UNIFORM CENTRAL LIMIT THEOREMS

RICHARD M. DUDLEY
Massachusetts Institute of Technology
## Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>Introduction: Donsker’s Theorem, Metric Entropy, and Inequalities</td>
<td>xiii</td>
</tr>
<tr>
<td>1</td>
<td>Empirical processes: the classical case</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Metric entropy and capacity</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>Inequalities</td>
<td>10</td>
</tr>
<tr>
<td>1.3</td>
<td>Problems</td>
<td>12</td>
</tr>
<tr>
<td>1.4</td>
<td>Notes</td>
<td>18</td>
</tr>
<tr>
<td>1.5</td>
<td>References</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>Gaussian Measures and Processes; Sample Continuity</td>
<td>23</td>
</tr>
<tr>
<td>2.1</td>
<td>Some definitions</td>
<td>23</td>
</tr>
<tr>
<td>2.2</td>
<td>Gaussian vectors are probably not very large</td>
<td>24</td>
</tr>
<tr>
<td>2.3</td>
<td>Inequalities and comparisons for Gaussian distributions</td>
<td>31</td>
</tr>
<tr>
<td>2.4</td>
<td>Gaussian measures and convexity</td>
<td>40</td>
</tr>
<tr>
<td>2.5</td>
<td>The isonormal process: sample boundedness and continuity</td>
<td>43</td>
</tr>
<tr>
<td>2.6</td>
<td>A metric entropy sufficient condition for sample continuity</td>
<td>52</td>
</tr>
<tr>
<td>2.7</td>
<td>Majorizing measures</td>
<td>59</td>
</tr>
<tr>
<td>2.8</td>
<td>Sample continuity and compactness</td>
<td>74</td>
</tr>
<tr>
<td>2.9</td>
<td>Volumes, mixed volumes, and ellipsoids</td>
<td>78</td>
</tr>
<tr>
<td>2.10</td>
<td>Convex hulls of sequences</td>
<td>82</td>
</tr>
<tr>
<td>3</td>
<td>Foundations of Uniform Central Limit Theorems:</td>
<td>91</td>
</tr>
<tr>
<td>3.1</td>
<td>Definitions: convergence in law</td>
<td>91</td>
</tr>
<tr>
<td>3.2</td>
<td>Measurable cover functions</td>
<td>95</td>
</tr>
</tbody>
</table>
Contents

3.3 Almost uniform convergence and convergence in outer probability 100
3.4 Perfect functions 103
3.5 Almost surely convergent realizations 106
3.6 Conditions equivalent to convergence in law 111
3.7 Asymptotic equicontinuity and Donsker classes 117
3.8 Unions of Donsker classes 121
3.9 Sequences of sets and functions 122

Problems 127
Notes 130
References 132

4 Vapnik-Červonenkis Combinatorics 134
4.1 Vapnik-Červonenkis classes 134
4.2 Generating Vapnik-Červonenkis classes 138

*4.3 Maximal classes 142
*4.4 Classes of index 1 145
*4.5 Combining VC classes 152
4.6 Probability laws and independence 156
4.7 Vapnik-Červonenkis properties of classes of functions 159
4.8 Classes of functions and dual density 161

**4.9 Further facts about VC classes 165
Problems 166
Notes 167
References 168

5 Measurability 170
5.1 Sufficiency 171
5.2 Admissibility 179

5.3 Suslin properties, selection, and a counterexample 185

Problems 191
Notes 193
References 194

6 Limit Theorems for Vapnik-Červonenkis and Related Classes 196
6.1 Koltchinskii-Pollard entropy and Glivenko-Cantelli theorems 196
6.2 Vapnik-Červonenkis-Steele laws of large numbers 203
6.3 Pollard’s central limit theorem 208
6.4 Necessary conditions for limit theorems 215

