UNIFORM CENTRAL LIMIT THEOREMS

RICHARD M. DUDLEY
Massachusetts Institute of Technology

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1999

Printed in the United States of America

Typeface Times 10/13 pt. *System* LaTeX [RW]

A catalog record of this book is available from the British Library

Library of Congress cataloging in publication data

Dudley, R. M. (Richard M.)
Uniform central limit theorems / R. M. Dudley.
p. cm. – (Cambridge studies in advanced mathematics: 63)
Includes bibliographical references.
ISBN 0 521 46102 2
1. Central limit theorem. I. Title. II. Series.
QA 273.67.D84 1999
519.2–DC21 98-35582
CIP

ISBN 0 521 46102 2 hardback

Contents

1

Introduction: Donsker's Theorem, Metric Entropy, and Inequalities

Let P be a probability measure on the Borel sets of the real line $\mathbb{R}$ with distribution function $F(x) := P((-\infty, x])$. Here and throughout, ":=" means "equals by definition." Let $X_1, X_2, \cdots$ be i.i.d. (independent, identically distributed) random variables with distribution P. For each $n = 1, 2, \cdots$ and any Borel set $A \subset \mathbb{R}$, let $P_n(A) := \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j}(A)$, where $\delta_x(A) = 1_A(x)$. Then P_n is a probability measure for each $X_1, \cdots, X_n$ and is called the *empirical measure*. Let F_n be the distribution function of P_n. Then F_n is called the *empirical distribution function*.

The developments to be described in this book began with the Glivenko-Cantelli theorem, a uniform law of large numbers, which says that with probability 1, F_n converges to F as $n \to \infty$, *uniformly* on $\mathbb{R}$, meaning that $\sup_x |(F_n - F)(x)| \to 0$ as $n \to \infty$ (RAP, Theorem 11.4.2); as mentioned in the Note at the end of the Preface, "RAP" refers to the author's book *Real Analysis and Probability*.

The next step was to consider the limiting behavior of $\alpha_n := n^{1/2}(F_n - F)$ as $n \to \infty$. For any fixed t, the central limit theorem in its most classical form, for binomial distributions, says that $\alpha_n(t)$ converges in distribution to $N(0, F(t)(1 - F(t)))$, in other words a normal (Gaussian) law, with mean 0 and variance $F(t)(1 - F(t))$. Here a *law* is a probability measure defined on the Borel sets.

For any finite set T of values of t, the multidimensional central limit theorem (RAP, Theorem 9.5.6) tells us that $\alpha_n(t)$ for t in T converges in distribution as $n \to \infty$ to a normal law $N(0, C_F)$ with mean 0 and covariance $C_F(s, t) = F(s)(1 - F(t))$ for $s \leq t$.

The *Brownian bridge* (RAP, Section 12.1) is a stochastic process $y_t(\omega)$ defined for $0 \leq t \leq 1$ and ω in some probability space Ω, such that for any finite set $S \subset [0, 1]$, y_t for t in S have distribution $N(0, C)$, where $C = C_U$ for the uniform distribution function $U(t) = t$, $0 \leq t \leq 1$, and $t \mapsto y_t(\omega)$ is

continuous for almost all ω. So the empirical process α_n converges in distribution to the Brownian bridge composed with F, namely $t \mapsto y_{F(t)}$, at least when restricted to finite sets.

It was then natural to ask whether this convergence extends to infinite sets or the whole interval or line. Kolmogorov (1933) showed that when F is continuous, the supremum $\sup_t \alpha_n(t)$ and the supremum of absolute value, $\sup_t |\alpha_n(t)|$, converge in distribution to the laws of the same functionals of y_F. Then, these functionals of y_F have the same distributions as for the Brownian bridge itself, since F takes $\mathbb{R}$ onto an interval including $(0, 1)$ and which may or may not contain 0 or 1; this makes no difference to the suprema since $y_0 \equiv y_1 \equiv 0$. Also, $y_t \to 0$ almost surely as $t \downarrow 0$ or $t \uparrow 1$ by sample continuity; the suprema can be restricted to a countable dense set such as the rational numbers in $(0, 1)$ and are thus measurable. Kolmogorov evaluated the distributions of $\sup_t y_t$ and $\sup_t |y_t|$ explicitly (see RAP, Propositions 12.3.3 and 12.3.4).

Doob (1949) asked whether the convergence in distribution held for more general functionals. Donsker (1952) stated and proved (not quite correctly) a general extension. This book will present results proved over the past few decades by many researchers, where the collection of half-lines $(-\infty, x]$, $x \in \mathbb{R}$, is replaced by much more general classes of sets in, and functions on, general sample spaces, for example the class of all ellipsoids in $\mathbb{R}^3$.

To motivate and illustrate the general theory, the first section will give a revised formulation and proof of Donsker's theorem. Then the next two sections, on metric entropy and inequalities, provide concepts and facts to be used in the rest of the book.

1.1 Empirical processes: the classical case

In this section, the aim is to treat an illuminating and historically basic special case. There will be plenty of generality later on. Here let P be the uniform distribution (Lebesgue measure) on the unit interval $[0, 1]$. Let U be its distribution function, $U(t) = t,\ 0 \le t \le 1$. Let U_n be its empirical distribution functions and $\alpha_n := n^{1/2}(U_n - U)$ on $[0, 1]$.

It will be proved that as $n \to \infty,\ \alpha_n$ converges in law (in a sense to be made precise below) to a Brownian bridge process $y_t,\ 0 \le t \le 1$ (RAP, before Theorem 12.1.5). Recall that y_t can be written in terms of a Wiener process (Brownian motion) x_t, namely $y_t = x_t - tx_1, 0 \le t \le 1$. Or, y_t is x_t conditioned on $x_1 = 0$ in a suitable sense (RAP, Proposition 12.3.2). The Brownian bridge (like the Brownian motion) is sample-continuous, that is, it can be chosen such that for all ω, the function $t \mapsto y_t(\omega)$ is continuous on $[0, 1]$ (RAP, Theorem 12.1.5).

Donsker in 1952 proved that the convergence in law of α_n to the Brownian bridge holds, in a sense, with respect to uniform convergence in t on the whole interval $[0, 1]$. How to define such convergence in law correctly, however, was not clarified until much later. General definitions will be given in Chapter 3. Here, a more special approach will be taken in order to state and prove an accessible form of Donsker's theorem.

For a function f on $[0, 1]$, we have the sup norm

$$\|f\|_\infty := \sup\{|f(t)|: 0 \le t \le 1\}.$$

Here is the form of Donsker's theorem that will be the main result of this section.

1.1.1 Theorem *For $n = 1, 2, \cdots$, there exist probability spaces Ω_n such that:*

(a) *On Ω_n, there exist n i.i.d. random variables $X_1, \cdots, X_n$ with uniform distribution in $[0, 1]$. Let α_n be the nth empirical process based on these X_i;*

(b) *On Ω_n a sample-continuous Brownian bridge process Y_n: $(t, \omega) \mapsto Y_n(t, \omega)$ is also defined;*

(c) *$\|\alpha_n - Y_n\|_\infty$ is measurable, and for all $\varepsilon > 0$, $\Pr(\|\alpha_n - Y_n\|_\infty > \varepsilon) \to 0$ as $n \to \infty$.*

Notes. (i) Part (c) gives a sense in which the empirical process α_n converges in distribution to the Brownian bridge with respect to the sup norm $\|\cdot\|_\infty$.

