by \geq c\}$ for real a, b, c with a and b not both 0. Define a wedge as an intersection of two half-planes. Let $\mathcal{C}$ be the collection of all wedges in $\mathbb{R}^2$. Show that $S(\mathcal{C}) \geq 5$. Also find an upper bound for $S(\mathcal{C})$.

7. Let $\mathcal{C}$ be the set of all interiors of triangles in $\mathbb{R}^2$. Show that $S(\mathcal{C}) \geq 7$. Also give an upper bound for $S(\mathcal{C})$.

8. Show that the lower bounds for $S(\mathcal{C})$ in problems 6 and 7 are the values of $S(\mathcal{C})$. Hint: for a convex polygon, the set F of vertices can be arranged in cyclic order, say clockwise around the boundary of the polygon, $v_1, v_2, ..., v_n, v_1$. Show that if a half-plane J contains v_i and v_j with $i < j$ then it includes either $\{v_i, v_{i+1}, ..., v_j\}$ or $\{v_j, v_{j+1}, ..., v_n, v_1, ..., v_i\}$. Thus find what kind of set the intersection of J and F must be. From that, find what occurs if two or three half-planes are intersected (or unioned, via complements).

9. In the example at the end of section 4.3, for each set $A \subset X$ with 3 elements, find a specific subset of A not in $A \sqcap \mathcal{C}$.

10. Let $\mathcal{F}$ be the class of all probability distribution functions on $\mathbb{R}$. Show that $\mathcal{F}$ is a VC major class but not a VC subgraph class. Hint: show that the subgraphs of functions in $\mathcal{F}$ shatter all sets $\{(x_j, y_j)\}_{j=1}^n$ with $x_1 < \cdots < x_n$ and $0 < y_1 < \cdots < y_n < 1$.

11. Let $\mathcal{C}(j)$ be the class of all unions of j intervals in $\mathbb{R}$ for $j = 1, 2, \cdots$. Show that $S(\mathcal{C}(j)) = 2j$ for all j and that for any finite set $F \subset \mathbb{R}$ with $|F| = n$ we have $\Delta^{\mathcal{C}(j)}(F) = {}_nC_{\leq 2j}$ (the largest possible value by Sauer's Lemma). Hints: one can take $F = \{1, 2, \cdots, n\}$. For $A \subset F$ and $x, y \in F$ let $x =_A y$ mean that $A \cap \{x, y\} = \emptyset$ or $\{x, y\}$, otherwise $x \neq_A y$. If $A \neq \emptyset$ let $j_1 := j_1(A)$ be the least element of A. If $j_1(A), \cdots, j_k(A)$ are defined let $j_{k+1}(A)$ be the least $j > j_k(A)$ such that $j \neq_A j_k(A)$ and $j \leq n$, if there is such a j. Show that there is a 1-1 correspondence between subsets $A \subset F$ and finite sequences $(j_1, j_2, \cdots, j_r)$ for $r = r(A) = 1, \cdots, n$ where $r(\emptyset) := 0$. Show that $A \in \mathcal{C}(j) \sqcap F$ if and only if $r(A) \leq 2j$.

REFERENCES FOR CHAPTER 4

Alexander, Kenneth S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.* **12**, 1041-1067.

Alexander, K. S. (1987). The central limit theorem for empirical processes on Vapnik-Červonenkis classes. *Ann. Probab.* **15**, 178-203.

Alon, N., and Tarsi, M. (1992). Colorings and orientations of graphs. *Combinatorica* **12**, 125-134.

Assouad, Patrice (1981). Sur les classes de Vapnik-Červonenkis. *C. R. Acad. Sci. Paris* Sér. I **292**, 921-924.

Assouad, P. (1983). Densité et dimension. *Ann. Inst. Fourier* (Grenoble) **33** no. 3, 233-282.

Danzer, L., Grünbaum, B. and Klee, V. L. (1963). Helly's theorem and its relatives. *Proc. Symp. Pure Math. (Amer. Math. Soc.)* **7**, 101-180.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6**, 899-929; Correction **7** (1979), 909-911.

Dudley, R. M. (1984). A course on empirical processes. *Ecole d'été de probabilités de St.-Flour, 1982. Lecture Notes in Math.* (Springer) **1097**, 1-142.

Dudley, R. M. (1985). The structure of some Vapnik-Červonenkis classes. In *Proc. Berkeley Conf. in honor of J. Neyman and J. Kiefer* **2**, 495-508. Wadsworth, Belmont, CA.

Dudley, R. M. (1987). Universal Donsker classes and metric entropy. *Ann. Probab.* **15**, 1306-1326.

Harary, F. (1969). *Graph Theory.* Addison-Wesley, Reading, Mass.

Haussler, D. (1995). Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *J. Combinatorial Theory (A)* **69**, 217-232.

Laskowski, M. C. (1992). Vapnik-Chervonenkis classes of definable sets. *J. London Math. Soc.* (Ser. 2) **45**, 377-384.

Pollard, David (1982). A central limit theorem for k-means clustering. *Ann. Probab.* **10**, 919-926.

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

Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory* **1**, 295-314.

Radon, J. (1921). Mengen konvexer Körper, die einen gemeinsamen Punkt enthalten. *Math. Ann.* **83**, 113-115.

Sauer, N. (1972). On the density of families of sets. *J. Combin. Theory* Ser. A **13**, 145-147.

Shelah, S. (1972). A combinatorial problem: stability and order for models and theories in infinitary languages. *Pacific J. Math.* **41**, 247-261.

Smoktunowicz, A. (1997). A remark on Vapnik-Chervonienkis classes. *Colloq. Math.* **74**, 93-98.

Stengle, G., and Yukich, J. E. (1989). Some new Vapnik-Chervonenkis classes. *Ann. Statist.* **17**, 1441-1446.

van der Vaart, A. W., and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer, New York.

Vapnik, V. N., and Červonenkis, A. Ya. (1968). Uniform convergence of frequencies of occurrence of events to their probabilities. *Dokl. Akad. Nauk SSSR* **181**, 781-783 (Russian) = *Sov. Math. Doklady* **9**, 915-918 (English).

Vapnik, V. N., and Červonenkis, A. Ya. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theor. Probability Appls.* **16**, 264-280 = *Teor. Verojatnost. i Primenen.* **16**, 264-279.

Vapnik, V. N., and Červonenkis, A. Ya. (1974). *Teoriya Raspoznavaniya Obrazov: Statisticheskie problemy obucheniya* [Theory of Pattern Recognition; Statistical problems of learning; in Russian]. Nauka, Moscow. German ed.: *Theorie der Zeichenerkennung*, by W. N. Wapnik and A. J. Tscherwonenkis, transl. by K. G. Stöckel and B. Schneider, ed. S. Unger and K.

Fritzsch. Akademie-Verlag, Berlin, 1979 (*Elektronisches Rechnen und Regeln*, Sonderband).

Wenocur, R. S., and Dudley, R. M. (1981). Some special Vapnik-Červonenkis classes. *Discrete Math.* **33**, 313-318.

Chapter 5

Measurability

Let A be the set of all possible empirical distribution functions F_1 for one observation $x \in [0,1]$, namely $F_1(t) = 0$ for $t < x$ and $F_1(t) = 1$ for $t \geq x$. We noted previously that A in the supremum norm is non-separable: it is an uncountable set, in which any two points are at a distance 1 apart. Thus A and all its subsets are closed. If $x := X_1$ has a continuous distribution such as the uniform distribution $U[0,1]$ on $[0,1]$, then $x \mapsto (t \mapsto 1_{t \geq x})$ takes $[0,1]$ onto A, but it is not continuous for the supremum norm. Also, it is not measurable for the Borel σ-algebra on the range space. So, in Chapter 3, functions f^* and upper expectations E^* were used to get around measurability problems.

Here is a different kind of example. It is related to the basic "ordinal triangle" counter-example in integration theory, showing why measurability is needed in the Tonelli-Fubini theorem on cartesian product integrals. Let $(\Omega, \leq)$ be an uncountable well-ordered set such that for each $x \in \Omega$, the initial segment $I_x := \{y : y \leq x\}$ is countable. (In terms of ordinals, Ω is, or is order-isomorphic to, the least uncountable ordinal.) Let $\mathcal{S}$ be the σ-algebra of subsets of Ω consisting of sets that are countable or have countable complement. Let P be the probability measure on $\mathcal{S}$ which is 0 on countable sets and 1 on sets with countable complement. Then

$$\int \int 1_{y \leq x} dP(y) dP(x) \;=\; 0 \;<\; \int \int 1_{y \leq x} dP(x) dP(y) \;=\; 1.$$

Since all other conditions of the Tonelli-Fubini theorem hold, the function $(x, y) \mapsto 1_{y \leq x}$ must not be measurable for the product σ-algebra, even if $\mathcal{S}$ is replaced by any larger σ-algebra of subsets of Ω to which P can be extended. For example, according to the continuum hypothesis, we could take Ω to be $[0,1]$ (where the well-ordering is unrelated to the usual ordering), and P to be Lebesgue measure or any other nonatomic law on $[0,1]$.

Now, consider the class $\mathcal{C}$ of sets I_x for each $x \in \Omega$. Each of these sets is countable by assumption. The sets are linearly ordered by inclusion since $\leq$ is a linear ordering. Thus $S(\mathcal{C}) = 1$ by Theorem 4.10. But, $\mathcal{C}$ is not a weak or strong Glivenko-Cantelli class as defined in Section 3.3 (still less a Donsker class), since for any possible $X_1, \cdots, X_n$, a maximum $x := \max(X_1, \cdots, X_n)$ for the well-ordering exists, so $P_n(I_x) = 1$ while $P(I_x) = 0$, so $\sup_{A \in \mathcal{C}} |(P_n - P)(A)| \;\equiv\; 1$.

In sections 5.2 and 5.3, some measurability conditions will be developed, which will hold for classes of sets encountered in practice. It will be seen in the next chapter that these conditions, together with the Vapnik-Červonenkis or related properties, are enough to imply Glivenko-Cantelli and Donsker properties.

5.1 Sufficiency

Sufficiency is a concept from mathematical statistics. Suppose that a probability measure P is known to be in a certain family $\mathcal{P}$ of laws and we have observed $X_1, ..., X_n$ i.i.d. (P), but nothing else is known about P. A *statistic*, T, which is a measurable function of $X_1, ..., X_n$, will roughly speaking be said to be sufficient for $\mathcal{P}$ if, given T, no further information about $X_1, ..., X_n$ is useful in making decisions or inferences about $P \in \mathcal{P}$. A precise definition is given below. In this section it will be seen that the empirical measure P_n is sufficient even when $\mathcal{P}$ is the family of all probability measures on a measurable space.

Note that P_n is a symmetric function of the X_i in the sense that it is preserved by any permutation of the indices $1, ..., n$. Once P_n is given, knowing that the X_i were observed in a certain order will not help in making inferences about P.

Here is the formal definition of sufficiency: let $(Y, \mathcal{S})$ be a measurable space (a set Y and a σ-algebra $\mathcal{S}$ of subsets of Y). Let $\mathcal{Q}$ be a set of probability laws on $(Y, \mathcal{S})$. A sub-σ-algebra $\mathcal{B}$ of $\mathcal{S}$ is called *sufficient* for $\mathcal{Q}$ iff for every $C \in \mathcal{S}$ there is some $\mathcal{B}$-measurable function g_C such that for every $Q \in \mathcal{Q}$, the conditional probability

$$Q(C|\mathcal{B}) \;=\; g_C \;\text{ almost surely for }\; Q. \tag{5.1}$$

The essential point is that g_C does not depend on Q in $\mathcal{Q}$.

Most often, there will be some $n > 1$, a measurable space $(X, \mathcal{A})$ and a family $\mathcal{P}$ of laws on $(X, \mathcal{A})$ such that Y is the n-fold Cartesian product X^n with the product σ-algebra $\mathcal{S} = \mathcal{A}^n$ and $\mathcal{Q} = \mathcal{P}^n := \{P^n : P \in \mathcal{P}\}$, where P^n is the n-fold Cartesian product $P \times P \times \cdots \times P$ (RAP, Theorem 4.4.6).

The meaning of sufficiency is clarified by the factorization theorem, to be stated next. A family $\mathcal{P}$ of probability measures on a measurable space $(S, \mathcal{B})$ is said to be *dominated* by a σ-finite measure μ if every $P \in \mathcal{P}$ is absolutely continuous with respect to μ. Then we have the density (Radon-Nikodym derivative) $dP/d\mu$ (RAP, Section 5.5).

If there is a nonatomic law on $(X, \mathcal{A})$, the family $\mathcal{P}$ of all laws on $(X, \mathcal{A})$ is not dominated. Factorization is still useful in that case, in the proof of Theorem 5.3 below.

Theorem 5.1. *(Factorization theorem). Let $(S, \mathcal{B})$ be a measurable space, $\mathcal{A}$ a sub-σ-algebra of $\mathcal{B}$, and $\mathcal{P}$ a family of probability measures on $\mathcal{B}$, dominated by a σ-finite measure μ. Then $\mathcal{A}$ is sufficient for $\mathcal{P}$ if and only if there is a $\mathcal{B}$-measurable function $h \geq 0$ such that for all $P \in \mathcal{P}$, there is an $\mathcal{A}$-measurable function f_P with $dP/d\mu = f_P h$ almost everywhere for μ. We can take $h \in \mathcal{L}^1(S, \mathcal{B}, \mu)$.*

Given a statistic T, i.e. a measurable function, from S into Y for measurable spaces $(S, \mathcal{B})$ and $(Y, \mathcal{F})$, let $\mathcal{A} := T^{-1}(\mathcal{F}) := \{T^{-1}(A) : A \in \mathcal{F}\}$, a σ-algebra. For a family $\mathcal{Q}$ of laws on $(S, \mathcal{B})$, T is called a *sufficient statistic* for $\mathcal{Q}$ iff $\mathcal{A}$ is sufficient for $\mathcal{Q}$. If T is sufficient we can write $f_P = g_P \circ T$ for some $\mathcal{F}$-measurable function g_P by RAP, Theorem 4.2.8. Sufficiency, defined in terms of conditional probabilities of measurable sets, can be extended to suitable conditional expectations:

Theorem 5.2. *Let $\mathcal{A}$ be sufficient for a family $\mathcal{P}$ of laws on a measurable space $(S, \mathcal{B})$. Then for any measurable real-valued function f on $(S, \mathcal{B})$ which is integrable for each $P \in \mathcal{P}$, there is an $\mathcal{A}$-measurable function g such that g such that $g = E_P(f|\mathcal{A})$ a.s. for all $P \in \mathcal{P}$.*

Let μ and ν be two probability measures on the same measurable space $(V, \mathcal{U})$. Take the Lebesgue decomposition (RAP, Theorem 5.5.3) $\nu = \nu_{ac} + \nu_s$ where ν_{ac} is absolutely continuous, and ν_s is singular, with respect to μ. Let $A \in \mathcal{U}$ with $\nu_s(A) = \mu(V \setminus A) = 0$, so $\nu_{ac}(V \setminus A) = 0$. Then the *likelihood ratio* $R_{\nu/\mu}$ is defined as the Radon-Nikodym derivative $d\nu_{ac}/d\mu$ on A and $+\infty$ on $V \setminus A$. By uniqueness of the Hahn decomposition of V for $\nu_s - \mu$ (RAP, Theorem 5.6.1), $R_{\nu/\mu}$ is defined up to equality $(\mu + \nu_s)$- and so $(\mu + \nu)$-almost everywhere.

Theorem 5.3. *For any family $\mathcal{P}$ of laws on a measurable space $(S, \mathcal{B})$ and sub-σ-algebra $\mathcal{A} \subset \mathcal{B}$, $\mathcal{A}$ is sufficient for $\mathcal{P}$ if and only if for all $P, Q \in \mathcal{P}$, $R_{Q/P}$ can be taken to be $\mathcal{A}$-measurable, i.e. is equal $(P + Q)$-almost everywhere to an $\mathcal{A}$-measurable function.*

Suppose we observe $X_1, \cdots, X_n$ i.i.d. with law P or Q but we do not know which and want to decide. Suppose we have no *a priori* reason to favor a choice of P or Q, only the data. Then it is natural to evaluate the likelihood ratio R_{Q^n/P^n} and choose Q if $R_{Q^n/P^n} > 1$ and P if $R_{Q^n/P^n} < 1$, while if $R_{Q^n/P^n} = 1$ we still have no basis to prefer P or Q. More generally, decisions between P and Q can be made optimally in terms of minimizing error probabilities or expected losses by way of the likelihood ratio $R_{Q/P}$ or R_{Q^n/P^n} as appropriate (the Neyman-Pearson Lemma, Lehmann, 1991, pp. 74, 125; Ferguson, 1967, Section 5.1). By Theorem 5.3, if $\mathcal{B}$ is sufficient for $\mathcal{P}^n$ for some $\mathcal{P} \supset \{P, Q\}$, then R_{Q^n/P^n} is $\mathcal{B}$-measurable. Specifically, if T is a sufficient statistic, then by Theorem 5.3 and RAP, Theorem 4.2.8, R_{Q^n/P^n} is a measurable function of T. Thus, no information in $(X_1, \cdots, X_n)$ beyond T is helpful in choosing P or Q. In this sense, the definition of sufficiency fits with the informal notion of sufficiency given at the beginning of the section.

It will be seen that empirical measures are sufficient in a sense to be defined. Let $\mathcal{S}_n$ be the sub-σ-algebra of $\mathcal{A}^n$ consisting of sets invariant under all permutations of the coordinates.

Theorem 5.4. *$\mathcal{S}_n$ is sufficient for $\mathcal{P}^n := \{P^n : P \in \mathcal{P}\}$ where $\mathcal{P}$ is the set of all laws on $(X, \mathcal{A})$.*

For example, if X has just two points, say $X = \{0, 1\}$, and $S := \sum_{i=1}^n x_i$, then $\mathcal{S}_n$ is the smallest σ-algebra for which S is measurable. In this case no σ-algebra strictly smaller than $\mathcal{S}_n$ is sufficient ($\mathcal{S}_n$ is "minimal sufficient").

For each $B \in \mathcal{A}$ and $x = (x_1, ..., x_n) \in X^n$, let

$$P_n(B)(x) := \frac{1}{n} \sum_{j=1}^n 1_B(x_j).$$

So P_n is the usual empirical measure, except that in this section, $x \mapsto P_n(B)(x)$ is a measurable function, or statistic, on a measurable space, rather than a probability space, since no particular law P or P^n has been specified as yet. Here, $P_n(B)(x)$ is just a function of B and x.

For a collection $\mathcal{F}$ of measurable functions on $(X^n, \mathcal{A}^n)$, let $\mathcal{S}_\mathcal{F}$ be the smallest σ-algebra making all functions in $\mathcal{F}$ measurable. Then $\mathcal{F}$ will be called *sufficient* if and only if $\mathcal{S}_\mathcal{F}$ is sufficient.

Theorem 5.5. *For any measurable space $(X, \mathcal{A})$ and for each $n = 1, 2, ...$, the empirical measure P_n is sufficient for $\mathcal{P}^n$ where $\mathcal{P}$ is the set of all laws on $(X, \mathcal{A})$. In other words the set $\mathcal{F}$ of functions $x \mapsto P_n(B)(x)$, for all $B \in \mathcal{A}$, is sufficient. In fact the σ-algebra $\mathcal{S}_\mathcal{F}$ is exactly $\mathcal{S}_n$.*

For some subclasses $\mathcal{C} \subset \mathcal{A}$, the restriction of P_n to $\mathcal{C}$ may be sufficient, and handier than the values of P_n on the whole σ-algebra $\mathcal{A}$. Recall that a class $\mathcal{C}$ included in a σ-algebra $\mathcal{A}$ is called a *determining class* if any two measures on $\mathcal{A}$, equal and finite on $\mathcal{C}$, are equal on all of $\mathcal{A}$. If $\mathcal{C}$ generates the σ-algebra $\mathcal{A}$, $\mathcal{C}$ is not necessarily a determining class unless for example it is an algebra (RAP, Theorem 3.2.7 and the example after it).

Sufficiency of $P_n(A)$ for $A \in \mathcal{C}$ can depend on n. Let $X = \{1,2,3,4,5\}$, $\mathcal{A} = 2^X$, and $\mathcal{C} = \{\{1,2,3\},\{2,3,4\},\{3,4,5\}\}$. Then $\mathcal{C}$ is sufficient for $n = 1$, but not for $n = 2$ since for example $(\delta_1 + \delta_4)/2 \equiv (\delta_2 + \delta_5)/2$ on $\mathcal{C}$. In this case $\mathcal{C}$ generates $\mathcal{A}$ but is not a determining class.

Theorem 5.6. *Let (X,d) be a separable metric space which is a Borel subset of its completion, with Borel σ-algebra. Suppose $\mathcal{C} = \{C_k\}_{k=1}^{\infty}$ is a countable determining class for $\mathcal{A}$. Then for each $n = 1,2,\cdots$, the sequence $\{P_n(C_k)\}_{k=1}^{\infty}$ is sufficient for the class $\mathcal{P}^n$ of all laws P^n on $(X^n, \mathcal{A}^n)$ where $P \in \mathcal{P}$, the class of all laws on $(X, \mathcal{A})$.*

In the real line $\mathbb{R}$, the closed half-lines $(-\infty, x]$ form a determining class. In other words, as is well known, a probability measure P on the Borel σ-algebra of $\mathbb{R}$ is uniquely determined by its distribution function F (RAP, Theorem 3.2.6). It follows that the half-lines $(-\infty, q]$ for q rational are a determining class: for any real x, take rational $q_k \downarrow x$, then $F(q_k) \downarrow F(x)$. Thus we have:

Corollary 5.7. *In $\mathbb{R}$, the empirical distribution functions $F_n(x) := P_n((-\infty, x])$ are sufficient for the family $\mathcal{P}^n$ of all laws P^n on $\mathbb{R}^n$ where P varies over all laws on the Borel σ-algebra in $\mathbb{R}$.*

5.2 Admissibility

Let $\mathcal{F}$ be a family of real-valued functions on a set X, measurable for a σ-algebra $\mathcal{A}$ on X. Then there is a natural function, called here the *evaluation map*, $\mathcal{F} \times X \mapsto \mathbb{R}$ given by $(f,x) \mapsto f(x)$. It turns out that for general $\mathcal{F}$ there may not exist any σ-algebra of subsets of $\mathcal{F}$ for which the evaluation map is jointly measurable. The possible existence of such a σ-algebra and its uses will be the subject of this section.

Let $(X, \mathcal{B})$ be a measurable space. Then $(X, \mathcal{B})$ will be called *separable* if $\mathcal{B}$ is generated by some countable subclass $\mathcal{C} \subset \mathcal{B}$ and $\mathcal{B}$ contains all singletons $\{x\}$, $x \in X$. In this section $(X, \mathcal{B})$ will be assumed to be such a space. Let $\mathcal{F}$ be a collection of real-valued functions on X. (The following definition is unrelated to the usage of "admissible" for estimators in statistics.)

Definition. $\mathcal{F}$ is called *admissible* iff there is a σ-algebra $\mathcal{T}$ of subsets of $\mathcal{F}$ such that the evaluation map $(f,x) \mapsto f(x)$ is jointly measurable from $(\mathcal{F}, \mathcal{T}) \times (X, \mathcal{B})$ (with product σ-algebra) to $\mathbb{R}$ with Borel sets. Then $\mathcal{T}$ will be called an *admissible structure* for $\mathcal{F}$.

$\mathcal{F}$ will be called *image admissible via* $(Y, \mathcal{S}, T)$ if $(Y, \mathcal{S})$ is a measurable space and T is a function from Y onto $\mathcal{F}$ such that the map $(y,x) \mapsto T(y)(x)$ is jointly measurable from $(Y, \mathcal{S}) \times (X, \mathcal{B})$ with product σ-algebra to $\mathbb{R}$ with Borel sets.

To apply these definitions to a family $\mathcal{C}$ of sets let $\mathcal{F} = \{1_A : A \in \mathcal{C}\}$.

Remarks. For one example, let (K,d) be a compact metric space and let $\mathcal{F}$ be a set of continuous real-valued functions on K, compact for the supremum norm. Then the functions in $\mathcal{F}$ are uniformly equicontinuous on K by the Arzelà-Ascoli theorem (RAP, 2.4.7). It follows

that the map $(f, x) \mapsto f(x)$ is jointly continuous for the supremum norm on $f \in \mathcal{F}$ and d on K. Since both spaces are separable metric spaces, the map is also jointly measurable, so that $\mathcal{F}$ is admissible.

If a family $\mathcal{F}$ is admissible, then it is image admissible, taking T to be the identity. In regard to the converse direction here is an example. Let $X = [0, 1]$ with usual Borel σ-algebra $\mathcal{B}$. Let $(Y, \mathcal{S})$ be a countable product of copies of $(X, \mathcal{B})$. For $y = \{y_n\}_{n=1}^{\infty} \in Y$ let $T(y)(x) := 1_J(x, y)$ where $J := \{(x, y) : \ x = y_n \text{ for some } n\}$. Let $\mathcal{C}$ be the class of all countable subsets of X and $\mathcal{F}$ the class of indicator functions of sets in $\mathcal{C}$. Then it is easy to check that $\mathcal{F}$ is image admissible via $(Y, \mathcal{S}, T)$. If a σ-algebra $\mathcal{T}$ is defined on $\mathcal{F}$ by setting $\mathcal{T} := \{F \subset \mathcal{F} : \ T^{-1}(F) \in \mathcal{S}\}$ then it can be shown that $\mathcal{T}$ is not countably generated (Freedman, 1966, Lemma (5)) although $\mathcal{S}$ is. This example shows how sometimes image admissibility may work better than admissibility.

Theorem 5.8. *For any separable measurable space $(X, \mathcal{B})$, there is a subset Y of $[0, 1]$ and a 1-1 function M from X onto Y which is a measurable isomorphism (is measurable and has measurable inverse) for the Borel σ-algebra on Y.*

Remark. Note that Y is not necessarily a measurable subset of $[0, 1]$. On the other hand if $(X, \mathcal{B})$ is given as a separable metric space which is a Borel subset of its completion, with Borel σ-algebra, then $(X, \mathcal{B})$ is measurably isomorphic either to a countable set, with the σ-algebra of all its subsets, or to all of $[0, 1]$ by the Borel isomorphism theorem (RAP, Theorem 13.1.1).

Let $(X, \mathcal{B})$ be a separable measurable space where $\mathcal{B}$ is generated by a sequence $\{A_i\}$. By taking the union of the finite algebras generated by $A_1, ..., A_n$ for each n, we can and do take $\mathcal{A} := \{A_i\}_{i \geq 1}$ to be an algebra.

Let $\mathcal{F}_0$ be the class of all finite sums $\sum_{i=1}^{n} c_i 1_{A_i}$ for rational $c_i \in \mathbb{R}$ and $n = 1, 2, \cdots$. Then "Borel classes" or "Banach classes" are defined as follows by transfinite recursion (RAP, 1.3.2). Let $(\Omega, <)$ be an uncountable well-ordered set such that for each $\beta \in \Omega$, $\{\alpha \in \Omega : \ \alpha < \beta\}$ is countable. (Specifically, one can take Ω to be the set of all countable ordinals with their usual ordering.) For any countable set $A \subset \Omega$, $\{y : \ y \leq x \text{ for some } x \in A\}$ is countable, so there is a $z \in \Omega$ with $x < z$ for all $x \in A$. For each $\alpha \in \Omega$ there is a next larger element called $\alpha + 1$. Let 0 be the smallest element of Ω. For each $\alpha \in \Omega$, given $\mathcal{F}_\alpha$, let $\mathcal{F}_{\alpha+1}$ be the set of all limits of everywhere pointwise convergent sequences of functions in $\mathcal{F}_\alpha$. If $\beta \in \Omega$ is not of the form $\alpha + 1$ (β is a "limit ordinal"), $\beta > 0$ and $\mathcal{F}_\alpha$ is defined for all $\alpha < \beta$ let $\mathcal{F}_\beta$ be the union of all $\mathcal{F}_\alpha$ for $\alpha < \beta$. Note that $\mathcal{F}_\alpha \subset \mathcal{F}_\beta$ whenever $\alpha < \beta$. Let $U := \bigcup_{\alpha \in \Omega} \mathcal{F}_\alpha$.

Theorem 5.9. *U is the set of all measurable real functions on X.*

On admissibility there is the following main theorem:

Theorem 5.10. (Aumann). *Let $I := [0, 1]$ with usual Borel σ-algebra. Given a separable measurable space $(X, \mathcal{B})$ and a class $\mathcal{F}$ of measurable real-valued functions on X, the following are equivalent:*
(i) $\mathcal{F} \subset \mathcal{F}_\alpha$ for some $\alpha \in \Omega$;
(ii) there is a jointly measurable function $G : \ I \times X \mapsto \mathbb{R}$ such that for each $f \in \mathcal{F}$, $f = G(t, \cdot)$ for some $t \in I$;
(iii) there is a separable admissible structure for $\mathcal{F}$;
(iv) $\mathcal{F}$ is admissible;
(v) $2^{\mathcal{F}}$ is an admissible structure for $\mathcal{F}$;
(vi) $\mathcal{F}$ is image admissible via some $(Y, \mathcal{S}, T)$.

Remarks. The specific classes $\mathcal{F}_\alpha$ depend on the choice of the countable family $\mathcal{A}$ of generators, but condition (i) does not: if $\mathcal{C}$ is another countable set of generators of $\mathcal{B}$ with corresponding classes $\mathcal{G}_\alpha$, then for any $\alpha \in \Omega$ there are $\beta \in \Omega$ and $\gamma \in \Omega$ with $\mathcal{F}_\alpha \subset \mathcal{G}_\beta$ and $\mathcal{G}_\alpha \subset \mathcal{F}_\gamma$.

For $0 \le p < \infty$ and a probability law Q on $(X, \mathcal{B})$ we have the space $\mathcal{L}^p(X, \mathcal{B}, Q)$ of measurable real-valued functions f on X such that $\int |f|^p dQ < \infty$, with the pseudo-metric

$$d_{p,Q}(f,g) := (\int |f-g|^p dQ)^{1/p}, \ 1 \le p < \infty;$$

$$\int |f-g|^p dQ, 0 < p < 1;$$

$$\inf\{\varepsilon > 0 : \ Q(|f-g| > \varepsilon) < \varepsilon\}, \quad p = 0.$$

In admissible classes, $d_{p,Q}$-open sets are measurable, as follows:

Theorem 5.11. *Let $(X, \mathcal{B})$ be a separable measurable space, $0 \le p < \infty$, and $\mathcal{F} \subset \mathcal{L}^p(X, \mathcal{B}, Q)$ where $\mathcal{F}$ is admissible. Then if $\mathcal{F}$ is image admissible via $(Y, \mathcal{S}, T)$, $U \subset \mathcal{F}$ and U is relatively $d_{p,Q}$-open in $\mathcal{F}$, we have $T^{-1}(U) \in \mathcal{S}$.*

Corollary 5.12. *If $\mathcal{F} \subset \mathcal{L}^1(X, \mathcal{B}, Q)$ where $(X, \mathcal{B})$ is a separable measurable space, and $\mathcal{F}$ is image admissible via $(Y, \mathcal{S}, T)$ then $y \mapsto \int T(y) dQ$ is $\mathcal{S}$-measurable.*

Proof. For any real u, $\{f : \ \int f dQ > u\}$ is open for $d_{1,Q}$. □

If $1 \le p < \infty$ and $f, g \in \mathcal{L}^p(X, \mathcal{B}, Q)$ let $\rho_{p,Q}(f,g) := d_{p,Q}(f_{0,Q}, g_{0,Q})$ where for $h \in \mathcal{L}^1(X, \mathcal{B}, Q)$, $h_{0,Q} := h - \int h dQ$. Thus for ρ_Q as defined in Section 3.1, $\rho_Q \equiv \rho_{2,Q}$.

Corollary 5.13. *If $(X, \mathcal{B})$ is a separable measurable space, $1 \le p < \infty$, $\mathcal{F} \subset \mathcal{L}^p(X, \mathcal{B}, Q)$, $\mathcal{F}$ is image admissible via $(Y, \mathcal{S}, T)$, $\mathcal{S}$ is separable, $U \subset \mathcal{F}$ and U is $\rho_{p,Q}$-open, then $T^{-1}(U) \in \mathcal{S}$.*

5.3 Suslin properties, selection, and a counterexample

Here is another counterexample on measurability, to add to the two at the beginning of the chapter. Let $X = [0,1]$ with Borel σ-algebra and uniform (Lebesgue) probability measure $P := U[0,1]$. Let A be a non-Lebesgue measurable subset of $[0,1]$, e.g. RAP, Theorem 3.4.4. Let $\mathcal{C} := \{\{x\} : \ x \in A\}$. Then $\mathcal{C}$ is a collection of disjoint sets, so $S(\mathcal{C}) = 1$ by Theorem 4.10. Also $\mathcal{C}$, being a class of singletons, is admissible, e.g. by Theorem 5.10(ii) with $G(t,s) = 1$ for $t = s$, $G(t,s) = 0$ otherwise. But, $\|P_1\|_\mathcal{C}$ is non-measurable, being 1 if and only if $X_1 \in A$, and likewise any $\|P_n\|_\mathcal{C}$ and $\|P_n - P\|_\mathcal{C}$ is non-measurable. So some measurability condition beyond admissibility is needed for $\|P_n - P\|_\mathcal{F}$ to be measurable. A sufficient condition will be provided by Suslin properties, as follows.