**6.5 Inequalities for empirical processes 220
**6.6 Glivenko-Cantelli properties and random entropy 223
**6.7 Classification problems and learning theory 226
Problems 227
## Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notes</td>
<td>228</td>
</tr>
<tr>
<td>References</td>
<td>230</td>
</tr>
<tr>
<td><strong>7 Metric Entropy, with Inclusion and Bracketing</strong></td>
<td>234</td>
</tr>
<tr>
<td>7.1 Definitions and the Blum-DeHardt law of large numbers</td>
<td>234</td>
</tr>
<tr>
<td>7.2 Central limit theorems with bracketing</td>
<td>238</td>
</tr>
<tr>
<td>7.3 The power set of a countable set: the Borisov-Durst theorem</td>
<td>244</td>
</tr>
<tr>
<td><strong>7.4 Bracketing and majorizing measures</strong></td>
<td>246</td>
</tr>
<tr>
<td>Problems</td>
<td>247</td>
</tr>
<tr>
<td>Notes</td>
<td>248</td>
</tr>
<tr>
<td>References</td>
<td>248</td>
</tr>
<tr>
<td><strong>8 Approximation of Functions and Sets</strong></td>
<td>250</td>
</tr>
<tr>
<td>8.1 Introduction: the Hausdorff metric</td>
<td>250</td>
</tr>
<tr>
<td>8.2 Spaces of differentiable functions and sets with differentiable boundaries</td>
<td>252</td>
</tr>
<tr>
<td>8.3 Lower layers</td>
<td>264</td>
</tr>
<tr>
<td>8.4 Metric entropy of classes of convex sets</td>
<td>269</td>
</tr>
<tr>
<td>Problems</td>
<td>281</td>
</tr>
<tr>
<td>Notes</td>
<td>282</td>
</tr>
<tr>
<td>References</td>
<td>283</td>
</tr>
<tr>
<td><strong>9 Sums in General Banach Spaces and Invariance Principles</strong></td>
<td>285</td>
</tr>
<tr>
<td>9.1 Independent random elements and partial sums</td>
<td>286</td>
</tr>
<tr>
<td>9.2 A CLT implies measurability in separable normed spaces</td>
<td>291</td>
</tr>
<tr>
<td>9.3 A finite-dimensional invariance principle</td>
<td>293</td>
</tr>
<tr>
<td>9.4 Invariance principles for empirical processes</td>
<td>301</td>
</tr>
<tr>
<td><strong>9.5 Log log laws and speeds of convergence</strong></td>
<td>306</td>
</tr>
<tr>
<td>Problems</td>
<td>309</td>
</tr>
<tr>
<td>Notes</td>
<td>310</td>
</tr>
<tr>
<td>References</td>
<td>311</td>
</tr>
<tr>
<td><strong>10 Universal and Uniform Central Limit Theorems</strong></td>
<td>314</td>
</tr>
<tr>
<td>10.1 Universal Donsker classes</td>
<td>314</td>
</tr>
<tr>
<td>10.2 Metric entropy of convex hulls in Hilbert space</td>
<td>322</td>
</tr>
<tr>
<td>**10.3 Uniform Donsker classes</td>
<td>328</td>
</tr>
<tr>
<td>Problems</td>
<td>330</td>
</tr>
<tr>
<td>Notes</td>
<td>330</td>
</tr>
<tr>
<td>References</td>
<td>330</td>
</tr>
<tr>
<td><strong>11 The Two-Sample Case, the Bootstrap, and Confidence Sets</strong></td>
<td>332</td>
</tr>
<tr>
<td>11.1 The two-sample case</td>
<td>332</td>
</tr>
<tr>
<td>11.2 A bootstrap central limit theorem in probability</td>
<td>335</td>
</tr>
<tr>
<td>11.3 Other aspects of the bootstrap</td>
<td>357</td>
</tr>
</tbody>
</table>
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 $X_1, X_2, \ldots$ be i.i.d. (independent, identically distributed) random variables with distribution $P$. For each $n = 1, 2, \ldots$ 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 $X_1, \ldots, X_n$ 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 $\omega$ in some probability space $\Omega$, 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 $\omega$. So the empirical process $\alpha_n$ 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, $U(t) = t$, $0 \leq t \leq 1$. 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 $y_t$, $0 \leq t \leq 1$ (RAP, before Theorem 12.1.5). Recall that $y_t$ can be written in terms of a Wiener process (Brownian motion) $x_t$, namely $y_t = x_t - tx_1$, $0 \leq t \leq 1$. 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 $\omega$, the function $t \mapsto y_t(\omega)$ is continuous on $[0, 1]$ (RAP, Theorem 12.1.5).
1.1 Empirical processes: the classical case