(ii) It is actually possible to use one probability space on which $X_1, X_2, \cdots$ are i.i.d., while $Y_n = (B_1 + \cdots + B_n)/\sqrt{n}$, B_j being independent Brownian bridges. This is an example of an *invariance principle*, to be treated in Chapter 9, not proved in this section.

(iii) One can define all α_n and Y_n on one probability space and make Y_n all equal some Y, although here the joint distributions of α_n for different n will be different from their original ones. Then α_n will converge to Y in probability and moreover can be defined so that $\|\alpha_n - Y\|_\infty \to 0$ almost surely, as will be shown in Section 3.5.

Proof For a positive integer k, let L_k be the set of $k + 1$ equally spaced points,

$$L_k := \{0, 1/k, 2/k, \cdots, 1\} \subset [0, 1].$$

It will first be shown that both processes α_n and y_t, for large enough n and k, can be well approximated by step functions and then by piecewise-linear interpolation of their values on L_k.

Given $0 < \varepsilon \leq 1$, take $k = k(\varepsilon)$ large enough so that

$$4k \cdot \exp\left(-k\varepsilon^2/648\right) < \varepsilon/6. \tag{1.1.2}$$

Let $I_{jk} := [j/k, (j+1)/k]$, $j = 0, \cdots, k-1$. By the representation $y_t = x_t - tx_1$, we have

$$\Pr\{|y_t - y_{j/k}| > \varepsilon/6 \text{ for some } t \in I_{jk}\} \leq p_1 + p_2,$$

where

$$p_1 := \Pr\{|x_1| > k\varepsilon/18\}, \qquad p_2 := \Pr\{|x_t - x_{j/k}| > \varepsilon/9 \text{ for some } t \in I_{jk}\}.$$

Then $p_1 \leq 2 \cdot \exp(-k^2\varepsilon^2/648)$ (RAP, Lemma 12.1.6(b)). For p_2, via a reflection principle (RAP, 12.3.1) and the fact that $\{x_{u+h} - x_u\}_{h \geq 0}$ has the same distribution as $\{x_h\}_{h \geq 0}$ (applied to $u = j/k$), we have $p_2 \leq 4\exp(-k\varepsilon^2/162)$. Thus by (1.1.2),

$$\Pr\{|y_t - y_{j/k}| > \varepsilon/6 \text{ for some } j = 0, \cdots, k-1 \text{ and some } t \in I_{jk}\} < \varepsilon/3. \tag{1.1.3}$$

Next, we need a similar bound for α_n when n is large. The following will help:

1.1.4 Lemma *Given the uniform distribution U on $[0, 1]$:*

(a) *For $0 \leq u \leq 1$ and any finite set $S \subset [0, 1-u]$, the joint distribution of $\{U_n(u+s) - U_n(u)\}_{s \in S}$ is the same as for $u = 0$.*
(b) *The same holds for α_n in place of U_n.*
(c) *The distribution of $\sup\{|\alpha_n(t + j/k) - \alpha_n(j/k)|: 0 \leq t \leq 1/k\}$ is the same for all j.*

Proof (a) Let $S = \{s_i\}_{i=1}^m$ where we can assume $s_0 = 0$. It's enough to consider $\{U_n(u+s_j) - U_n(u+s_{j-1})\}_{j=1}^m$, whose partial sums give the desired quantities. Multiplying by n, we get m random variables from a multinomial distribution for n observations for the first m of $m+1$ categories, which have probabilities $\{s_j - s_{j-1}\}_{j=1}^{m+1}$, where $s_{m+1} = 1$ (Appendix B, Theorem B.2). This distribution doesn't depend on u.

(b) Since $\alpha_n(u+s) - \alpha_n(u) = n^{1/2}(U_n(u+s) - U_n(u) - s)$, (b) follows from (a).

(c) The statement holds for finite subsets of I_{jk} by (b). By monotone convergence, we can let the finite sets increase up to the countable set of rational numbers in I_{jk}. Since U_n is right-continuous, suprema over the rationals in I_{jk} equal suprema over the whole interval (the right endpoint is rational), and Lemma 1.1.4 is proved. □

So in bounding the supremum in Lemma 1.1.4(c) we can take $j = 0$, and we need to bound $\Pr\{n^{1/2}|U_n(t) - t| > \varepsilon$ for some $t \in [0, 1/k]\}$. Suppose given a multinomial distribution of numbers $n_1, \cdots, n_r$ of sample size $n = n_1 + \cdots + n_r$ in r bins with probabilities $p_1, \cdots, p_r$. Then for each j, the conditional distribution of n_{j+1} given $n_1, \cdots, n_j$ is the same as that given $n_1 + \cdots + n_j$, namely a binomial distribution for $n - n_1 - \cdots - n_j$ trials with probability $p_{j+1}/(p_{j+1} + \cdots + p_n)$ of success on each (see Appendix B, Theorem B.3(c)). It follows that the empirical distribution function U_n has the following *Markov* property: if $0 < t_1 < \cdots < t_j < t < u$, then the conditional distribution of $U_n(u)$ given $U_n(t_1), \cdots, U_n(t_j), U_n(t)$ is the same as that given $U_n(t)$. Specifically, given that $U_n(t) = m/n$, the conditional distribution of $U_n(u)$ is that of $(m + X)/n$ where X has a binomial distribution for $n - m$ trials with success probability $(u - t)/(1 - t)$. To be given $U_n(t) = m/n$ is equivalent to being given $\alpha_n(t) = n^{1/2}(\frac{m}{n} - t)$, and α_n also has the Markov property. So the conditional distribution of $\alpha_n(u)$ given $m = nU_n(t)$ has mean

$$\mu_m := n^{1/2}\left\{\frac{m}{n}\left[1 - \frac{u-t}{1-t}\right] + \frac{u-t}{1-t} - u\right\} = n^{1/2}\left(\left(\frac{m}{n} - t\right)\left(\frac{1-u}{1-t}\right)\right)$$

and variance

$$\frac{(n-m)(u-t)(1-u)}{n(1-t)^2} \le \frac{u-t}{1-t} \le u.$$

So, by Chebyshev's inequality,

$$\Pr\left\{\left|\alpha_n(u) - \mu_m\right| \ge 2u^{1/2} \,\middle|\, m\right\} \le 1/4.$$

If $u \le 1/2$, then $\frac{1-u}{1-t} \ge \frac{1}{2}$. Let $0 < \delta < 1$. If $\alpha_n(t) > \delta$, then $\frac{m}{n} - t > \delta/n^{1/2}$ and $\mu_m > \delta(\frac{1-u}{1-t}) \ge \delta/2$, so for any $\gamma > \delta$ (such that $\Pr\{\alpha_n(t) = \gamma\} > 0$),

$$\Pr\left\{\alpha_n(u) > \frac{\delta}{2} - 2u^{1/2} \,\middle|\, \alpha_n(t) = \gamma\right\} \ge 3/4.$$