A *Polish space* is a topological space metrizable as a complete separable metric space. A separable measurable space $(Y, \mathcal{S})$ will be called a *Suslin space* iff there is a Polish space X and a Borel measurable map from X onto Y. If $(Y, \mathcal{S})$ is a measurable space, a subset $Z \subset Y$ will be called a *Suslin set* iff it is a Suslin space with the relative σ-algebra $Z \sqcap \mathcal{S}$.

Given a measurable space $(X, \mathcal{B})$ and $M \subset X$, M is called *universally measurable* or *u. m.* iff for every probability law P on $\mathcal{B}$, M is measurable for the completion of P, in other

words for some $A, B \in \mathcal{B}$, $A \subset M \subset B$ and $P(A) = P(B)$. In a Polish space, all Suslin sets are universally measurable (RAP, Theorems 13.2.1 and 13.2.6). A function f from X into Z, where $(Z, \mathcal{A})$ is a measurable space, will be called *universally measurable* or *u. m.* iff for each set $B \in \mathcal{A}$, $f^{-1}(B)$ is universally measurable.

If $(\Omega, \mathcal{A})$ is a measurable space and $\mathcal{F}$ a set, then a real-valued function $X : (f, \omega) \mapsto X(f, \omega)$ will be called *image admissible Suslin via* $(Y, \mathcal{S}, T)$ iff $(Y, \mathcal{S})$ is a Suslin measurable space, T is a function from Y onto $\mathcal{F}$, and $(y, \omega) \mapsto X(T(y), \omega)$ is jointly measurable on $Y \times \Omega$. Equivalently, Y could be taken to be Polish with $\mathcal{S}$ its Borel σ-algebra .

As the notation suggests, a main case of interest will be where $\mathcal{F}$ is a set of functions on Ω and $X(f, \omega) \equiv f(\omega)$. Also, X or $\mathcal{F}$ will be called *image admissible Suslin* if X is image admissible Suslin via some $(Y, \mathcal{S}, T)$ as above.

Recall the notion of separable measurable space defined in the last section. Note that any separable metric space with its Borel σ-algebra is a separable measurable space, as follows from RAP, Proposition 2.1.4. We have:

Theorem 5.14. *A measurable space $(X, \mathcal{B})$, where $\mathcal{B}$ is countably generated, is separable if and only if it separates the points of X, so that for any $x \neq y$ in X, there is some $A \in \mathcal{B}$ containing just one of x, y.*

Theorem 5.15. *Selection Theorem (*Sainte-Beuve*). Let $(\Omega, \mathcal{A})$ be any measurable space and let $X : \mathcal{F} \times \Omega \mapsto \mathbb{R}$ be image admissible Suslin via $(Y, \mathcal{S}, T)$. Then for any Borel set $B \subset \mathbb{R}$,*

$$\Pi_X(B) := \{\omega : X(f, \omega) \in B \text{ for some } f \in \mathcal{F}\}$$

is u. m. in Ω, and there is a u. m. function H from $\Pi_X(B)$ into Y such that $X(T(H(\omega)), \omega) \in B$ for all $\omega \in \Pi_X(B)$.

Note. Here $(\Omega, \mathcal{A})$ need not be Suslin or even separable.

Some possibilities for the set $B \subset \mathbb{R}$ are the sets $\{x : x > t\}$ or $\{x : |x| > t\}$ for any real t. These choices give:

Corollary 5.16. *Let $(f, \omega) \mapsto X(f, \omega)$ be real-valued and image admissible Suslin via some $(Y, \mathcal{S}, T)$. Then $\omega \mapsto \sup\{X(f, \omega) : f \in \mathcal{F}\}$ and $\omega \mapsto \sup\{|X(f, \omega)| : f \in \mathcal{F}\}$ are u. m. functions.*

The image admissible Suslin property is preserved by composing with a measurable function:

Theorem 5.17. *Let $X^1, ..., X^k$ be image admissible Suslin real-valued functions on $\mathcal{F}_i \times \Omega$, $i = 1, ..., k$, for one measurable space $(\Omega, \mathcal{A})$, via $(Y_i, \mathcal{S}_i, T_i)$, $i = 1, ..., k$. Let g be a Borel measurable function from $\mathbb{R}^k$ into $\mathbb{R}$. Then $(\omega, f_1, ..., f_k) \mapsto g(X^1(f_1, \omega), ..., X^k(f_k, \omega))$ is image admissible Suslin via some $(Y, \mathcal{S}, T)$. Specifically, we can let $Y = Y_1 \times ... \times Y_k$ with product σ-algebra $\mathcal{S} = \mathcal{S}_1 \otimes ... \otimes \mathcal{S}_k$ and let $T(y_1, ..., y_k) := (T_1(y_1), ..., T_k(y_k))$.*

Proof. Clearly $(Y, \mathcal{S})$ is Suslin and the joint measurability holds. □

Next, here are examples showing that if the Suslin assumption on Y is removed from Theorem 5.15 it may fail. Let $(\Omega, <)$ and the class $\mathcal{C}$ of countable initial segments I_x be as in the second example at the beginning of this chapter. We can take Ω to be (in 1-1 correspondence

with) a subset of $[0,1]$, by the axiom of choice (or, by the continuum hypothesis, all of $[0,1]$). Here the ordering $<$ has no relation to any usual structure on $[0,1]$. Then $\mathcal{C}$ is admissible by Theorem 5.10, since the collection all countable sets is of bounded Borel class: all finite unions of open intervals (q,r) with rational endpoints are in some $\mathcal{F}_\alpha$, then all finite sets are in $\mathcal{F}_{\alpha+1}$ and all countable sets in $\mathcal{F}_{(\alpha+1)+1} =: \mathcal{F}_{\alpha+2}$.

Let P be a law on Ω which is 0 on countable sets and 1 on sets with countable complement. Such a law, on the σ-algebra generated by singletons, exists on any uncountable set. Under the continuum hypothesis with $\Omega = [0,1]$ we can take P to be Lebesgue measure or any nonatomic law.

Let $P_1 = \delta_x$, $P_2 = (\delta_x + \delta_y)/2$ and $Q_1 = \delta_z$ where x, y, z are coordinates on Ω^3 with law P^3, so x, y and z are i.i.d. (P). Then $\sup_{A\in\mathcal{C}}(P_1 - Q_1)(A) = 1$ if and only if $x < z$. Thus as seen at the beginning of this chapter, $\sup_{A\in\mathcal{C}}(P_1 - Q_1)(A)$ is non-measurable.

If we let $B := \{(x,y,z) : \sup_{A\in\mathcal{C}} |(P_2 - Q_1)(A)| = 1\}$, it can be seen likewise that B must not be measurable. So "Suslin" cannot simply be removed from Corollary 5.16 or Theorem 5.15.

Returning to positive results, an admissible structure can be put on spaces of closed sets. Let (X,d) be a separable metric space and $\mathcal{F}_0$ the collection of all non-empty closed subsets of X. Then $\mathcal{F}_0$ is admissible: there is a countable base for the topology of X, so for some α, all finite unions of sets in the base are in $\mathcal{F}_\alpha$, so all open sets are in $\mathcal{F}_{\alpha+1}$. Then all closed sets are in $\mathcal{F}_{\alpha+2}$ since any closed set F is a countable intersection of open sets

$$U_n := \{x : d(x,F) := \inf_{y\in F} d(x,y) < 1/n\}.$$

The topology of X can be metrized by a metric d for which (X,d) is totally bounded (RAP, Theorem 2.8.2). Assume d is such a metric. Define the Hausdorff metric h_d by

$$h_d(A,B) := \max\{\sup_{x\in A} d(x,B), \sup_{y\in B} d(y,A)\}$$

for any two closed sets A, B.

Since d is totally bounded, it's easily seen that $(\mathcal{F}_0, h_d)$ is separable, since the finite subsets of a countable dense set in X are dense in $\mathcal{F}_0$ for h_d.

The Borel σ-algebra of h_d will be called an *Effros* Borel structure on $\mathcal{F}_0$. (Effros, 1965, proved that for d totally bounded this Borel structure is unique.)

Proposition 5.18. *For any separable metric space X with totally bounded metric d, the Effros Borel structure (of h_d) is admissible on $\mathcal{F}_0$. Also, for any law P on the Borel sets of X, $P(\cdot)$ is measurable for the Effros Borel structure.*

For families of functions, we have the following.

Theorem 5.19. *(a) Let S be a topological space and $\mathcal{F}$ a family of bounded real functions on S, equicontinuous at each point of S. Then $(f,x) \mapsto f(x)$ is jointly continuous $\mathcal{F}\times S \mapsto \mathbb{R}$, with the supremum norm $\|f\|_\infty := \sup_x |f(x)|$ on $\mathcal{F}$.*
(b) If in addition $\mathcal{F}$ is separable for $\|\cdot\|_\infty$ and S is metrizable as a separable metric space (S,d), then $(f,x) \mapsto f(x)$ is jointly measurable for the Borel σ-algebras of $\|\cdot\|_\infty$ on $\mathcal{F}$ and d on S. Thus $\mathcal{F}$ is admissible. So is $\mathcal{G}$, the collection of all subgraphs $\{(x,y) : 0 \le y \le f(x)$ or $f(x) \le y \le 0\}$ for $f \in \mathcal{F}$.
(c) If, moreover, $\mathcal{F}$ with $\|\cdot\|_\infty$ distance and its Borel σ-algebra is a Suslin set, then $\mathcal{F}$ and $\mathcal{G}$ are image admissible Suslin.

Strobl (1994, 1995) and Ziegler (1994, 1997a,b) have given special attention to measurability issues for empirical processes.

Example (Adamski and Gaenssler). Let $H \subset [0,1]$ be a non-measurable set with $\lambda_*(H) = 0$ and $\lambda^*(H) = 1$ where λ is Lebesgue measure (e.g. RAP, Theorem 3.4.4). Then for each Borel set $B \subset [0,1]$, letting $\mu(B \cap H) := \lambda^*(B \cap H)$ defines a countably additive probability measure on the Borel subsets of H, as a metric space with the usual metric from $\mathbb{R}$ (RAP, Theorem 3.3.6). Likewise, λ^* gives a probability measure ν on the Borel sets of $H^c := [0,1] \setminus H$. Take the countable product of probability spaces $(\Omega, \rho) := \Pi_{n=1}^{\infty}(A_n, \mathcal{B}_n, \rho_n)$ where for n odd, $A_n = H$ and $\rho_n = \mu$ while for n even, $A_n = H^c$ and $\rho_n = \nu$. Such a product of probability spaces always exists, e.g. RAP, Theorem 8.2.2. Let X_n be the nth coordinate on Ω, viewed as a map from Ω into $[0,1]$. Then each X_n is measurable and has law $U[0,1]$, the uniform distribution on $[0,1]$. Thus, the X_i are i.i.d. $U[0,1]$. Let $\mathcal{C}$ be the collection of all finite subsets of H. Let P_n be the empirical measures defined by the given X_i. Then $\|P_n\|_{\mathcal{C}} \equiv 1/2$ for n even and $\|P_n\|_{\mathcal{C}} \equiv (n+1)/(2n)$ for n odd. On the other hand if X_i are coordinates on a countable product of copies of $([0,1], \lambda)$, then $\|P_n\|_{\mathcal{C}}$ is non-measurable. This illustrates that $\mathcal{C}$ is a pathological class of sets, but the pathology can be obscured if one doesn't use the standard model.

REFERENCES FOR CHAPTER 5

Aumann, R. J. (1961). Borel structures for function spaces. *Illinois J. Math.* **5**, 614-630.

Bahadur, R. R. (1954). Sufficiency and statistical decision functions. *Ann. Math. Statist.* **25**, 423-462.

Dudley, R. M. (1984). A course on empirical processes. *Ecole d'été de probabilités de St.-Flour*, 1982. *Lecture Notes in Math.* (Springer) **1097**, 1-142.

Durst, Mark, and Dudley, R. M. (1981). Empirical processes, Vapnik- Chervonenkis class and Poisson processes. *Probab. Math. Statist.* (Wroclaw) **1** no. 2, 109-115.

Effros, E. G. (1965). Convergence of closed subsets in a topological space. *Proc. Amer. Math. Soc.* **16**, 929-931.

Ferguson, T. S. (1967). *Mathematical Statistics: A decision theoretic approach.* Academic Press, New York.

Fisher, Ronald Aylmer (1922). IX. On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London* Ser. A **222**, 309-368.

Freedman, David A. (1966). On two equivalence relations between measures. *Ann. Math. Statist.* **37**, 686-689.

Gutmann, S. (1981). Unpublished manuscript.

Halmos, Paul R. and Savage, L. J. (1949). Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Statist.* **20**, 225-241.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, 2d. ed. Repr. 1997, Springer, New York.

Natanson, I. P. (1957). *Theory of Functions of a Real Variable*, vol. 2, 2d ed. Transl. by L. F. Boron, Ungar, New York, 1961.

Neveu, Jacques (1977). Processus ponctuels. *Ecole d'été de probabilités de St.-Flour VI*, 1976, *Lecture Notes Math.* (Springer) **598**, 249-445.

*Neyman, Jerzy (1935). Su un teorema concernente le cosidette statistiche sufficienti.

Giorn. Ist. Ital. Attuari **6**, 320-334.

Rao, B. V. (1971). Borel structures for function spaces. *Colloq. Math.* **23**, 33-38.

Sainte-Beuve, M.-F. (1974). On the extension of von Neumann-Aumann's theorem. *J. Functional Analysis* **17**, 112-129.

Strobl, F. (1994). *Zur Theorie empirischer Prozesse.* Dissertation, Mathematik, Univ. München.

Strobl, F. (1995). On the reversed sub-martingale property of empirical discrepancies in arbitrary sample spaces. *J. Theoret. Probab.* **8**, 825-831.

Ziegler, K. (1994). *On functional central limit theorems and uniform laws of large numbers for sums of independent processes.* Dissertation, Mathematik, Univ. München.

*Ziegler, K. (1997a). A maximal inequality and a functional central limit theorem for set-indexed empirical processes. *Results Math.* **31**, 189-194.

*Ziegler, K. (1997b). On Hoffmann-Jørgensen-type inequalities for outer expectations with applications. *Results Math.* **32**, 179-192.

* - I learned of these papers from secondary sources and have not seen them in the original.

Chapter 6

Limit theorems for Vapnik-Červonenkis and related classes

6.1 Koltchinskii-Pollard entropy and Glivenko-Cantelli theorems

For central limit theorems over Vapnik-Červonenkis and certain related classes some good sufficient conditions were proved by Pollard (1982), using the following form of "entropy" or "capacity".

Let $(X, \mathcal{A})$ be a measurable space and $\mathcal{F} \subset \mathcal{L}^0(X, \mathcal{A})$, the space of all real-valued measurable functions on X. Recall that $F_{\mathcal{F}}(x) := \sup\{|f(x)| : f \in \mathcal{F}\}$ (Section 4.8). Then $F_{\mathcal{F}}(x) \equiv \|\delta_x\|_{\mathcal{F}}$. A measurable function $F \in \mathcal{L}^0(X, \mathcal{A})$ with $F \geq F_{\mathcal{F}}$ will be called an *envelope function* for $\mathcal{F}$. If $F_{\mathcal{F}}$ is $\mathcal{A}$-measurable it will be called *the* envelope function of $\mathcal{F}$. If a law P is given on $(X, \mathcal{A})$, then $F^*_{\mathcal{F}}$ for P will be called *the* envelope function of $\mathcal{F}$ for P, defined up to equality P-a.s.

Let , be the set of all laws on X of the form $n^{-1}\sum_{j=1}^n \delta_{x(j)}$ for some $x(j) \in X$, $j = 1, \ldots, n$, and $n = 1, 2, \ldots$, where the $x(j)$ need not be distinct. For $\delta > 0$, $0 < p < \infty$, and $\gamma \in$, recall (Section 4.8) that if F is an envelope function of $\mathcal{F}$,

$$D_F^{(p)}(\delta, \mathcal{F}, \gamma) := \sup\Big\{m : \text{ for some } f_1, \ldots, f_m \in \mathcal{F}, \text{ and all } i \neq j,$$

$$\int |f_i - f_j|^p \, d\gamma > \delta^p \int F^p d\gamma\Big\}. \tag{6.1}$$

Let $D_F^{(p)}(\delta, \mathcal{F}) := \sup_{\gamma \in \Gamma} D_F^{(p)}(\delta, \mathcal{F}, \gamma)$. Here $D_F^{(p)}(\delta, \mathcal{F}, \gamma)$ is a kind of packing number, involving the envelope function F. The corresponding "entropy" will be the logarithm of $D^{(p)}$. Such logarithms will appear in Section 6.3.

Let $\mathcal{G} := \{f/F\colon\ f \in \mathcal{F}\}$, where $0/0$ is replaced by 0. Then $|g(x)| \leq 1$ for all $g \in \mathcal{G}$ and $x \in X$. Given F, p, and $\gamma \in$, let $Q(B) := \int_B F^p d\gamma / \gamma(F^p)$, if $\gamma(F^p) > 0$. Then $Q := Q_\gamma$ is a law and for $1 \leq p < \infty$, $D_F^{(p)}(\delta, \mathcal{F}, \gamma) = D(\delta, \mathcal{G}, d_{p,Q})$, where (as defined in section 5.2) $d_{p,Q}(f, g) := (\int |f - g|^p \, dQ)^{1/p}$.

For example, if $\mathcal{C}$ is a collection of measurable sets, whose union is all of X, and $\mathcal{F} := \{1_A : A \in \mathcal{C}\}$ the envelope function is $F \equiv 1$ and $D_F^{(p)}(\delta, \mathcal{F}, \gamma) = D(\delta, \mathcal{F}, d_{p,\gamma})$. The next few results will connect $D_F^{(p)}(\delta, \mathcal{F})$ with other ways of measuring the size of certain classes $\mathcal{F}$. First we have Vapnik-Červonenkis classes $\mathcal{C}$ of sets with $\text{dens}(\mathcal{C}) \leq S(\mathcal{C}) < +\infty$ (Corollary 4.4).

Theorem 6.1. *If $\mathcal{C} \subset \mathcal{A}$, $dens(\mathcal{C}) < +\infty$, $1 \leq p < \infty$, $F \in \mathcal{L}^p(X, \mathcal{A}, P)$, $F \geq 0$, and $\mathcal{F} := \{F1_A : A \in \mathcal{C}\}$ then for any $w > dens(\mathcal{C})$ there is a $K < \infty$ such that*

$$D_F^{(p)}(\delta, \mathcal{F}) \leq K\delta^{-pw}, \quad 0 < \delta \leq 1.$$

Next, here is a kind of converse to Theorem 6.1:

Proposition 6.2. *Suppose $1 \leq p < \infty$, $\mathcal{F} = \{1_B : B \in \mathcal{C}\}$ for some collection $\mathcal{C}$ of sets, $F \equiv 1$, and for some δ with $0 < \delta^p < 1/2$, $D_F^{(p)}(\delta, \mathcal{F}) < \infty$. Then $dens(\mathcal{C}) \leq S(\mathcal{C}) < \infty$.*

Next let us consider families of functions of the form $\mathcal{F} = \{Fg : g \in \mathcal{G}\}$ where $\mathcal{G}$ is a family of functions totally bounded in the supremum norm

$$\|g\|_{\sup} := \sup_{x \in X} |g(x)|,$$

with the associated metric $d_{\sup}(g, h) := \|g - h\|_{\sup}$ and with $\|g\|_{\sup} \leq 1$ for all $g \in \mathcal{G}$.

Proposition 6.3. *For $1 \leq p < \infty$ and $0 < \varepsilon \leq 1$ then for any such $\mathcal{F}$,*

$$D_F^{(p)}(\varepsilon, \mathcal{F}) \leq D(\varepsilon, \mathcal{G}, d_{\sup}).$$

Proof. If $d_{\sup}(g, h) \leq \delta$ then for all x, $|Fg - Fh|^p(x) \leq \delta^p F(x)^p$, so the result follows from the definition (6.1). □

Next we come to the Koltchinskii-Pollard method of symmetrization of empirical measures. Given $n = 1, 2, \ldots$, let $x_1, \ldots, x_{2n}$ be coordinates on $(X^{2n}, \mathcal{A}^{2n}, P^{2n})$, hence i.i.d. P. Let $\sigma(1), \ldots, \sigma(n)$ be random variables independent of each other and the x_i with $\Pr(\sigma(i) = 2i) = \Pr(\sigma(i) = 2i - 1) = \frac{1}{2}$, $i = 1, \ldots, n$. Let $\tau(i) = 2i$ if $\sigma(i) = 2i - 1$ and $\tau(i) = 2i - 1$ if $\sigma(i) = 2i$. Let $x(i) := x_i$. Then the $x(\sigma(j))$ are i.i.d. P. Let

$$P_n' := n^{-1} \sum_{j=1}^n \delta_{x(\sigma(j))}, \quad P_n'' := n^{-1} \sum_{j=1}^n \delta_{x(\tau(j))},$$

$$\nu_n' := n^{1/2}(P_n' - P), \quad \nu_n'' := n^{1/2}(P_n'' - P),$$

$$P_n^0 := P_n' - P_n'', \quad \nu_n^0 := n^{1/2} P_n^0.$$

Note that $P_n'' = 2P_{2n} - P_n'$ and that ν_n' and ν_n'' are two independent copies of ν_n. Here is a symmetrization fact:

Lemma 6.4. *(Symmetrization) Let $\zeta > 0$ and $\mathcal{F} \subset \mathcal{L}^2(X, \mathcal{A}, P)$ with $\int |f|^2 dP \leq \zeta^2$ for all $f \in \mathcal{F}$. Assume $\mathcal{F}$ is image admissible Suslin via some $(Y, \mathcal{S}, T)$. Then for any $\eta > 0$*

$$\Pr\{\|\nu_n^0\|_{\mathcal{F}} > \eta\} \geq (1 - \zeta^2\eta^{-2}) \Pr\{\|\nu_n\|_{\mathcal{F}} > 2\eta\}.$$

Some reverse martingale and submartingale properties of the empirical measures P_n will be stated. Recall that $Q(f) := \int f\, dQ$ for any $f \in \mathcal{L}^1(Q)$, and that in defining empirical measures $P_n := \frac{1}{n}\sum_{j=1}^n \delta_{X_j}$, the X_j are always (in these notes) taken as coordinates on a product of copies of a probability space $(X, \mathcal{A}, P)$, so that the underlying probability measure for P_n is P^n, a product of n copies of P.

The definitions of (reversed) (sub)martingale given e.g. in RAP, Sections 10.3 and 10.6, for random variables Y_n and σ-algebras $\mathcal{B}_n$, will be slightly extended here by allowing Y_n to be measurable for the completion of $\mathcal{B}_n$.

Theorem 6.5. *Let $(X, \mathcal{A}, P)$ be a probability space, $\mathcal{F} \subset \mathcal{L}^1(P)$, and P_n empirical measures for P. Let $\mathcal{S}_n$ be the smallest σ-algebra for which $P_k(f)$ are measurable for all $k \geq n$ and $f \in \mathcal{L}^1(X, \mathcal{A}, P)$. Then: (a) For any $f \in \mathcal{F}$, $\{P_n(f), \mathcal{S}_n\}_{n\geq 1}$ is a reversed martingale, in other words*

$$E\left(P_{n-1}(f)|\mathcal{S}_n\right) = P_n(f) \text{ a.s., if } n \geq 2.$$

(b) (F. Strobl) *Suppose that $\mathcal{F}$ has an envelope function $F \in \mathcal{L}^1(X, \mathcal{A}, P)$ and that for each n, $\|P_n - P\|_{\mathcal{F}}$ is measurable for the completion of P^n. Then $\left(\|P_n - P\|_{\mathcal{F}}, \mathcal{S}_n\right)_{n\geq 1}$ is a reversed submartingale, in other words*

$\|P_n - P\|_{\mathcal{F}} \leq E(\|P_k - P\|_{\mathcal{F}}|\mathcal{S}_n)$ *a.s. for $k \leq n$.*

Remark. $\|P_n - P\|_{\mathcal{F}}$ will be completion measurable if $\mathcal{F}$ is image admissible Suslin, by Corollaries 5.16 and 5.12.

Here is a law of large numbers (generalized Glivenko-Cantelli theorem):

Theorem 6.6. *Let $(X, \mathcal{A}, P)$ be a probability space, $F \in \mathcal{L}^1(X, \mathcal{A}, P)$, and $\mathcal{F}$ a collection of measurable functions on X having F as an envelope function. Suppose $\mathcal{F}$ is image admissible Suslin via $(Y, \mathcal{S}, T)$. Assume that*

$$D_F^{(1)}(\delta, \mathcal{F}) < \infty \quad \text{for all} \quad \delta > 0. \tag{6.2}$$

Then $\lim_{n\to\infty} \|P_n - P\|_{\mathcal{F}} = 0$ a.s.

Corollary 6.7. *Let $(X, \mathcal{A})$ be a measurable space and $\mathcal{C} \subset \mathcal{A}$ where $S(\mathcal{C}) < \infty$ and $\mathcal{C}$ is image admissible Suslin. Then for any probability law P on $\mathcal{A}$, we have*

$$\lim_{n\to\infty} \sup_{A\in\mathcal{C}} |(P_n - P)(A)| \;=\; 0 \text{ a.s.}$$

Proof. In Theorem 6.6, take $F \equiv 1$ and apply Theorem 6.1 with $p = 1$. □

6.2 Vapnik-Červonenkis-Steele laws of large numbers.

Let $(X, \mathcal{A}, P)$ be a probability space and $\mathcal{C} \subset \mathcal{A}$. Let $\{x_n\}_{n\geq 1}$ be coordinates in $(X^\infty, \mathcal{A}^\infty, P^\infty)$ so that x_j are i.i.d. P. For certain classes $\mathcal{C}$ with $S(\mathcal{C}) = +\infty$ one will have $\Delta^{\mathcal{C}}(\{x_1, \ldots, x_n\}) < 2^n$ with P^n-probability converging to 1. For such classes, with sufficient measurability properties, a law of large numbers will still hold. Here a main result is as follows:

Theorem 6.8. *If $(X, \mathcal{A}, P)$ is any probability space, $\mathcal{C} \subset \mathcal{A}$, and $\mathcal{C}$ is image admissible Suslin, then the following are equivalent:*
(a) $\|P_n - P\|_{\mathcal{C}} \to 0$ a.s. as $n \to \infty$;
(b) $\|P_n - P\|_{\mathcal{C}} \to 0$ in probability as $n \to \infty$;
(c) $\lim_{n\to\infty} n^{-1} E \log \Delta^{\mathcal{C}}(\{x_1, \ldots, x_n\}) = 0$.

For a finite set $F \subset X$ and collection $\mathcal{C} \subset 2^X$, let $k^{\mathcal{C}}(F) := S(\mathcal{C} \sqcap F)$.

Lemma 6.9. *Under the hypotheses of Theorem 6.8, $\Delta^{\mathcal{C}}(x_1, \cdots, x_n)$ and $k^{\mathcal{C}}(\{x_1, \cdots, x_n\})$ are universally measurable.*

The main step in Steele's proof uses Kingman's subadditive ergodic theorem (the equivalent superadditive ergodic theorem is RAP, Theorem 10.7.1). To state it, here is some terminology. Let $\mathbb{N}$ denote the set of nonnegative integers. A *subadditive process* is a doubly indexed set $\{x_{mn}\}_{0 \le m < n < \infty}$ of real random variables, m, $n \in \mathbb{N}$, $m < n$, such that

$$x_{kn} \le x_{km} + x_{mn} \quad \text{whenever} \quad k < m < n. \tag{6.3}$$

Let $x_{nn} := 0$ for all $n \in \mathbb{N}$. If instead of (6.3),

$$x_{kn} \ge x_{km} + x_{mn}, \quad k < m < n, \tag{6.4}$$

then $\{x_{mn}\}_{0 \le m < n}$ is called *superadditive*.

A process which is both subadditive and superadditive is called *additive* and can clearly be written as $x_{kn} = \sum_{k < j \le n} x_j$, where $x_j := x_{j-1,j}$, i.e., one has just partial sums of a sequence of random variables.

A subadditive process $\{x_{mn}\}_{0 \le m < n}$ defined on a probability space $(\Omega, \mathcal{B}, \Pr)$ will be called *stationary* if there is a measure-preserving transformation V of Ω onto itself such that for any integers $0 \le m < n$, $x_{mn}(V(\omega)) = x_{m+1,n+1}(\omega)$. Recall that $(f \circ g)(x) := f(g(x))$. Let $V^k := V \circ (V \circ (\cdots \circ V) \cdots)$ to k terms. Then for $k = 1, 2, \cdots$, $x_{mn} \circ V^k = x_{m+k,n+k}$. Let $\mathcal{S}$ be the σ-algebra of all $B \in \mathcal{B}$ such that $V^{-1}(B) = B$.

Another useful hypothesis for subadditive processes is:

$$\text{For each } n \in \mathbb{N},\ E|x_{0n}| < +\infty, \text{ and } \kappa := \inf_{n \ge 1} E x_{0n}/n > -\infty. \tag{6.5}$$

A σ-algebra $\mathcal{D} \subset \mathcal{B}$ will be called *degenerate* if $\Pr(D) = 0$ or 1 for all $D \in \mathcal{D}$.

Theorem 6.10. *(Kingman's subadditive ergodic theorem). Let $\{x_{mn}\}_{0 \le m < n}$ be a stationary subadditive process satisfying (6.5). Then as $n \to \infty$, x_{0n}/n converges a.s. and in $\mathcal{L}^1$ to a random variable $y := \lim_{n\to\infty} n^{-1} E(x_{0n}|\mathcal{S})$ with $Ey = \kappa$. If $\mathcal{S}$ is degenerate, then $y = \kappa$ a.s.*

Proof. RAP, Theorem 10.7.1 applies with f_n there defined as $-x_{0n}$. From near the end of the proof (RAP, p. 296), we have $Ey \le E x_{0n}/n$ for all n and $E x_{0n}/n \to Ey$ as $n \to \infty$, so $Ey = \kappa$. □

To apply Theorem 6.10 in proving Theorem 6.8 we have:

Theorem 6.11. *Let $(\Omega, \mathcal{T}, Pr)$ be a probability space, $(X, \mathcal{A})$ a measurable space, X_1 a measurable function from Ω into X, and V a measure-preserving transformation of Ω onto itself. Let $X_j := X_1 \circ V^{j-1}$ for $j = 2, 3, \cdots$. Let $\mathcal{C}$ be image admissible Suslin, $\emptyset \neq \mathcal{C} \subset \mathcal{A}$. Then each of the following is a stationary subadditive process satisfying (6.5):*

(a) $D_{mn}^{\mathcal{C}} := \sup_{A \in \mathcal{C}} \left| \sum_{m<i\leq n} (1_A(X_i) - P(A)) \right|$;

(b) $\log \Delta_{mn}^{\mathcal{C}} := \log \Delta^{\mathcal{C}} (\{X_{m+1}, \ldots, X_n\})$;

(c) $k_{mn}^{\mathcal{C}} := k^{\mathcal{C}} (\{X_{m+1}, \ldots, X_n\})$.

Proof. We have measurability by Corollary 5.16 in (a) and Lemma 6.9 in (b) and (c). Stationarity clearly holds for the same V in each case. Subadditivity is clear in (a) and not difficult for (b) and (c). All three processes are nonnegative: in (b), $\mathcal{C}$ non-empty implies $\Delta^{\mathcal{C}} \geq 1$, so (6.5) holds. □

6.3 Pollard's central limit theorem

By way of the Koltchinskii-Pollard kind of entropy and law of large numbers (Section 6.1 above) the following can be proved:

Theorem 6.12. (Pollard) *Let $(X, \mathcal{A}, P)$ be a probability space and $\mathcal{F} \subset \mathcal{L}^2(X, \mathcal{A}, P)$. Let $\mathcal{F}$ be image admissible Suslin via $(Y, \mathcal{S}, T)$ and have an envelope function $F \in \mathcal{L}^2(X, \mathcal{A}, P)$. Suppose that*

$$\int_0^1 \left(\log D_F^{(2)}(x, \mathcal{F}) \right)^{1/2} dx < \infty. \tag{6.6}$$

Then $\mathcal{F}$ is a Donsker class for P.

Here is a consequence:

Theorem 6.13. (Jain and Marcus). *Let (K, d) be a compact metric space. Let $C(K)$ be the space of continuous real functions on K with supremum norm. Let $X_1, X_2, \ldots$ be i.i.d. random variables in $C(K)$. Suppose $EX_1(t) = 0$ and $EX_1(t)^2 < \infty$ for all $t \in K$. Assume that for some random variable M with $EM^2 < \infty$,*

$$|X_1(s) - X_1(t)|(\omega) \leq M(\omega) d(s, t) \quad \text{for all} \quad \omega \quad \text{and} \quad s, t \in K.$$

Suppose that

$$\int_0^1 (\log D(\varepsilon, K, d))^{1/2} \, d\varepsilon < \infty.$$

Then the central limit theorem holds, in other words, in $C(K)$, $\mathcal{L}\left(n^{-1/2}(X_1 + \cdots + X_n)\right)$ converges to some Gaussian law.