Donsker in 1952 proved that the convergence in law of \( \alpha_n \) 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

\[
\| f \|_\infty := \sup \{|f(t)| : 0 \leq t \leq 1\}.
\]

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

1.1.1 Theorem  
For \( n = 1, 2, \ldots \), there exist probability spaces \( \Omega_n \) such that:

(a) On \( \Omega_n \), there exist \( n \) i.i.d. random variables \( X_1, \ldots, X_n \) with uniform distribution in \([0, 1]\). Let \( \alpha_n \) be the \( n \)th empirical process based on these \( X_i \);

(b) On \( \Omega_n \) a sample-continuous Brownian bridge process \( Y_n \colon (t, \omega) \mapsto Y_n(t, \omega) \) 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 \( \alpha_n \) 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 \( X_1, X_2, \ldots \)
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 \( \alpha_n \) and \( Y_n \) on one probability space and make \( Y_n \) all
equal some \( Y \), although here the joint distributions of \( \alpha_n \) for different \( n \) will
be different from their original ones. Then \( \alpha_n \) 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,

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

It will first be shown that both processes \( \alpha_n \) and \( y_{ij} \), for large enough \( n \) and \( k \), can be well approximated by step functions and then by piecewise-linear
interpolation of their values on \( L_k \).
4 Introduction: Donsker’s Theorem, Metric Entropy, and Inequalities

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

\[(1.1.2)\quad 4k \cdot \exp \left(-k\varepsilon^2 / 648\right) < \varepsilon / 6.\]

Let $I_{jk} := [j/k, (j+1)/k]$, $j = 0, \ldots, k-1$. 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

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

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),

\[(1.1.3)\quad \Pr(|y_t - y_{j/k}| > \varepsilon / 6 \text{ for some } j = 0, \ldots, k-1 \text{ and some } t \in I_{jk}) < \varepsilon / 3.\]

Next, we need a similar bound for $n$ 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 $\alpha_n$ in place of $U_n$.

(c) The distribution of $\sup_{j=0}^{k-1} |\alpha_n(t+j/k) - \alpha_n(j/k)| : 0 \leq t \leq 1/k$ 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. \qed
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 \( n_1, \ldots, n_r \) of sample size \( n = n_1 + \cdots + n_r \) in \( r \) bins with probabilities \( p_1, \ldots, p_r \). Then for each \( j \), the conditional distribution of \( n_{j+1} \) given \( n_1, \ldots, n_j \) 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 \( p_{j+1}/(p_j + \cdots + p_n) \) 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 \( U_n(t_1), \ldots, U_n(t_j), U_n(t) \) 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 \( \alpha_n(t) = n^{1/2}(m/n - t) \), and \( \alpha_n \) also has the Markov property. So the conditional distribution of \( \alpha_n(u) \) given \( m = nU_n(t) \) has mean \( \mu_m := n^{1/2} \left\{ \frac{m}{n} \left( 1 - \frac{u - t}{1 - t} \right) + \frac{u - t}{1 - t} \right\} = n^{1/2} \left( \left( \frac{m}{n} - t \right) \left( \frac{1 - u}{1 - t} \right) \right) \) and variance \( \frac{(n - m)(u - t)(1 - u)}{n(1 - t)^2} \leq \frac{u - t}{1 - t} \leq u. \)