(For such a γ, $\gamma = n^{1/2}(\frac{m}{n} - t)$ for some integer m.) If $u < \delta^2/64$, then $u < 1/2$ and

$$\Pr\{\alpha_n(u) > \delta/4 \mid \alpha_n(t) = \gamma\} \ge 3/4.$$

Let $u = 1/k$ and $\delta = \varepsilon/4$. Then by (1.1.2), since $e^{-x} < 1/24$ implies $x > 2$, we have $u < \delta^2/64$, so

$$\Pr\{\alpha_n(1/k) > \varepsilon/16 \mid \alpha_n(t) = \gamma\} \ge 3/4 \text{ for } \gamma > \varepsilon/4.$$

Now take a positive integer r and let τ be the smallest value of $j/(kr)$, if any, for $j = 1, \cdots, r$, for which $\alpha_n(\tau) > \varepsilon/4$. Let A_r be the event that such a j exists. Let $A_{rj} := \{\tau = j/(kr)\}$. Then A_r is the union of the

disjoint sets A_{rj} for $j = 1, \cdots, r$. For each such j, by the Markov property, $\Pr\{\alpha_n(1/k) > \varepsilon/16 \mid A_{rj}\} \geq 3/4$. Thus

$$\Pr\{\alpha_n(1/k) > \varepsilon/16 \mid A_r\} \ \geq \ 3/4.$$

Let $r \to \infty$. Then by right continuity of U_n and α_n, we get

$$\Pr\{\alpha_n(t) > \varepsilon/4 \text{ for some } t \in [0, 1/k]\} \ \leq \ \tfrac{4}{3} \Pr\{\alpha_n(1/k) > \varepsilon/16\}.$$

Likewise,

$$\Pr\{\alpha_n(t) < -\varepsilon/4 \text{ for some } t \in [0, 1/k]\} \ \leq \ \tfrac{4}{3} \Pr\{\alpha_n(1/k) < -\varepsilon/16\}.$$

Thus by Lemma 1.1.4(c),

$$\begin{aligned} (1.1.5) \quad \Pr\{|\alpha_n(t) - \alpha_n(j/k)| > \varepsilon/4 \text{ for some } t \in I_{jk} \\ \text{and } j = 0, 1, \cdots, k-1\} \ \leq \ (4k/3) \Pr(|\alpha_n(1/k)| > \varepsilon/16). \end{aligned}$$

As $n \to \infty$, for our fixed k, by the central limit theorem and RAP, Lemma 12.1.6(b),

$$\Pr\{|\alpha_n(1/k)| > \varepsilon/16\} \ \to \ \Pr\{|y_{1/k}| > \varepsilon/16\} \ \leq \ 2 \cdot \exp\big(-k\varepsilon^2/512\big).$$

So for n large enough, say $n \geq n_0 = n_0(\varepsilon)$, recalling that $k = k(\varepsilon)$,

$$\Pr\{|\alpha_n(1/k)| > \varepsilon/16\} \ < \ 3 \cdot \exp\big(-k\varepsilon^2/512\big).$$

Then by (1.1.5) and (1.1.2), for $n \geq n_0$,

$$\begin{aligned} (1.1.6) \qquad \Pr\{|\alpha_n(t) - \alpha_n(j/k)| > \varepsilon/4 \\ \text{for some } j = 0, \cdots, k-1 \text{ and } t \in I_{jk}\} \ \leq \ \varepsilon/6. \end{aligned}$$

As mentioned previously, the law, say $\mathcal{L}_k(\alpha_n)$, of $\{\alpha_n(i/k)\}_{i=0}^k$ converges by the central limit theorem in $\mathbb{R}^{k+1}$ to that of $\{y_{i/k}\}_{i=0}^k$, say $\mathcal{L}_k(y)$. On $\mathbb{R}^{k+1}$, put the metric $d_\infty(x, y) := |x - y|_\infty := \max_i |x_i - y_i|$, which of course metrizes the usual topology. Since convergence of laws is metrized by Prokhorov's metric ρ (RAP, Theorem 11.3.3), for n large enough, say $n \geq n_1(\varepsilon) \geq n_0(\varepsilon)$, we have $\rho(\mathcal{L}_k(\alpha_n), \mathcal{L}_k(y)) < \varepsilon/6$. Then by Strassen's theorem (RAP, Corollary 11.6.4), there is a probability measure μ_n on $\mathbb{R}^{k+1} \times \mathbb{R}^{k+1}$ such that for (X, Y) with $\mathcal{L}(X, Y) = \mu_n$, we have

$$\begin{aligned} (1.1.7) \qquad \mathcal{L}(X) = \mathcal{L}_k(\alpha_n), \qquad \mathcal{L}(Y) = \mathcal{L}_k(y), \\ \text{and } \mu_n\{(x, y) : \ |x - y|_\infty > \varepsilon/6\} \ \leq \ \varepsilon/6 \end{aligned}$$

(RAP, Section 9.2).

Let Lib_k (for "linear in between") be the function from $\mathbb{R}^{k+1}$ into the space $C[0,1]$ of all continuous real functions on $[0,1]$ such that $Lib_k(x)(j/k) = x_j$, $j = 0, \cdots, k$, and $Lib_k(x)(\cdot)$ is linear (affine) on each closed interval $I_{jk} = [j/k, (j+1)/k]$, $j = 0, \cdots, k-1$. For any $x, y \in \mathbb{R}^{k+1}$, $Lib_k(x) - Lib_k(y)$ is also linear on each I_{jk}, so it attains its maximum, minimum, and maximum absolute value at endpoints. So for the supremum norm $\|f\|_\infty := \sup_{0\le x\le 1} |f(x)|$ on $C[0,1]$, Lib_k is an isometry into:

$$\|Lib_k(x) - Lib_k(y)\|_\infty = |x - y|_\infty \quad \text{for all } x, y \in \mathbb{R}^{k+1}.$$

Since $C[0,1]$ is a separable metric space (e.g., RAP, Corollary 11.2.5), $(f, g) \mapsto \|f - g\|_\infty$ is jointly measurable in $f, g \in C[0,1]$ (RAP, Proposition 4.1.7). So, from (1.1.7) we get

$$\mu_n\{\|Lib_k(x) - Lib_k(y)\|_\infty > \varepsilon/6\} \le \varepsilon/6. \tag{1.1.8}$$

For any real-valued function f on $[0,1]$, let $\pi_k(f) = \{f(j/k)\}_{j=0}^k \in \mathbb{R}^{k+1}$. Then $\pi_k(Lib_k(x)) = x$ for all $x \in \mathbb{R}^{k+1}$.

For a sample-continuous Brownian bridge process $(\omega, t) \mapsto y_t(\omega)$, $0 \le t \le 1$, the map

$$\omega \mapsto \{t \mapsto y_t(\omega) \colon 0 \le t \le 1\} \in C[0,1]$$

is measurable for the Borel σ-algebra on $C[0,1]$ (by a simple adaptation of RAP, Proposition 12.2.2). (Recall that in any topological space, the Borel σ-algebra is the one generated by the open sets.) If $|x_{j+1} - x_j| < \varepsilon$ then $|Lib_k(x)(t) - x_j| < \varepsilon$ for all $t \in I_{jk}$. It follows from (1.1.3) that

$$\Pr\{\|y - Lib_k(\pi_k(y))\|_\infty > \varepsilon/3\} < \varepsilon/3, \tag{1.1.9}$$

where for each ω, we have a function $y \colon t \mapsto y_t(\omega)$, $0 \le t \le 1$.