Remark. In the situation of Theorem 6.13, K may be given originally with a metric e. The metric d may be chosen, perhaps as a function $d = f(e)$, where $f(x)$ may approach 0 slowly as $x \downarrow 0$, e.g., $f(x) = x^{\varepsilon}$ for $\varepsilon > 0$ or $f(x) = 1/\max(|\log x|, 2)$. Thus one can increase the possibilities for obtaining the Lipschitz property of X_1 with respect to d, so long as (6.6) holds for d.

Proposition 6.14. *Let $(X, \mathcal{A}, P)$ be a probability space. Suppose a class $\mathcal{F}$ of measurable real-valued functions on X has an envelope function $F \in \mathcal{L}^2(X, \mathcal{A}, P)$. If $D_F^{(2)}(\delta, \mathcal{F}) < \infty$ for all $\delta > 0$, then $\mathcal{F}$ is totally bounded in $\mathcal{L}^2(X, \mathcal{A}, P)$.*

Corollary 6.15. (Pollard). *Let $(X, \mathcal{A}, P)$ be a probability space, and let $\mathcal{F}$ be an image admissible Suslin Vapnik-Červonenkis subgraph class of functions with envelope $F \in \mathcal{L}^2(X, \mathcal{A}, P)$. Then $\mathcal{F}$ is a Donsker class for P.*

Proof. This follows from Theorem 6.12 and Theorem 4.51 for $p = 2$. □

Corollary 6.16. *Let $(X, \mathcal{A}, P)$ be a probability space, $F \in \mathcal{L}^2(X, \mathcal{A}, P)$, and $\mathcal{F} = \{F1_C : C \in \mathcal{C}\}$ where $\mathcal{C}$ is an image admissible Suslin Vapnik-Červonenkis class of sets. Then $\mathcal{F}$ is a Donsker class for P.*

Proof. Since F is measurable, the image admissible Suslin property of $\mathcal{F}$ follows from that of $\mathcal{C}$. By Theorem 6.1 for $p = 2$, (6.6) holds and Theorem 6.12 applies. □

6.4 Necessary conditions for limit theorems

Theorems 6.1 and 6.12 imply that every class $\mathcal{C} \subset \mathcal{A}$ with $S(\mathcal{C}) < +\infty$, and which is image admissible Suslin, is a Donsker class, for an *arbitrary law* P on $\mathcal{A}$. In this section it will be seen that to obtain, for all P, such a central limit theorem (or even the pregaussian property), for a class $\mathcal{C}$ of sets, the condition $S(\mathcal{C}) < +\infty$ is necessary. Then it will be noted that some measurability, beyond that of $\|P_n - P\|_{\mathcal{C}}$, is needed to obtain even a law of large numbers for $S(\mathcal{C}) < +\infty$. Lastly, it will be seen that $S(\mathcal{C}) < \infty$ is necessary so that $\|P_n - P\|_{\mathcal{C}} \to 0$ in outer probability as $n \to \infty$, uniformly in P.

Theorem 6.17. *Let $(X, \mathcal{A})$ be a measurable space and $\mathcal{C} \subset \mathcal{A}$. Suppose that for all laws P on $\mathcal{A}$, $\{1_A : A \in \mathcal{C}\}$ is a pregaussian class (as defined in Section 3.1). Then $S(\mathcal{C}) < \infty$.*

Remark. Now let us recall the example at the beginning of Chapter 5, where $\mathcal{C}$ is the collection of all countable initial segments of an uncountable well-ordered set $(X, <)$ and P is a continuous law on some σ-algebra $\mathcal{A}$ containing all countable subsets of X. Then $S(\mathcal{C}) = 1$ but $\sup_{A \in \mathcal{C}} |(P_n - P)(A)| \equiv 1$ for all n. Thus the latter random variable is measurable. For this class the weak law of large numbers, hence the strong law and central limit theorem, all fail as badly as possible. This shows that in Theorem 6.6 and Corollary 6.7, the "image admissible Suslin" condition cannot simply be removed, nor replaced by simple measurability of random variables appearing in the statements of the results. Further, for all $A \in \mathcal{C}$, $1_A = 0$ a.s. P, so vanishing a.s. (P) even with $S(\mathcal{C}) = 1$ does not imply a law of large numbers.

Remark. If X is a countably infinite set and $\mathcal{A} = 2^X$, then for an arbitrary law P on $\mathcal{A}$, $\lim_{n \to \infty} \sup_{A \in \mathcal{A}} |(P_n - P)(A)| = 0$ a.s., but $S(\mathcal{A}) = +\infty$, so the hypothesis of Theorem 6.17 cannot be weakened to a law of large numbers for all P.

Next it will be seen that the Vapnik-Červonenkis property is also necessary for a law of large numbers to hold uniformly in P or that there exist an estimator of P based on $X_1, \cdots, X_n$

(which might or might not equal P_n) converging to P uniformly over $\mathcal{C}$ and uniformly in P. Here are some definitions.

Let $(X, \mathcal{B})$ be a measurable space. Let its n-fold Cartesian product be $(X^n, \mathcal{B}^n)$. Let $\mathcal{P}$ be the class of all probability measures on $(X, \mathcal{B})$. Let $\mathcal{C} \subset \mathcal{B}$ be any collection of measurable sets. A real-valued function T_n on $X^n \times \mathcal{C}$ will be called a *$\mathcal{C}$-estimator* if it is a stochastic process indexed by $\mathcal{C}$, in other words, for each $A \in \mathcal{C}$, $x \mapsto T_n(x, A)$ is measurable on X^n.

A $\mathcal{C}$-estimator T_n will be called an *estimator* if for each x, there is a probability measure μ on $\mathcal{A}$ which equals $T_n(x, \cdot)$ on $\mathcal{C}$.

For any probability measure P on $(X, \mathcal{B})$ and product law P^n on $(X^n, \mathcal{B}^n)$, for $x = (X_1, \cdots, X_n)$ so that X_i are i.i.d. (P), we would like T_n to be a good approximation to P with probability $\to 1$ as $n \to \infty$. The goodness of the approximation will be measured by the *loss function* $L(T_n, P) := \|T_n - P\|_{\mathcal{C}}$. From it we get the *risk* $r(T_n, P, \mathcal{C}) := E_P L(T_n, P)^*$ where E_P denotes expectation with respect to P^n. For any class $\mathcal{Q}$ of laws, let $r(T_n, \mathcal{Q}, \mathcal{C}) := \sup\{r(T_n, P, \mathcal{C}) : P \in \mathcal{Q}\}$, and let $r_n(\mathcal{Q}, \mathcal{C})$ be the *minimax risk*, i.e. the infimum of $r(T_n, \mathcal{Q}, \mathcal{C})$ over all $\mathcal{C}$-estimators T_n.

In finding minimax risks for $\mathcal{C}$-estimators we can assume T_n takes values in $[0, 1]$ since $\max(0, \min(T_n, 1))$ will clearly have risks no larger than those of T_n.

For any $\mathcal{C}$ and $\mathcal{Q}$, clearly $0 \le r_n(\mathcal{Q}, \mathcal{C}) \le 1$ and r_n is non-increasing in n. The following theorem holds for $\mathcal{C}$-estimators and so *a fortiori* for estimators, whose values are probability measures. The following fact is mainly due to P. Assouad:

Theorem 6.18. *Let $\mathcal{P}$ be the class of all probability measures on a sample space $(X, \mathcal{B})$ and $\mathcal{C} \subset \mathcal{B}$. If the minimax risk $r_n(\mathcal{P}, \mathcal{C}) < 1/2$ for some n, then $\mathcal{C}$ is a Vapnik-Červonenkis class.*

A corollary of Theorem 6.18, taking $T_n = P_n$, is:

Theorem 6.19. (Assouad) *If $(X, \mathcal{B})$ is a measurable space and $\mathcal{C} \subset \mathcal{B}$ is a uniform Glivenko-Cantelli class of sets, that is, $\sup_P E_P \|P_n - P\|_{\mathcal{C}}^* \to 0$ as $n \to \infty$, where the supremum is over all probability laws on $(X, \mathcal{B})$, then $\mathcal{C}$ is a Vapnik-Červonenkis class.*

Now we'll see that the constant $1/2$ in Theorem 6.18 is sharp:

Proposition 6.20. *For the class $\mathcal{P}$ of all probability measures on a sample space $(X, \mathcal{B})$ there is always a $\mathcal{B}$-estimator T, not depending on n or x, with $r(T, \mathcal{P}, \mathcal{B}) \le 1/2$, so that for any $\mathcal{Q} \subset \mathcal{P}$, $\mathcal{C} \subset \mathcal{B}$ and n we have $r_n(\mathcal{Q}, \mathcal{C}) \le 1/2$. Moreover when $(X, \mathcal{B})$ is the unit interval $[0, 1]$ with the Borel σ-algebra, there is a class $\mathcal{C}$ which is not a Vapnik-Červonenkis class, and for which T on $\mathcal{C}$ is given by a probability measure, so T is an estimator, not only a $\mathcal{C}$-estimator.*

Proof. A $\mathcal{C}$-estimator T (not depending on n or x) is defined by $T(x, A) \equiv 1/2$ for all $A \in \mathcal{B}$....

Next, it will be seen that one of the hypotheses of Theorem 6.6, existence of an integrable envelope function, is essentially necessary for a law of large numbers (Glivenko-Cantelli property). This is related to the fact that that for i.i.d. real random variables $Y_1, \cdots, Y_n, \cdots$, $(Y_1 + \cdots Y_n)/n$ converges a.s. to a finite limit if and only if $E|Y_1| < \infty$ (RAP, Theorem 8.3.5).

Theorem 6.21. *If $(S, \mathcal{B}, P)$ is a probability space, $\mathcal{F} \subset \mathcal{L}^1(S, \mathcal{B}, P)$, $\|P\|_{\mathcal{F}} < \infty$, and if $\mathcal{F}$ is a strong Glivenko-Cantelli class for P, i.e. $\|P_n - P\|_{\mathcal{F}}^* \to 0$ a.s., then $\mathcal{F}$ has an integrable envelope function: $E\|P_1\|_{\mathcal{F}}^* < \infty$.*

REFERENCES

Adler, R. J. (1990). *An Introduction to Continuity, Extrema and Related Topics for General Gaussian Processes. IMS Lecture Note and Monograph Series* **12**.

Adler, R. J. and Brown, L. D. (1986). Tail behaviour for suprema of empirical processes. *Ann. Probab.* **14**, 1-30.

Adler, R. J. and Samorodnitsky, G. (1987). Tail behaviour for the suprema of Gaussian processes with applications to empirical processes. *Ann. Probab.* **15**, 1339-1351.

Alexander, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.* **12**, 1041-1067; Correction **15** (1987), 428-430.

Alon, N., Ben-David, S., Cesa-Bianchi, N., and Haussler, D. (1997). Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM* **44**, 615-631.

Assouad, P. (1982). Classes de Vapnik-Červonenkis et vitesse d'estimation (unpublished manuscript)

Assouad, P. (1985). Observations sur les classes de Vapnik-Červonenkis et la dimension combinatoire de Blei. In *Séminaire d'analyse harmonique, 1983-84, Publ. Math. Orsay*, Univ. Paris XI, Orsay, 92-112.

Assouad, P., and Dudley, R. M. (1990). Minimax nonparametric estimation over classes of sets (unpublished manuscript).

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. (1989). Learnability and the Vapnik-Chervonenkis dimension. *J. ACM* **36**, 929-965.

Cabaña, E. M. (1984). On the transition density of multidimensional parameter Wiener process with one barrier. *J. Applied Prob.* **21**, 197-200.

Cabaña, E. M., and Wschebor, M. (1982). The two-parameter Brownian bridge: Kolmogorov inequalities and upper and lower bounds for the distribution of the maximum. *Ann. Probab.* **10**, 289-302.

Devroye, L. (1982). Bounds for the uniform deviation of empirical measures. *J. Multivar. Analysis* **12**, 72-79.

Devroye, L., Gyorfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, New York

Dudley, R. M. (1984). A course on empirical processes. In *Ecole d'été de probabilités de St.-Flour, 1982. Lecture Notes in Math.* (Springer) **1097**, 1-142.

Dudley, R. M., Giné, E., and Zinn, J. (1991). Uniform and universal Glivenko-Cantelli classes. *J. Theoret. Probab.* **4**, 485-510.

Dudley, R. M., Kulkarni, S. R., Richardson, T., and Zeitouni, O. (1994). A metric entropy bound is not sufficient for learnabiity. *IEEE Trans. Inform. Theory* **40**, 883-885.

Durst, Mark, and Dudley, R. M. (1981). Empirical processes, Vapnik-Chervonenkis Classes and Poisson processes. *Probab. Math. Statist. (Wrocław)* **1**, no. 2, 109-115.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **27**, 642-669.

Giné, E., and Zinn, J. (1986). Lectures on the central limit theorem for empirical processes. In *Probability and Banach Spaces* (Zaragoza, 1985), *Lecture Notes in Math.* (Springer) **1221**, 50-113.

Goodman, V. (1976). Distribution estimates for functionals of the two-parameter Wiener process. *Ann. Probab.* **4**, 977-982.

Haussler, David (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications *Inform. and Comput.* **100**, 78-150.

Jain, Naresh, and Marcus, Michael B. (1975). Central limit theorems for $C(S)$-valued random variables. *J. Funct. Analysis* **19**, 216-231.

Koltchinskii, V. I. (1981). On the central limit theorem for empirical measures. *Theor. Probab. Math. Statist.* **24**, 71-82. Transl. from *Teor. Verojatnost. i Mat. Statist. (1981)* **24**, 63-75.

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes.* Springer, New York

Ledoux, M. (1996). Isoperimetry and Gaussian analysis. In *Ecole d'été de probabilités de St.-Flour, 1994*, *Lecture Notes in Math.*
(Springer) **1648**, 165-294.

Massart, P. (1983). Vitesses de convergence dans le théorème central limite pour des processus empiriques. *C. R. Acad. Sci. Paris Sér. I* **296**, 937-940.

Massart, P. (1986). Rates of convergence in the central limit theorem for empirical processes. *Ann. Inst. H. Poincaré Probab.-Statist.* **22**, 381-423. See also *Lecture Notes in Math.* (Springer) **1193**, 73-109.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* **18**, 1269-1283.

Pickands, James III (1969). Upcrossing probabilities for stationary Gaussian processes. *Trans. Amer. Math. Soc.* **145**, 51-73.

Pollard, David B. (1982). A central limit theorem for empirical processes. *J. Austral. Math. Soc. Ser. A* **33**, 235-248.

Samorodnitsky, G. (1991). Probability tails of Gaussian extrema. *Stochastic Process. Appl.* **38**, 55-84.

Shortt, Rae M. (1984). Universally measurable spaces: an invariance theorem and diverse characterizations. *Fund. Math.* **121**, 169-176.

Smith, D. L., and Dudley, R. M. (1992). Exponential bounds in Vapnik-Červonenkis classes of index 1. In *Probability in Banach Spaces* **8**, ed. R. M. Dudley, M. G. Hahn, J. Kuelbs, Birkhäuser, Boston 451-465.

Steele, J. Michael (1978). Empirical discrepancies and subadditive processes. *Ann. Probab.* **6**, 118-127.

Strobl, F. (1995). On the reversed sub-martingale property of empirical discrepancies in arbitrary sample spaces. *J. Theoret. Probab.* **8**, 825-831.

Talagrand, M. (1987a). The Glivenko-Cantelli problem. *Ann. Probab.* **15**, 837-870.

Talagrand, M. (1987b). Donsker classes and random geometry. *Ann. Probab.* **15**, 1327-1338.

Talagrand, M. (1988). Donsker classes of sets. *Probab. Theory Rel. Fields* **78**, 169-191.

Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann. Probab.* **22**, 28-76.

Talagrand, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publs. I.H.E.S.* **81**, 73-205.

Talagrand, M. (1996a). New concentration inequalities in product spaces. *Invent. Math.* **126**, 505-563.

Talagrand, M. (1996b). The Glivenko-Cantelli problem, ten years later. *J. Theoret. Probab.* **9**, 371-384.

Valiant, L. G. (1984). A theory of the learnable. *C. ACM* **27**, 1134-1142.

Vapnik, V. N. (1982). *Estimation of Dependences Based on Empirical Data.* Springer, New York

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* Springer, New York

Vapnik, V. N., and Červonenkis, A. Ya. (1968). Uniform convergence of frequencies of occurrence of events to their probabilities. *Dokl. Akad. Nauk SSSR* **181**, 781-783. Engl. transl. *Sov. Math. Doklady* **9**, 915-918.

Vapnik, V. N., and Červonenkis, A. Ya. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16**, 264-280.

Vapnik, V. N., and Červonenkis, A. Ya. (1974). *Teoriya Raspoznavaniya Obrazov* (Theory of Pattern Recognition) (in Russian). Nauka, Moscow. German ed.: *Theorie der Zeichenerkennung*, by W. N. Wapnik and A. Ja. Tscherwonenkis, transl. by K. G. Stöckel and B. Schneider, ed. S. Unger and K. Fritzsch. Akademie-Verlag, Berlin, 1979 (*Elektronisches Rechnen und Regeln*, Sonderband).

Vapnik, V. N., and Červonenkis, A. Ya. (1981). Necessary and sufficient conditions for the uniform convergence means to their expectations. *Theory Probab. Appl.* **26**, 532-553.

Wolfowitz, J. (1954). Generalization of the theorem of Glivenko-Cantelli. *Ann. Math. Statist.* **25**, 131-138.

Chapter 7

Metric Entropy, With Inclusion And Bracketing

7.1 Definitions and the Blum-DeHardt law of large numbers

Definitions. Given a measurable space $(A, \mathcal{A})$, let $\mathcal{L}^0(A, \mathcal{A})$ denote the set of all real-valued $\mathcal{A}$-measurable functions on A. Given $f, g \in \mathcal{L}^0(A, \mathcal{A})$ let $[f, g] := \{h \in \mathcal{L}^0(A, \mathcal{A}) : f \le h \le g\}$ (empty unless $f \le g$). A set $[f, g]$ will be called a *bracket*. Given a probability space $(A, \mathcal{A}, P)$, $1 \le q \le \infty$, $\mathcal{F} \subset \mathcal{L}^q(A, \mathcal{A}, P)$ with usual seminorm $\|\cdot\|_q$, and $\epsilon > 0$, let $N_{[\,]}^{(q)}(\epsilon, \mathcal{F}, P)$ denote the smallest m such that for some $f_1, \ldots, f_m$ and $g_1, \cdots, g_m$ in $\mathcal{L}^q(A, \mathcal{A}, P)$, with $\|g_i - f_i\|_q \le \epsilon$ for $i = 1, \cdots, m$,

$$\mathcal{F} \subset \bigcup_{i=1}^{m} [f_i, g_i]\,. \tag{7.1}$$

Here $\log N_{[\,]}^{(q)}(\epsilon, \mathcal{F}, P)$ will be called a *metric entropy with bracketing*.

Note that the f_j and g_j are not required to be in $\mathcal{F}$. For example, if $\mathcal{F}$ is the set of indicators of half-planes in $\mathbb{R}^2$, then $f \le h \le g$ for f, g, h in $\mathcal{F}$ would require the boundary lines of all three half-planes to be parallel. If instead we let f be the indicator of an intersection of two half-planes and g that of a union, then there can be a non-degenerate set of $h \in \mathcal{F}$ with $f \le h \le g$.

Also note that an individual bracket $[f, g]$ has the envelope function $\max(-f, g) = \max(|f|, |g|)$ and so if (7.1) holds, for some ε, then $\mathcal{F}$ has an envelope function given by $\max_{1 \le j \le m} \max(-f_j, g_j)$. The set of all differences $h - H$ for h and H in $[f, g]$ has an envelope function $g - f$. So, in this chapter, unlike the last, envelope functions will not be singled out for special attention.

If $\mathcal{F} \subset \mathcal{L}^r$ then for $q \le r \le \infty$, $\mathcal{F} \subset \mathcal{L}^q$ and

$$N_{[\,]}^{(q)}(\epsilon, \mathcal{F}, P) \le N_{[\,]}^{(r)}(\epsilon, \mathcal{F}, P) \quad \text{for all} \quad \epsilon > 0. \tag{7.2}$$

For $r = \infty$, more is true. Let $d_{\sup}(f, g) := \sup_x |(f - g)(x)|$, $\|f\|_{\sup} := d_{\sup}(f, 0)$. It is easily seen using brackets $[f_j - \epsilon, f_j + \epsilon]$ that for any law P,

$$N_{[\,]}^{(\infty)}(2\epsilon, \mathcal{F}, P) \le D(\epsilon, \mathcal{F}, d_{\sup})\,. \tag{7.3}$$

Thus, for example, a set $\mathcal{F}$ of continuous functions, totally bounded in the usual supremum norm with given bounds $D(\epsilon, \mathcal{F}, d_{\sup})$ will have the same bounds on all $N_{[\,]}^{(q)}(2\epsilon, \mathcal{F}, P)$, $1 \leq q \leq \infty$.

If $\mathcal{F}$ consists of indicator functions of measurable sets then in finding brackets $[f_i, g_i]$ to cover $\mathcal{F}$, it is no loss to assume $0 \leq f_i \leq g_i \leq 1$ for all i. Next, if $C(i) := \{x : f_i(x) > 0\}$, $D(i) := \{x : g_i(x) = 1\}$, and $f_i \leq 1_C \leq g_i$ then

$$f_i \;\leq\; 1_{C(i)} \;\leq\; 1_C \;\leq\; 1_{D(i)} \;\leq\; g_i.$$

So, $[f_i, g_i]$ can be replaced by $[1_{C(i)}, 1_{D(i)}]$. If $\mathcal{C}$ is a collection of measurable sets and $\epsilon > 0$, let $N_I(\epsilon, \mathcal{C}, P) := \inf\{m :$ for some $C_1, \ldots, C_m$ and $D_1, \cdots, D_m$ in $\mathcal{A}$, for all $C \in \mathcal{C}$ there is an i with $C_i \subset C \subset D_i$ and $P(D_i \setminus C_i) \leq \epsilon\}$. Here the I in N_I indicates "inclusion." Then it follows that

$$N_I(\epsilon, \mathcal{C}, P) \;=\; N_{[\,]}^{(1)}(\epsilon, \mathcal{F}, P) \quad \text{where} \quad \mathcal{F} = \{1_C :\; C \in \mathcal{C}\}. \tag{7.4}$$

We have the following law of large numbers:

Theorem 7.1. (Blum-DeHardt) *Suppose $\mathcal{F} \subset \mathcal{L}^1(A, \mathcal{A}, P)$ and for all $\epsilon > 0$, $N_{[\,]}^{(1)}(\epsilon, \mathcal{F}, P) < \infty$. Then $\mathcal{F}$ is a strong Glivenko-Cantelli class, that is,*

$$\lim_{n\to\infty} \|P_n - P\|_{\mathcal{F}}^* = 0 \;\; a.s.$$

The sufficient condition in Theorem 7.1 is not necessary. In fact, the following holds.

Proposition 7.2. *There is a probability space $(A, \mathcal{A}, P)$ and a strong Glivenko-Cantelli class $\mathcal{F} := \{1_C :\; C \subset \mathcal{C}\}$ for P, where $\mathcal{C} \subset \mathcal{A}$ is such that for all $\varepsilon < 1/2$, we have $N_{[\,]}^{(1)}(\epsilon, \mathcal{F}, P) = +\infty$.*

Proof. Let $A = [0,1]$ with $P = U[0,1]$ the uniform (Lebesgue) law. Let $C_m := C(m)$ be independent sets with $P(C_m) = 1/m$. One can show that $\mathcal{C} := \{C(m)\}_{m\geq 1}$ has the stated properties. □

On the other hand let $\mathcal{C}$ be the collection of all finite subsets of $[0,1]$ with Lebesgue law P. Then $\|P_n - P\|_{\mathcal{C}} \equiv 1 \not\to 0$ although $1_A = 0$ a.s. for all $A \in \mathcal{C}$. This shows that in Theorem 7.1, $N_I < \infty$ cannot be replaced by $N(\epsilon, \mathcal{F}, d_p) \equiv 1$ for any $\mathcal{L}^p$ distance d_p.

A Banach space $(S, \|\cdot\|)$ has a dual space $(S', \|\cdot\|')$ of continuous linear forms $f : S \mapsto \mathbb{R}$ with $\|f\|' := \sup\{|f(x)| :\; x \in S,\; \|x\| \leq 1\} < \infty$ (RAP, Section 6.1). One way to apply Theorem 7.1 is via the following:

Proposition 7.3. *Let $(S, \|\cdot\|)$ be a separable Banach space and P a law on the Borel sets of S such that $\int \|x\|\, dP(x) < \infty$. Let $\mathcal{F}$ be the unit ball of the dual space S', $\mathcal{F} := \{f \in S' : \|f\|' \leq 1\}$. Then for every $\epsilon > 0$, $N_{[\,]}^{(1)}(\epsilon, \mathcal{F}, P) < \infty$.*

Corollary 7.4. *(Mourier's strong law of large numbers) Let $(S, \|\cdot\|)$ be a separable Banach space, P a law on S such that $\int \|x\|\, dP(x) < \infty$, and $X_1, X_2, \ldots$ i.i.d. P. Let $S_n := X_1 + \cdots + X_n$. Then S_n/n converges a.s. in $(S, \|\cdot\|)$ to some $x_0 \in S$.*

Corollary 7.5. *If $\mathcal{F} \subset \mathcal{L}^1(A, \mathcal{A}, P)$ and $\{\delta_x : x \in A\}$ is separable for $\|\cdot\|_{\mathcal{F}}$, then $\|P_n - P\|_{\mathcal{F}}$ $= \|P_n - P\|^*_{\mathcal{F}} \to 0$ a.s.*

Proof. This follows from Corollary 7.4, since finite linear combinations of δ_x, $x \in A$, with rational coefficients, are dense in their completion for $\|\cdot\|_{\mathcal{F}}$, a Banach space. □

The proof of Proposition 7.3 and Corollary 7.4 together from Theorem 7.1 is no shorter than a direct proof. On the other hand if $\mathcal{F} = \left\{1_{[0,t]} : 0 < t < 1\right\}$ and P is Lebesgue measure on $[0, 1]$ then Theorem 7.1 applies but Corollary 7.5 does not.

7.2 Central limit theorems with bracketing

In this section the bracketing will be in L^2. The following main theorem will be stated. Then, Corollary 7.7 gives a hypothesis on $N^{(1)}_{[\,]}$ for uniformly bounded classes of functions.

Theorem 7.6. (M. Ossiander) *Let $(X, \mathcal{A}, P)$ be a probability space and let $\mathcal{F} \subset \mathcal{L}^2(X, \mathcal{A}, P)$ be such that*

$$\int_0^1 \left(\log N^{(2)}_{[\,]}(x, \mathcal{F}, P)\right)^{1/2} dx < \infty.$$

Then $\mathcal{F}$ is a P-Donsker class.

Theorem 7.6 implies the following for L^1 entropy with bracketing:

Corollary 7.7. *Let $(X, \mathcal{A}, P)$ be a probability space and $\mathcal{F}$ a uniformly bounded set of measurable functions on X. Suppose that*

$$\int_0^1 \left(\log N^{(1)}_{[\,]}(x^2, \mathcal{F}, P)\right)^{1/2} dx < \infty.$$

Then $\mathcal{F}$ is a Donsker class for P.

Proof. Suppose $|f(x)| \le M < \infty$ for all $f \in \mathcal{F}$ and $x \in X$. Since multiplication by a constant preserves the Donsker property (by Theorem 3.28), we can assume $M = 1/2$. Then for any $f, g \in \mathcal{F}$ and $\varepsilon > 0$, $|f - g| \le 1$ everywhere. So if $\int |f - g| dP \le \varepsilon^2$ then $(\int |f - g|^2 dP)^{1/2} \le \varepsilon$. So $N^{(2)}_{[\,]}(\varepsilon, \mathcal{F}, P) \le N^{(1)}_{[\,]}(\varepsilon^2, \mathcal{F}, P)$ and the result follows from Theorem 7.6. □

It will be seen in the next section that Corollary 7.7, and thus Theorem 7.6, are best possible (provide a characterization of the Donsker property) in some cases.

7.3 The power set of a countable set: Borisov-Durst theorem

Let P be a law on the set $\mathbb{N}$ of nonnegative integers. The next theorem gives a criterion for the Donsker property of the collection $2^{\mathbb{N}}$ of all subsets of $\mathbb{N}$, for P, in terms of the numbers $p_m := P(\{m\})$ for $m \ge 0$. We also find that the sufficient condition given in Corollary 7.7 is necessary for $2^{\mathbb{N}}$. Recall N_I as defined above Theorem 7.1.

Theorem 7.8. *The following are equivalent:*

(a) $2^{\mathbb{N}}$ *is a Donsker class for* P*;*

(b) $\sum_m p_m^{1/2} < \infty$*;*

(c) $\int_0^1 \left(\log N_I\left(x^2, 2^{\mathbb{N}}, P\right)\right)^{1/2} dx < \infty$.

Recall the Remark after Theorem 6.17, which implies that if $\mathcal{C} = 2^{\mathbb{N}}$ then $\sup_{A\in\mathcal{C}} |(P_n - P)(A)| \to 0$ a.s., $n \to \infty$, for any law P on $\mathcal{C}$.

REFERENCES

Andersen, Niels Trolle, Giné, E., Ossiander, M., and Zinn, J. (1988). The central limit theorem and the law of the iterated logarithm for empirical processes under local conditions. *Probab. Theory Related Fields* **77**, 271–305.

Arcones, Miguel A., and Giné, E. (1993). Limit theorems for U-processes. *Ann. Probab.* **21**, 1494-1542.

Blum, J. R. (1955). On the convergence of empiric distribution functions. *Ann. Math. Statist.* **26**, 527–529.

Borisov, I. S. (1981). Some limit theorems for empirical distributions (in Russian). *Abstracts of Reports, Third Vilnius Conf. Probability Th. Math. Statist.* **1**, 71–72.

DeHardt, J. (1971). Generalizations of the Glivenko-Cantelli theorem. *Ann. Math. Statist.* **42**, 2050–2055.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6**, 899-929; Correction, *ibid.* **7** (1979), 909-911.

Dudley, R. M. (1984). A course on empirical processes. Ecole d'été de probabilités de St.-Flour, 1982. *Lecture Notes in Math. (Springer)* **1097**, 1-142.

Durst, Mark, and Dudley, R. M. (1981). Empirical processes, Vapnik-Chervonenkis classes and Poisson processes. *Probab. Math. Statist.* (Wrocław) **1**, no. 2, 109-115.

Gaenssler, Peter, and Stute, Winfried (1979). Empirical processes: a survey of results for independent and identically distributed random variables. *Ann. Probab.* **7**, 193–243.

Mourier, Edith (1951), Lois de grands nombres et théorie ergodique. *C. R. Acad. Sci. Paris* **232**, 923–925.

Mourier, E. (1953). Éléments aléatoires dans un espace de Banach. *Ann. Inst. H. Poincaré* **13**, 161–244.

Ossiander, Mina (1987). A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab.* **15**, 897–919.

Chapter 8

Approximation of functions and sets

8.1 Introduction: the Hausdorff metric

In this chapter upper and lower bounds will be stated for the metric entropies of various concrete classes of functions on Euclidean spaces and sets in such spaces. Some metric entropies with bracketing are treated, and some without. Metrics for functions are in $\mathcal{L}^p$, $1 \leq p \leq \infty$. For sets we use d_P metrics $d_P(B, C) := P(B \Delta C)$ or the Hausdorff metric, defined as follows.

For any metric space (S, d), $x \in S$, and a non-empty $B \subset S$, let

$$d(x, B) \quad := \quad \inf \{d(x, y) : \ y \in B\} .$$

For non-empty $B, C \subset S$ the Hausdorff pseudo-metric is defined by

$$h(B, C) \quad := \quad \max \left(\sup_{x \in B} d(x, C), \sup_{x \in C} d(x, B) \right) .$$

Then h is a metric on the collection of bounded, closed, non-empty sets. Let $h(\emptyset, C) := h(C, \emptyset) := +\infty$ for $C \neq \emptyset$, and $h(\emptyset, \emptyset) := 0$.