So, by Chebyshev’s inequality,

\[
\Pr \left\{ \left| \alpha_n(u) - \mu_m \right| \geq 2u^{1/2} \mid m \right\} \leq 1/4.
\]

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

\[
\Pr \left\{ \alpha_n(u) > \frac{\delta}{2} - 2u^{1/2} \mid \alpha_n(t) = \gamma \right\} \geq 3/4.
\]

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

\[
\Pr(\alpha_n(u) > \delta/4 \mid \alpha_n(t) = \gamma) \geq 3/4.
\]

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

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

Now take a positive integer \( r \) and let \( \tau \) 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
Introduction: Donsker’s Theorem, Metric Entropy, and Inequalities

disjoint sets $A_{rj}$ for $j = 1, \ldots, r$. For each such $j$, by the Markov property,

\[
\Pr(\alpha_n(1/k) > \epsilon/16 \mid A_{rj}) \geq 3/4.
\]

Thus

\[
\Pr(\alpha_n(1/k) > \epsilon/16 \mid A_r) \geq 3/4.
\]

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

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

Likewise,

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

Thus by Lemma 1.1.4(c),

\[
(1.1.5) \quad \Pr(|\alpha_n(t) - \alpha_n(j/k)| > \epsilon/4 \text{ for some } t \in I_{jk}
\]

and $j = 0, 1, \ldots, k-1 \leq (4k/3) \Pr(|\alpha_n(1/k)| > \epsilon/16).

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

\[
12.1.6(b),
\]

\[
\Pr(|\alpha_n(1/k)| > \epsilon/16) \to \Pr(|\gamma_{1/k}| > \epsilon/16) \leq 2 \cdot \exp \left( -k\epsilon^2/512 \right).
\]

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

\[
\Pr(|\alpha_n(1/k)| > \epsilon/16) < 3 \cdot \exp \left( -k\epsilon^2/512 \right).
\]

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

\[
(1.1.6) \quad \Pr(|\alpha_n(t) - \alpha_n(j/k)| > \epsilon/4
\]

for some $j = 0, \ldots, k-1$ and $t \in I_{jk} \leq \epsilon/6$.

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 $\{\gamma_{i/k}\}_{i=0}^{k}$, 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 $\rho$ (RAP, Theorem 11.3.3), for $n$ large enough, say $n \geq n_1(\epsilon) \geq n_0(\epsilon)$, we have

\[
\rho(\mathcal{L}_k(\alpha_n), \mathcal{L}_k(y)) < \epsilon/6. \text{ Then by Strassen’s theorem (RAP, Corollary 11.6.4),}
\]

there is a probability measure $\mu_n$ on $\mathbb{R}^{k+1} \times \mathbb{R}^{k+1}$ such that for $(X, Y)$ with $\mathcal{L}(X, Y) = \mu_n$, we have

\[
\mathcal{L}(X) = \mathcal{L}_k(\alpha_n), \quad \mathcal{L}(Y) = \mathcal{L}_k(y),
\]

and $\mu_n(x, y) : |x - y|_\infty > \epsilon/6 \leq \epsilon/6$.

(RAP, Section 9.2).
1.1 Empirical processes: the classical case

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

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

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

$$\mu_\varepsilon(\|\text{Lib}_k(x) - \text{Lib}_k(y)\|_\infty > \varepsilon/6) \leq \varepsilon/6. \quad (1.1.8)$$

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 $\pi_k(\text{Lib}_k(x)) = x$ for all $x \in \mathbb{R}^{k+1}$.