We can take the probability space for each α_n process as the unit cube I^n, where the n i.i.d. uniform variables in defining U_n and α_n are $x_1, \cdots, x_n$, with $x = (x_1, \cdots, x_n) \in I^n$. Then

$$A_k \colon x \mapsto \{\alpha_n(j/k)\}_{j=0}^k$$

is measurable from I^n into $\mathbb{R}^{k+1}$ and has distribution $\mathcal{L}_k(\alpha_n)$ on $\mathbb{R}^{k+1}$. Also, $x \mapsto Lib_k(A_k(x))$ is measurable from I^n into $C[0,1]$.

The next theorem will give a way of linking up or "coupling" processes. Recall that a *Polish* space is a topological space metrizable by a complete separable metric.

1.1.10 Theorem (Vorob'ev-Berkes-Philipp) *Let X, Y and Z be Polish spaces with Borel σ-algebras. Let α be a law on $X \times Y$ and let β be a law on $Y \times Z$.*

Let $\pi_Y(x, y) := y$ and $\tau_Y(y, z) := y$ for all $(x, y, z) \in X \times Y \times Z$. Suppose the marginal distributions of α and β on Y are equal, in other words $\eta := \alpha \circ \pi_Y^{-1} = \beta \circ \tau_Y^{-1}$ on Y. Let $\pi_{12}(x, y, z) := (x, y)$ and $\pi_{23}(x, y, z) := (y, z)$. Then there exists a law γ on $X \times Y \times Z$ such that $\gamma \circ \pi_{12}^{-1} = \alpha$ and $\gamma \circ \pi_{23}^{-1} = \beta$.

Proof There exist conditional distributions α_y for α on X given $y \in Y$, so that for each $y \in Y$, α_y is a probability measure on X, for any Borel set $A \subset X$, the function $y \mapsto \alpha_y(A)$ is measurable, and for any integrable function f for α,

$$\int f d\alpha \;=\; \iint f(x, y)\, d\alpha_y(x)\, d\eta(y)$$

(RAP, Section 10.2). Likewise, there exist conditional distributions β_y on Z for β. Let x and z be conditionally independent given y. In other words, define a set function γ on $X \times Y \times Z$ by

$$\gamma(C) \;=\; \iiint 1_C(x, y, z)\, d\alpha_y(x)\, d\beta_y(z)\, d\eta(y).$$

The integral is well-defined if

(a) $C = U \times V \times W$ for Borel sets U, V, and W in X, Y, and Z, respectively;

(b) C is a finite union of such sets, which can be taken to be disjoint (RAP, Proposition 3.2.2 twice); or

(c) C is any Borel set in $X \times Y \times Z$, by RAP, Proposition 3.2.3 and the monotone class theorem (RAP, Theorem 4.4.2).

Also, γ is countably additive by monotone convergence (for all three integrals). So γ is a law on $X \times Y \times Z$. Clearly $\gamma \circ \pi_{12}^{-1} = \alpha$ and $\gamma \circ \pi_{23}^{-1} = \beta$. □

Now, let's continue the proof of Theorem 1.1.1. The function $(x, f) \mapsto \|\alpha_n - f\|_\infty$ is jointly Borel measurable for $x \in I^n$ and $f \in C[0, 1]$. Also, $u \mapsto Lib_k(u)$ is continuous and thus Borel measurable from $\mathbb{R}^{k+1}$ into $C[0, 1]$. So $(x, u) \mapsto \|\alpha_n - Lib_k(u)\|_\infty$ is jointly measurable on $I^n \times \mathbb{R}^{k+1}$. (This is true even though $\alpha_n \notin C[0, 1]$ and the functions $t \mapsto \alpha_n(t)$ for different ω form a nonseparable space for $\|\cdot\|_\infty$.) Thus $x \mapsto \|\alpha_n - Lib_k(A_k(x))\|_\infty$ is measurable on I^n. From (1.1.6), we then have

$$\Pr\{\|\alpha_n - Lib_k(A_k(x))\|_\infty > \varepsilon/2\} \;\le\; \varepsilon/6. \tag{1.1.11}$$

Apply Theorem 1.1.10 to $(X, Y, Z) = (I^n, \mathbb{R}^{k+1}, \mathbb{R}^{k+1})$, with the law of $(x, A_k(x))$ on $I^n \times \mathbb{R}^{k+1}$, and μ_n from (1.1.7) on $\mathbb{R}^{k+1} \times \mathbb{R}^{k+1}$, both of which induce the law $\mathcal{L}_k(\alpha_n)$ on $Y = \mathbb{R}^{k+1}$, to get a law γ_n.

Then apply Theorem 1.1.10, this time to $(X, Y, Z) = (I^n \times \mathbb{R}^{k+1}, \mathbb{R}^{k+1}, C[0,1])$, with γ_n on $X \times Y$ and the law of $(\pi_k(y), y)$ on $Y \times Z$, where y is the Brownian bridge.

We see that there is a probability measure ζ_n on $I^n \times C[0,1]$ such that if $\mathcal{L}(V_n, Y_n) = \zeta_n$, then $\mathcal{L}(V_n)$ is uniform on I^n, $\mathcal{L}(Y_n)$ is the law of the Brownian bridge, and if we take $\alpha_n = \alpha_n(\cdot)(V_n)$, then for $n \geq n_1(\varepsilon)$ defined after (1.1.6),

$$\begin{aligned} \|\alpha_n - Y_n\|_\infty &\leq \|\alpha_n - Lib_k(\pi_k(\alpha_n))\|_\infty \\ &\quad + \|Lib_k(\pi_k(\alpha_n)) - Lib_k(\pi_k(Y_n))\|_\infty \\ &\quad + \|Lib_k(\pi_k(Y_n)) - Y_n\|_\infty \\ &\leq \varepsilon/2 + \varepsilon/6 + \varepsilon/3 \leq \varepsilon \end{aligned}$$

except on a set with probability at most $\frac{\varepsilon}{6} + \frac{\varepsilon}{6} + \frac{\varepsilon}{3} < \varepsilon$, by (1.1.11), (1.1.8), and (1.1.9), respectively.