On $\mathbb{R}^d$ we have the usual Euclidean metric $d(x, y) := |x - y|$ where $|u| := \left(u_1^2 + \cdots + u_d^2\right)^{1/2}$, $u \in \mathbb{R}^d$. For any set $H \subset \mathbb{R}^{d-1}$ and function f from H into $[0, \infty]$ let

$$J_f := J(f) := \left\{ x \in \mathbb{R}^d : 0 \leq x_d \leq f\left(x_{(d)}\right), x_{(d)} \in H \right\}$$

where $x_{(d)} := (x_1, \ldots, x_{d-1})$. Then $J(f)$ is the subgraph of f. For any other function $g \geq 0$ on H, clearly $h\left(J_f, J_g\right) \leq d_{\sup}(f, g)$. Thus for any collection $\mathcal{F}$ of real functions ≥ 0 on H, and any $\epsilon > 0$, with $D(\varepsilon, \cdot, \cdot)$ as defined in Appendix K,

$$D\left(\epsilon, \left\{J_f : \ f \in \mathcal{F}\right\}, h\right) \quad \leq \quad D\left(\epsilon, \mathcal{F}, d_{\sup}\right) . \tag{8.1}$$

If $d_{\sup}(f, g) \leq \epsilon$ and $j := \max(f - \epsilon, 0)$, where $g \geq 0$, then $0 \leq j \leq g \leq f + \epsilon$, so $J_j \subset J_g \subset J_{f+\epsilon}$. If P is a law on $H \times [0, \infty[$ having a density p with respect to Lebesgue measure on $\mathbb{R}^d$ with $p(x) \leq M < \infty$ for all x, then

$$N_I\left(2M\epsilon, \left\{J_f : \ f \in \mathcal{F}\right\}, P\right) \quad \leq \quad D\left(\epsilon, \mathcal{F}, d_{\sup}\right) . \tag{8.2}$$

In the converse direction there are corresponding estimates for Lipschitz functions. Recall that $\|f\|_L := \sup_{x \neq y} |f(x) - f(y)| / |x - y|$. Then we have:

Lemma 8.1. *If* $\|f\|_L \le K$ *and* $\|g\|_L \le K$ *on* $H \subset \mathbb{R}^{d-1}$, *with* $f \ge 0$ *and* $g \ge 0$, *then* $h\,(J_f, J_g) \ge \min(1, 1/K) d_{\sup}(f,g)/2$.

Proof. Let $t := d_{\sup}(f,g)$. Let $0 < s < t$. By symmetry, assume that for some $x \in H$, $f(x) \ge g(x) + s$. Then for any $y \in H$, either $|x-y| \ge s/(2K)$ or $g(y) \le g(x) + s/2 \le f(x) - s/2$. In either case, if $z \le g(y)$ then $|\langle x, f(x)\rangle - \langle y, z\rangle| \ge \min(1, 1/K)s/2$. Letting $s \uparrow t$, the result follows. □

Recall that a bounded number of Boolean operations preserve the Vapnik-Červonenkis property (Theorem 4.7). The same holds for classes of sets satisfying bounds on metric entropy (with inclusion). For any families $\mathcal{C}_j$ of subsets of a set X, extending the notation in Section 4.5, let

$$\sqcap_{j=1}^k \mathcal{C}_j \;:=\; \{\cap_{j=1}^k A_j :\; A_j \in \mathcal{C}_j \;\text{ for all }\; j\},$$

$$\sqcup_{j=1}^k \mathcal{C}_j \;:=\; \{\cup_{j=1}^k A_j :\; A_j \in \mathcal{C}_j \;\text{ for all }\; j\}.$$

Theorem 8.2. *Let* $(X, \mathcal{A}, P)$ *be a probability space and* $\mathcal{C}_j \subset \mathcal{A}$ *for* $j = 1, \cdots, k$. *Let* $\mathcal{C}_0 := \sqcap_{j=1}^k \mathcal{C}_j$. *Let* $f_{1j}(\epsilon) := \log N_I(\epsilon, \mathcal{C}_j, P)$, $f_{2j}(\epsilon) := \log D(\epsilon, \mathcal{C}_j, d_P)$ *for* $j = 0, 1, \cdots, k$. *If for* $i = 1$ *or* 2, *there are a* $\gamma > 0$ *and constants* $M_1, \cdots, M_k$ *such that* $f_{ij}(\epsilon) \le M_j \epsilon^{-\gamma}$ *for* $0 < \epsilon < 1$ *and* $j = 1, \cdots, k$, *then the same holds for* $j = 0$. *The statements also hold for* $i = 1$ *or 2 for* $\mathcal{C}_0 := \sqcup_{j=1}^k \mathcal{C}_j$.

8.2 Spaces of differentiable functions and sets with differentiable boundaries

For any $\alpha > 0$, spaces of functions will be defined having "bounded derivatives through order α". If β is the largest integer $< \alpha$, the functions will have partial derivatives through order β bounded, and the derivatives of order β will satisfy a uniform Hölder condition of order $\alpha - \beta$. Still more specifically: for $x := (x_1, \ldots, x_d) \in \mathbb{R}^d$ and $p = (p_1, \ldots, p_d) \in \mathbb{N}^d$ (where $\mathbb{N}$ is the set of nonnegative integers) let $[p] := p_1 + \cdots + p_d$,

$$x^p := x_1^{p_1} x_2^{p_2} \cdots x_d^{p_d}, \quad D^p := \partial^{[p]} / \partial x_1^{p_1} \cdots \partial x_d^{p_d}.$$

For a function f on an open set $U \subset \mathbb{R}^d$ having all partial derivatives $D^p f$ of orders $[p] \le \beta$ defined everywhere on U, let

$$\|f\|_\alpha := \|f\|_{\alpha, U} \;:=\; \max_{[p] \le \beta} \sup \{|D^p f(x)| :\; x \in U\}$$

$$+ \max_{[p]=\beta} \sup_{x \ne y,\; x,y \in U} \left\{ |D^p f(x) - D^p f(y)| \,/\, |x-y|^{\alpha - \beta} \right\}.$$

Here if $0 < \alpha \le 1$, so $\beta = 0$, $D^0 f := D^{(0,0,\cdots,0)} f := f$. Let I^d denote the unit cube $\{x \in \mathbb{R}^d : 0 \le x_j \le 1,\; j = 1, \ldots, d\}$, and $x_{(d)} := (x_1, \ldots, x_{d-1})$, $x \in \mathbb{R}^d$. Let $F \subset \mathbb{R}^d$ be a closed set which is the closure of its interior U. Let $\mathcal{F}_{\alpha,K}(F)$ denote the set of all continuous $f :\; F \mapsto \mathbb{R}$ with $\|f\|_{\alpha, U} \le K$. For $\alpha = 1$, $\mathcal{F}_{1,K}(F)$ is the set of bounded Lipschitz functions f on F with $\max(\|f\|_{\sup}, \|f\|_L) \le K$. Then recalling the bounded Lipschitz norm $\|f\|_{BL} := \|f\|_L + \|f\|_{\sup}$ we have

$$\{f :\; \|f\|_{BL} \le K\} \subset \mathcal{F}_{1,K}(F) \le \{f :\; \|f\|_{BL} \le 2K\}.$$

Let $\mathcal{G}_{\alpha,K,d} := \mathcal{F}_{\alpha,K}\left(I^d\right)$. Let $\mathcal{C}(\alpha,K,d)$ be the collection of all sets

$$J_f \;=\; J(f) \;=\; \left\{x \in I^d :\; 0 \le x_d \le f\left(x_{(d)}\right)\right\}, \quad f \in \mathcal{G}_{\alpha,K,d-1}, \quad f \ge 0.$$

If g and h and two functions defined for (small enough) $y > 0$, then $g \asymp h$ (as $y \downarrow 0$) means that

$$0 \;<\; \liminf_{y\downarrow 0}(g/h)(y) \;\le\; \limsup_{y\downarrow 0}(g/h)(y) \;<\; +\infty.$$

Clearly, if $f \in \mathcal{G}_{\alpha,K,d}$ and $[p] < \alpha$ then $D^p f \in \mathcal{G}_{\alpha-[p],K,d}$. Let $B_d := \{x \in \mathbb{R}^d :\; |x| < 1\}$ and let $\overline{B}_d = \{x \in \mathbb{R}^d :\; |x| \le 1\}$ be the open and closed unit balls respectively in $\mathbb{R}^d$. Here are some bounds on metric entropies, which by Ossiander's theorem (7.6) will give that if $\alpha > d/2$, $\mathcal{G}_{\alpha,K,d}$ is a Donsker class for any law P, and $\mathcal{C}(\alpha,K,d+1)$ is for laws with bounded densities.

Theorem 8.3. *For $0 < K < \infty$, $0 < \alpha < \infty$ and $d \ge 1$, as $\epsilon \downarrow 0$*

$$\log D\left(\epsilon, \mathcal{G}_{\alpha,K,d}, d_{\sup}\right) \asymp \epsilon^{-d/\alpha}.$$

For some $T := T(\alpha,K,d)$, any law P on I^d, $1 \le r \le \infty$ and $0 < \epsilon < 1$,

$$\begin{aligned} \log N_{[\,]}^{(r)}\left(\epsilon, \mathcal{G}_{\alpha,K,d}, P\right) &\le T\epsilon^{-d/\alpha}, \quad \text{and} \\ \log D\left(\epsilon, \mathcal{C}(\alpha,K,d+1), h\right) &\le T\epsilon^{-d/\alpha}. \end{aligned}$$

If Q is a law on I^{d+1} having a density with respect to Lebesgue measure bounded by M, then for some $M_1 = M_1(M,d,K,\alpha)$,

$$\log N_I(\epsilon, \mathcal{C}(\alpha,K,d+1), Q) \;\le\; M_1\epsilon^{-d/\alpha}, \quad 0 < \epsilon \le 1.$$

The same statements all hold for $\overline{B}_d$ in place of I^d and so $\mathcal{F}_{\alpha,K}(\overline{B}_d)$ in place of $\mathcal{G}_{\alpha,K,d}$, with possibly larger constants T and M_1.

Notes For $\alpha \ge 1$, in the statement about h, the order $\epsilon^{-d/\alpha}$ is precise, see Corollary 8.7. The statement about $N_{[\,]}^{(r)}$ for $r = 1$ implies that $\mathcal{G}_{\alpha,K,d}$ is a Glivenko-Cantelli class for any $\alpha > 0$, $K < \infty$ and d, by the Blum-DeHardt theorem 7.1.

Next, some lower bounds for metric entropies in the L^1 norm will be given. For a collection $\mathcal{F} \subset \mathcal{L}^1(A,\mathcal{A},P)$, we have the $\mathcal{L}^1$ distance $d_{1,P}(f,g) := P(|f-g|)$.

Theorem 8.4. *Let P be a law on I^d having a density with respect to Lebesgue measure bounded below by $\gamma > 0$. Then for some $C = C(\gamma,\alpha,K,d) > 0$, and $1 \le r \le \infty$, $N_{[\,]}^{(r)}\left(\epsilon, \mathcal{G}_{\alpha,K,d}, P\right) \ge N_{[\,]}^{(1)}\left(\epsilon, \mathcal{G}_{\alpha,K,d}, P\right) \ge D\left(\epsilon, \mathcal{G}_{\alpha,K,d}, d_{1,P}\right) \ge \exp\left(C\epsilon^{-d/\alpha}\right)$ for ϵ small enough, and if $d \ge 2$, for small enough $\epsilon > 0$, and $M := C(\gamma,\alpha,K,d-1)$,*

$$N_I\left(\epsilon, \mathcal{C}(\alpha,K,d), P\right) \;\ge\; D\left(\epsilon, \mathcal{C}(\alpha,K,d), d_P\right) \;\ge\; \exp\left(M\epsilon^{-(d-1)/\alpha}\right).$$

In the proof, the following combinatorial fact is used:

Lemma 8.5. *Let B be a set with n elements, $n = 0, 1, \cdots$. Then there exist subsets $E_i \subset B$, $i = 1, \ldots, k$, where $k \geq e^{n/6}$, such that for $i \neq j$, the symmetric difference $E_i \Delta E_j$ has at least $n/5$ elements.*

To get lower bounds for the Hausdorff metric, the following will help:

Lemma 8.6. *If $\alpha \geq 1$ and $f, g \in \mathcal{G}_{\alpha,K,d}$, then $h\left(J_f, J_g\right) \geq d_{\sup}(f,g)/(2\max(1, Kd))$.*

Corollary 8.7. *If $\alpha \geq 1$ and $d = 1, 2, \cdots$, then as $\epsilon \downarrow 0$,*

$$\log D\left(\epsilon, \mathcal{C}\left(\alpha, K, d+1\right), h\right) \asymp \epsilon^{-d/\alpha}.$$

Remark. For $m = 1, 2, \cdots$, let I^d be decomposed into a grid of m^d sub-cubes of side $1/m$. Let E be the set of centers of the cubes. For any $A \subset I^d$ let $B \subset E$ be the set of centers of the cubes in the grid that A intersects. Then $h(A, B) \leq d^{1/2}/(2m)$, which includes the possibility that $A = B = \emptyset$. For $0 < \epsilon < 1$ there is a least $m = 1, 2, \cdots$ such that $d^{1/2}/m \leq \epsilon$, namely $m = \lceil d^{1/2}/\epsilon \rceil$. It follows that

$$D\left(\epsilon, 2^{I^d}, h\right) \quad \leq \quad 2^{\left(1 + d^{1/2}/\epsilon\right)^d}.$$

Hence for $\alpha < d/(d+1)$, Corollary 8.7 cannot hold, nor can the upper bound for h in Theorem 8.3 be sharp.

The classes $\mathcal{C}(\alpha, K, d)$ considered so far contain sets with flat faces except for one curved face. There are at least two ways to form more general classes of sets with piecewise differentiable boundaries, still satisfying the bounds in Theorem 8.3. One is to take a bounded number of Boolean operations. Let $v_1, \cdots, v_k$ be non-zero vectors in $\mathbb{R}^d$ where $d \geq 2$. For constants $c_1, \cdots, c_k$ let $H_j := \{x \in \mathbb{R}^d : (x, v_j) = c_j\}$, a hyperplane. Let π_j map each $x \in \mathbb{R}^d$ to its nearest point in H_j, $\pi_j(x) := x - ((x, v_j) - c_j)v_j/|v_j|^2$. Let T_j be a cube in H_j and $\alpha, K > 0$. Let f_j be a linear transformation taking T_j onto I^{d-1}. For $g \in \mathcal{G}_{\alpha,K,d-1}$ with $g \geq 0$, let

$$J_j(g) := \{x \in \mathbb{R}^d : \pi_j(x) \in T_j,\ c_j \leq (v_j, x) \leq c_j + g(f_j(\pi_j(x)))\}.$$

Let

$$\mathcal{C}_j := \mathcal{C}_j(\alpha, K, v_j, c_j) := \{J_j(g) : g \in \mathcal{G}_{\alpha,K,d-1}\}.$$

Then Theorem 8.3 implies that if C is a compact set in $\mathbb{R}^d$ (e.g., a cube) including all sets in $\mathcal{C}_j$, and P is a law on C having bounded density with respect to Lebesgue measure λ, then for some $M_j < \infty$,

$$\log D(\epsilon, \mathcal{C}_j, d_P) \ \leq \ \log N_I(\epsilon, \mathcal{C}_j, P) \ \leq \ M_j \epsilon^{(1-d)/\alpha}.$$

We then have by Theorem 8.2 the following:

Theorem 8.8. *Let $d \geq 2$ and let $\mathcal{C}_0 := \sqcap_{j=1}^k \mathcal{C}_j$ or $\mathcal{C}_0 := \sqcup_{j=1}^k \mathcal{C}_j$, for $\mathcal{C}_j$ as just defined. Then for some $M < \infty$,*

$$\log D(\epsilon, \mathcal{C}_0, d_P) \ \leq \ \log N_I(\epsilon, \mathcal{C}_0, P) \ \leq \ M \epsilon^{(1-d)/\alpha}.$$

By intersections or unions of k sets in classes $\mathcal{C}_j$ (with k depending on d), one can obtain sets with smooth boundaries (through order α) such as ellipsoids. One can also get more general sets, since, e.g. for $\alpha > 1$, the minimum or maximum of two functions in $\mathcal{G}_{\alpha,K,d}$ need not have first derivatives everywhere and then will not be in $\mathcal{G}_{\gamma,\kappa,d}$ for any $\gamma > 1$ and $\kappa < \infty$.

Recall that a C^∞ real-valued function on an open set on $\mathbb{R}^d$ is one such that the partial derivatives $D^p f$ exist for all $p \in \mathbb{N}^d$ and are continuous. For functions $f := (f_1, \cdots, f_k)$ into $\mathbb{R}^k$, for f to be C^∞ means that each f_j is. Another way to generate sets with boundaries differentiable of order α is as follows. The unit sphere $S^{d-1} := \{x \in \mathbb{R}^d : |x| = 1\}$ is a C^∞ manifold, specifically as follows. S^{d-1} is the union of two sets $A := \{x \in S^{d-1} : x_1 > -1/2\}$ and $C := \{x \in S^{d-1} : x_1 < 1/2\}$. There is a 1-1, C^∞ function ψ from $\{x \in \mathbb{R}^{d-1} : |x| < 9/8\}$ into $\mathbb{R}^d$, with derivative matrix $\{\partial\psi_i/\partial x_j\}_{i=1,j=1}^{d,d-1}$ of maximum rank $d-1$ everywhere, such that ψ takes $B_{d-1} := \{x \in \mathbb{R}^{d-1} : |x| < 1\}$ onto A. Let $\eta(y) := (-\psi_1(y), \psi_2(y), \cdots, \psi_d(y))$. Then the above statements for ψ and A also hold for η and C.

For $0 < \alpha, K < \infty$ let $\mathcal{F}_{\alpha,K}(S^{d-1})$ be the set of functions $h : S^{d-1} \mapsto \mathbb{R}$ such that for $\overline{B}_{d-1} := \{x \in \mathbb{R}^{d-1} : |x| \le 1\}$, $h \circ \psi$ and $h \circ \eta \in \mathcal{F}_{\alpha,K}(\overline{B}_{d-1})$, recalling that $f \circ g(y) := f(g(y))$. Let $\mathcal{F}^{(d)}_{\alpha,K}(S^{d-1})$ be the set of functions $h = (h_1, \cdots, h_d)$ such that $h_j \in \mathcal{F}_{\alpha,K}(S^{d-1})$ for each $j = 1, \cdots, d$.

Two continuous functions F, G from one topological space X to another, Y, are called *homotopic* iff there exists a jointly continuous function H from $X \times [0,1]$ into Y such that $H(\cdot, 0) \equiv F$ and $H(\cdot, 1) \equiv G$. H is then called a *homotopy* of F and G. Let $I(F)$ be the set of all $y \in Y$, not in the range of F, such that among mappings of X into $Y \setminus \{y\}$, F is not homotopic to any constant map $G(x) \equiv z \ne y$.

For a function F let $R(F) := \mathrm{ran}(F) := \mathrm{range}(F)$ and $C(F) := I(F) \cup R(F)$.

For example, if F is the identity from S^{d-1} onto itself in $\mathbb{R}^d$, then $I(F) = \{y : |y| < 1\}$ by well known facts in algebraic topology, e. g. Eilenberg and Steenrod (1952, Chapter 11, Theorem 3.1).

Let $I(d, \alpha, K) := \{I(F) : F \in \mathcal{F}^{(d)}_{\alpha,K}(S^{d-1})\}$ and $\mathcal{K}(d, \alpha, K) := \{C(F) : F \in \mathcal{F}^{(d)}_{\alpha,K}(S^{d-1})\}$. Then $I(d, \alpha, K)$ is a collection of open sets and $\mathcal{K}(d, \alpha, K)$ of compact sets each of which, in a sense, have boundaries differentiable of order α. (For functions F that are not one-to-one, the boundaries may not be differentiable in some other senses.) For $\mathcal{K}(d, \alpha, K)$ and to some extent for $I(d, \alpha, K)$ there are bounds as for other classes of sets with α times differentiable boundaries (Theorem 8.8):

Theorem 8.9. *For each $d = 2, 3, \cdots$, $K \ge 1$ and $\alpha \ge 1$,*
(a) there is a constant $H_{d,\alpha,K} < \infty$ such that for $0 < \epsilon \le 1$, and the Hausdorff metric h,

$$\log D(\epsilon, \mathcal{K}(d, \alpha, K), h) \le H_{d,\alpha,K}/\epsilon^{(d-1)/\alpha}.$$

(b) For any $\zeta < \infty$ there is a there is a constant $A_{d,\alpha,K,\zeta} < \infty$ such that for any law P on $\mathbb{R}^d$ having density with respect to λ^d bounded above by ζ, for $0 < \epsilon \le 1$,

$$\max(\log N_I(\epsilon, \mathcal{K}(d, \alpha, K), P), \log N_I(\epsilon, I(d, \alpha, K), P)) \le$$

$$A_{d,\alpha,K,\zeta}/\epsilon^{(d-1)/\alpha}.$$

Corollary 8.10. *For any law P on $\mathbb{R}^d$, $d \geq 2$, having bounded density with respect to Lebesgue measure, and $K < \infty$,*
(a) (Tze-Gong Sun and R. Pyke) *$I(d, \alpha, K)$ is a Donsker class for P if $\alpha > d - 1$.*
(b) $I(d, \alpha, K)$ is a Glivenko-Cantelli class for P whenever $\alpha \geq 1$.

8.3 Lower layers

A set $B \subset \mathbb{R}^d$ is called a *lower layer* if and only if for all $x = (x_1, \ldots, x_d) \in B$ and $y = (y_1, \ldots, y_d)$ with $y_j \leq x_j$ for $j = 1, \ldots, d$, we have $y \in B$. Let $\mathcal{LL}_d$ denote the collection of all non-empty lower layers in $\mathbb{R}^d$ with non-empty complement. Let $\emptyset$ be the empty set and

$$\mathcal{LL}_{d,1} \;:=\; \left\{ L \cap I^d :\; L \in \mathcal{LL}_d,\; L \cap I^d \neq \emptyset \right\}.$$

Let $\lambda := \lambda_I^d$ denote Lebesgue measure on I^d. Recall that $f \sim g$ means $f/g \to 1$. The size of $\mathcal{LL}_{d,1}$ will be bounded first when $d = 1$ and 2. Let $\lceil x \rceil$ be the smallest integer $\geq x$.

Theorem 8.11. *For $d = 1$,*

$$D\left(\epsilon, \mathcal{LL}_{1,1}, h\right) \;=\; D\left(\epsilon, \mathcal{LL}_1, d_\lambda\right) \;=\; N_I\left(\epsilon, \mathcal{LL}_{1,1}, \lambda\right) \;=\; \lceil 1/\epsilon \rceil.$$

For $d = 2$, any $m = 1, 2, \cdots$, and $0 < t < 2^{1/2}/m$, we have

$$\max\left(N_I\left(2/m, \mathcal{LL}_2, \lambda_I^2\right), D\left(2^{1/2}/m, \mathcal{LL}_{2,1}, h\right)\right) \leq \binom{2m-2}{m-1} \leq D\left(t, \mathcal{LL}_{2,1}, h\right).$$

For $0 < \epsilon \leq 1$, $N_I\left(\epsilon, \mathcal{LL}_{2,1}, \lambda_I^2\right) \leq 4^{2/\epsilon}$ and

$$D\left(\epsilon, \mathcal{LL}_{2,1}, h\right) \leq \exp\left(\left(2^{1/2} \log 4\right)/\epsilon\right).$$

8.4 Metric entropy of classes of convex sets

Let $\mathcal{C}_d$ denote the class of all non-empty closed convex subsets of the open unit ball $B(0,1) := \{x\colon\ |x| < 1\}$ in $\mathbb{R}^d$. Let λ be the uniform Lebesgue measure on $\mathbb{R}^d$. Upper and lower bounds will be given for the metric entropy of $\mathcal{C}_d$ for the metric d_λ and for the Hausdorff metric h.

Theorem 8.12. (E. M. Bronštein) *For each $d \geq 2$ we have*

$$\log D(\epsilon, \mathcal{C}_d, d_\lambda) \asymp \log D(\epsilon, \mathcal{C}_d, h) \asymp \epsilon^{(1-d)/2} \quad \text{as} \quad \epsilon \downarrow 0.$$

The proof of Bronštein's theorem is very long. A natural idea is to approximate convex sets by polyhedra. Polyhedra can be difficult to approximate because of sharp edges and vertices. Moreover it turns out that sometimes convex sets are approximated by non-convex polyhedra. One of Bronštein's main ideas is to use the transformation $B \mapsto B^1$ where $B^1 := \{y :\ |x - y| < 1 \text{ for some } x \in B\}$. It can be shown that $B \mapsto B^1$ is an isometry for the Hausdorff metric.

A set B^1 has a boundary no more curved than a sphere of radius 1. The sets B^1 can thus be approximated by polyhedra more easily than the original sets B can.

Remark. For $d = 1$, $\mathcal{C}_1$ is just the class of subintervals of the open interval $(-1, 1)$. Then it's rather easy to see that

$$D(\epsilon, \mathcal{C}_1, d_\lambda) \asymp D(\epsilon, \mathcal{C}_1, h) \asymp \epsilon^{-2}.$$

From Bronštein's theorem one can prove:

Corollary 8.13. *In $\mathbb{R}^d$, for $d \geq 2$, if P is a law whose restriction to $B(0,1)$ has a bounded density f with respect to Lebesgue measure, then*
(a) $\log N_I(\epsilon, \mathcal{C}_d, P) = O(\epsilon^{(1-d)/2})$ *as* $\epsilon \downarrow 0$.
(b) (E. Bolthausen) *For $d = 2$, $\mathcal{C}_2$ is a Donsker class for P.*
(c) For any d, $\mathcal{C}_d$ is a Glivenko-Cantelli class.
(d) If also $f \geq v$ on $B(0,1)$ for some constant $v > 0$, then

$$\log N_I(\epsilon, \mathcal{C}_d, P) \asymp \epsilon^{(1-d)/2} \quad \text{as} \quad \epsilon \downarrow 0.$$

REFERENCES

Bolthausen E. (1978). Weak convergence of an empirical process indexed by the closed convex subsets of I^2. *Z. Wahrscheinlichkeitsth. verw. Gebiete* **43**, 173-181.

Bonnesen, Tommy, and Fenchel, Werner (1934). *Theorie der Konvexen Körper*. Springer, Berlin; repub. Chelsea, New York, 1948.

Bronshtein [Bronštein], E. M. (1976). ϵ-entropy of convex sets and functions. *Siberian Math. J.* **17**, 393-398, transl. from *Sibirsk. Mat. Zh.* **17**, 508–514.

Clements, G. F. (1963). Entropies of several sets of real valued functions. *Pacific J. Math* **13**, 1085–1095.

Dudley, R. M. (1974). Metric entropy of some classes of sets with differentiable boundaries. *J. Approx. Theory* **10**, 227–236; Correction **26** (1979), 192–193.

Eggleston, H. G. (1958). *Convexity*. Cambridge University Press. Reprinted with corrections, 1969.

Eilenberg, S., and Steenrod, N. (1952). *Foundations of Algebraic Topology*. Princeton University Press.

Gruber, P. M. (1983). Approximation of convex bodies. In *Convexity and its Applications*, ed. P. M. Gruber, J. M. Wills. Birkhäuser, Basel, pp. 131-162.

Hausdorff, Felix (1914). *Mengenlehre*, transl. by J. P. Aumann et al. as *Set Theory*. 3d English ed. of transl. of 3d German edition (1937). Chelsea, New York, 1978.

Hoffman, K. (1975). *Analysis in Euclidean Space*. Prentice-Hall, Englewood Cliffs, NJ.

Kolmogorov, A. N. (1955). Bounds for the minimal number of elements of an ε-net in various classes of functions and their applications to the question of representability of functions of several variables by functions of fewer variables (in Russian). *Uspekhi Mat. Nauk* (N.S.) **10** no. 1 (63), 192-194.

Kolmogorov, A. N., and Tikhomirov, V. M. (1959). ϵ-entropy and ϵ-capacity of sets in function spaces. *Uspekhi Mat. Nauk* **14** no. 2, 3–86=*Amer. Math. Soc. Transl.* (Ser. 2) **17** (1961), 277–364.

Lorentz, George C. (1966). Metric entropy and approximation. *Bull. Amer. Math. Soc.* **72**, 903–937.

Pyke, R. (1983). The Haar-function construction of Brownian motion indexed by sets. *Z. Wahrscheinlichkeitsth. verw. Geb.* **64**, 523-539. Rudin, Walter (1976). *Principles of Mathematical Analysis*, 3d ed. McGraw-Hill, New York.

Sun, Tze-Gong and Pyke, R. (1982). Weak convergence of empirical measures. Technical Report no. 19, Dept. of Statistics, University of Washington, Seattle.

Wright, F. T. (1981). The empirical discrepancy over lower layers and a related law of large numbers. *Ann. Probab.* **9**, 323–329.

Chapter 9

Sums in General Banach Spaces and Invariance Principles

Let $(S, \|\cdot\|)$ be a Banach space (in general non-separable). A subset $\mathcal{F}$ of the unit ball $\left\{f \in S' : \|f\|' \leq 1\right\}$ is called a *norming* subset if and only if $\|s\| = \sup_{f\in\mathcal{F}} |f(s)|$ for all $s \in S$. The whole unit ball in S' is always a norming subset by the Hahn-Banach theorem (RAP, Corollary 6.1.5).

Conversely, given *any* set $\mathcal{F}$, let $S := \ell^\infty(\mathcal{F})$ be the set of all bounded real functions on $\mathcal{F}$, with the supremum norm

$$\|s\| = \|s\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |s(f)|, \quad s \in S.$$

Then the natural map $f \mapsto (s \mapsto s(f))$ takes $\mathcal{F}$ one-to-one onto a norming subset of S'.

So, limit theorems for empirical measures, uniformly over a class $\mathcal{F}$ of functions, can be viewed as limit theorems in a Banach space S with norm $\|\cdot\|_{\mathcal{F}}$. Conversely, limit theorems in a general Banach space S with norm $\|\cdot\|$ can be viewed as limit theorems for empirical measures on S, uniformly over a class $\mathcal{F}$ of functions, such as the unit ball of S', since for $f \in S'$ and $x_1, \ldots, x_n \in S$, $(\delta_{x_1} + \cdots + \delta_{x_n})(f) = f(x_1 + \cdots + x_n)$.

Suppose that X_j are i.i.d. real random variables with mean 0 and variance 1. Let $S_n := \sum_{j\leq n} X_j$. One form of "invariance principle" will say that on some probability space, there exist such X_j and also i.i.d. N(0,1) variables $Y_1, Y_2, \cdots$, with $T_n := \sum_{j\leq n} Y_j$, such that as $n \to \infty$, $\max_{k\leq n} |S_k - T_k|/n^{1/2} \to 0$ in probability. Since $T_n/n^{1/2}$ also has a $N(0,1)$ distribution for each n, the invariance principle implies that $S_n/n^{1/2}$ is close to $T_n/n^{1/2}$, which implies the central limit theorem. Although it is not as obvious, central limit theorems generally imply invariance principles. The phrase "invariance principle" means that such quantities as $\max_{k\leq n} |S_k|/n^{1/2}$ have an asymptotic distribution as $n \to \infty$, invariant under the choice of $\mathcal{L}(X_1)$ with mean 0 and variance 1. Section 9.3 treats invariance principles for variables with finite-dimensional values. Then the main result of the chapter will be Theorem 9.4, saying that the Donsker property is equivalent to an invariance principle for empirical processes.

9.1 Independent random elements and partial sums

We need a notion of independence for functions which may not be measurable. Let $(A_j, \mathcal{A}_j, P_j)$, $j = 1, 2, \ldots$, be probability spaces, and form a product $\prod_{j=1}^n (A_j, \mathcal{A}_j, P_j) = (B, \mathcal{B}, P)$ with

points $x := \{x_j\}_{j=1}^n$. If X_j are functions on B of the form $X_j = h_j(x_j)$, $j = 1, \ldots, n$, where each h_j is a function on A_j (not necessarily measurable) then we call X_j *independent random elements*. If the h_j are measurable this implies independence in the usual sense.

[A set of lemmas about independent random elements is omitted.]

9.2 A CLT implies measurability in separable normed spaces

Here "CLT" abbreviates "central limit theorem." If F_n is an empirical distribution function, $\Pr(F_n \in A)$ need not be defined if A is complete and discrete, hence Borel, for the (non-separable) supremum norm. Thus the variables $X_j := \delta_{x(j)} - P$ need not be Borel measurable in general for a norm $\|\cdot\|_{\mathcal{C}}$ or $\|\cdot\|_{\mathcal{F}}$. But in separable Banach spaces one usually assumes that variables are Borel measurable. It will be seen here that in a separable normed space, a form of central limit theorem can hold only if X_1 is measurable. If this holds in $\mathbb{R}^1$ it extends easily to separable normed spaces.

Define the inner measure $\Pr_*(B) := \sup\{\Pr(C) : C \subset B\}$ and let

$$f_* := -((-f)^*) = \operatorname{ess.sup}\{g : g \le f,\ g \text{ measurable}\}.$$

Theorem 9.1. *Let $(A, \mathcal{A}, P)$ be a probability space, x_n coordinates on $(A^\infty, \mathcal{A}^\infty, P^\infty)$, $n = 1, 2, \cdots$. Let $h : A \to \mathbb{R}$ (where h is not assumed measurable), $X_i := h(x_i)$. Let $S_n := X_1 + \cdots + X_n$. If for all t,*

$$\lim_{n\to\infty} \Pr^*\left(S_n/n^{1/2} \le t\right) = \lim_{n\to\infty} \Pr_*\left(S_n/n^{1/2} \le t\right)$$
$$= N(0,1)((-\infty, t]),$$

then h is measurable for the completion of P, so that X_i are measurable, $EX_i = 0$ and $EX_i^2 = 1$.