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

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

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

$$\Pr\{\|y - \text{Lib}_k(\pi_k(y))\|_\infty > \varepsilon/3\} < \varepsilon/3, \quad (1.1.9)$$

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

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

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

is measurable from $I^n$ into $\mathbb{R}^{k+1}$ and has distribution $L_k(\alpha_n)$ on $\mathbb{R}^{k+1}$. Also, $x \mapsto \text{Lib}_k(A_{\alpha_n}(x))$ 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 $\sigma$-algebras. Let $\alpha$ be a law on $X \times Y$ and let $\beta$ be a law on $Y \times Z$. 

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

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 \( \alpha \) and \( \beta \) 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 \( \gamma \) 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 \( \alpha_y \) for \( \alpha \) on \( X \) given \( y \in Y \), so that for each \( y \in Y \), \( \alpha_y \) is a probability measure on \( X \), with \( \int f(x) \, d\alpha_y(x) \leq ||f||_\infty \). Let \( x \) and \( z \) be conditionally independent given \( y \). In other words, define a set function \( \gamma \) on \( X \times Y \times Z \) by

\[
\gamma(C) = \int \int \int 1_C(x, y, z) \, d\alpha_y(x) \, d\beta_y(z) \, d\eta(y).
\]

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

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 \( \omega \) 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

\[
(1.1.11) \quad \Pr(\|\alpha_n - Lib_k(A_k(x))\|_\infty > \varepsilon/2) \leq \varepsilon/6.
\]

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 \( \mu_n \) 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 \( \gamma_n \).
1.1 Empirical processes: the classical case

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 \(\gamma_n\) 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 \(\zeta_n\) on \(I^n \times C\{0, 1\}\) such that if \(L(V_n, Y_n) = \zeta_n\), then \(L(V_n)\) is uniform on \(I^n\), \(L(Y_n)\) 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),

\[
\|\alpha_n - Y_n\|_\infty \leq \|\alpha_n - Lib_k(\pi_k(\alpha_n))\|_\infty \\
+ \|Lib_k(\pi_k(\alpha_n)) - Lib_k(\pi_k(Y_n))\|_\infty \\
+ \|Lib_k(\pi_k(Y_n)) - Y_n\|_\infty \\
\leq \varepsilon/2 + \varepsilon/6 + \varepsilon/3 \leq \varepsilon
\]

except on a set with probability at most \(\varepsilon/6 + \varepsilon/7 < \varepsilon\), 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 \(r = 1, 2, \ldots\), let \(n_r := n_1(1/r)\). Let \(N_r\) be an increasing sequence with \(n_r \geq n_r\) for all \(r\). For \(n < N_r\), define \(\mu_n\) as in (1.1.7) but with \(1\) in place of \(\varepsilon/6\) (both times), so that it always holds: one can take \(\mu_n\) as the product measure \(L_k(\alpha_n) \times L_k(\gamma)\). Define \(\zeta_n\) on \(\Omega_n\) as above, but with \(1\) in place of \(\varepsilon/m\) for \(m = 2, 4, 6\) in (1.1.6) and (1.1.11). For \(N_r \leq n < N_{r+1}\), define \(\mu_n\) and \(\zeta_n\) 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 \(\alpha_n\) 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 \( \sigma \)-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 \( y = -x, z = -y, \) and \( x = -z \). 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 \( C \) of sets is the class of half-lines \((-\infty, x], x \in \mathbb{R}, \) so that \( C \) 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 \( \varepsilon \)-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 \( \varepsilon \)-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 \( \varepsilon \) and centers in \( S \) needed to cover \( A \).