Let $\Omega_n := I^n \times C[0,1]$, $n \geq 1$. For $r = 1, 2, \cdots$, let $n_r := n_1(1/r)$. Let N_r be an increasing sequence with $N_r \geq n_r$ for all r. For $n < N_1$, define μ_n as in (1.1.7) but with 1 in place of $\varepsilon/6$ (both times), so that it always holds: one can take μ_n as the product measure $\mathcal{L}_k(\alpha_n) \times \mathcal{L}_k(y)$. Define ζ_n on Ω_n as above, but with 1 in place of ε/m for $m = 2, 4$, or 6 in (1.1.6) and (1.1.11). For $N_r \leq n < N_{r+1}$, define μ_n and ζ_n as for $\varepsilon = 1/r$. Then $\Pr(\|\alpha_n - Y_n\|_\infty > 1/r) \leq 1/r$ for $n \geq N_r$, $r \geq 1$, and Theorem 1.1.1 is proved. □

Remarks. It would be nice to be able to say that α_n converges to the Brownian bridge y in law in some space S of functions with supremum norm. The standard definition of convergence in law, at least if S is a separable metric space, would say that $Eh(\alpha_n) \to Eh(y)$ for all bounded continuous real functions h on S (RAP, Section 9.3). Donsker (1952) stated this when continuity is assumed only at almost all values of y in $C[0,1]$. But then, h could be nonmeasurable away from the support of y, and $Eh(\alpha_n)$ is not necessarily defined. Perhaps more surprisingly, $Eh(\alpha_n)$ may not be defined even if h is bounded and continuous everywhere. Consider for example $n = 1$. Then in the set of all possible functions $U_1 - U$, any two distinct functions are at distance 1 apart for $\|\cdot\|_\infty$. So the set and all its subsets are complete, closed, and discrete for $\|\cdot\|_\infty$. If the image of Lebesgue (uniform) measure on $[0,1]$ by the function $x \mapsto (t \mapsto 1_{\{x \geq t\}} - t)$ were defined on all Borel sets for $\|\cdot\|_\infty$ in its range, or specifically on all complete, discrete sets, it would give an extension of Lebesgue measure to a countably additive measure on all subsets of $[0,1]$. Assuming the continuum hypothesis, which is consistent with the other axioms of set theory, such an extension is not possible (RAP, Appendix C).

So in a nonseparable metric space, such as a space of empirical distribution functions with supremum norm, the Borel σ-algebra may be too large. In Chapter 3 it will be shown how to get around the lack of Borel measurability.

Here is an example relating to the Vorob'ev theorem (1.1.10). Let $X = Y = Z = \{-1, 1\}$. In $X \times Y \times Z$ let each coordinate x, y, z have the uniform distribution giving probability $1/2$ each to $-1,\ 1$. Consider the laws on the products of two of the three spaces such that $y \equiv -x,\ z \equiv -y$, and $x \equiv -z$. There exist such laws having the given marginals on X, Y and Z. But there is no law on $X \times Y \times Z$ having the given marginals on $X \times Y,\ Y \times Z$, and $Z \times X$, since the three equations together yield a contradiction.

1.2 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 the last section, 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. Here $N(\varepsilon, A, S, d)$ is sometimes called a *covering number*. It's the number of closed balls of radius ε and centers in S needed to cover A.

For any set $C \subset S$, define the *diameter* of C by

$$\operatorname{diam} C \ := \ \sup\{d(x, y)\colon\ x, y \in C\}.$$

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. Then, in a Banach space, $D(2\varepsilon, A, d)$ is the largest number of disjoint closed balls of radius ε that can be "packed" into A and is sometimes called a "packing number."

The three quantities just defined are related by the following inequalities:

1.2.1 Theorem *For any $\varepsilon > 0$ and set A in a metric space S with metric d,*

$$\begin{aligned} D(2\varepsilon, A, d) &\le N(\varepsilon, A, d) \le N(\varepsilon, A, S, d) \\ &\le N(\varepsilon, A, A, d) \le D(\varepsilon, A, d). \end{aligned}$$

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 $\overline{B}(x, \varepsilon) := \{y : d(x, y) \le \varepsilon\}$ 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 $N(\varepsilon, A, d)$. $N(\varepsilon, A, S, d)$ 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 $N(\varepsilon, B, B, d) = 1 < N(\varepsilon, A, A, d)$ for $1 < \varepsilon < 2$. This was the reason, apparently, that Kolmogorov chose to use $N(\varepsilon, A, d)$.

In this book I 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 *$\varepsilon$-capacity* as $\log D(\varepsilon, A, d)$. Some other authors take logarithms to the base 2, by analogy with information-theoretic entropy. In this book 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 $1/\varepsilon$.

1.3 Inequalities

This section collects several inequalities bounding the probabilities that random variables, and specifically sums of independent random variables, are large. Many of these follow from a basic inequality of S. Bernstein and P. L. Chebyshev.

1.3.1 Theorem *For any real random variable X and $t \in \mathbb{R}$,*

$$\Pr\{X \geq t\} \leq \inf_{u \geq 0} e^{-tu} E e^{uX}.$$

Proof For any fixed $u \geq 0$, the indicator function of the set where $X \geq t$ satisfies $1_{\{X \geq t\}} \leq e^{u(X-t)}$, so the inequality holds for a fixed u; then take $\inf_{u \geq 0}$. □

For any independent real random variables $X_1, \cdots, X_n$, let $S_n := X_1 + \cdots + X_n$.

1.3.2 Bernstein'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 $\sigma_j^2 = \operatorname{Var}(X_j)$ and $\tau_n^2 := \operatorname{Var}(S_n) = \sigma_1^2 + \cdots + \sigma_n^2$. Then for any $K > 0$,*

$$\Pr\left\{|S_n| \geq K n^{1/2}\right\} \leq 2 \cdot \exp\left(-nK^2/\left(2\tau_n^2 + 2Mn^{1/2}K/3\right)\right). \tag{1.3.3}$$

Proof We can assume that $\tau_n^2 > 0$, since otherwise $S_n = 0$ a.s. (where a.s. means almost surely) and the inequality holds. For any $u \geq 0$ and $j = 1, \cdots, n$,

$$E \exp(uX_j) = 1 + u^2 \sigma_j^2 F_j/2 \leq \exp\left(\sigma_j^2 F_j u^2/2\right), \tag{1.3.4}$$

where $F_j := 2\sigma_j^{-2} \sum_{r=2}^{\infty} u^{r-2} E X_j^r / r!$, or $F_j = 0$ if $\sigma_j^2 = 0$. For $r \geq 2$, $|X_j|^r \leq X_j^2 M^{r-2}$ a.s., so $F_j \leq 2 \sum_{r=2}^{\infty} (Mu)^{r-2}/r! \leq \sum_{r=2}^{\infty} (Mu/3)^{r-2} = 1/(1 - Mu/3)$ for all $j = 1, \cdots, n$ if $0 < u < 3/M$.

Let $v := Kn^{1/2}$ and $u := v/(\tau_n^2 + Mv/3)$, so that $v = \tau_n^2 u/(1 - Mu/3)$. Then $0 < u < 3/M$. Thus, multiplying the factors on the right side of (1.3.4) by independence, we have

$$E \exp(uS_n) \leq \exp\left(\tau_n^2 u^2/2(1 - Mu/3)\right) = \exp(uv/2).$$

So by Theorem 1.3.1, $\Pr\{S_n \geq v\} \leq e^{-uv/2}$ and

$$\begin{aligned} e^{-uv/2} &= \exp\big(-v^2/\big(2\tau_n^2 + 2Mv/3\big)\big) \\ &= \exp\big(-nK^2/\big(2\tau_n^2 + 2MKn^{1/2}/3\big)\big). \end{aligned}$$ □

Here are some remarks on Bernstein's inequality. Note that 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 $2 \cdot \exp(-K^2/(2\sigma^2))$, 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 $M_n/n^{1/2} \to 0$. Sometimes the inequality can be applied to unbounded variables, replacing them by "truncated" ones, say replacing an unbounded f by f_M where $f_M(x) := f(x)1_{\{|f(x)| \leq M\}}$.