For the measurable case we need the following converse of the usual 1-dimensional central limit theorem.

Theorem 9.2. *Let $X_1, X_2, \ldots,$ be i.i.d. real random variables and $S_n := X_1 + \ldots + X_n$. If the law of $S_n/n^{1/2}$ converges to $N(0,1)$ as $n \to \infty$ then $EX_1 = 0$ and $EX_1^2 = 1$.*

Corollary 9.3. *Suppose $(S, |\cdot|)$ is a separable normed space and $X_n = h(x_n)$ where x_n are independent, identically distributed random variables with values in some measurable space $(A, \mathcal{A})$ and h is any function from A into S (not assumed measurable).*

Suppose Y_n are i.i.d. Gaussian variables in S with mean 0 and

$$\lim_{n\to\infty} n^{-1/2} \left|\sum_{j\le n} X_j - Y_j\right|^* = 0 \tag{9.1}$$

in probability. Then the X_j are completion measurable for the Borel σ-algebra on S.

9.3 A finite-dimensional invariance principle

This section, with no details given here, treats the special case of the theorem in the next section where $\mathcal{F}$ is a finite set, and is used in the proof of that theorem. Note that any finite set of $\mathcal{L}^2$ functions is a Donsker class.

9.4 Invariance principles for empirical processes

Recall the notion of coherent G_P process (Section 3.1). Let $(A, \mathcal{A}, P)$ be a probability space and $\mathcal{F} \subset \mathcal{L}^2(A, \mathcal{A}, P)$. Let Ω be the product of $([0,1], \mathcal{B}, \lambda)$, where $\mathcal{B}$ is the Borel σ-algebra, $\lambda =$ Lebesgue measure, and a countable product of copies of $(A, \mathcal{A}, P)$, with coordinates x_i. Then $\mathcal{F}$ will be called a *functional Donsker class* for P iff it is pregaussian for P and there are independent coherent G_P processes $Y_j(f, \omega)$, $f \in \mathcal{F}$, $\omega \in \Omega$, such that $f \mapsto Y_j(f, \omega)$ is bounded and ρ_P-uniformly continuous for each j and almost all ω, and such that in outer probability,

$$n^{-1/2} \max_{m \le n} \left\| \sum_{j=1}^{m} \delta_{x_j} - P - Y_j \right\|_{\mathcal{F}} \to 0. \tag{9.2}$$

Recalling the notion of Donsker class (as defined in Section 3.1), we have an equivalence:

Theorem 9.4. *Given any probability space $(A, \mathcal{A}, P)$ and $\mathcal{F} \subset \mathcal{L}^2(A, \mathcal{A}, P)$, $\mathcal{F}$ is a functional Donsker class if and only if it is a Donsker class.*

REFERENCES

Alexander, K. (1987). The central limit theorem for empirical processes on Vapnik-Červonenkis classes. *Ann. Probab.* **15**, 178-203.

Alexander, K., and Talagrand, M. (1989). The law of the iterated logarithm for empirical processes on Vapnik-Červonenkis classes. *J. Multivariate Analysis* **30**, 155-166.

Beck, József (1985). Lower bounds on the approximation of the multivariate empirical process. *Z. Wahrscheinlichkeitsth. verw. Gebiete* **70**, 289-306.

Borovkov, A. A., and Korolyuk, V. S. (1965). On the results of asymptotic analysis in problems with boundaries. *Theory Probab. Appl.* **10**, 236–246 (English), 255–266 (Russian).

Breiman, Leo (1968). *Probability*. Addison-Wesley, Reading, MA.

Bretagnolle, J., and Massart, P. (1989). Hungarian constructions from the nonasymptotic viewpoint. *Ann. Probab.* **17**, 239-256.

Donsker, Monroe D. (1951). An invariance principle for certain probability limit theorems. *Memoirs Amer. Math. Soc.* **6**.

Dudley, R. M., and Koltchinskii, V. I. (1994). Envelope moment conditions and Donsker classes. *Theor. Probab. Math. Statist.* (Kiev) **51**, 39-48.

Dudley, R. M., and Philipp, Walter (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Z. Wahrscheinlichkeitsth. verw. Gebiete* **62**, 509–552.

Freedman, David (1971). *Brownian Motion and Diffusion*. Holden-Day, San Francisco.

Giné, E., and Zinn, J. (1986). Lectures on the central limit theorem for empirical processes. In *Probability and Banach Spaces*, (Proc. Zaragoza, 1985), *Lecture Notes in Math.* (Springer) **1221**, Springer, Heidelberg, pp. 50-113.

Gnedenko, B. V., and Kolmogorov, A. N. (1949). Limit Theorems for Sums of Independent Random Variables. Transl. and ed. by K. L. Chung, Addison-Wesley, Reading, MA, 1968.

Goodman, V., Kuelbs, J., and Zinn, J. (1981). Some results on the LIL in Banach space with applications to weighted empirical processes. *Ann. Probab.* **9**, 713–752.

Heinkel, B. (1979). Relation entre théorème central-limite et loi du logarithme itéré dans les espaces de Banach. *Z. Wahrscheinlichkeitsth. verw. Gebiete* **49**, 211-220.

Hoffmann-Jørgensen, Jørgen (1974). Sums of independent Banach space valued random elements. *Studia Math.* **52**, 159–186.

Jain, Naresh (1976). An example concerning CLT and LIL in Banach space. *Ann. Probab.* **4**, 690-694.

Jain, Naresh C., and Marcus, Michael B. (1975b), Integrability of infinite sums of independent vector-valued random elements. *Trans. Amer. Math. Soc.* **212**, 1–36.

Kahane, J.-P. (1985). *Some Random Series of Functions*, 2d. ed. Cambridge University Press, New York.

Kolmogorov, A. N. (1931). Eine Verallgemeinerung des Laplace- Liapounoffschen Satzes. *Izv. Akad. Nauk SSSR Otdel. Mat. Estest. Nauk* (Ser. 7) no. 7, 959–962.

Kolmogorov, A. N. (1933). Über die Grenzwertsätze der Wahrscheinlichtkeitsrechnung. *Izv. Akad. Nauk SSSR (Bull. Acad. Sci. URSS)* (Ser. 7) no. 3, 363–372.

Koltchinskii, V.I. (1981). On the law of the iterated logarithm in Strassen's form for empirical measures. *Teor. Veroiatnost. i Mat. Statist.* (Kiev) **1981** no. 25, 40–47 (Russian); *Theory Probab. Math. Statist.* no. 25, 43–49 (English).

Koltchinskii, V.I. (1994). Komlos-Major-Tusnady approximation for the general empirical process and Haar expansions of classes of functions. *J. Theoret. Probab.* **7**, 73-118.

Komlós, J., Major, P., and Tusnády, G. (1975, 1976). An approximation of partial sums of independent RV's and the sample DF. I, II. *Z. Wahrscheinlichkeitsth. verw. Gebiete* **32**, 111–131; **34**, 33–58.

Kuelbs, J. (1977). Kolmogorov's law of the iterated logarithm for Banach space valued random variables. *Illinois J. Math.* **21**, 784–800.

Kuelbs, J., and Dudley, R.M. (1980). Log log laws for empirical measures. *Ann. Probab.* **8**, 405–418.

Kuelbs, J., and Zinn, J. (1979). Some stability results for vector valued random variables. *Ann. Probab.* **7**, 75–84.

Ledoux, M., and Talagrand, M. (1991). Probability in Banach Spaces. Springer, Berlin.

Major, Peter (1976). Approximation of partial sums of i.i.d. r.v.s when the summands have only two moments. *Z. Wahrscheinlichkeitsth. verw. Gebiete* **35**, 221–229.

Massart, P. (1989). Strong approximations for multidimensional empirical and related processes, via KMT constructions. *Ann. Probab.* **17**, 266-291.

Nagaev, S. V. (1970). On the speed of convergence in a boundary problem I, II. *Theor. Probab. Appl.* **15**, 163–186, 403–429 (English), 179–199, 419–441 (Russian).

Philipp, Walter (1980). Weak and L^p-invariance principles for sums of B-valued random variables. *Ann. Probab.* **8**, 68–82. Correction, *Ann. Probab.* **14** (1986), 1095–1101.

Pisier, G. (1976). Le théorème de la limite centrale et la loi du logarithme itéré dans les espaces de Banach. *Séminaire Maurey-Schwartz* 1975–76, Exposés III, IV, École Polytechnique, Paris.

Révész, P. (1976). On strong approximation of the multidimensional empirical process. *Ann. Probab.* **4**, 729-743.

Rio, Emmanuel (1993a,b). Strong approximation for set-indexed partial sum processes, via K.M.T. constructions, I, II. *Ann. Probab.* **21**, 759-790, 1706-1727.

Rio, Emmanuel (1994). Local invariance principles and their application to density estimation. *Probab. Theory Related Fields* **98**, 21-45.

Tusnády, G. (1977). A remark on the approximation of the sample DF in the multidimensional case. *Period. Math. Hungar.* **8**, 53-55.

Chapter 10

Universal and uniform central limit theorems

In this chapter we look at cases where a class $\mathcal{F}$ of measurable functions is a Donsker class for all probability laws on the underlying space (universal Donsker classes) and such classes where convergence in the central limit theorem holds uniformly in P (uniform Donsker classes).

10.1 Universal Donsker classes

Let X be a set and $\mathcal{A}$ a σ-algebra of subsets of X. Then a class $\mathcal{F}$ of measurable functions on X will be called a *universal Donsker class* if it is a P-Donsker class for every probability measure P on $(X, \mathcal{A})$.

Recall that every universal Donsker class of sets is a Vapnik-Cervonenkis class (Theorem 6.17), and the converse holds under measurability conditions (Corollary 6.16).

For a real-valued function f let $\operatorname{diam}(f) := \sup f - \inf f$. The following says that a universal Donsker class is uniformly bounded up to additive constants:

Proposition 10.1. *If $\mathcal{F}$ is a universal Donsker class, then*

$$\sup_{f \in \mathcal{F}} diam(f) < \infty.$$

A function h on $\mathcal{F}$ will be said to *ignore additive constants* if $h(f) = h(f + c)$ whenever $f \in \mathcal{F}$, c is a constant and $f + c \in \mathcal{F}$. Recall (Section 3.1) that a G_P process on $\mathcal{F}$ is called *coherent* if each sample function $G_P(\cdot)(\omega)$ is prelinear, bounded and uniformly continuous on $\mathcal{F}$ with respect to ρ_P. Here $\mathcal{F}$ is P-pregaussian if and only if a coherent G_P process on it exists (Theorem 3.1). If Z is a coherent G_P process then if f and $f + c$ are both in $\mathcal{F}$ for a constant c, we have $\rho_P(f, f + c) = 0$ so $Z(f) = Z(f + c)$ and $Z(\cdot)(\omega)$ ignores additive constants for all ω. Such a Z can be consistently extended to the set $\mathcal{F} + \mathbb{R}$ of all functions $f + c$, $f \in \mathcal{F}$, $c \in \mathbb{R}$, letting $Z(f + c) = Z(f)$. Then Z is a coherent G_P process on $\mathcal{F} + \mathbb{R}$, so $\mathcal{F} + \mathbb{R}$ is pregaussian.

Next, suppose $\mathcal{F}$ is a Donsker class for P. Then each function $P_n - P$ ignores additive constants on $\mathcal{F}$ and is well defined on $\mathcal{F} + \mathbb{R}$. If $\rho_P(f + c, g + d) < \delta$ for some $f, g \in \mathcal{F}$, constants c, d and $\delta > 0$, then $\rho_P(f, g) < \delta$. Since the total boundedness for ρ_P and asymptotic equicontinuity condition both extend directly from $\mathcal{F}$ to $\mathcal{F} + \mathbb{R}$, it follows by Theorem 3.28 that $\mathcal{F} + \mathbb{R}$ is a Donsker class for P. Thus if $\mathcal{F}$ is a universal Donsker class, so is $\mathcal{F} + \mathbb{R}$.

For a given law P, any subset of a Donsker class for P is also a P-Donsker class, e.g. by asymptotic equicontinuity (Theorem 3.28), so any subset of a universal Donsker class is also a universal Donsker class. If $c(\cdot)$ is any real-valued function on $\mathcal{F}$, then $\mathcal{F}$ is a universal Donsker class if and only if $\mathcal{G} := \{f - c(f) : f \in \mathcal{F}\}$ is. Taking $c(f) := \inf f$, we obtain from any class $\mathcal{F}$ of bounded functions a class $\mathcal{G}$ of nonnegative functions such that $\mathcal{G}$ is Donsker for a given P, or universal Donsker, if and only if $\mathcal{F}$ has the same property. If $\mathcal{F}$ and $\mathcal{G}$ are universal Donsker classes then $\mathcal{G}$ is uniformly bounded by Proposition 10.1.

Since multiplication by a positive constant is easily seen to preserve the Donsker property, in finding conditions for a class to be universal Donsker it will be enough to consider classes $\mathcal{F}$ of functions f with $0 \le f \le 1$.

The Vapnik-Červonenkis properties of classes of functions treated in Sections 4.7 and 4.8 (VC subgraph, VC major, VC hull) will all be seen to imply the universal Donsker property for uniformly bounded classes of functions. So the relations among these different VC properties are of interest here. Recall (Section 4.8) that $D^{(p)}(\varepsilon, \mathcal{F}, Q)$ is the largest m such that for some $f_1, \cdots, f_m \in \mathcal{F}$, $\int |f_i - f_j|^p dQ > \varepsilon^p$ for all $i \neq j$. Also, $D^{(2)}(\varepsilon, \mathcal{F})$ is the supremum over all laws Q with finite support of $D^{(2)}(\varepsilon, \mathcal{F}, Q)$.

The following is related to Theorem 4.52.

Proposition 10.2. *There exist uniformly bounded VC major (thus VC hull) classes which do not satisfy (4.7), thus are not VC subgraph classes.*

If

$$\int_0^1 (\log D^{(2)}(\varepsilon, \mathcal{F}))^{1/2} d\varepsilon < \infty, \tag{10.1}$$

as in Theorem 6.12 for $F \equiv 1$, then $\mathcal{F}$ will be said to satisfy *Pollard's entropy condition.*

Theorem 10.3. *If $\mathcal{F}$ is a uniformly bounded, image admissible Suslin class of measurable functions and satisfies Pollard's entropy condition then $\mathcal{F}$ is a universal Donsker class.*

Corollary 10.4. *A uniformly bounded, image admissible Suslin VC subgraph class is a universal Donsker class.*

Proof. This follows from Theorems 4.52 and 10.3. □

Specializing further, the set of indicators of an image admissible Suslin VC class of sets is a universal Donsker class (Corollary 6.16 for $F = 1$).

For a class $\mathcal{F}$ of real-valued functions on a set X, recall from Section 4.7 the class $H(\mathcal{F}, M)$ which is M times the symmetric convex hull of $\mathcal{F}$, and $\overline{H}_s(\mathcal{F}, M)$ which is the closure of $H(\mathcal{F}, M)$ for sequential pointwise convergence. Note that for any uniformly bounded class $\mathcal{F}$ of measurable functions for a σ-algebra $\mathcal{A}$ and any law Q defined on $\mathcal{A}$, $H(\mathcal{F}, M)$ is dense in $\overline{H}_s(\mathcal{F}, M)$ for the $L^2(Q)$ distance (or any $L^p(Q)$ distance, $1 \le p < \infty$).

Theorem 10.5. *If $\mathcal{F}$ is a Donsker class for a law P such that $\mathcal{F}$ has an envelope function in $\mathcal{L}^2(P)$, or a universal Donsker class, then for any $M < \infty$ $\overline{H}_s(\mathcal{F}, M)$ is a P-Donsker (resp. universal Donsker) class.*

By Theorem 10.3 above, for any $\delta > 0$, if $\log D^{(2)}(\varepsilon, \mathcal{F}) = O(1/\varepsilon^{2-\delta})$ as $\varepsilon \downarrow 0$, and if $\mathcal{F}$ satisfies a measurability condition (specifically, if $\mathcal{F}$ is image admissible Suslin) then $\mathcal{F}$ is a universal Donsker class. In the converse direction we have:

Theorem 10.6. *For a uniformly bounded class $\mathcal{F}$ to be a universal Donsker class it is necessary that*

$$\log D^{(2)}(\varepsilon, \mathcal{F}) = O(\varepsilon^{-2}) \quad as \quad \varepsilon \downarrow 0.$$

Theorem 10.6 is optimal, as the following shows:

Proposition 10.7. *There exists a universal Donsker class $\mathcal{E}$ such that*

$$\liminf_{\delta \downarrow 0} \delta^2 \log D^{(2)}(\delta, \mathcal{E}) \ > \ 0.$$

Proof. Let $A_j := A(j)$ be disjoint, nonempty measurable sets for $j = 1, 2, \cdots$. Let $\| \cdot \|_2$ be the ℓ^2 norm, $\|x\|_2 = (\sum_j x_j^2)^{1/2}$ for $x = \{x_j\}_{j=1}^{\infty}$. Let

$$\mathcal{E} := \left\{ \sum_j x_j 1_{A(j)} : \ \|x\|_2 \le 1 \right\}.$$

(So, $\mathcal{E}$ is an ellipsoid with center 0 and semiaxes $1_{A(j)}$.) The proof that $\mathcal{E}$ has the stated properties is omitted. □

Proposition 10.8. *There is a uniformly bounded class $\mathcal{F}$ of measurable functions, which is not a universal Donsker class, such that*

$$\log D^{(2)}(\varepsilon, \mathcal{F}) \ \le \ \frac{2}{\varepsilon^2 \log(1/\varepsilon)} \quad as \quad \varepsilon \downarrow 0.$$

Proof. Let $B_j := B(j)$ be disjoint nonempty measurable sets. Recall that $Lx := \max(1, \log x)$. Let $\alpha_j := 1/(jLj)^{1/2}$, $j \ge 1$, and

$$\mathcal{F} \ = \ \{\textstyle\sum_{j=1}^{\infty} x_j 1_{B(j)} : \ x_j = \pm \alpha_j \text{ for all } j\}.$$

Take c such that $\sum_{j=1}^{\infty} p_j = 1$, where $p_j := c(\alpha_j / LLj)^2$. Take a probability measure P with $P(B_j) = p_j$ for all j. The rest of the proof is omitted. □

Theorems 10.3 and 10.6 show that Pollard's entropy condition (10.1) comes close to characterizing the universal Donsker property, but Propositions 10.7 and 10.8 show that there is no characterization of the universal Donsker property in terms of $D^{(2)}$.

10.2 Metric entropy of convex hulls in Hilbert space

Let H be a real Hilbert space and for any subset B of H let $\mathrm{co}(B)$ be its convex hull,

$$\mathrm{co}(B) \ := \ \{\textstyle\sum_{j=1}^{k} t_j x_j : \ t_j \ge 0, \ \sum_{j=1}^{k} t_j = 1, \ x_j \in B, \ k = 1, 2, \cdots\}.$$

Recall that $D(\varepsilon, B)$ is the maximum number of points in B more than ε apart.

Theorem 10.9. *Suppose that B is an infinite subset of a Hilbert space H, $\|x\| \le 1$ for all $x \in B$ and that for some $K < \infty$ and $0 < \gamma < \infty$, we have $D(\varepsilon, B) \le K\varepsilon^{-\gamma}$ for $0 < \varepsilon \le 1$. Let $s := 2\gamma/(2+\gamma)$. Then for any $t > s$, there are constants C_1 and C_2, which depend only on K, γ and t, such that*

$$D(\varepsilon, co(B)) \ \le \ C_1 \exp(C_2 \varepsilon^{-t}) \ \ \textit{for} \ \ 0 < \varepsilon \le 1.$$

Note van der Vaart and Wellner (1996), Theorem 2.6.9, and Carl (1997) give the sharper bound with $t = s$.

Example. The exponent $2\gamma/(2+\gamma)$ in Theorem 10.9 is sharp for the following set B. Let $\{e_n\}_{n\ge1}$ be an orthonormal basis of H and for $0 < \gamma < \infty$ let $B := \{n^{-1/\gamma} e_n\}_{n\ge1} \cup \{-n^{-1/\gamma} e_n\}_{n\ge1}$.

Recall the definitions of $D^{(2)}$ from Section 4.8 and $\overline{H}_s$ from after Corollary 10.4.

Corollary 10.10. *If $\mathcal{G}$ is a uniformly bounded class of measurable functions and for some $K < \infty$ and $0 < \gamma < \infty$, $D^{(2)}(\varepsilon, \mathcal{G}) \le K\varepsilon^{-\gamma}$ for $0 < \varepsilon < 1$, then for any $t > r := 2\gamma/(2+\gamma)$, and for the constants $C_i = C_i(2K, \gamma, t)$, $i = 1, 2$, of Theorem 10.9,*

$$D^{(2)}(\varepsilon, \overline{H}_s(\mathcal{G}, 1)) \ \le \ C_1 \exp(C_2 \varepsilon^{-t}) \ \ \textit{for} \ \ 0 < \varepsilon < 1.$$

Now, recall the notions of VC subgraph and VC subgraph hull class from Section 4.7.

Corollary 10.11. *If $\mathcal{G}$ is a uniformly bounded VC subgraph class and $M < \infty$ then the VC subgraph hull class $\overline{H}_s(\mathcal{G}, M)$ satisfies Pollard's entropy condition (10.1).*

Proof. Let $M\mathcal{G} := \{Mg : g \in \mathcal{G}\}$. Then $M\mathcal{G}$ is a uniformly bounded VC subgraph class. By Theorem 4.52(a), $M\mathcal{G}$ satisfies the hypothesis of Corollary 10.10, so $r < 2$ and we can take $t < 2$. □

Remark. It follows by Theorem 6.12 that for a uniformly bounded VC subgraph class $\mathcal{G}$, if $\mathcal{F} \subset \overline{H}_s(\mathcal{G}, M)$ and $\mathcal{F}$ satisfies the image admissible Suslin measurability condition, then $\mathcal{F}$ is a universal Donsker class. This also follows from Corollary 10.4 and Theorem 10.5.

Example. Let $\mathcal{C}$ be the set of all intervals $(a, b]$ for $0 \le a \le b \le 1$. Let G be the set of all real functions f on $[0, 1]$ such that $|f(x)| \le 1/2$ for all x, $|f(x) - f(y)| \le |x - y|$ for $0 < x, y < 1$, and $f(x) = 0$ for $x \le 0$ or $x \ge 1$. Each f in G has total variation at most 2 (at most 1 on the open interval $0 < x < 1$ and $1/2$ at each endpoint $0, 1$). By the Jordan decomposition we have, for each $f \in G$, $f = g - h$ where g and h are both nondecreasing functions, 0 for $x \le 0$. Then g and h have equal total variations ≤ 1 and $G \subset \overline{H}_s(\mathcal{C}, 2)$ by the proof of Theorem 4.50(b). Let P be Lebesgue measure on $[0, 1]$. By Theorem 8.4, and since $(\int |f|^2 dP)^{1/2} \ge \int |f| dP$, there is a $c > 0$ such that $D^{(2)}(\varepsilon, G) \ge e^{c/\varepsilon}$ as $\varepsilon \downarrow 0$ (consider laws with finite support which approach P). Since $S(\mathcal{C}) = 2$, the exponent γ can be taken as 2 by Corollary 4.4 and Haussler's Theorem 4.47. Thus t in Corollary 10.10 can be any number > 1, and we saw above that it cannot be < 1 in this case, so again the exponent is sharp.

10.3 Uniform Donsker classes

A class $\mathcal{F}$ of measurable functions on a measurable space $(X, \mathcal{A})$ is a uniform Donsker class if it is a universal Donsker class and the convergence in law of ν_n to G_P is also uniform in P. To formulate this notion precisely we will follow Giné and Zinn (1991) in using the dual-bounded-Lipschitz metric β as defined just before Theorem 3.23.

Let $\mathcal{P}(X)$ be the set of all probability measures on $(X, \mathcal{A})$ and let $\mathcal{P}_f(X)$ be the set of all laws in $\mathcal{P}(X)$ with finite support. For $\delta > 0$, a class $\mathcal{F}$ of measurable real-valued functions on X and a pseudo-metric d on $\mathcal{F}$ let

$$\mathcal{F}'(\delta, d) \ := \ \{f - g : \ f, g \in \mathcal{F}, \ d(f, g) \leq \delta\}.$$

Definitions. A class $\mathcal{F}$ is *uniformly pregaussian* if it is pregaussian for all $P \in \mathcal{P}(X)$, and if, for a coherent version of G_P for each P, we have both

$$\sup_{P \in \mathcal{P}(X)} E\|G_P\|_{\mathcal{F}} < \infty$$

and

$$\lim_{\delta \downarrow 0} \sup_{P \in \mathcal{P}(X)} E\|G_P\|_{\mathcal{F}'(\delta, \rho_P)} = 0.$$

The class $\mathcal{F}$ is *finitely uniformly pregaussian* if the same holds with $\mathcal{P}_f(X)$ in place of $\mathcal{P}(X)$.

The class $\mathcal{F}$ is a *uniform Donsker class* if it is uniformly pregaussian and

$$\lim_{n \to \infty} \sup_{P \in \mathcal{P}(X)} \beta_{\mathcal{F}}(\nu_n, G_P) \ = \ 0$$

where $\beta_{\mathcal{F}}$ is the dual-bounded-Lipschitz metric β based on $\|\cdot\|_{\mathcal{F}}$ as in Section 3.6.

Giné and Zinn (1991) proved:

Theorem. Let $(X, \mathcal{A})$ be a measurable space and $\mathcal{F}$ a class of real-valued measurable functions on X. Then $\mathcal{F}$ is a uniform Donsker class if and only if it is finitely uniformly pregaussian and thus, if and only if it is uniformly pregaussian.

The theorem is a very useful characterization since it's easier to check the finitely uniformly pregaussian property than to check the uniform Donsker property directly.

A uniform Donsker class is clearly a universal Donsker class. Thus it is uniformly bounded up to additive constants (Proposition 10.1).

Giné and Zinn show (as a corollary of their Proposition 3.1) that the hypotheses of Theorem 10.3 (Pollard's entropy condition (10.1), together with suitable boundedness and measurability) actually imply that $\mathcal{F}$ is uniformly Donsker. Thus most of the examples of universal Donsker classes treated in Sections 10.1 and 10.2 are uniformly Donsker. An exception is the "ellipsoid" universal Donsker class of Proposition 10.7.

Some uniform Donsker classes of functions on $\mathbb{R}$ will be defined. For any function f from $\mathbb{R}$ into $\mathbb{R}$ and $0 < p < \infty$ the *p-variation* of f is defined by

$$v_p(f) \ := \ \sup\{\sum_{j=1}^{n} |f(x_j) - f(x_{j-1})|^p :$$

$$-\infty < x_0 < x_1 < \cdots < x_n < +\infty, \ n = 1, 2, \cdots\}.$$

For $p = 1$ this is the ordinary total variation.

Theorem (Dudley, 1992). For each $M < \infty$ and p with $0 < p < 2$, the class $\mathcal{F} := \{f : \mathbb{R} \mapsto \mathbb{R},\ v_p(f) \leq M\}$ is a uniform Donsker class.

REFERENCES

Carl, B. (1982). On a characterization of operators from ℓ_q into a Banach space of type p with some applications to eigenvalue problems. *J. Funct. Anal.* **48**, 394-407.

Carl, B. (1997). Metric entropy of convex hulls in Hilbert space. *Bull. London Math. Soc.* **29**, 452-458.

Dudley, R. M. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Funct. Anal.* **1**, 290-330.

Dudley, R. M. (1987). Universal Donsker classes and metric entropy. *Ann. Probab.* **15**, 1306-1326.

Dudley, R. M. (1992). Fréchet differentiability, p-variation and uniform Donsker classes. *Ann. Probab.* **20**, 1968-1982.

Giné, Evarist, and Zinn, Joel (1991). Gaussian characterization of uniform Donsker classes of functions. *Ann. Probab.* **19**, 758-782.

Pisier, G. (1981). Remarques sur un resultat non publié de B. Maurey. *Séminaire d'Analyse Fonctionelle 1980-1981* V.1-V.12. Ecole Polytechnique, Centre de Mathématiques, Palaiseau.

Chapter 11

The two-sample case, the bootstrap and confidence sets

11.1 The two-sample case

Let $X_1, \cdots, X_m, \cdots, Y_1, \cdots, Y_n, \cdots$, be some random variables taking values in a set A where $(A, \mathcal{A})$ is a measurable space. Thus $(X_1, \cdots, X_m)$ and $(Y_1, \cdots, Y_n)$ are "samples," of which we have two. Let

$$P_m := \frac{1}{m}\sum_{i=1}^{m} \delta_{X_i}, \quad Q_n := \frac{1}{n}\sum_{j=1}^{n} \delta_{Y_j}.$$

The object of two-sample tests in statistics is to decide whether P_m and Q_n are empirical measures from the same, but unknown, law (probability measure) on $(A, \mathcal{A})$. Since P is unknown, we cannot directly compare P_m or Q_n to it by forming $m^{1/2}(P_m - P)$ or $n^{1/2}(Q_n - P)$. Instead, P_m and Q_n can be compared to each other, setting

$$\nu_{m,n} := \left(\frac{mn}{m+n}\right)^{1/2}(P_m - Q_n).$$

The *basic hypothesis* will be that there are two laws P, Q on $(A, \mathcal{A})$ and a product of two countable products of copies of $(A, \mathcal{A})$ with factor laws P and Q respectively, namely

$$(\Omega, \mathcal{D}, \Pr) = (\Omega_1, \mathcal{B}_1, \Pr_1) \times (\Omega_2, \mathcal{B}_2, \Pr_2)$$

where

$$(\Omega_1, \mathcal{B}_1, \Pr_1) = \prod_{i=1}^{\infty}(A_i, \mathcal{A}, P), \qquad (\Omega_2, \mathcal{B}_2, \Pr_2) = \prod_{j=1}^{\infty}(B_j, \mathcal{A}, Q),$$

and each A_i and B_j is a copy of A. On these products let X_i be the A_i coordinate and Y_j the B_j coordinate. If $\mathcal{P}$ is a class of laws on $(A, \mathcal{A})$, the $(\mathcal{P})$ *null hypothesis* is that in addition, $P = Q \in \mathcal{P}$. A class $\mathcal{F}$ of measurable functions on $(A, \mathcal{A})$ will be called a *$\mathcal{P}$-universal Donsker class* if it is a P-Donsker class for every $P \in \mathcal{P}$.

Theorem 11.1. *Suppose $\mathcal{F}$ is a $\mathcal{P}$-universal Donsker class of functions on $(A, \mathcal{A})$. Then for each $P \in \mathcal{P}$, under the $(\mathcal{P})$ null hypothesis, $\nu_{m,n} \Rightarrow G_P$ as $m, n \to \infty$.*

The classical two-sample situation is the special case where $A = \mathbb{R}$, $\mathcal{A}$ is the Borel σ-algebra, $\mathcal{F}$ is the set of all indicator functions of half-lines $(-\infty, x]$, and $\mathcal{P}$ is the set of all continuous (nonatomic) laws on $\mathbb{R}$. Thus $m^{1/2}(P_m - P)((-\infty, x]) = m^{1/2}(F_m - F)(x)$ where F is the distribution function of P and F_m an empirical distribution function. Here $\mathcal{F}$ is a universal Donsker class by any of several previous results, for example Corollary 6.16, and $G_P(1_{(-\infty,x]}) = y_{F(x)}$ where y is the Brownian bridge. (Actually, $\mathcal{F}$ is a uniform Donsker class.) Since F is continuous, it takes all values in the open interval $(0,1)$, $y_0 \equiv y_1 \equiv 0$, and $y_t \to 0$ as $t \downarrow 0$ or $t \uparrow 1$. Thus the distribution of $\sup_x y_{F(x)}$ and $\sup_x |y_{F(x)}|$ and the joint distribution of $(\inf_x y_{F(x)}, \sup_x y_{F(x)})$ do not depend on F, for $P \in \mathcal{P}$. Let F_m and G_n be independent empirical distribution functions for the same F. Let $H_{mn} := (mn/(m+n))^{1/2}(F_m - G_n)$. By Theorems 3.23 and 3.26, which extend straightforwardly to limits as $m, n \to \infty$, the distributions of the supremum, supremum of absolute value, and supremum minus infimum of H_{mn} converge to those of the same functionals for y_t. Thus we get:

Corollary 11.2. *If F_m and G_n are independent empirical distribution functions for a continuous distribution on $\mathbb{R}$, then for any $u > 0$,*
(a) $\lim_{m,n\to\infty} P(\sup_x H_{mn}(x) > u) = \exp(-2u^2)$,
(b) $\lim_{m,n\to\infty} P(\sup_x |H_{mn}(x)| > u) = 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp(-2k^2u^2)$,
(c) $\lim_{m,n\to\infty} P(\sup_x H_{mn}(x) - \inf_y H_{mn}(y) > u) = 2\sum_{k=1}^{\infty}(4k^2u^2 - 1)\exp(-2k^2u^2)$.

Proof. The distributions of the given functionals of the Brownian bridge y_t are given in RAP, Propositions 12.3.3, 12.3.4, and 12.3.6. All three are continuous in u for $u > 0$. Thus convergence follows from RAP, Theorem 9.3.6. □

11.2 A bootstrap central limit theorem in probability

Iterating the operation by which we get an empirical measure P_n from a law P, we form the bootstrap empirical measure P_n^B by sampling n independent points whose distribution is the empirical measure P_n. The bootstrap was first introduced in nonparametric statistics, where the law P is unknown and we want to make inferences about it from the observed P_n. This can be done by way of bootstrap central limit theorems, which say that under some conditions, $n^{1/2}(P_n^B - P_n)$ behaves like $n^{1/2}(P_n - P)$ and both behave like G_P.

Let $(S, \mathcal{S}, P)$ be a probability space and $\mathcal{F}$ a class of real-valued measurable functions on S. Let as usual $X_1, X_2, \cdots$, be coordinates on a countable product of copies of $(S, \mathcal{S}, P)$. Then let $X_{n1}^B, \cdots, X_{nn}^B$ be independent with distribution P_n. Let

$$P_n^B := \frac{1}{n}\sum_{j=1}^{n} \delta_{X_{nj}^B}.$$

Then P_n^B will be called a *bootstrap empirical measure.*

A statistician has a data set, represented by a fixed P_n, and estimates the distribution of P_n^B by repeated resampling from the same P_n. So we are interested not so much in the unconditional distribution of P_n^B as P_n varies, but rather in the conditional distribution of P_n^B given P_n or $(X_1, \cdots, X_n)$. Let $\nu_n^B := n^{1/2}(P_n^B - P_n)$.

The limit theorems will be formulated in terms of dual-bounded-Lipschitz "metric" β of Section 3.6, which metrizes convergence in distribution for not necessarily measurable random

elements of a possibly nonseparable metric space (S, d), to a limit which is a measurable random variable with separable range. Let $\beta_{\mathcal{F}}$ be the β distance where d is the metric defined by the norm $\|\cdot\|_{\mathcal{F}}$.

Definition. Let $(S, \mathcal{S}, P)$ be a probability space and $\mathcal{F}$ a class of measurable real-valued functions on S. Then the *bootstrap central limit theorem holds in probability* (respectively, *almost surely*) for P and $\mathcal{F}$ if and only if $\mathcal{F}$ is pregaussian for P and $\beta_{\mathcal{F}}(\nu_n^B, G_P)$, conditional on $X_1, \cdots, X_n$, converges to 0 in outer probability (resp., almost uniformly) as $n \to \infty$.

In other words, the bootstrap central limit theorem holds in probability if and only if for any $\varepsilon > 0$ there is an n_0 large enough so that for any $n \geq n_0$, there is a set A of values of $X^{(n)} := (X_1, \cdots, X_n)$ with $P^n(A) < \varepsilon$ such that if $X^{(n)} \notin A$, and $\nu_n^B(X^{(n)})(\cdot)$ is ν_n^B conditional on $X^{(n)}$, we have $\beta_{\mathcal{F}}(\nu_n^B(X^{(n)})(\cdot), G_P) < \varepsilon$.

A main bootstrap limit theorem will be stated.

Theorem 11.3. (Giné and Zinn) *Let $(X, \mathcal{A}, P)$ be any probability space. Then the bootstrap central limit theorem holds in probability for P and $\mathcal{F}$ if $\mathcal{F}$ is a Donsker class for P.*

Remarks Giné and Zinn (1990), see also Giné (1997), also proved "only if" under a measurability condition, and proved a corresponding almost sure form of the theorem where $\mathcal{F}$ has an $\mathcal{L}^2$ envelope up to additive constants.

The proof of Theorem 11.3 is quite long. This will be illustrated just by giving the lemmas and other theorems used in its proof, without their proofs. E^B, Pr^B and $\mathcal{L}^B$ will denote the conditional expectation, probability and law given the sample $X^{(n)} := (X_1, \cdots, X_n)$. Given the sample, ν_n^B has only finitely many possible values.

First, a finite-dimensional bootstrap central limit theorem is needed.

Theorem 11.4. *Let $X_1, X_2, \cdots$ be i.i.d. random variables with values in $\mathbb{R}^d$ and let $X_{n,i}^B$, $i = 1, \cdots, n$, be i.i.d. (P_n), where $P_n := \frac{1}{n}\sum_{j=1}^n \delta_{X_j}$. Let $\overline{X}_n := \frac{1}{n}\sum_{j=1}^n X_j$. Assume that $E|X_1|^2 < \infty$. Let C be the covariance matrix of X_1, $C_{rs} := E(X_{1r}X_{1s}) - E(X_{1r})E(X_{1s})$. Then for the usual convergence of laws in $\mathbb{R}^d$, almost surely as $n \to \infty$,*

$$\mathcal{L}^B(n^{-1/2}\sum_{j=1}^n (X_{n,j}^B - \overline{X}_n)) \to N(0, C). \tag{11.1}$$

Next is a desymmetrization fact:

Lemma 11.5. *Let T be a set and for any real-valued function f on T let $\|f\|_T := \sup_{t\in T} |f(t)|$. Let X and Y be two stochastic processes indexed by $t \in T$ defined on a probability space $(\Omega \times \Omega', \mathcal{S} \otimes \mathcal{S}', P \times P')$, where $X(t)(\omega, \omega')$ depends only on $\omega \in \Omega$ and $Y(t)(\omega, \omega')$ only on $\omega' \in \Omega'$. Then*
(a) for any $s > 0$ and any $u > 0$ such that $\sup_{t\in T} \Pr\{|Y(t)| \geq u\} < 1$, we have

$$\Pr^*(\|X\|_T > s) \leq \Pr^*\{\|X - Y\|_T > s - u\}/[1 - \sup_{t\in T} \Pr^*\{|Y(t)| \geq u\}].$$

(b) If $\theta > \sup_{t\in T} E(Y(t)^2)$ then for any $s > 0$,

$$P^*(\|X\|_T > s) \leq 2\Pr^*(\|X - Y\|_T > s - (2\theta)^{1/2}).$$

Now ε_i will denote i.i.d. Rademacher variables, i.e. variables with the distribution $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2$, defined on a probability space $(\Omega_\varepsilon, \mathcal{S}_\varepsilon, P_\varepsilon)$ which we can take as $[0,1]$ with Lebesgue measure. We also take the product of $(\Omega_\varepsilon, \mathcal{S}_\varepsilon, P_\varepsilon)$ with the probability space on which other random variables and elements are defined.

Next, some hypotheses will be given for later reference.

Let $(S, \mathcal{S}, P)$ be a probability space and $(S^n, \mathcal{S}^n, P^n)$ a Cartesian product of n copies of $(S, \mathcal{S}, P)$. Let $\mathcal{F} \subset \mathcal{L}^2(S, \mathcal{S}, P)$. Let $\varepsilon_1, \cdots, \varepsilon_n$ be i.i.d. Rademacher variables defined on a probability space $(\Omega', \mathcal{A}, P')$. Then take the probability space

$$(S^n \times \Omega', \mathcal{S}^n \otimes \mathcal{A}, P^n \times P'). \tag{11.2}$$

References to (11.2) will also be to the preceding paragraph.

Here is another fact on symmetrization and desymmetrization.

Lemma 11.6. *Under (11.2), for any $t > 0$ and $n = 1, 2, \cdots$,*
(a) $\Pr^* \left(\| \sum_{i=1}^n \varepsilon_i f(X_i) \|_{\mathcal{F}} > t \right) \le 2 \max_{k \le n} \Pr^* \left(\| \sum_{i=1}^k f(X_i) \|_{\mathcal{F}} > t/2 \right)$.
(b) Suppose that $\alpha^2 := \sup_{f \in \mathcal{F}} \int (f - Pf)^2 dP < \infty$. *Then for* $t > 2^{1/2} \alpha n^{1/2}$ *and all* $n = 1, 2, \cdots$,

$$C_t := \Pr^* \left(\| \sum_{i=1}^n (f(X_i) - Pf) \|_{\mathcal{F}} > t \right)$$

$$\le 4 \Pr^* \left\{ \| \sum_{i=1}^n \varepsilon_i f(X_i) \|_{\mathcal{F}} > (t - (2n)^{1/2} \alpha)/2 \right\}.$$

Next, we have some consequences or forms of Jensen's inequality.

Lemma 11.7. *(a) Let $(S, \mathcal{S}, P)$ be a probability space and $\mathcal{F} \subset \mathcal{L}^1(S, \mathcal{S}, P)$. Then $\| Ef \|_{\mathcal{F}} \le E \| f \|_{\mathcal{F}}^*$.*
(b) On a product space $(A, \mathcal{A}, P) \times (A, \mathcal{A}, Q)$ let X and Y be coordinate functions. Let $\mathcal{F}$ be a class of real-valued measurable functions on $(A, \mathcal{A})$. If $Qf = 0$ for all $f \in \mathcal{F}$, then

$$E^* \| f(X) \|_{\mathcal{F}} \le E^* \| f(X) + f(Y) \|_{\mathcal{F}}. \tag{11.3}$$

The following lemma and theorem are known as Hoffmann-Jørgensen inequalities, see Hoffmann-Jørgensen (1974, p. 164).

Lemma 11.8. *Let $X_1, \cdots, X_n$ be coordinates on a product probability space $(S^n, \mathcal{S}^n, \Pi_{j=1}^n P_j)$. Let $\mathcal{F}$ be a class of measurable real-valued functions on $(S, \mathcal{S})$. Let $S_k(f) := \sum_{j=1}^k f(X_j)$ for $k = 1, \cdots, n$. Then for any $s > 0$ and $t > 0$,*

$$\Pr \left(\max_{k \le n} \| S_k(f) \|_{\mathcal{F}}^* > 3t + s \right)$$

$$\le \left(P \left\{ \max_{k \le n} \| S_k \|_{\mathcal{F}}^* > t \right\} \right)^2 + P \left(\max_{j \le n} \| X_j \|_{\mathcal{F}}^* > s \right). \tag{11.4}$$

Theorem 11.9. *Let $0 < p < \infty$, $n = 1, 2, \cdots$, let $X_1, \cdots, X_n$ be coordinates on a product probability space $(S^n, \mathcal{S}^n, \Pi_{j=1}^n P_j)$. Let $\mathcal{F}$ be a class of measurable real-valued functions on $(S, \mathcal{S})$ such that for $i = 1, \cdots, n$, $E(\|f(X_i)\|_{\mathcal{F}}^{*p}) < \infty$. Let*

$$u := \inf\left\{t > 0 : \Pr\left[\max_{k \le n} \left\|\sum_{i=1}^k f(X_i)\right\|_{\mathcal{F}}^* > t\right] \le 1/(2 \cdot 4^p)\right\}.$$

Then

$$E \max_{k \le n} \left(\left\|\sum_{i=1}^k f(X_i)\right\|_{\mathcal{F}}^{*p}\right) \le 2 \cdot 4^p E(\max_{j \le n}(\|f(X_j)\|_{\mathcal{F}}^*)^p) + 2(4u)^p.$$

Next is another symmetrization-desymmetrization inequality.

Lemma 11.10. *Under (11.2),*

$$\begin{aligned} \tfrac{1}{2} E^* \| \textstyle\sum_{j=1}^n \varepsilon_j (f(X_j) - Pf)\|_{\mathcal{F}} &\le E^* \| \textstyle\sum_{j=1}^n (f(X_j) - Pf)\|_{\mathcal{F}} \\ &\le 2E^* \| \textstyle\sum_{j=1}^n \varepsilon_j (f(X_j) - Pf)\|_{\mathcal{F}}, \end{aligned} \tag{11.5}$$

which also holds if the Pf in the last expression is deleted.

Next will be some Poissonization facts. Recall that any real-valued stochastic process $X(t)$, $t \in T$, indexed by a set T, has a law P_X defined on the space $\mathbb{R}^T$ of all real-valued functions on T, on the smallest σ-algebra for which the coordinate projections $f \mapsto f(t)$ are measurable for each $t \in T$. The process is called *centered* if $EX(t) = 0$ for all $t \in T$. Recall that Y has a Poisson distribution with parameter $\lambda > 0$ if and only if $P(Y = k) = e^{-\lambda}\lambda^k/k!$ for $k = 1, 2, \cdots$.

Lemma 11.11. *Let T be any set. For each $i = 1, \cdots, n$, let $X(i) := X_i$ be a centered, real-valued stochastic process indexed by T. Let $X_{i,j}$, $j = 1, 2, \cdots$, be independent copies of X_i, taken as coordinates on a product, say A, of copies of $(\mathbb{R}^T, P_{X(i)})$. Let $N_i := N(i)$, $i = 1, 2, \cdots$, be i.i.d. Poisson variables with parameter $\lambda = 1$, defined on a probability space (Ω', P'), and take the product of this space with A, so that $N_1, N_2, \cdots$ are independent of $X_{i,j}$. Let $\|f\|_T := \sup_{t \in T} |f(t)|$ and $x \wedge y := \min(x, y)$. Then*

$$E^* \| \sum_{i=1}^n X_i \|_T \le \frac{e}{e-1} E^* \| \sum_{i=1}^n \sum_{j=1}^{N(i)} X_{i,j} \|_T. \tag{11.6}$$

For any two finite signed measures μ and ν on the Borel sets of a separable Banach space B recall that the convolution $\mu * \nu$ is defined by $(\mu * \nu)(A) := \int \mu(A - x) d\nu(x)$ for any Borel set A. Here convolution is commutative and associative. For any finite signed measure μ on B and $k = 1, 2, \cdots$, let μ^k be the kth convolution power $\mu * \cdots * \mu$ to k factors. Let $e^\mu := \exp(\mu) := \sum_{k=0}^\infty \mu^k/k!$. Let $\mu^0 := \delta_0$. If $\mu \ge 0$ let $\mathrm{Pois}(\mu) := e^{-\mu(B)} e^\mu$. If μ and ν are two finite measures on B it is straightforward to check that $\mathrm{Pois}(\mu + \nu) = \mathrm{Pois}(\mu) * \mathrm{Pois}(\nu)$. If X is a measurable function from a probability space $(\Omega, \mathcal{S}, P)$ into a measurable space $(S, \mathcal{A})$, recall that the *law* of X is the image measure $\mathcal{L}(X) := P \circ X^{-1}$ on $\mathcal{A}$. For any $c > 0$ and $x \in B$, $\mathrm{Pois}(c\delta_x) = \mathcal{L}(N_c x)$ where N_c is a Poisson random variable with parameter c.

If $\mathcal{L}(X_{i,j}) = \mu_i$ for $j = 1, 2, \cdots$ and $N_i := N(i)$ are Poisson with parameter 1, where all $X_{i,j}$ and N_r are jointly independent, $i, r = 1, \cdots, k$, $j = 1, 2, \cdots$, then by induction on k,

$$\mathcal{L}\left(\sum_{i=1}^{k} \sum_{j=1}^{N(i)} X_{i,j}\right) = \mathrm{Pois}\left(\sum_{r=1}^{k} \mu_r\right).$$

If X_i are random variables with values in a separable Banach space with $EX_i = 0$ for each $i = 1, \cdots, n$, then the depoissonization inequality (11.6) gives

$$E\|\sum_{j=1}^{n} X_j\| \leq \frac{e}{e-1} \int \|x\| \, d\mathrm{Pois}\left(\sum_{j=1}^{n} \mathcal{L}(X_j)\right). \tag{11.7}$$

Here is another depoissonization inequality.

Lemma 11.12. *Let $(B, \|\cdot\|)$ be a normed space. For each $n = 1, 2, \cdots$, let $v_1, \cdots, v_n$ be n distinct points of B and $v := (v_1 + \cdots v_n)/n$. Let $V_1, \cdots, V_n$ be i.i.d. B-valued random variables with $\Pr(V_i = v_j) = 1/n$ for $i, j = 1, \cdots, n$. Let $N_1, \cdots, N_n$ be Poisson variables with parameter 1. Let all V_i and N_j be jointly independent. Then*

$$E\|\sum_{j=1}^{n} (V_j - v)\| \leq \frac{e}{e-1} E\|\sum_{j=1}^{n} (N_j - 1)(v_j - v)\|. \tag{11.8}$$

The next fact is about triangular arrays with i.i.d. summands.

Theorem 11.13. *Let (T, d) be a totally bounded pseudo-metric space. For each $n = 1, 2, \cdots$, suppose given a product of n copies of a probability space, $\Omega_n^n := (\Omega_n, \mathcal{A}_n, P^{(n)})^n$. For each n, let Y_n be a measurable real-valued function on Ω_n. For $\omega := \{\omega_j\}_{j=1}^n \in \Omega^n$ let $X_{n,j}(\omega) := Y_n(\omega_j)$. Thus $X_{n,j}$ are i.i.d. copies of Y_n. Suppose that each Y_n has bounded sample paths a.s. Assume that*
(i) For all t in a dense subset $D \subset T$ for d and for all $\beta > 0$,

$$\lim_{n\to\infty} n\Pr^*\{|X_{n,1}(t)| > \beta n^{1/2}\} = 0, \tag{11.9}$$

(ii) For any $\delta > 0$, $\sup_{t\in T} \Pr\{|Y_n(t)| > \delta n^{1/2}\} \to 0$ as $n \to \infty$, and
(iii) for all $\varepsilon > 0$, as $\delta \downarrow 0$,

$$\limsup_{n\to\infty} \Pr^*\Big\{n^{-1/2} \sup_{d(s,t)\leq\delta} |\sum_{i=1}^{n} X_{n,i}(t) - EX_{n,i}(t)$$

$$-X_{n,i}(s) + EX_{n,i}(s)| > \varepsilon\Big\} \to 0.$$

Then for any $\gamma > 0$,

$$\lim_{n\to\infty} n\Pr^*\left\{\|X_{n,1}\|_T > \gamma n^{1/2}\right\} = 0. \tag{11.10}$$

If the hypotheses only hold along a subsequence n_k, then the conclusion also holds along the subsequence.

Next will be a characterization of Donsker classes, in which the asymptotic equicontinuity condition (Theorem 3.28) is put into symmetrized forms. For a class $\mathcal{F}$ of functions let

$$\mathcal{F}'_\delta := \mathcal{F}_\delta := \{f - g : f, g \in \mathcal{F}, \ \rho_P(f, g) < \delta\}.$$

Let $\|\cdot\|_{\delta,\mathcal{F}} := \|\cdot\|_{\mathcal{G}}$ where $\mathcal{G} := \mathcal{F}'_\delta$.

Theorem 11.14. *Let $(X, \mathcal{A}, P)$ be a probability space and $\mathcal{F}$ a class of functions included in $\mathcal{L}^2(X, \mathcal{A}, P)$. Assume (11.2), so that ε_i are i.i.d. Rademacher functions, independent of $X_1, X_2, \cdots$. Suppose that for each $x \in X$,*

$$F_c(x) \ := \ \sup_{f \in \mathcal{F}} |f(x) - Pf| < \infty. \tag{11.11}$$

Then the following are equivalent:
(a) $\mathcal{F}$ is a Donsker class for P;
(b) $\mathcal{F}$ is totally bounded for ρ_P and for any $\varepsilon > 0$, as $\delta \to 0$,

$$\limsup_{n \to \infty} \Pr^* \left\{ n^{-1/2} \| \textstyle\sum_{i=1}^n \varepsilon_i (f(X_i) - Pf) \|_{\delta, \mathcal{F}} > \varepsilon \right\} \ \to \ 0.$$

(c) $(\mathcal{F}, \rho_P)$ is totally bounded and as $\delta \to 0$,

$$\limsup_{n \to \infty} n^{-1/2} E^* \| \textstyle\sum_{i=1}^n \varepsilon_i (f(X_i) - Pf) \|_{\delta, \mathcal{F}} \ \to \ 0;$$

(d) $(\mathcal{F}, \rho_P)$ is totally bounded and for $\nu_n \ := \ n^{1/2}(P_n - P)$, as $\delta \to 0$, $\limsup_{n \to \infty} E^* \|\nu_n\|_{\delta, \mathcal{F}} \to 0$.

In proving (c) the following will be helpful.

Lemma 11.15. *Let ξ_i, $i = 1, 2, \cdots$, be i.i.d. nonnegative random variables such that*

$$M \ := \ \sup_{t > 0} t^2 \Pr\{\xi_i > t\} < \infty. \tag{11.12}$$

Then, for all r such that $0 < r < 2$,

$$\sup_n n^{-r/2} E \max_{1 \le i \le n} \xi_i^r < \infty. \tag{11.13}$$

Next, Theorem 11.14 extends to multipliers other than Rademacher variables. For any real random variable Y let

$$\Lambda_{2,1}(Y) \ := \ \int_0^\infty [\Pr(|Y| > t)]^{1/2} dt.$$

Then $\Lambda_{2,1}(Y) < \infty$ implies $E(Y^2) < \infty$, and for any $\delta > 0$, $E|Y|^{2+\delta} < \infty$ implies $\Lambda_{2,1}(Y) < \infty$.

Lemma 11.16. *Let $(\Omega, \mathcal{S}, P)$ be a probability space and $\mathcal{F} \subset \mathcal{L}^1(P)$. Assume the usual hypotheses (11.2) and that furthermore, by another product space, there are i.i.d. symmetric real random variables $\xi_i \ := \ \xi(i)$ independent of the X_j and ε_j. Then for integers $0 \le m < n < \infty$ we have*

$$\begin{aligned} n^{-1/2}(E|\xi_1|) E^* \| \textstyle\sum_{i=1}^n \varepsilon_i f(X_i) \|_{\mathcal{F}} \ &\le \ n^{-1/2} E^* \| \textstyle\sum_{i=1}^n \xi_i f(X_i) \|_{\mathcal{F}} \\ &\le \ m n^{-1/2} (E^* \|f(X_1)\|_{\mathcal{F}}) E(\max_{i \le n} |\xi_i|) \\ &+ \ \Lambda_{2,1}(\xi_1) \max_{m < k \le n} k^{-1/2} E^* \| \textstyle\sum_{i=m+1}^k \varepsilon_i f(X_i) \|_{\mathcal{F}}. \end{aligned} \tag{11.14}$$

If the variables ξ_i have $E\xi_1 = 0$ but are not necessarily symmetric, then (11.14) holds with the following changes: $E|\xi_1|$ at the left is replaced by $E|\xi_1 - \xi_2|/2$, the first summand on the right is multiplied by 2 and the second by 3.

Next, here is a characterization of Donsker classes in terms of multipliers ξ_i.

Theorem 11.17. *Let $\mathcal{F}$ be a class such that hypotheses (11.2) hold and $\{f - Pf : f \in \mathcal{F}\}$ has a finite envelope function (11.11). Let ξ_j be i.i.d. centered real random variables, independent of $X_1, X_2, \cdots$, specifically, defined on a different factor of a product probability space, such that $E|\xi_1| > 0$ and $\Lambda_{2,1}(\xi_1) < \infty$. Then $\mathcal{F}$ is Donsker for P if and only if both $(\mathcal{F}, \rho_P)$ is totally bounded and*

$$\limsup_{n\to\infty} E^* \left\{ n^{-1/2} \| \sum_{i=1}^n \xi_i (f(X_i) - Pf) \|_{\delta,\mathcal{F}} \right\} = 0. \tag{11.15}$$

11.3 Other aspects of the bootstrap

B. Efron (1979) invented the bootstrap and by now there is a very large literature about it. This section will address some aspects of the application of the Giné-Zinn theorems. These do not cover the entire field by any means. For example, some statistics of interest, such as $\max(X_1, \cdots, X_n)$, are not averages $\frac{1}{n}(f(X_1) + \ldots f(X_n))$ as f ranges over a class $\mathcal{F}$.

Some bootstrap limit theorems are stated in probability, and others for almost sure convergence. To compare their usefulness, first note that almost sure convergence is not always preferable to convergence in probability:

Example. Let X_n be a sequence of real-valued random variables converging to some X_0 in probability but not almost surely. Then some subsequences X_{n_k} converge to X_0 almost surely. Suppose this occurs whenever $n_k \geq k^2$ for all k. Let $Y_n := X_{2^k}$ for $2^k \leq n < 2^{k+1}$ where $k = 0, 1, \cdots$. Then $Y_n \to X_0$ almost surely, but in a sense, $X_n \to X_0$ faster although it only converges in probability.

Another point is that almost sure convergence is applicable in statistics when inferences will be made from data sets with increasing values of n, in other words, in the part of statistics called *sequential analysis*. But suppose one has a fixed value of the sample size n, as has generally been the case with the bootstrap. Then the probability of an error of a given size, for a given n, which relates to convergence in probability, may be more relevant than the question of what *would* happen for values of $n \to \infty$, as in almost sure convergence.

The rest of this section will be devoted to confidence sets. A basic example of a confidence set is a confidence interval. As an example, suppose $X_1, \cdots, X_n$ are i.i.d. with distribution $N(\mu, \sigma^2)$ where σ^2 is known but μ is not. Then $\overline{X} := (X_1 + \cdots + X_n)/n$ has a distribution $N(\mu, \sigma^2/n)$. Thus

$$P(\overline{X} \leq \mu - 1.96\sigma/n^{1/2}) \doteq .025 \doteq P(\overline{X} \geq \mu + 1.96\sigma/n^{1/2}).$$

So we have 95 percent confidence that the unknown μ belongs to the interval $[\overline{X} - 1.96\sigma/n^{1/2}, \overline{X} + 1.96\sigma/n^{1/2}]$, which is then called a 95% confidence interval for μ.

Next, suppose $X_1, \cdots, X_n$ are i.i.d. in $\mathbb{R}^k$ with a normal (Gaussian) distribution $N(\mu, \sigma^2 I)$ where I is the identity matrix. Suppose $\alpha > 0$ and $M_\alpha = M_\alpha(k)$ is such that $N(0, I)\{x : |x| \geq M_\alpha\} = \alpha$. Then $n^{1/2}(\overline{X} - \mu)/\sigma$ has distribution $N(0, I)$ so $P(|\overline{X} - \mu| \geq M_\alpha \sigma/n^{1/2}) = \alpha$. Thus, the ball with center $\overline{X}$ and radius $M_\alpha \sigma/n^{1/2}$ is called a $100(1-\alpha)\%$ confidence set for the unknown μ.

When the distribution of the X_i is not necessarily normal, but has finite variance, then the distribution of $\overline{X}$ will be approximately normal by the central limit theorem for n large, so we get some approximate confidence sets.

Now let's extend these ideas to the bootstrap. Let $X_1, \cdots, X_n$ be i.i.d. from an otherwise unknown distribution P. Let P_n be the empirical measure formed from $X_1, \cdots, X_n$. Let $\mathcal{F}$ be a universal Donsker class (Section 10.1). Then we know from the Giné-Zinn theorem in the last section that ν_n^B and ν_n have asymptotically the same distribution on $\mathcal{F}$. By repeated resampling, given a small $\alpha > 0$ such as $\alpha = .05$, one can find $M = M(\alpha)$ such that approximately $\Pr(\|\nu_n^B\|_{\mathcal{F}} > M) \doteq \alpha$. Then

$$\{Q : \ \|Q - P_n\|_{\mathcal{F}} \ \leq \ M/n^{1/2}\}$$

is an approximate $100(1-\alpha)\%$ confidence set for P.

REFERENCES

Araujo, A. and Giné, E. (1980). *The Central Limit Theorem for Real and Banach Valued Random Variables.* Wiley, New York.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9**, 1196-1217.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6**, 899-929; Correction, *ibid.* **7** (1979), 909-911.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1-26.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the bootstrap.* Chapman and Hall, New York.

Fernique, X. (1975). Régularité des trajectoires des fonctions aléatoires gaussiennes. *Ecole d'été de probabilités de St.-Flour, 1974. Lecture Notes in Math.* **480**, 1-96. Springer, Berlin.

Fernique, X. (1985). Sur la convergence étroite des mesures gaussiennes. *Z. Wahrscheinlichkeitsth. verw. Gebiete* **68**, 331-336.

Gaenssler, P. (1986). Bootstrapping empirical measures indexed by Vapnik-Chervonenkis classes of sets. In *Probability Theory and Mathematical Statistics* (Vilnius, 1985), Yu. V. Prohorov, V. A. Statulevicius, V. V. Sazonov and B. Grigelionis, Eds., VNU Science Press, Utrecht, 467-481.

Giné, E. (1997). Lectures on some aspects of the bootstrap. In *Lectures on Probability Theory and Statistics*, Ecole d'été de probabilités de Saint-Flour (1996), ed. P. Bernard. *Lecture Notes in Math.* **1665**, 37-151. Springer, Berlin.

Giné, E. and Zinn, J. (1984). Some limit theorems for empirical processes. *Ann. Probab.* **12**, 929-989.

Giné, E. and Zinn, J. (1986). Lectures on the central limit theorem for empirical processes. In *Probability and Banach Spaces* (Zaragoza, 1985). *Lecture Notes in Math.* **1221**, 50-113. Springer, New York

Giné, E. and Zinn, J. (1990). Bootstrapping general empirical measures. *Ann. Probab.* **18**, 851-869.

Hall, Peter (1992). *The Bootstrap and Edgeworth Expansion.* Springer, New York.

Hoffman-Jørgensen, J. (1974). Sums of independent Banach space valued random variables. *Studia Math.* **52**, 159-186.

Jain, N. C. and Marcus, M. B. (1978). Continuity of subgaussian processes. In *Probability on Banach Spaces*, Ed. J. Kuelbs, *Advances in Probability and Related Topics* **4**, 81-196.

Dekker, New York.

Kahane, J.-P. (1968). *Some Random Series of Functions.* D. C. Heath, Lexington, Mass. 2d. ed. Cambridge Univ. Press, New York, 1985.

Ledoux, M. and Talagrand, M. (1988). Characterization of the law of the iterated logarithm in Banach spaces. *Ann. Probab.* **16**, 1242-1264.

Shao, Jun, and Tu, Dongsheng (1995). *The Jackknife and Bootstrap.* Springer, New York.

Strobl, F. (1994). *Zur Theorie empirischer Prozesse.* Dissertation, University of Munich, Faculty of Mathematics.

van der Vaart, A. W., and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer, Berlin.

Chapter 12

Classes of Sets or Functions Too Large for Central Limit Theorems

12.1 Universal lower bounds.

This chapter is primarily about asymptotic lower bounds for $\|P_n - P\|_{\mathcal{F}}$ on certain classes $\mathcal{F}$ of functions, as treated in Chapter 8, mainly classes of indicators of sets. Section 12.2 will give some upper bounds which indicate the sharpness of some of the lower bounds. Section 12.4 gives some relatively difficult lower bounds on classes such as the convex sets in $\mathbb{R}^3$ and lower layers in $\mathbb{R}^2$. In preparation for this, Section 12.3 treats Poissonization and random "stopping sets" analogous to stopping times. The present section gives lower bounds in some cases which hold not only with probability converging to 1, but for all possible P_n. Definitions are as in Sections 3.1 and 8.2, with $P := U(I^d) = \lambda^d =$ Lebesgue measure on I^d. Specifically, recall the classes $\mathcal{G}(\alpha, K, d) := \mathcal{G}_{\alpha,K,d}$ of functions on the unit cube $I^d \subset \mathbb{R}^d$ with derivatives through αth order bounded by K, and the related families $\mathcal{C}(\alpha, K, d)$ of sets, both defined early in Section 8.2.

Theorem 12.1. (Bakhvalov) *For $P = U(I^d)$, any $d = 1, 2, \ldots$ and $\alpha > 0$, there is a $\gamma = \gamma(d, \alpha) > 0$ such that for all $n = 1, 2, \cdots$, and all possible values of P_n, we have $\|P_n - P\|_{\mathcal{G}(\alpha,1,d)} \geq \gamma n^{-\alpha/d}$.*

Remarks. When $\alpha < d/2$, this shows that $\mathcal{G}(\alpha, K, d)$, $K > 0$, is not a Donsker class. For $\alpha > d/2$ the lower bound in Theorem 12.1 is not useful, since it is smaller than the average size of $\|P_n - P\|_{\mathcal{G}(\alpha,1,d)}$, which is at least of order $n^{-1/2}$: even for one function f not constant a.e. P, $E|(P_n - P)(f)| \geq cn^{-1/2}$ for some $c > 0$.

Theorem 12.1 gives information about accuracy of possible methods of numerical integration in several dimensions, or "cubature," using the values of a function $f \in \mathcal{G}(\alpha, K, d)$ at just n points chosen in advance (actually, one has the same lower bound even if one can use any partial derivatives of f at the n points). It was in this connection that Bakhvalov (1959) proved the theorem.

Theorem 12.2. *For $P = U(I^d)$, any $K > 0$ and $0 < \alpha < d - 1$ there is a $\delta = \delta(\alpha, K, d) > 0$ such that for all $n = 1, 2, \cdots$ and all possible values of P_n, $\|P_n - P\|_{\mathcal{C}(\alpha,K,d)} > \delta n^{-\alpha/(d-1+\alpha)}$.*

Remark. Since $\alpha/(d-1+\alpha) < 1/2$ for $\alpha < d-1$, the classes $\mathcal{C}(\alpha, K, d)$ are then not Donsker classes. For $\alpha > d-1$, $\mathcal{C}(\alpha, K, d)$ is a Donsker class by Theorem 8.3 and Corollary 7.7. For $\alpha = d-1$, it is not a Donsker class (Theorem 12.10 below).