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.

1.3.5 Proposition (Hoeffding) *For any $t \geq 0$ and real a_j,*

$$\Pr\left\{\sum_{j=1}^n a_j s_j \geq t\right\} \leq \exp\left(-t^2 \Big/ \left(2\sum_{j=1}^n a_j^2\right)\right).$$

Proof Since $1/(2n)! \leq 2^{-n}/n!$ for $n = 0, 1, \cdots$, we have $\cosh x \equiv (e^x + e^{-x})/2 \leq \exp(x^2/2)$ for all x. Apply Theorem 1.3.1, where by calculus, $\inf_u \exp(-ut + \sum_{j=1}^n a_j^2 u^2/2)$ is attained at $u = t/\sum_{j=1}^n a_j^2$, and the result follows. □

Here are some remarks on Proposition 1.3.5. 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 $Y_1, \cdots, Y_n$,

$$\Pr\left\{\sum_{j=1}^n s_j Y_j > K \;\middle|\; Y_1, \cdots, Y_n\right\} \leq \exp\left(-K^2 \Big/ \left(2\sum_{j=1}^n Y_j^2\right)\right)$$

by 1.3.5. 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 Bernstein inequality (1.3.2) 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.

1.3.6 Theorem (Hoeffding) *Let $X_1, \cdots, X_n$ be independent variables with $0 \le X_j \le 1$ for all j. Let $\overline{X} := (X_1 + \cdots + X_n)/n$ and $\mu := E\overline{X}$. Then for $0 < t < 1 - \mu$,*

$$\begin{aligned} \Pr\left\{ \overline{X} - \mu \ge t \right\} &\le \left\{ \left(\frac{\mu}{\mu + t} \right)^{\mu + t} \left(\frac{1 - \mu}{1 - \mu - t} \right)^{1 - \mu - t} \right\}^n \\ &\le e^{-nt^2 g(\mu)} \le e^{-2nt^2}, \end{aligned}$$

where

$$\begin{aligned} g(\mu) &:= (1 - 2\mu)^{-1} \log((1 - \mu)/\mu) \qquad \text{for } 0 < \mu < 1/2, \text{ or} \\ &:= 1/(2\mu(1 - \mu)) \qquad\qquad\quad \text{for } 1/2 \le \mu \le 1. \end{aligned}$$

Remarks. For $t > 1 - \mu$, $\Pr(\overline{X} - \mu > t) \le \Pr(\overline{X} > 1) = 0$. 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.

Proof For any $v > 0$, the function $f(x) := e^{vx}$ is convex (its second derivative is positive), so any chord lies above its graph (RAP, Section 6.3). Specifically, if $0 < x < 1$, then $e^{vx} \le 1 - x + xe^v$. Taking expectations gives $E \exp(vX_j) \le 1 - \mu_j + \mu_j e^v$, where $\mu_j := EX_j$. (Note that the latter inequality becomes an equation for a Bernoulli variable X_j, taking only the values $0, 1$.) Let $S_n := X_1 + \cdots + X_n$. Then

$$\begin{aligned} \Pr\left(\overline{X} - \mu \ge t \right) &= \Pr(S_n - ES_n \ge nt) \\ &\le E \exp(v(S_n - ES_n - nt)) \\ &= e^{-vn(t+\mu)} \prod_{j=1}^n E \exp(vX_j) \\ &\le e^{-nv(t+\mu)} \prod_{j=1}^n \left(1 - \mu_j + \mu_j e^v\right). \end{aligned}$$

The following fact is rather well known (e.g., RAP (5.1.6) and p. 276):

1.3.7 Lemma *For any nonnegative real numbers $t_1, \cdots, t_n$, the geometric mean is less than or equal to the arithmetic mean, in other words*

$$(t_1 t_2 \cdots t_n)^{1/n} \le (t_1 + \cdots + t_n)/n.$$

Applying the lemma gives

$$\begin{aligned}\Pr\left\{\overline{X}-\mu>t\right\} &\le e^{-nv(t+\mu)}\left(\frac{1}{n}\sum_{j=1}^{n}1-\mu_j+\mu_j e^{v}\right)^{n}\\ &\le e^{-nv(t+\mu)}\left(1-\mu+\mu e^{v}\right)^{n}.\end{aligned}$$

To find the minimum of this for $v>0$, note that it becomes large as $v\to\infty$ since $t+\mu<1$, while setting the derivative with respect to v equal to 0 gives a single solution, where $1-\mu+\mu e^{v}=\mu e^{v}/(t+\mu)$ and $e^{v}=1+t/(\mu(1-\mu-t))$. Substituting these values into the bounds gives the first, most complicated bound in the statement of the theorem. This bound can be written as $\Pr(\overline{X}-\mu\ge t)\le\exp(-nt^2G(t,\mu))$, where

$$G(t,\mu):=\frac{\mu+t}{t^2}\log\left(\frac{\mu+t}{\mu}\right)+\left(\frac{1-\mu-t}{t^2}\right)\log\left(\frac{1-\mu-t}{1-\mu}\right).$$

The next step is to show that $\min_{0<t<1}G(t,\mu)=g(\mu)$ as defined in the statement of the theorem. For $0<x<1$ let

$$H(x):=\left(1-\frac{2}{x}\right)\log(1-x).$$

In $\partial G(t,\mu)/\partial t$, the terms not containing logarithms cancel, giving

$$\begin{aligned}t^2\frac{\partial G(t,\mu)}{\partial t} &= \left[1-\frac{2}{t}(1-\mu)\right]\log\left(1-\frac{t}{1-\mu}\right)\\ &\quad-\left[1-\frac{2}{t}(\mu+t)\right]\log\left(1-\frac{t}{\mu+t}\right)\\ &= H\left(\frac{t}{1-\mu}\right)-H\left(\frac{t}{\mu+t}\right).\end{aligned}$$

To see that H is increasing in x for $0<x<1$, take the Taylor series of $\log(1-x)$ and multiply by $1-\frac{2}{x}$ to get the Taylor series of H around 0, all whose coefficients are nonnegative. Only the first order term is 0. Thus $\partial G/\partial t>0$ if and only if $t/(1-\mu)>t/(\mu+t)$, or equivalently $t>1-2\mu$. So if $\mu<1/2$, then $G(t,\mu)$, for fixed μ, has a minimum with respect to $t>0$ at $t=1-2\mu$, giving $g(\mu)$ for that case as stated. Or if $\mu\ge1/2$, then $G(t,\mu)$ is increasing in $t>0$, with $\lim_{t\downarrow0}G(t,\mu)=g(\mu)$ as stated for that case, using the first two terms of the Taylor series around $t=0$ of each logarithm.