Theorem 12.3. (W. Schmidt) *Let $d = 2, 3, \cdots$. For the collection $\mathcal{C}_d$ of closed convex subsets of a bounded non-empty open set U in $\mathbb{R}^d$ there is a constant $b := b(d, U) > 0$ such that for $P =$ Lebesgue measure normalized on U, and all P_n,*

$$\sup\{|(P_n - P)(C)|: \ C \in \mathcal{C}_d\} \quad \geq \quad bn^{-2/(d+1)}.$$

Thus for $d \geq 4$, $\mathcal{C}_d$ is not a Donsker class for P. If $d = 3$ it is not either, see Dudley (1982). $\mathcal{C}_2$ is a Donsker class for λ^2 on I^2 by Theorem 8.12 and Corollary 7.7.

12.2 An upper bound.

Here, using metric entropy with bracketing N_I as in Section 7.1, is an upper bound for $\|\nu_n\|_{\mathcal{C}} := \sup_{B \in \mathcal{C}} |\nu_n(B)|$ which applies in many cases where the hypotheses of Corollary 7.7 fail. Let $(X, \mathcal{A}, Q)$ be a probability space, $\nu_n := n^{1/2}(Q_n - Q)$, and recall N_I as defined before (7.4).

Theorem 12.4. *Let $\mathcal{C} \subset \mathcal{A}$, $1 \leq \zeta < \infty$, $\eta > 2/(\zeta + 1)$ and $\Theta := (\zeta - 1)/(2\zeta + 2)$. If for some $K < \infty$, $N_I(\varepsilon, \mathcal{C}, Q) \leq \exp(K\varepsilon^{-\zeta})$, $0 < \varepsilon \leq 1$, then*

$$\lim_{n \to \infty} \Pr^*\{\|\nu_n\|_{\mathcal{C}} > n^{\Theta}(\log n)^{\eta}\} \quad = \quad 0.$$

Remarks. The classes $\mathcal{C} = \mathcal{C}(\alpha, M, d)$ satisfy the hypothesis of Theorem 12.4 for $\zeta = (d-1)/\alpha \geq 1$, i.e. $\alpha \leq d-1$, by the last inequality in Theorem 8.3. Then $\Theta = \frac{1}{2} - \frac{\alpha}{d-1+\alpha}$. Thus Theorem 12.2 shows that the exponent Θ is sharp for $\zeta > 1$. Conversely, Theorem 12.4 shows that the exponent on n in Theorem 12.2 cannot be improved. In Theorem 12.4 we cannot take $\zeta < 1$, for then $\Theta < 0$, which is impossible even for a single set, $\mathcal{C} = \{C\}$, with $0 < P(C) < 1$.

12.3 Poissonization and random sets.

Section 12.4 will give some lower bounds $\|\nu_n\|_{\mathcal{C}} \geq f(n)$ with probability converging to 1 as $n \to \infty$ where f is a product of powers of logarithms or iterated logarithms. Such an f has the following property. A real-valued function f defined for large enough $x > 0$ is called *slowly varying* (in the sense of Karamata) iff for every $c > 0$, $f(cx)/f(x) \to 1$ as $x \to +\infty$.

Lemma 12.5. *If f is continuous and slowly varying then for every $\varepsilon > 0$ there is a $\delta = \delta(\varepsilon) > 0$ such that whenever $x > 1/\delta$ and $|1 - \frac{y}{x}| < \delta$ we have $\left|1 - \frac{f(y)}{f(x)}\right| < \varepsilon$.*

Recall the Poisson law P_c on $\mathbb{N}$ with parameter $c \geq 0$, so that $P_c(k) := e^{-c}c^k/k!$ for $k = 0, 1, \cdots$. Given a probability space $(X, \mathcal{A}, P)$, let U_c be a Poisson point process on $(X, \mathcal{A})$ with intensity measure cP. That is, for any disjoint $A_1, \cdots, A_m$ in $\mathcal{A}$, $U_c(A_j)$ are independent random variables, $j = 1, \cdots, m$, and for any $A \in \mathcal{A}$, $U_c(A)(\cdot)$ has law $P_{cP(A)}$.

Let $Y_c(A) := (U_c - cP)(A)$, $A \in \mathcal{A}$. Then Y_c has mean 0 on all A and still has independent values on disjoint sets.

Let $x(1), x(2), \cdots$ be coordinates for the product space $(X^\infty, \mathcal{A}^\infty, P^\infty)$. For $c > 0$ let $n(c)$ be a random variable with law P_c, independent of the $x(i)$. Then for $P_n := n^{-1}(\delta_{x(1)} + \cdots + \delta_{x(n)})$, $n \geq 1$, $P_0 := 0$ we have:

Lemma 12.6. *The process $Z_c := n(c)P_{n(c)}$ is a Poisson process with intensity measure cP.*

From here on, the version $U_c \equiv Z_c$ will be used. Thus for each ω, $U_c(\cdot)(\omega)$ is a countably additive integer-valued measure of total mass $U_c(X)(\omega) = n(c)(\omega)$. Then

$$\begin{aligned} Y_c &= n(c)P_{n(c)} - cP = n(c)(P_{n(c)} - P) + (n(c) - c)P, \\ Y_c/c^{1/2} &= (n(c)/c)^{1/2}\nu_{n(c)} + (n(c) - c)c^{-1/2}P. \end{aligned} \tag{12.1}$$

The following says that the empirical process ν_n is asymptotically "as large" as a corresponding Poisson process.

Lemma 12.7. *Let $(X, \mathcal{A}, P)$ be a probability space and $\mathcal{C} \subset \mathcal{A}$. Assume that for each n and constant t, $\sup_{A\in\mathcal{C}} |(P_n - tP)(A)|$ is measurable. Let f be a continuous, slowly varying function such that as $x \to +\infty$, $f(x) \to +\infty$. For $b > 0$ let*

$$g(b) := \liminf_{x\to+\infty} \Pr\{\sup_{A\in\mathcal{C}} |Y_x(A)| \geq bf(x)x^{1/2}\}.$$

Then for any $a < b$,

$$\liminf_{n\to\infty} \Pr\{\sup_{A\in\mathcal{C}} |\nu_n(A)| \geq af(n)\} \geq g(b).$$

Next, the Poisson process's independence property on disjoint sets will be extended to suitable random sets. Let $(X, \mathcal{A})$ be a measurable space, and $(\Omega, \mathcal{B}, \Pr)$ a probability space. A collection $\{\mathcal{B}_A : A \in \mathcal{A}\}$ of sub-σ-algebras of $\mathcal{B}$ will be called a *filtration* if $\mathcal{B}_A \subset \mathcal{B}_B$ whenever $A \subset B$ in $\mathcal{A}$. A stochastic process Y indexed by $\mathcal{A}$, $(A, \omega) \mapsto Y(A)(\omega)$, will be called *adapted* to $\{\mathcal{B}_A : A \in \mathcal{A}\}$ if for every $A \in \mathcal{A}$, $Y(A)(\cdot)$ is $\mathcal{B}_A$ measurable. Then the process and filtration will be written $\{Y(A), \mathcal{B}_A\}_{A\in\mathcal{A}}$. A stochastic process $Y : \langle A, \omega\rangle \to Y(A)(\omega)$, $A \in \mathcal{A}$, $\omega \in \Omega$, will be said to have *independent pieces* iff for any disjoint $A_1, \cdots, A_m \in \mathcal{A}$, $Y(A_j)$ are independent, $j = 1, \cdots, m$, and $Y(A_1 \cup A_2) = Y(A_1) + Y(A_2)$ almost surely. Clearly each Y_c has independent pieces. If in addition the process is adapted to a filtration $\{\mathcal{B}_A : A \in \mathcal{A}\}$, the process $\{Y(A), \mathcal{B}_A\}_{A\in\mathcal{A}}$ will be said to have *independent pieces* iff for any disjoint sets $A_1, \cdots, A_n$ in $\mathcal{A}$, the random variables $Y(A_2), \cdots, Y(A_n)$ and any random variable measurable for the σ-algebra $\mathcal{B}_{A_1}$ are jointly independent.

For example, for any $C \in \mathcal{A}$ let $\mathcal{B}_C$ be the smallest σ-algebra for which every $Y(A)(\cdot)$ is measurable for $A \subset C$, $A \in \mathcal{A}$. This is clearly a filtration, and the smallest filtration to which Y is adapted.

A function G from Ω into $\mathcal{A}$ will be called a *stopping set* for a filtration $\{\mathcal{B}_A : A \in \mathcal{A}\}$ iff for all $C \in \mathcal{A}$, $\{\omega : G(\omega) \subset C\} \in \mathcal{B}_C$. Given a stopping set $G(\cdot)$, let $\mathcal{B}_G$ be the σ-algebra of all sets $B \in \mathcal{B}$ such that for every $C \in \mathcal{A}$, $B \cap \{G \subset C\} \in \mathcal{B}_C$. (Note that if G is not a stopping set, then $\Omega \notin \mathcal{B}_G$, so $\mathcal{B}_G$ would not be a σ-algebra.) If $G(\omega) \equiv H \in \mathcal{A}$ then it is easy to check that G is a stopping set and $\mathcal{B}_G = \mathcal{B}_H$.

Lemma 12.8. *Suppose $\{Y(A), \mathcal{B}_A\}_{A\in\mathcal{A}}$ has independent pieces and for all $\omega \in \Omega$, $G(\omega) \in \mathcal{A}$, $A(\omega) \in \mathcal{A}$ and $E(\omega) \in \mathcal{A}$.*

Assume that:

(i) *$G(\cdot)$ is a stopping set;*

(ii) *For all ω, $G(\omega)$ is disjoint from $A(\omega)$ and from $E(\omega)$;*

(iii) *$G(\omega)$, $A(\omega)$ and $E(\omega)$ each have just countably many possible values $G(j) := G_j \in \mathcal{A}$, $C(i) := C_i \in \mathcal{A}$ and $D(j) := D_j \in \mathcal{A}$ respectively;*
(iv) *For all i, j, $\{A(\cdot) = C_i\} \in \mathcal{B}_G$ and $\{E(\cdot) = D_j\} \in \mathcal{B}_G$.*

Then the conditional probability law (joint distribution) of $Y(A)$ and $Y(E)$ given $\mathcal{B}_G$ satisfies

$$\mathcal{L}\{(Y(A), Y(E))|\mathcal{B}_G\} \;=\; \sum_{i,j} 1_{\{A(\cdot)=C(i), E(\cdot)=D(j)\}} \mathcal{L}(Y(C_i), Y(D_j))$$

where $\mathcal{L}(Y(C_i), Y(D_j))$ is the unconditional joint distribution of $Y(C_i)$ and $Y(D_j)$. If this unconditional distribution is the same for all i, j, then $(Y(A), Y(E))$ is independent of $\mathcal{B}_G$.

Here is another fact about stopping sets, which corresponds to a known fact about nonnegative real-valued stopping times or Markov times (e.g. RAP, Lemma 12.2.5):

Lemma 12.9. *If G and H are stopping sets and $G \subset H$ then $\mathcal{B}_G \subset \mathcal{B}_H$.*

Proof. For any measurable set D and $A \in \mathcal{B}_G$, we have

$$A \cap \{H \subset D\} \;=\; (A \cap \{G \subset D\}) \cap \{H \subset D\} \;\in\; \mathcal{B}_D.$$

□ .

12.4 Lower bounds in borderline cases.

Recall the classes $\mathcal{C}(\alpha, K, d)$ of subgraphs of functions with bounded derivatives through order α in $\mathbb{R}^d$, defined in Section 8.2. We had lower bounds for $P_n - P$ on $\mathcal{C}(\alpha, K, d)$ in Theorem 12.2 which imply that for $\alpha < d-1$, $\|\nu_n\|_{\mathcal{C}(\alpha,K,d)} \to \infty$ surely as $n \to \infty$. For $\alpha > d-1$, $\mathcal{C}(\alpha, K, d)$ is a Donsker class by Theorem 8.3 and Corollary 7.7, so $\|\nu_n\|_{\mathcal{C}(\alpha,K,d)}$ is bounded in probability. Thus $\alpha = d-1$ is a borderline case. Other such cases are given by the class $\mathcal{LL}_2$ of lower layers in $\mathbb{R}^2$ (Section 8.3) and the class $\mathcal{C}_3$ of convex sets in $\mathbb{R}^3$ (Section 8.4), for $\lambda^d =$ Lebesgue measure on the unit cube I^d, where $I := [0, 1]$.

Any lower layer A has a closure $\overline{A}$ which is also a lower layer, with $\lambda^d(\overline{A} \setminus A) = 0$, where in the present case $d = 2$. It is easily seen that suprema of our processes over all lower layers are equal to suprema over closed lower layers, so it will be enough to consider closed lower layers. Let $\overline{\mathcal{LL}}_2$ be the class of all closed lower layers in $\mathbb{R}^2$.

Let $P = \lambda^d$ and $c > 0$. Recall the centered Poisson process Y_c from Section 12.3. Let $N_c := U_c - V_c$ where U_c and V_c are independent Poisson processes, each with intensity measure cP. Equivalently, we could take U_c and V_c to be centered. The following lower bound holds for all the above borderline cases:

Theorem 12.10. *For any $K > 0$ and $\delta > 0$ there is a $\gamma = \gamma(d, K, \delta) > 0$ such that*

$$\lim_{x \to +\infty} \Pr\left\{ \|Y_x\|_{\mathcal{C}} > \gamma (x \log x)^{1/2} (\log\log x)^{-\delta - 1/2} \right\} \;=\; 1$$

and

$$\lim_{n \to \infty} \Pr\left\{ \|\nu_n\|_{\mathcal{C}} > \gamma (\log n)^{1/2} (\log\log n)^{-\delta - 1/2} \right\} \;=\; 1$$

where $\mathcal{C} = \mathcal{C}(d-1, K, d)$, $d \geq 2$, or $\mathcal{C} = \overline{\mathcal{LL}}_2$, or $\mathcal{C} = \mathcal{C}_3$.

A larger lower bound with probability close to 1, of order $(\log n)^{3/4}$ has been found in the lower layer case ($\mathcal{C} = \mathcal{LL}_2, d = 2$). Shor (1986) first showed that $E\|Y_x\|_{\mathcal{C}} > \gamma x^{1/2}(\log x)^{3/4}$ for some $\gamma > 0$ and x large enough. Shor's lower bound also applies to $\mathcal{C}(1, K, 2)$ by a 45° rotation as in Section 8.3. For an upper bound with a 3/4 power of the log also for convex subsets of a fixed bounded set in $\mathbb{R}^3$ see Talagrand (1994, Theorem 1.6).

To see that the supremum of N_c, Y_c or an empirical process ν_n over $\mathcal{LL}_2$ is measurable, note first for P_n that for each $F \subset \{1, \ldots, n\}$ and each ω, there is a smallest, closed lower layer $L_F(\omega)$ containing the x_j for $j \in F$, with $L_F(\omega) := \emptyset$ for $F = \emptyset$. For any $c > 0$, $\omega \mapsto (P_n - cP)\,(L_F(\omega))(\omega)$ is measurable. The supremum of $P_n - cP$ over $\mathcal{LL}_2$, as the maximum of these 2^n measurable functions, is measurable. Letting $n = n(c)$ as in Lemma 12.6 and (12.1) then shows $\sup\{Y_c(A) \colon A \in \mathcal{LL}_2\}$ is measurable. Likewise, there is a largest, open lower layer not containing x_j for any $j \in F$, so $\sup\{|Y_c(A)| : A \in \mathcal{LL}_2\}$ and $\sup\{|\nu_n(A)| : A \in \mathcal{LL}_2\}$ are measurable.

For N_c, taking non-centered Poisson processes U_c and V_c, their numbers of points $m(\omega)$ and $n(\omega)$ are measurable, as are the m-tuple and n-tuple of points occurring in each. For each $i = 0, 1, \cdots, m$ and $k = 0, 1, \cdots, n$, it is a measurable event that there exists a lower layer containing exactly i of the m points and k of the n, and so the supremum of N_c over all lower layers, as a measurable function of the indicators of these finitely many events, is measurable.

Theorem 12.11. *For every $\varepsilon > 0$ there is a $\delta > 0$ such that for the uniform distribution P on the unit square I^2, and n large enough,*

$$Pr\Big(\sup\{|\nu_n(A)| : A \in \mathcal{LL}_2\} \ \geq \ \delta(\log n)^{3/4}\Big) \ \geq \ 1 - \varepsilon,$$

and the same holds for $\mathcal{C}(1, 2, 2)$ in place of $\mathcal{LL}_2$. Also, ν_n can be replaced by $N_c/c^{1/2}$ or $Y_c/c^{1/2}$ if $\log n$ is replaced by $\log c$, for c large enough.

Remark. The order $(\log n)^{3/4}$ of the lower bound is best possible, as there is an upper bound in expectation of the same order, see Rhee and Talagrand (1988), Leighton and Shor (1989), and Coffman and Shor (1991).

REFERENCES

Bakhvalov, N. S. (1959). On approximate calculation of multiple integrals. *Vestnik Moskov. Univ. Ser. Mat. Mekh. Astron. Fiz. Khim.* **1959**, No. 4, 3–18.

Coffman, E. G., and Lueker, G. S. (1991). *Probabilistic Analysis of Packing and Partitioning Algorithms.* Wiley, New York.

Coffman, E. G., and Shor, P. W. (1991). A simple proof of the $O(\sqrt{n}\log^{3/4} n)$ upright matching bound. *SIAM J. Discrete Math.* **4**, 48-57.

Dudley, R. M. (1982). Empirical and Poisson processes on classes of sets or functions too large for central limit theorems. *Z. Wahrscheinlichkeitsth. verw. Gebiete* **61**, 355–368.

Dudley, R. M. (1984). A course on empirical processes. *Ecole d'été de probabilités de St.-Flour. Lecture Notes in Math.* **1097**, 1–142. Springer, New York.

Evstigneev, I.V. (1977). "Markov times" for random fields. *Theor. Probab. Appl.* **22**, 563–569; *Teor. Veroiatnost. i Primenen.* **22**, 575–581.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications* Vol. II, 2nd ed., Wiley, New York.

Kac, M. (1949). On deviations between theoretical and empirical distributions. *Proc. Nat. Acad. Sci. USA* **35**, 252–257.

Leighton, T., and Shor, P. (1989). Tight bounds for minimax grid matching, with applications to the average case analysis of algorithms. *Combinatorica* **9**, 161-187.

Pyke, R. (1968). The weak convergence of the empirical process with random sample size. *Proc. Cambridge Philos. Soc.* **64**, 155–160.

Rhee, W. T., and Talagrand, M. (1988). Exact bounds for the stochastic upward matching problem. *Trans. Amer. Math. Soc.* **307**, 109-125.

Schmidt, W. M. (1975). Irregularities of distribution IX. *Acta Arith.* **27**, 385–396.

Shor, P. W. (1986). The average-case analysis of some on-line algorithms for bin packing. *Combinatorica* **6**, 179–200.

Talagrand, M. (1994). Matching theorems and empirical discrepancy computations using majorizing measures. *J. Amer. Math. Soc.* **7**, 455-537.

APPENDIX A. DIFFERENTIATING UNDER AN INTEGRAL SIGN

There are various sufficient conditions for the equation

$$\frac{d}{dt}\int f(x,t)d\mu(x) \;=\; \int \frac{\partial f(x,t)}{\partial t}d\mu(x).$$

Here $x \in X$, where $(X, \mathcal{S}, \mu)$ is a measure space, and t is real-valued. The derivatives with respect to t will be taken at some point $t = t_0$. The function f will be defined for $x \in X$ and t in an interval J containing t_0 in its interior. We assume that $\partial f/\partial t$ exists at $t = t_0$ for μ-almost all x.

For interchanging two integrals, there is a standard theorem, the Tonelli-Fubini theorem (e.g. RAP, Theorem 4.4.5). But for interchanging a derivative and an integral there is apparently not such a handy single theorem.

One sufficient condition is that the difference-quotients $[f(x, t_0 + h) - f(x, t_0)]/h$ for h small enough are dominated in absolute value by some μ-integrable function. Details of other sufficient conditions are omitted.

REFERENCES

Brown, Lawrence D. (1986). *Fundamentals of Statistical Exponential Families.* Inst. Math. Statist. Lecture Notes-Monograph Ser. **9**.

Hobson, E. W. (1926). *The Theory of Functions of a Real Variable and the Theory of Fourier's Series*, vol. 2, 2d ed. Repr. Dover, New York, 1957.

Il'in, V. A., and Pozniak, E. G. (1982). *Fundamentals of Mathematical Analysis.* Transl. from the 1980 4th Russian edition by V. Shokurov. Mir, Moscow.

Kartashev, A. P., and Rozhdestvenskiĭ, B. L. (1984). *Matematicheskiĭ Analiz* (Mathematical Analysis; in Russian). Nauka, Moscow. French transl.: Kartachev, A., and Rojdestvenski, B. *Analyse mathématique*, transl. by Djilali Embarek. Mir, Moscow, 1988.

Lang, Serge (1993). *Real and Functional Analysis*, 3d ed. of *Real Analysis*, Springer, New York.

APPENDIX B. MULTINOMIAL DISTRIBUTIONS

[classical properties of multinomial probabilities]

APPENDIX C. MEASURES ON NONSEPARABLE METRIC SPACES

Let (S, d) be a metric space. Under fairly general conditions, to be given, any probability measure on the Borel sets will be concentrated in a separable subspace.

The problem reduces to one about discrete spaces. An *open cover* of S will be a family $\{U_\alpha\}_{\alpha\in I}$ of open subsets U_α of S, where I is any set, here called an index set, such that $S = \bigcup_{\alpha\in I} U_\alpha$. An open cover $\{V_\beta\}_{\beta\in J}$ of S, with some index set J, will be called a *refinement* of $\{U_\alpha\}_{\alpha\in I}$ iff for all $\beta \in J$ there exists an $\alpha \in I$ with $V_\beta \subset U_\alpha$. An open cover $\{V_\alpha\}_{\alpha\in I}$ will be called *σ-discrete* if I is the union of a sequence of sets I_n such that for each n and $\alpha \neq \beta$ in I_n, U_α and U_β are disjoint. Recall that the ball $B(x, r)$ is defined as $\{y \in S : \; d(x, y) < r\}$. For two sets A, B we have $d(A, B) := \inf\{d(x, y) : \; x \in A, \; y \in B\}$, and $d(y, B) := d(\{y\}, B)$ for a point y.

C.1 Theorem. *For any metric space (S, d), any open cover $\{U_\alpha\}_{\alpha \in I}$ of S has an open σ-discrete refinement.*

Cardinal numbers are defined in RAP, in the last part of Appendix A (as smallest ordinals with given cardinality). A cardinal number ζ is said to be *measurable* if for a set S of cardinality ζ, there exists a probability measure P defined on all subsets of S which is nonatomic, in other words $P(\{x\}) = 0$ for all $x \in S$. If there is no such P, ζ is said to be *of measure* 0.

The continuum hypothesis implies that the cardinality c of the continuum (that is, of $[0, 1]$) is of measure 0 (RAP, Appendix C).

The *separability character* of a metric space is the smallest cardinality of a dense subset. We have:

C.2 Theorem. *Let (S, d) be a metric space. Let P be a probability measure on the σ-algebra of Borel sets, generated by the open sets. Then either there is a separable subspace T with $P(T) = 1$, or the separability character ζ of S is measurable.*

It is consistent with the usual axioms of set theory (including the axiom of choice) that there are no measurable cardinals, in other words all cardinals are of measure 0, e.g. Drake (1974, pp. 67-68, 177-178). It is apparently unknown whether existence of measurable cardinals is consistent (Drake, 1974, pp. 185-186). So, for practical purposes, a probability measure defined on the Borel sets of a metric space is always concentrated in some separable subspace.

Here is another fact giving separability:

C.3 Theorem. *Let f be a Borel measurable function from a separable metric space S into a metric space T. Then, assuming the continuum hypothesis, f has separable range.*

REFERENCES

Drake, F. R. (1974). *Set Theory: An Introduction to Large Cardinals.* North-Holland, Amsterdam.

Kelley, John L. (1955). *General Topology.* Van Nostrand, Princeton. Repr. Springer, New York, 1975.

Marczewski, E., and Sikorski, R. (1948). Measures in non-separable metric spaces. *Colloq. Math.* **1**, 133-139.

Stone, Arthur H. (1948). Paracompactness and product spaces. *Bull. Amer. Math. Soc.* **54**, 631-632.

APPENDIX D. AN EXTENSION OF LUSIN'S THEOREM

Lusin's theorem says that for any measurable real-valued function f, on $[0, 1]$ with Lebesgue measure λ for example, and $\varepsilon > 0$, there is a set A with $\lambda(A) < \varepsilon$ such that restricted to the complement of A, f is continuous. Here $[0, 1]$ can be replaced by any normal topological space and λ by any finite measure μ which is *closed regular*, meaning that for each Borel measurable set B, $\mu(B) = \sup\{\mu(F) : F \text{ closed}, F \subset B\}$ (RAP, Theorem 7.5.2). Recall that any finite Borel measure on a metric space is closed regular (RAP, Theorem 7.1.3).

Proofs of Lusin's theorem are often based on Egorov's theorem (RAP, Theorem 7.5.1), which says that if measurable functions f_n from a finite measure space to a metric space converge pointwise, then for any $\varepsilon > 0$ there is a set of measure less than ε outside of which the f_n converge uniformly.

Here, the aim will be to extend Lusin's theorem to functions having values in any separable metric space. The proof of Lusin's theorem in RAP, however, also relied on the Tietze-Urysohn extension theorem, which says that a continuous real-valued function on a closed subset of a normal space can be extended to be continuous on the whole space. Such an extension may not exist for some range spaces: for example, the identity from $\{0,1\}$ onto itself doesn't extend to a continuous function from $[0,1]$ onto $\{0,1\}$, in fact there is no such function since $[0,1]$ is connected.

It turns out, however, that the Tietze-Urysohn extension and Egorov's theorem are both unnecessary in proving Lusin's theorem.

D.1 Theorem. *Let $(X,\mathcal{T})$ be a topological space and μ a finite, closed regular measure defined on the Borel sets of X. Let f be a Borel measurable function from x into S where (S,d) is a separable metric space. Then for any $\varepsilon > 0$ there is a closed set F with $\mu(X \setminus F) < \varepsilon$ such that f restricted to F is continuous.*

Proof. Let $\{s_n\}_{n\geq 1}$ be a countable dense set in S. For $m = 1,2,...,$ and any $x \in X$, let $f_m(x) = s_n$ for the least n such that $d(f(x),s_n) < 1/m$. Then f_m is measurable and defined on all of X. For each m, let $n(m)$ be large enough so that

$$\mu\{x:\ d(f(x),s_n) \geq 1/m \text{ for all } n \leq n(m)\} \ \leq\ 1/2^m.$$

For $n = 1,...,n(m)$, take a closed set $F_{mn} \subset f_m^{-1}\{s_n\}$ with

$$\mu(f_m^{-1}\{s_n\} \setminus F_{mn}) \ <\ \frac{1}{2^m n(m)}$$

by closed regularity. For each fixed m, the sets F_{mn} are disjoint for different values of n. Let $F_m := \bigcup_{n=1}^{n(m)} F_{mn}$. Then f_m is continuous on F_m. By choice of $n(m)$ and F_{mn}, $\mu(F_m) > 1 - 2/2^m$.

Since $d(f_m,f) < 1/m$ everywhere, clearly $f_m \to f$ uniformly (so Egorov's theorem is not needed). For $r = 1,2,...,$ let $H_r := \bigcap_{m=r}^{\infty} F_m$. Then H_r is closed and $\mu(H_r) \geq 1 - 4/2^r$. Take r large enough so that $4/2^r < \varepsilon$. Then f restricted to H_r is continuous as the uniform limit of continuous functions f_m on $H_r \subset F_m,\ m \geq r$, so we can let $F = H_r$ to finish the proof. $\square$

REFERENCES

Schaerf, H. M. (1947). On the continuity of measurable functions in neighborhood spaces. *Portugal. Math.* **6**, 33-44.

Schaerf, H. M. (1948). On the continuity of measurable functions in neighborhood spaces II. *Portugal. Math.* **7**, 91-92.

Zakon, Elias (1965). On "essentially metrizable" spaces and on measurable functions with values in such spaces. *Trans. Amer. Math. Soc.* **119**, 443-453.

APPENDIX E. BOCHNER AND PETTIS INTEGRALS

Let $(X,\mathcal{A},\mu)$ be a measure space and $(S,\|\cdot\|)$ a separable Banach space. A function f from X into S will be called *simple* or *μ-simple* if it is of the form $f = \sum_{i=1}^k 1_{A_i} y_i$ for some $y_i \in S,\ k < \infty$ and measurable A_i with $\mu(A_i) < \infty$. For a simple function, the Bochner integral is defined by

$$\int f\ d\mu\ :=\ \textstyle\sum_{i=1}^k \mu(A_i) y_i \in S.$$

E.1 Theorem. *The Bochner integral is well-defined for μ-simple functions, the μ-simple functions form a real vector space, and for any μ-simple functions f, g and real constant c, $\int cf + g\, d\mu = c \int f\ d\mu + \int g\ d\mu$.*

Proof. These facts are proved just as they are for real-valued functions (RAP, Proposition 4.1.4). □

For any measurable function g from X into S, where measurability is defined for the Borel σ-algebra generated by the open sets of S, $x \mapsto \|g(x)\|$ is a measurable, nonnegative real-valued function on S.

E.2 Lemma. *For any two μ-simple functions f and g from X into S, $\| \int f d\mu \| \leq \int \|f\| d\mu$ and $\| \int f\ d\mu - \int g\ d\mu \| \leq \int \|f - g\|\ d\mu$.*

Let $\mathcal{L}^1(X, \mathcal{A}, \mu, S) := \mathcal{L}^1(X, \mu, S)$ be the space of all measurable functions f from X into S such that $\int \|f\|\ d\mu < \infty$. By the triangle inequality, it's easily seen that $\mathcal{L}^1(X, \mathcal{A}, \mu, S)$ is a vector space. Also, all μ-simple functions belong to $\mathcal{L}^1(X, \mathcal{A}, \mu, S)$. Define $\|\cdot\|_1$ on $\mathcal{L}^1(X, \mathcal{A}, \mu, S)$ by $\|f\|_1 := \int \|f\|\ d\mu$. It is easily seen that $\|\cdot\|_1$ is a seminorm on $\mathcal{L}^1(X, \mathcal{A}, \mu, S)$. On the vector space of μ-simple functions, the Bochner integral is linear (by Theorem E.1) and continuous (also Lipschitz) for $\|\cdot\|_1$ by Lemma E.2.

E.3 Theorem. *For any separable Banach space $(S, \|\cdot\|)$ and any measure space $(X, \mathcal{A}, \mu)$, the μ-simple functions are dense in $\mathcal{L}^1(X, \mathcal{A}, \mu, S)$ for $\|\cdot\|_1$ and the Bochner integral extends uniquely to a linear, real-valued function on $\mathcal{L}^1(X, \mathcal{A}, \mu, S)$, continuous for $\|\cdot\|_1$.*

So the Bochner integral is well-defined for a function f if and only if f is in $\mathcal{L}^1(X, \mathcal{A}, \mu, S)$. A function from X into S will be called *Bochner integrable* if and only if it belongs to $\mathcal{L}^1(X, \mathcal{A}, \mu, S)$. The extension of the Bochner integral to $\mathcal{L}^1(X, \mathcal{A}, \mu, S)$ will also be written as $\int \cdot\, d\mu$. Thus Theorem E.3 implies that

$$\int cf + g\ d\mu \ = \ c \int f\ d\mu + \int g\ d\mu$$

for any Bochner integrable functions f, g and real constant c. Also, by taking limits in Lemma E.2 it follows that

$$\| \int f\ d\mu \| \ \leq \ \int \|f\|\ d\mu$$

for any Bochner integrable function f.

Although monotone convergence is not defined in general Banach spaces, a form of dominated convergence holds:

E.4 Theorem. *Let $(X, \mathcal{A}, \mu)$ be a measure space. Let f_n be measurable functions from X into a Banach space S such that for all n, $\|f_n\| \leq g$ where g is an integrable real-valued function. Suppose f_n converge almost everywhere to a function f. Then f is Bochner integrable and $\| \int f_n - f\ d\mu \| \ \leq \ \int \|f_n - f\|\ d\mu \ \to \ 0$ as $n \to \infty$.*

A Bochner integral $\int g\ d\mu = \int f\ d\mu$ can be defined when g is only defined almost everywhere for μ, f is Bochner integrable and $g = f$ where g is defined, just as for real-valued functions (RAP, Section 4.3). It's easy to check that when $S = \mathbb{R}$, the Bochner integral equals the usual Lebesgue integral.

A Tonelli-Fubini theorem holds for the Bochner integral:

E.5 Theorem. *Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be σ-finite measure spaces. Let f be a measurable function from $X \times Y$ into a Banach space S such that $\int \|f(x,y)\| d\mu(x) d\nu(y) < \infty$. Then for μ-almost all x, $f(x, \cdot)$ is Bochner integrable from Y into S; for ν-almost all y, $f(\cdot, y)$ is Bochner integrable from X into S, and*

$$\int f d(\mu \times \nu) \;=\; \int \int f(x,y) d\mu(x) d\nu(y) \;=\; \int \int f(x,y) d\nu(y) d\mu(x).$$

Now let $(S, \mathcal{T})$ be any topological vector space, in other words S is a real vector space, $\mathcal{T}$ is a topology on S, and the operation $(c, f, g) \mapsto cf + g$ is jointly continuous from $\mathbb{R} \times S \times S$ into S. Then the *dual space* S' is the set of all continuous linear functions from S into $\mathbb{R}$. Let $(X, \mathcal{A}, \mu)$ be a measure space. Then a function f from X into S is called *Pettis integrable* with *Pettis integral* $y \in S$ if and only if for every $t \in S'$, $\int t(f)\, d\mu$ is defined and finite and equals $t(y)$.

The Pettis integral is also due to Gelfand and Dunford and might be called the Gelfand-Dunford-Pettis integral. The Pettis integral may lack interest unless S' separates points of S, as is true for normed linear spaces by the Hahn-Banach theorem (RAP, Corollary 6.1.5).

E.6 Theorem. *For any measure space $(X, \mathcal{A}, \mu)$ and separable Banach space $(S, \|\cdot\|)$, each Bochner integrable function f from X into S is also Pettis integrable, and the values of the integrals are the same.*

Proof. The equation $t(\int f\, d\mu) = \int t(f)\, d\mu$ is easily seen to hold for simple functions and then, by Theorem E.3, for Bochner integrable functions. □

Example. A function can be Pettis integrable without being Bochner integrable. Let H be a separable, infinite-dimensional Hilbert space with orthonormal basis $\{e_n\}$. Let $\mu(f = ne_n) := \mu(f = -ne_n) := p_n := n^{-7/4}$, and $f = 0$ otherwise. Then $\int \|f\|\, d\mu = 2\sum_n n^{-3/4} = +\infty$, so f is not Bochner integrable. On the other hand for any x_n with $\sum_n x_n^2 < \infty$ we have $|\sum 2nx_n p_n| \le (\sum x_n^2)^{1/2} 2(\sum n^{-3/2})^{1/2} < \infty$ and by symmetry the Pettis integral of f is 0.

Example. The Tonelli-Fubini theorem, which holds for the Bochner integral (Theorem E.5), can fail for the Pettis integral. Let H be an infinite-dimensional Hilbert space and let $\xi_{i,j}$ be orthonormal for all positive integers i and j. Let $U(x,y) := 2^i \xi_{i,j}$ for $(j-1)/2^i \le x < j/2^i$ and $2^{-i} \le y < 2^{1-i}$, $j = 1, ..., 2^i$, and 0 elsewhere. Then it can be checked that U is Pettis integrable on $[0,1] \times [0,1]$ for Lebesgue measure but is not integrable with respect to y for fixed x.

REFERENCES

An asterisk (*) indicates a work I have seen discussed in secondary sources but have not seen in the original.

Birkhoff, Garrett (1935). Integration of functions with values in a Banach space. *Trans. Amer. Math. Soc.* **38**, 357-378.

Bochner, Salomon (1933). Integration von Funktionen, deren Werte die Elemente eines Vektorraumes sind. *Fund. Math.* **20**, 262-276.

Cohn, Donald L. (1980). *Measure Theory.* Birkhäuser, Boston.

Dunford, Nelson (1936). Integration and linear operations. *Trans. Amer. Math. Soc.* **40**, 474-494.

Dunford, Nelson, and Schwartz, J. T. (1958). *Linear Operators. Part I: General Theory.* Interscience, New York. Repr. 1988.

(*) Gelfand, Izrail Moiseevich (1936). Sur un lemme de la théorie des espaces linéaires. *Communications de l'Institut des Sciences Math. et Mécaniques de l'Université de Kharkoff et de la Societé Math. de Kharkov (= Zapiski Khark. Mat. Obshchestva)* (Ser. 4) **13**, 35-40.

Gelfand, I. M. (1938). Abstrakte Funktionen und lineare Operatoren. *Mat. Sbornik* (N. S.) **4**, 235-286.

Graves, L. M. (1927). Riemann integration and Taylor's theorem in general analysis. *Trans. Amer. Math. Soc.* **29**, 163-177.

Pettis, Billy Joe (1938). On integration in vector spaces. *Trans. Amer. Math. Soc.* **44**, 277-304.

Price, G. B. (1940). The theory of integration. *Trans. Amer. Math. Soc.* **47**, 1-50.

APPENDIX F
NON-EXISTENCE OF TYPES OF LINEAR FORMS ON SOME SPACES

Recall that a real vector space V with a topology $\mathcal{T}$ is called a *topological vector space* if addition is jointly continuous from $V \times V$ to V for $\mathcal{T}$, and scalar multiplication $(c, v) \mapsto cv$ is jointly continuous from $\mathbb{R} \times V$ into V for $\mathcal{T}$ on V and the usual topology on $\mathbb{R}$. If d is a metric on V, then (V, d) is called a *metric linear space* iff it is a topological vector space for the topology of d. Recall that for a given σ-algebra A of measurable sets, in our case the Borel σ-algebra, a *universally measurable set* is one measurable for the completion of every probability measure on A (RAP, Section 11.5). The universally measurable sets form a σ-algebra, and a function f is called universally measurable if for every Borel set B in its range, $f^{-1}(B)$ is universally measurable. Thus any Borel measurable function is universally measurable.

F.1 Theorem. *Let (E, d) be a complete metric real linear space. Let u be a universally measurable linear form: $E \mapsto \mathbb{R}$. Then u is continuous.*

Note that Theorem F.1 fails if the completeness assumption is omitted: let H be an infinite-dimensional Hilbert space and let h_n be an infinite orthonormal sequence in H. Let E be the set of all finite linear combinations of the h_n. Then there exists a linear form u on E with $u(h_n) = n$ for all n, and u is Borel measurable on E but not continuous.

F.2 Proposition. *For $0 < p < 1$ there are no non-zero continuous real linear forms on $L^p[0, 1]$ for the metric $\rho_p(f, g) := \int_0^1 |f - g|^p(t)dt$ and so, no Borel or universally measurable such forms.*

REFERENCES

Schaefer, H. H. (1966). *Topological Vector Spaces.* MacMillan, New York. 3d printing, corrected, Springer, New York, 1971.

Schwartz, L. (1966). Sur le théorème du graphe fermé. *C. R. Acad. Sci. Paris Sér. A* **263**, A602-605.

APPENDIX G
SEPARATION OF ANALYTIC SETS; BOREL INJECTIONS

Recall that a *Polish space* is a topological space S metrizable by a metric for which S is complete and separable. Also, in any topological space X, the σ-algebra of *Borel* sets is generated by the open sets. Two disjoint subsets A, C of X are said to be *separated by Borel sets* if there is a Borel set $B \subset X$ such that $A \subset B$ and $C \subset X \setminus B$. Recall that a set A in a Polish space Y is called *analytic* iff there is another Polish space X, a Borel subset B of X, and a Borel measurable function f from B into Y such that $A = f[B] := \{f(x) : x \in B\}$ (e.g. RAP, Section 13.2). Equivalently, we can take f to be continuous and/or $B = X$ and/or $X = \mathbb{N}^\infty$, where $\mathbb{N}$ is the set of nonnegative integers with discrete topology and $\mathbb{N}^\infty$ the Cartesian product of an infinite sequence of copies of $\mathbb{N}$, with product topology (RAP, Theorem 13.2.1).

G.1 Theorem (Separation theorem for analytic sets). *Let X be a Polish space. Then any disjoint analytic subsets A, C of X can be separated by Borel sets.*

G.2 Lemma. *If X is a Polish space and A_j, $j = 1, 2, ...$, are analytic in X, then $\bigcup_{j=1}^\infty A_j$ is analytic.*

G.3 Corollary. *If X is a Polish space and A_j, $j = 1, 2, ...$, are disjoint analytic subsets of X, then there exist disjoint Borel sets B_j such that $A_j \subset B_j$ for all j.*

G.4 Theorem. *Let S be a Polish space, Y a separable metric space and A a Borel subset of S. Let f be a 1-1, Borel measurable function from A into Y. Then the range $f[A]$ is a Borel subset of Y, and f^{-1} is a Borel measurable function from $f[A]$ onto A.*

REFERENCE

Cohn, D. (1980). *Measure Theory.* Birkhäuser, Boston and Basel.

APPENDIX H. YOUNG-ORLICZ SPACES

A convex, increasing function g from $[0, \infty)$ onto itself will be called a *Young-Orlicz modulus*. Then g is continuous since it is increasing and onto. Let $(X, \mathcal{S}, \mu)$ be a measure space and g a Young-Orlicz modulus. Let $\mathcal{L}_g(X, \mathcal{S}, \mu)$ be the set of all real-valued measurable functions f on X such that

$$\|f\|_g := \inf\{c > 0 : \int g(|f(x)|/c) d\mu(x) \le 1\} < \infty.$$

Let L_g be the set of equivalence classes of functions in $\mathcal{L}_g(X, \mathcal{S}, \mu)$ for equality almost everywhere (μ). By monotone convergence, we have

H.1 Proposition. *For any Young-Orlicz modulus g and any $f \in \mathcal{L}_g(X, \mathcal{S}, \mu)$, if $0 < c := \|f\|_g < +\infty$, we have $\int (g(|f|/c)) d\mu = 1$, in other words the infimum in the definition of $\|f\|_g$ is attained. Also, $\|f\|_g = 0$ if and only if $f = 0$ almost everywhere for μ.*

Next, we have

H.2 Lemma. *For any Young-Orlicz modulus g, and any measurable functions f_n, if $|f_n| \uparrow |f|$, then $\|f_n\|_g \uparrow \|f\|_g \le +\infty$.*

Next is a fact stating that (not surprisingly) convergence in $\|\cdot\|_g$ norm implies convergence in measure (or probability):

H.3 Lemma. *For any Young-Orlicz modulus g, and any $\varepsilon > 0$, there is a $\delta > 0$ such that if $\|f\|_g \leq \delta$, then $\mu(|f| > \varepsilon) < \varepsilon$.*

H.4 Theorem. *For any measure space $(X, \mathcal{S}, \mu)$, $L_g(X, \mathcal{S}, \mu)$ is a Banach space.*

Let Φ be a Young-Orlicz modulus. Then it has one-sided derivatives as follows (RAP, Corollary 6.3.3): $\phi(x) := \Phi'(x+) := \lim_{y \downarrow x}(\Phi(y) - \Phi(x))/(y - x)$ exists for all $x \geq 0$, and

$$\phi(x-) := \Phi'(x-) := \lim_{y \uparrow x}(\Phi(x) - \Phi(y))/(x - y)$$

exists for all $x > 0$. As the notation suggests, for each $x > 0$, $\phi(x-) \equiv \lim_{y \uparrow x} \phi(y)$, and ϕ is a nondecreasing function on $[0, \infty)$. Thus, $\phi(x-) = \phi(x)$ except for at most countably many values of x, where ϕ may have jumps with $\phi(x) > \phi(x-)$. On any bounded interval, where ϕ is bounded, Φ is Lipschitz and so absolutely continuous. Thus since $\Phi(0) = 0$ we have $\Phi(x) = \int_0^x \phi(u)du$ for any $x > 0$ (e.g. Rudin, 1974, Theorem 8.18). For any $x > 0$, $\phi(x) > 0$ since Φ is strictly increasing.

If ϕ is unbounded, for $0 \leq y < \infty$ let $\psi(y) := \phi^{\leftarrow}(y) := \inf\{x \geq 0 : \phi(x) \geq y\}$. Then $\psi(0) = 0$ and ψ is nondecreasing. Let $\Psi(y) := \int_0^y \psi(t)dt$. Then Ψ is convex and $\Psi' = \psi$ except on the at most countable set where ψ has jumps. Thus for each $y > 0$ we have $\psi(y) > 0$ and Ψ is also strictly increasing.

For any nondecreasing function f from $[0, \infty)$ into itself, it's easily seen that for any $x > 0$ and $u > 0$, $f^{\leftarrow}(u) \geq x$ if and only if $f(t) < u$ for all $t < x$. It follows that $(f^{\leftarrow})^{\leftarrow}(x) = f(x-)$ for all $x > 0$. Since a change in ϕ or ψ on a countable set (of its jumps) doesn't change its indefinite integral Φ or Ψ respectively, the relation between Φ and Ψ is symmetric.

A Young-Orlicz modulus Φ such that ϕ is unbounded and $\phi(x) \downarrow 0$ as $x \downarrow 0$ will be called an *Orlicz modulus*. Then ψ is also unbounded and $\psi(y) > 0$ for all $y > 0$, so Ψ is also an Orlicz modulus. In that case Φ and Ψ will be called *dual Orlicz moduli*. For such moduli we have a basic inequality due to W. H. Young:

H.5 Theorem (W. H. Young). *Let Φ, Ψ be any two dual Young-Orlicz moduli from $[0, \infty)$ onto itself. Then for any $x, y \geq 0$ we have*

$$xy \leq \Phi(x) + \Psi(y),$$

with equality if $x > 0$ and $y = \phi(x-)$.

One of the main uses of Theorem H.5 is to prove an extension of the Rogers-Hölder inequality to Young-Orlicz spaces:

H.6 Theorem. *Let Φ and Ψ be dual Orlicz moduli, and for a measure space $(X, \mathcal{S}, \mu)$ let $f \in \mathcal{L}_\Phi(X, \mathcal{S}, \mu)$ and $g \in \mathcal{L}_\Psi(X, \mathcal{S}, \mu)$. Then $fg \in \mathcal{L}^1(X, \mathcal{S}, \mu)$ and $\int |fg| d\mu \leq 2\|f\|_\Phi \|g\|_\Psi$.*

Proof. By homogeneity we can assume $\|f\|_\Phi = \|g\|_\Psi = 1$. Then applying Proposition H.1 with $c = 1$ and Theorem H.5 we get $\int |fg| d\mu(x) \leq 2$ and the conclusion follows. □

REFERENCES TO APPENDIX H

Birnbaum, Z. W., and Orlicz, W. (1931). Über die Verallgemeinerung des Begriffes der zueinander konjugierten Potenzen. *Studia Math.* **3**, 1-67.

Krasnosel'skiĭ, M. A., and Rutitskii, Ia. B. (1961). *Convex Functions and Orlicz Spaces.* Transl. by L. F. Boron. Noordhoff, Groningen.

Luxemburg, W. A. J., and Zaanen, A. C. (1956). Conjugate spaces of Orlicz spaces. *Akad. Wetensch. Amsterdam Proc.* Ser. A **59** = *Indag. Math.* **18**, 217-228.

Rudin, Walter (1974). *Real and Complex Analysis*, 2d ed. McGraw-Hill, New York.

Young, W. H. (1912). On classes of summable functions and their Fourier series. *Proc. Roy. Soc. London* Ser. A **87**, 225-229.

APPENDIX I
MODIFICATIONS AND VERSIONS OF ISONORMAL PROCESSES

Let T be any set and $(\Omega, \mathcal{A}, P)$ a probability space. Recall that a real-valued stochastic process indexed by T is a function $(t, \omega) \mapsto X_t(\omega)$ from $T \times \Omega$ into $\mathbb{R}$ such that for each $t \in T$, $X_t(\cdot)$ is measurable from Ω into $\mathbb{R}$. A *modification* of the process is another stochastic process Y_t defined for the same T and Ω such that for each t, we have $P(X_t = Y_t) = 1$. A *version* of the process X_t is a process Z_t, $t \in T$, for the same T but defined on a possibly different probability space $(\Omega_1, \mathcal{B}, Q)$ such that X_t and Z_t have the same laws, i.e. for each finite subset F of T, $\mathcal{L}(\{X_t\}_{t\in F}) = \mathcal{L}(\{Z_t\}_{t\in F})$. Clearly, any modification of a process is also a version of the process, but a version, even if on the same probability space, may not be a modification. For example, for an isonormal process L on a Hilbert space H, the process $M(x) := L(-x)$ is a version, but not a modification, of L.

One may take a version or modification of a process in order to get better properties such as continuity. It turns out that for the isonormal process on subsets of Hilbert space, what can be done with a version can also be done by a modification, as follows.

I.1 Theorem. *Let L be an isonormal process restricted to a subset C of Hilbert space. For each of the following properties, if there exists a version M of L with the property, there also is a modification N with the property. For each ω, $x \mapsto M(x)(\omega)$ for $x \in C$ is:*
(a) bounded *(b) uniformly continuous.*
Also, if there is a version with (a) and another with (b) then there is a modification $N(\cdot)$ having both properties.

APPENDIX J. INEQUALITIES

This appendix collects several inequalities bounding the probabilities that random variables, and specifically sums of independent random variables, are large. In UCLT, these inequalities are given in Section 1.3 and are frequently referred to in proofs. In these notes, since so few proofs are given, the inequalities play a smaller role.

Many of the inequalities follow from a basic one of S. Bernštein and P. L. Chebyshev:

J.1. Theorem. *For any real random variable X and $t \in \mathbb{R}$,*

$$Pr\{X \geq t\} \leq \inf_{u\geq 0} e^{-tu} E e^{uX}.$$

For any independent real random variables $X_1, \cdots, X_n$, let $S_n := X_1 + \cdots + X_n$.

J.2 Bernštein's inequality. *Let $X_1, X_2, \cdots, X_n$ be independent real random variables with mean 0. Let $0 < M < \infty$ and suppose that $|X_j| \leq M$ almost surely for $j = 1, \cdots, n$. Let $\sigma_j^2 =$*

var(X_j) *and* $\tau_n^2 := var(S_n) = \sigma_1^2 + \cdots + \sigma_n^2$. *Then for any* $K > 0$,

$$Pr\{|S_n| \geq Kn^{1/2}\} \ \leq \ 2 \cdot \exp(-nK^2/(2\tau_n^2 + 2Mn^{1/2}K/3)).$$

Note that in Bernštein's inequality, for fixed K and M, if X_i are i.i.d. with variance σ^2, then as $n \to \infty$, the bound approaches the normal bound $2\exp(-K^2/(2\sigma^2))$, as given in RAP, Lemma 12.1.6. Moreover, this is true even if $M := M_n \to \infty$ as $n \to \infty$ while K stays constant, provided that $M_n/n^{1/2} \to 0$.

Next, let $s_1, s_2, \cdots$, be i.i.d. variables with $P(s_i = 1) = P(s_i = -1) = 1/2$. Such variables are called "Rademacher" variables. We have the following inequality:

J.3 Proposition (Hoeffding). *For any* $t \geq 0$ *and real* a_j,

$$Pr\{\textstyle\sum_{j=1}^n a_j s_j \geq t\} \ \leq \ \exp(-t^2/(2\sum_{j=1}^n a_j^2)).$$

Here are some remarks on Proposition J.3. Let $Y_1, Y_2, \cdots$, be independent variables which are symmetric, in other words Y_j has the same distribution as $-Y_j$ for all j. Let s_i be Rademacher variables independent of each other and of all the Y_j. Then the sequence $\{s_j Y_j\}_{\{j \geq 1\}}$ has the same distribution as $\{Y_j\}_{\{j \geq 1\}}$. Thus to bound the probability that $\sum_{j=1}^n Y_j > K$, for example, we can consider the conditional probability for each $Y_1, \cdots, Y_n$,

$$\Pr\{\textstyle\sum_{j=1}^n s_j Y_j > K | Y_1, \cdots, Y_n\} \ \leq \ \exp(-K^2/(2\sum_{j=1}^n Y_j^2))$$

by J.3. Then to bound the original probability, integrating over the distribution of the Y_j, one just needs to have bounds on the distribution of $\sum_{j=1}^n Y_j^2$, which may simplify the problem considerably.

The Bernštein inequality (J.2) used variances as well as bounds for centered variables. The following inequalities, also due to Hoeffding, use only bounds. They are essentially the best that can be obtained, under their hypotheses, by the moment generating function technique.

J.4 Theorem (Hoeffding). *Let* $X_1, \cdots, X_n$ *be independent variables with* $0 \leq X_j \leq 1$ *for all* j. *Let* $\overline{X} := (X_1 + \cdots + X_n)/n$ *and* $\mu := E\overline{X}$. *Then for* $0 < t < 1 - \mu$,

$$\begin{aligned} Pr\{\overline{X} - \mu \geq t\} \ &\leq \ \left\{ \left(\frac{\mu}{\mu + t} \right)^{\mu + t} \left(\frac{1 - \mu}{1 - \mu - t} \right)^{1 - \mu - t} \right\}^n \\ &\leq \ e^{-nt^2 g(\mu)} \ \leq \ e^{-2nt^2}, \end{aligned}$$

where

$$\begin{aligned} g(\mu) \ &:= \ (1 - 2\mu)^{-1} \log((1 - \mu)/\mu) \quad \textit{for} \ \ 0 < \mu < 1/2, \ \textit{or} \\ &:= \ 1/(2\mu(1 - \mu)) \qquad \textit{for} \ \ 1/2 \leq \mu \leq 1. \end{aligned}$$

Remarks. For $t > 1 - \mu$, $\Pr(\overline{X} - \mu > t) \leq \Pr(\overline{X} > 1) = 0$. For $t < 0$, the given probability would generally be of the order of $1/2$ or larger, so no small bound for it would be expected.

For the empirical measure P_n, if A is a fixed measurable set, $nP_n(A)$ is a binomial random variable, and in a multinomial distribution, each n_i has a binomial distribution. So we will have need of some inequalities for binomial probabilities, defined by

$$B(k, n, p) \ := \ \textstyle\sum_{0 \leq j \leq k} \binom{n}{j} p^j q^{n-j}, \ 0 \leq q := 1 - p \leq 1,$$

$$E(k,n,p) \;:=\; \textstyle\sum_{k\le j\le n}\binom{n}{j}p^j q^{n-j}.$$

Here k is usually, but not necessarily, an integer. Thus, in n independent trials with probability p of success on each trial, so that q is the probability of failure, $B(k,n,p)$ is the probability of at most k successes and $E(k,n,p)$ is the probability of at least k successes.

J.5 Chernoff-Okamoto inequalities. *We have*

$$E(k,n,p) \;\le\; \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k} \quad \textit{if} \quad k \ge np,$$

$$B(k,n,p) \;\le\; \exp(-(np-k)^2/(2npq)) \quad \textit{if} \;\; k \le np \le n/2.$$

Proof. These facts follow directly from the Hoeffding inequality J.4. For the second one, note that $B(k,n,p) \;=\; E(n-k,n,1-p)$ and apply the $g(\mu)$ case with $\mu = 1-p$. □

J.6 Proposition. $E(k,n,p) \;\le\; (np/k)^k\, e^{k-np}$ *if* $k \ge np$.

The next inequality is for the special value $p = 1/2$:

J.7 Proposition. *If* $k \le n/2$ *then* $2^n B(k,n,1/2) \;\le\; (ne/k)^k$.

A form of Stirling's formula with error bounds is:

J.8 Theorem. *For* $n = 1,2,\cdots$, $e^{1/(12n+1)} \;\le\; n!(e/n)^n(2\pi n)^{-1/2} \;\le\; e^{1/12n}$.

Proof. See Feller (1968), vol. 1, Section II.9, p. 54. □

For any real x let $x^+ := \max(x,0)$. A *Poisson* random variable z with parameter m has the distribution given by $\Pr(z=k) \;=\; e^{-m}m^k/k!$ for each nonnegative integer k.

J.9 Lemma. *For any Poisson variable* z *with parameter* $m \ge 1$,

$$E(z-m)^+ \;\ge\; m^{1/2}/8.$$

In the following two facts, let $X_1, X_2, \cdots, X_n$ be independent random variables with values in a separable normed space S with norm $\|\cdot\|$. Let $S_j := X_1 + \cdots + X_j$ for $j = 1,\cdots,n$.

J.10 Ottaviani's inequality. *If for some* $\alpha > 0$ *and* c *with* $0 < c < 1$, *we have* $P(\|S_n - S_j\| > \alpha) \le c$ *for all* $j = 1,\cdots,n$, *then*

$$P\{\max_{j\le n}\|S_j\| \ge 2\alpha\} \;\le\; P(\|S_n\| \ge \alpha)/(1-c).$$

Proof. The proof in RAP, 9.7.2, for $S = \mathbb{R}^k$, works for any separable normed S. Here $(x,y) \mapsto \|x-y\|$ is measurable: $S \times S \mapsto \mathbb{R}$ by RAP, Proposition 4.1.7. □

When the random variables X_j are symmetric, there is a simpler inequality:

J.11 P. Lévy's inequality. *Given a probability space* (Ω, P) *and a countable set* Y, *let* $X_1,\ X_2, \ldots,$ *be stochastic processes defined on* Ω *indexed by* Y, *in other words for each* j *and*

$y \in Y$, $X_j(y)(\cdot)$ is a random variable on Ω. For any bounded function f on Y let $\|f\|_Y := \sup\{|f(y)|\colon\ y \in Y\}$. Suppose that the processes X_j are independent with $\|X_j\|_Y < \infty$ a.s., and symmetric, in other words for each j, the random variables $\{-X_j(y)\colon y \in Y\}$ have the same joint distribution as $\{X_j(y)\colon\ y \in Y\}$. Let $S_n := X_1 + \cdots + X_n$. Then for each n, and $M > 0$,

$$P\left(\max_{j \le n} \|S_j\|_Y > M\right) \le 2P\left(\|S_n\|_Y > M\right).$$

Note. The norm on a separable Banach space $(X, \|\cdot\|)$ can always be written in the form $\|\cdot\|_Y$ for Y countable, via the Hahn-Banach theorem (apply RAP, Corollary 6.1.5, to a countable dense set in the unit ball of X to get a countable norming subset Y in the dual X' of X, although X' may not be separable). On the other hand, the above Lemma applies to some nonseparable Banach spaces: the space of all bounded functions on an infinite Y with supremum norm is itself nonseparable.

REFERENCES

*An asterisk indicates a work I have seen discussed in secondary sources but not in the original.

Bennett, George (1962). Probability inequalities for the sum of independent random variables. *J. Amer. Statist. Assoc.* **57**, 33-45.

Bernštein, Sergei N. (1924). Ob odnom vidoizmenenii neravenstva Chebysheva i o pogreshnosti formuly Laplasa (in Russian). *Uchen. Zapiski Nauchn.-issled. Kafedr Ukrainy, Otdel. Mat.*, vyp. 1, 38-48; reprinted in S. N. Bernštein, *Sobranie Sochinenii˘ [Collected Works], Tom IV, Teoriya Veroiatnostei, Matematicheskaya Statistika*, Nauka, Moscow, 1964, pp. 71-79.

*Bernštein, Sergei N. (1927). *Teoriya Veroiatnostei* (in Russian). Moscow. 2d. ed., 1934.

Chernoff, Herman (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.* **23**, 493-507.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6**, 899-929; Correction **7** (1979), 909-911.

Dudley, R. M. (1982). Empirical and Poisson processes on classes of sets or functions too large for central limit theorems. *Z. Wahrscheinlichkeitsth. verw. Gebiete* **61**, 355-368.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications,* 3d ed. Wiley, New York.

Giné, Evarist (1974). On the central limit theorem for sample continuous processes. *Ann. Probab.* **2**, 629-641.

Hoeffding, Wassily (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13-30.

Kahane, J.-P. (1985). *Some Random Series of Functions*, 2d. ed. Cambridge University Press.

Okamoto, Masashi (1958). Some inequalities relating to the partial sum of binomial probabilities. *Ann. Inst. Statist. Math.* **10**, 29-35.

Uspensky, J. V. (1937). *Introduction to Mathematical Probability.* McGraw-Hill, New York.

APPENDIX K. METRIC ENTROPY AND CAPACITY

The word "entropy" is applied to several concepts in mathematics. What they have in common is apparently that they give some measure of the size or complexity of some set or transformation and that their definitions involve logarithms. Beyond this rather superficial resemblance, there are major differences. What are here called "metric entropy" and "metric capacity" are measures of the size of a metric space, which must be totally bounded (have compact completion) in order for the metric entropy or capacity to be finite. Metric entropy will provide a useful general technique for dealing with classes of sets or functions in general spaces, as opposed to Markov (or martingale) methods. The latter methods apply, as in Chapter 1, when the sample space is $\mathbb{R}$ and the class $\mathcal{C}$ of sets is the class of half-lines $(-\infty, x]$, $x \in \mathbb{R}$, so that $\mathcal{C}$ with its ordering by inclusion is isomorphic to $\mathbb{R}$ with its usual ordering.

Let (S, d) be a metric space and A a subset of S. Let $\varepsilon > 0$. A set $F \subset S$ (not necessarily included in A) is called an *ε-net* for A if and only if for each $x \in A$, there is a $y \in F$ with $d(x, y) \leq \varepsilon$. Let $N(\varepsilon, A, S, d)$ denote the minimal number of points in an ε-net in S for A.

For any set $C \subset S$, define the *diameter* of C by

$$\operatorname{diam} C \ := \ \sup\{d(x, y) : x, y \in C\}.$$

Let $N(\varepsilon, C, d)$ be the smallest n such that C is the union of n sets of diameter at most 2ε.

Let $D(\varepsilon, A, d)$ denote the largest n such that there is a subset $F \subset A$ with F having n members and $d(x, y) > \varepsilon$ whenever $x \neq y$ for x and y in F.

The three quantities just defined are related by the following inequalities:

K.1 Theorem. *For any $\varepsilon > 0$ and set A in a metric space S with metric d,*

$$D(2\varepsilon, A, d) \leq N(\varepsilon, A, d) \leq N(\varepsilon, A, S, d) \leq N(\varepsilon, A, A, d) \leq D(\varepsilon, A, d).$$

Proof. The first inequality holds since a set of diameter 2ε can contain at most one of a set of points more than 2ε apart. The next holds because any ball $\overline{B}(x, \varepsilon) \ := \ \{y : \ d(x, y) \leq \varepsilon\}$ is a set of diameter at most 2ε. The third inequality holds since requiring centers to be in A is more restrictive. The last holds because a set F of points more than ε apart, with maximal cardinality, must be an ε-net, since otherwise there would be a point more than ε away from each point of F, which could be adjoined to F, a contradiction unless F is infinite, but then the inequality holds trivially. □

It follows that as $\varepsilon \downarrow 0$, when all the functions in the Theorem go to ∞ unless S is a finite set, they have the same asymptotic behavior up to a factor of 2 in ε. So it will be convenient to choose one of the four and make statements about it, which will then yield corresponding results for the others. The choice is somewhat arbitrary. Here are some considerations that bear on the choice.

The finite set of points, whether more than ε apart or forming an ε-net, are often useful, as opposed to the sets in the definition of $N(\varepsilon, A, d)$. The quantity $N(\varepsilon, A, S, d)$ depends not only on A but on the larger space S. Many workers, possibly for these reasons, have preferred $N(\varepsilon, A, A, d)$. But the latter may decrease when the set A increases. For example, let A be the surface of a sphere of radius ε around 0 in a Euclidean space S and let $B \ := \ A \cup \{0\}$. Then $N(\varepsilon, B, B, d) = 1 < N(\varepsilon, A, A, d)$. This was the reason, apparently, that Kolmogorov chose to use $N(\varepsilon, A, d)$.

I have chosen to adopt $D(\varepsilon, A, d)$ as basic. It depends only on A, not on the larger space S, and is nondecreasing in A. If $D(\varepsilon, A, d) = n$, then there are n points which are more than ε apart and at the same time form an ε-net.

Now, the ε-*entropy* of the metric space (A, d) is defined as $H(\varepsilon, A, d) := \log N(\varepsilon, A, d)$, and the ε-*capacity* as $\log D(\varepsilon, A, d)$. Some other authors take logarithms to the base 2, by analogy with information-theoretic entropy. In these notes logarithms will be taken to the usual base e, which fits for example with bounds coming from moment generating functions as in the next section, and with Gaussian measures as in Chapter 2. There are a number of interesting sets of functions where $N(\varepsilon, A, d)$ is of the order of magnitude $\exp(\varepsilon^{-r})$ as $\varepsilon \downarrow 0$, for some power $r > 0$, so that the ε-entropy, and likewise the ε-capacity, have the simpler order ε^{-r}. But in other cases below, $D(\varepsilon, A, d)$ is itself of the order of a power of $1/\varepsilon$.

REFERENCES

Kolmogorov, A. N. (1955). Bounds for the minimal number of elements of an ε-net in various classes of functions and their applications to the question of representability of functions of several variables by superpositions of functions of fewer variables (in Russian). *Uspekhi Mat. Nauk* **10**, no. 1 (63), 192-194.

Kolmogorov, A. N., and Tikhomirov, V. M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Amer. Math. Soc. Transls. (Ser. 2)* **17** (1961), 277-364 (*Uspekhi Mat. Nauk* **14,** vyp. 2 (86), 3-86).

Lorentz, G. G. (1966). Metric entropy and approximation. *Bull. Amer. Math. Soc.* **72**, 903-937.

Posner, Edward C., Rodemich, Eugene R., and Rumsey, Howard Jr. (1967). Epsilon entropy of stochastic processes. *Ann. Math. Statist.* **38**, 1000-1020.

Posner, Edward C., Rodemich, Eugene R., and Rumsey, Howard Jr. (1969). Epsilon entropy of Gaussian processes. *Ann. Math. Statist.* **40**, 1272-1296.