Now to get the final bound $\exp(-2nt^2)$, it needs to be shown that the minimum of $g(\mu)$ for $0<\mu<1$ is 2. For $\mu\ge1/2$, g is increasing (its denominator is decreasing), and $g(1/2)=2$. For $\mu<1/2$, letting $w:=1-2\mu$, we get $g(\mu)=\frac{1}{w}\log\left(\frac{1+w}{1-w}\right)$. From the Taylor series of $\log(1+w)$ and $\log(1-w)$

around $w = 0$, we see that g is increasing in w, and so decreasing in μ, and converges to 2 as $\mu \to 1/2$. Thus g has a minimum at $\mu = 1/2$, which is 2. The theorem is proved. □

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

$$B(k, n, p) := \sum_{0 \le j \le k} \binom{n}{j} p^j q^{n-j}, \quad 0 \le q := 1 - p \le 1,$$

$$E(k, n, p) := \sum_{k \le j \le n} \binom{n}{j} p^j q^{n-j}.$$

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.

1.3.8 Chernoff-Okamoto inequalities *We have*

$$E(k, n, p) \le \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k} \quad \text{if } k \ge np, \tag{1.3.9}$$

$$B(k, n, p) \le \exp\left(-(np-k)^2/(2npq)\right) \quad \text{if } k \le np \le n/2. \tag{1.3.10}$$

Proof These facts follow directly from the Hoeffding inequality 1.3.6. For (1.3.10), note that $B(k, n, p) = E(n-k, n, 1-p)$ and apply the $g(\mu)$ case with $\mu = 1 - p$. □

If in (1.3.9) we set $x := nq/(n-k) \le e^{x-1}$, it follows that

$$E(k, n, p) \le (np/k)^k e^{k-np} \quad \text{if } k \ge np. \tag{1.3.11}$$

The next inequality is for the special value $p = 1/2$.

1.3.12 Proposition *If $k \le n/2$ then $2^n B(k, n, 1/2) \le (ne/k)^k$.*

Proof By (1.3.9) and symmetry, $B(k, n, 1/2) \le (n/2)^n k^{-k} (n-k)^{k-n}$. Letting $y := n/(n-k) \le e^{y-1}$ then gives the result. □

A form of Stirling's formula with error bounds is:

1.3.13 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.

1.3.14 Lemma *For any Poisson variable z with parameter* $m \ge 1$,

$$E(z - m)^+ \ge m^{1/2}/8.$$

Proof We have $E(z - m)^+ = \sum_{k>m} e^{-m}m^k(k - m)/k!$. Let $j := [m]$, meaning j is the greatest integer with $j \le m$. Then by a telescoping sum (which is absolutely convergent), $E(z-m)^+ = e^{-m}m^{j+1}/j!$. Then by Stirling's formula with error bounds (Theorem 1.3.13),

$$\begin{aligned} E(z - m)^+ &\ge e^{-m}m^{j+1}(e/j)^j(2\pi j)^{-1/2}e^{-1/(12j)} \\ &\ge \left(m^{j+1}/j^{j+\frac{1}{2}}\right)e^{-13/12}(2\pi)^{-1/2} \ge m^{1/2}/8. \end{aligned}$$ □

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 $\|\cdot\|$. (Such spaces are defined, for example, in RAP, Section 5.2.) Let $S_j := X_1 + \cdots + X_j$ for $j = 1, \cdots, n$.

1.3.15 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*

$$P\left\{\max_{j\le n} \|S_j\| \ge 2\alpha\right\} \le P(\|S_n\| \ge \alpha)/(1 - c).$$

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:

1.3.16 P. Lévy's inequality *Given a probability space* (Ω, P) *and a countable set* Y, *let* $X_1, X_2, \cdots$ *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* $\|f\|_Y := \sup\{|f(y)| : y \in Y\}$. *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 $\{-X_j(y): y \in Y\}$ have the same joint distribution as $\{X_j(y): y \in Y\}$. Let $S_n := X_1 + \cdots + X_n$. Then for each n, and $M > 0$,

$$P\left(\max_{j\le n} \|S_j\|_Y > M\right) \le 2P\left(\|S_n\|_Y > M\right).$$

Notes. Each $\|S_j\|_Y$ is a measurable random variable because Y is countable. Lemma 9.1.9 treats uncountable Y. 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 preceding lemma applies to some nonseparable Banach spaces: the space of all bounded functions on an infinite Y with supremum norm is itself nonseparable.

Proof Let $M_k(\omega) := \max_{j\le k} \|S_j\|_Y$. Let C_k be the disjoint events $\{M_{k-1} \le M < M_k\}$, $k = 1, 2, \cdots$, where we set $M_0 := 0$. Then for $1 \le m \le n$, $2\|S_m\|_Y \le \|S_n\|_Y + \|2S_m - S_n\|_Y$. So if $\|S_m\|_Y > M$, then $\|S_n\|_Y > M$ or $\|2S_m - S_n\|_Y > M$ or both. The transformation which interchanges X_j and $-X_j$ just for $m < j \le n$ preserves probabilities, by symmetry and independence. Then S_n is interchanged with $2S_m - S_n$, while X_j are preserved for $j \le m$. So $P(C_m \cap \{\|S_n\|_Y > M\}) = P(C_m \cap \{\|2S_m - S_n\|_Y > M\}) \ge P(C_m)/2$, and $P(M_n > M) = \sum_{m=1}^n P(C_m) \le 2P(\|S_n\|_Y > M)$. □

Problems

1. Find the covariance matrix on $\{0, 1/4, 1/2, 3/4, 1\}$ of
   (a) the Brownian bridge process y_t;
   (b) $U_4 - U$. *Hint:* Recall that $n^{1/2}(U_n - U)$ has the same covariances as y_t.
2. Let $0 < t < u < 1$. Let α_n be the empirical process for the uniform distribution on $[0, 1]$.
   (a) Show that the distribution of $\alpha_n(t)$ is concentrated in some finite set A_t.
   (b) Let $f(t, y, u) := E(\alpha_n(u)|\alpha_n(t) = y)$. Show that for any y in A_t, $(u, f(t, y, u))$ is on the straight line segment joining (t, y) to $(1, 0)$.

3. Let (S, d) be a complete separable metric space. Let μ be a law on $S \times S$ and let $\delta > 0$ satisfy

$$\mu(\{(x, y)\colon\ d(x, y) > 2\delta\}) \ \leq\ 3\delta.$$

Let $\pi_2(x, y) := y$ and $P := \mu \circ \pi_2^{-1}$. Let Q be a law on S such that $\rho(P, Q) < \delta$ where ρ is Prokhorov's metric. On $S \times S \times S$ let $\pi_{12}(x, y, z) := (x, y)$ and $\pi_3(x, y, z) := z$. Show that there exists a law α on $S \times S \times S$ such that $\alpha \circ \pi_{12}^{-1} = \mu$, $\alpha \circ \pi_3^{-1} = Q$, and

$$\alpha(\{(x, y, z)\colon\ d(x, z) > 3\delta\}) \ \leq\ 4\delta.$$

Hint: Use Strassen's theorem, which implies that for some law ν on $S \times S$, if $\mathcal{L}(Y, Z) = \nu$, then $\mathcal{L}(Y) = P$, $\mathcal{L}(Z) = Q$, and $\nu(\{d(Y, Z) > \delta\}) < \delta$. Then the Vorob'ev-Berkes-Philipp theorem applies.

4. Let $A = B = C = \{0, 1\}$. On $A \times B$, let

$$\mu\ :=\ \left(\delta_{(0,0)} + 2\delta_{(1,0)} + 5\delta_{(0,1)} + \delta_{(1,1)}\right)/9.$$

On $B \times C$, let $\nu := [\delta_{(0,0)} + \delta_{(1,0)} + \delta_{(0,1)} + 3\delta_{(1,1)}]/6$. Find a law γ on $A \times B \times C$ such that if $\gamma = \mathcal{L}(X, Y, Z)$, then $\mathcal{L}(X, Y) = \mu$ and $\mathcal{L}(Y, Z) = \nu$.

5. Let $I = [0, 1]$ with usual metric d. For $\varepsilon > 0$, evaluate $D(\varepsilon, I, d)$, $N(\varepsilon, I, d)$, and $N(\varepsilon, I, I, d)$. *Hint:* The ceiling function $\lceil x \rceil$ is defined as the least integer $\geq x$. Answers can be written in terms of $\lceil \cdot \rceil$.

6. For a Poisson variable X with parameter $\lambda > 0$, that is, $P(X = k) = e^{-\lambda}\lambda^k/k!$ for $k = 0, 1, 2, \cdots$, evaluate the moment-generating function Ee^{tX} for all t. For $M > \lambda$, find the bound for $\Pr(X \geq M)$ given by the moment-generating function inequality (1.3.1).

Notes

Notes to Section 1.1. The contributions of Kolmogorov (1933), Doob (1949), and Donsker (1952) were mentioned in the text.

When it was realized that the formulation by Donsker (1952) was incorrect because of measurability problems, Skorokhod (1956) – see also Kolmogorov (1956) – defined a separable metric d on the space $D[0, 1]$ of right-continuous functions with left limits on $[0, 1]$, such that convergence for d to a continuous function is equivalent to convergence for the sup norm, and the empirical process α_n converges in law in $D[0, 1]$ to the Brownian bridge; see, for example, Billingsley (1968, Chapter 3).

The formulation of Theorem 1.1.1 avoids the need for the Skorokhod topology and deals with measurability. I don't know whether Theorem 1.1.1 has

been stated before explicitly, although it is within the ken of researchers on empirical processes.

In Theorem 1.1.10, the assumption that X, Y and Z are Polish can be weakened: they could instead be any Borel sets in Polish spaces (RAP, Section 13.1). Still more generally, since the proof of Theorem 10.2.2 in RAP depends just on tightness, it is enough to assume that X, Y and Z are universally measurable subsets of their completions, in other words, measurable for the completion of any probability measure on the Borel sets (RAP, Section 11.5). Shortt (1983) treats universally measurable spaces and considers just what hypotheses on X, Y and Z are necessary.

Vorob'ev (1962) proved Theorem 1.1.10 for finite sets. Then Berkes and Philipp (1977, Lemma A1) proved it for separable Banach spaces. Their proof carries over to the present case. Vorob'ev (1962) treated more complicated families of joint distributions on finite sets, as did Shortt (1984) for more general measurable spaces.

Notes to Section 1.2. Apparently the first publication on ε-entropy was the announcement by Kolmogorov (1955). Theorem 1.2.1, and the definitions of all the quantities in it, are given in the longer exposition by Kolmogorov and Tikhomirov (1959, Section 1, Theorem IV).

Lorentz (1966) proposed the name "metric entropy" rather than "ε-entropy," urging that functions should not be named after their arguments, as functions of a complex variable z are not called "z-functions." The name "metric entropy" emphasizes the purely metric nature of the concept. Actually, "ε-entropy" has been used for different quantities. Posner, Rodemich, and Rumsey (1967, 1969) define an ε, δ entropy, for a metric space S with a probability measure P defined on it, in terms of a decomposition of S into sets of diameter at most ε and one set of probability at most δ. Also, Posner et al. define ε-entropy as the infimum of entropies $-\sum_i P(U_i)\log(P(U_i))$ where the U_i have diameters at most ε. So Lorentz's term "metric entropy" seems useful and will be adopted here.

Notes to Section 1.3. Sergei Bernstein (1927, pp. 159–165) published his inequality. The proof given is based on Bennett (1962, p. 34) with some incorrect, but unnecessary steps (his (3), (4), ...) removed as suggested by Giné (1974). For related and stronger inequalities under weaker conditions, such as unbounded variables, see also Bernstein (1924, 1927), Hoeffding (1963), and Uspensky (1937, p. 205).

Hoeffding (1963, Theorem 2) implies Proposition 1.3.5. Chernoff (1952, (5.11)) proved (1.3.9). Okamoto (1958, Lemma 2(b′)) proved (1.3.10). Inequality (1.3.11) appeared in Dudley (1978, Lemma 2.7) and Lemma 1.3.12 in

Dudley (1982, Lemma 3.3). On Ottaviani's inequality (1.3.15) for real-valued functions, see (9.7.2) and the notes to Section 9.7 in RAP. The P. Lévy inequality (1.3.16) is given for Banach-valued random variables in Kahane (1985, Section 2.3). For the case of real-valued random variables, it was known much earlier; see the notes to Section 12.3 in RAP.

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.

Berkes, István, and Philipp, Walter (1977). An almost sure invariance principle for the empirical distribution function of mixing random variables. *Z. Wahrscheinlichkeitsth. verw. Gebiete* **41**, 115–137.

Bernstein, 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 Sochineniĭ [Collected Works], Tom IV, Teoriya Veroiatnostei, Matematicheskaya Statistika*, Nauka, Moscow, 1964, pp. 71–79.

*Bernstein, Sergei N. (1927). *Teoriya Veroiatnostei* (in Russian), 2d ed. Moscow, 1934.

Billingsley, Patrick (1968). *Convergence of Probability Measures.* Wiley, New York.

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.

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

Doob, J. L. (1949). Heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.* **20**, 393–403.

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, Cambridge.

*Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giorn. Ist. Ital. Attuari* **4**, 83–91.

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. (1956). On Skorokhod convergence. *Theory Probab. Appl.* **1**, 215–222.

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.

Okamoto, Masashi (1958). Some inequalities relating to the partial sum of binomial probabilities. *Ann. Inst. Statist. Math.* **10**, 29–35.

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.

Shortt, Rae M. (1983). Universally measurable spaces: an invariance theorem and diverse characterizations. *Fund. Math.* **121**, 169–176.

Shortt, Rae M. (1984). Combinatorial methods in the study of marginal problems over separable spaces. *J. Math. Anal. Appl.* **97**, 462–479.

Skorokhod, A. V. (1956). Limit theorems for stochastic processes. *Theory Probab. Appl.* **1**, 261–290.

Uspensky, J. V. (1937). *Introduction to Mathematical Probability.* McGraw-Hill, New York.

Vorob'ev, N. N. (1962). Consistent families of measures and their extensions. *Theory Probab. Appl.* **7**, 147–163 (English), 153–169 (Russian).