For any set \( C \subset S \), define the diameter of \( C \) by

\[
\text{diam } C := \sup\{d(x, y) : x, y \in C\}.
\]

Let \( N(\varepsilon, C, d) \) be the smallest \( n \) such that \( C \) is the union of \( n \) sets of diameter at most \( 2\varepsilon \). 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 \( \varepsilon \) 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 $d$,

$$D(2\varepsilon, A, d) \leq N(\varepsilon, A, d) \leq N(\varepsilon, A, S, d) \leq N(\varepsilon, A, A, d) \leq D(\varepsilon, A, d).$$

Proof The first inequality holds since a set of diameter $2\varepsilon$ can contain at most one of a set of points more than $2\varepsilon$ apart. The next holds because any ball $B(x, \varepsilon) := \{y : d(x, y) \leq \varepsilon\}$ is a set of diameter at most $2\varepsilon$. 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 $\varepsilon$ apart, with maximal cardinality, must be an $\varepsilon$-net, since otherwise there would be a point more than $\varepsilon$ 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 $\infty$ unless $S$ is a finite set, they have the same asymptotic behavior up to a factor of $2$ in $\varepsilon$. 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 $\varepsilon$ apart or forming an $\varepsilon$-net, are often useful, as opposed to the sets in the definition of $N(\varepsilon, A, A, d)$. $N(\varepsilon, A, S, d)$ 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 $\varepsilon$ 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 $N(\varepsilon, A, A, d)$.

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 $\varepsilon$ apart and at the same time form an $\varepsilon$-net.

Now, the $\varepsilon$-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 $\varepsilon$-entropy, and likewise the $\varepsilon$-capacity, have
the simpler order $e^{-r}$. 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 $t \in \mathbb{R}$,

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

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

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

### 1.3.2 Bernstein’s inequality

Let $X_1, X_2, \ldots, X_n$ be independent real random variables with mean 0. Let $0 < M < \infty$ and suppose that $|X_j| \leq M$ almost surely for $j = 1, \ldots, n$. Let $\sigma_j^2 = \text{Var}(X_j)$ and $\tau_n^2 := \text{Var}(S_n) = \sigma_1^2 + \cdots + \sigma_n^2$. Then for any $K > 0$,

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

**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 $j = 1, \ldots, n$,

$$E \exp(uX_j) = 1 + u^2 \sigma_j^2 F_j/2 \leq \exp(\sigma_j^2 F_j u^2/2),$$

where $F_j := 2\sigma_j^{-2} \sum_{r=2}^{\infty} u^{r-2}EX_r/r!$, 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 $j = 1, \ldots, n$ if $0 < u < 3/M$.

Let $v := Kn^{1/2}$ and $u := v/(\tau_n^2 + Mu/3)$, 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

$$E \exp(uS_n) \leq \exp(\tau_n^2 u^2/2(1 - Mu/3)) = \exp(uv/2).$$
So by Theorem 1.3.1, \( \Pr\{S_n \geq v\} \leq e^{-uv/2} \) and
\[
e^{-uv/2} = \exp\left( -v^2/(2\tau_n^2 + 2Mv/3) \right) = \exp\left( -nK^2/(2\tau_n^2 + 2MKn^{1/2}/3) \right).
\]

Here are some remarks on Bernstein’s inequality. Note that for fixed \( K \) and \( M \), if \( X_i \) are i.i.d. with variance \( \sigma^2 \), 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 \( f_M(x) := f(x) \mathbb{1}_{|f(x)| \leq M} \).

Next, let \( s_1, s_2, \ldots, \) 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 \( a_j \),
\[
\Pr\left\{ \sum_{j=1}^{n} a_js_j \geq t \right\} \leq \exp\left( -t^2 \left/ \left( 2 \sum_{j=1}^{n} a_j^2 \right) \right. \right).
\]

Proof Since \( 1/(2n)! \leq 2^{-n}/n! \) for \( n = 0, 1, \ldots \), we have \( \cosh x = (e^x + e^{-x})/2 \leq \exp(x^2/2) \) 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 \( Y_1, Y_2, \ldots, \) be independent variables which are symmetric, in other words \( Y_j \) has the same distribution as \(-Y_j\) for all \( j \). Let \( s_j \) be Rademacher variables independent of each other and of all the \( Y_j \). Then the sequence \( \{s_jY_j\}_{j \geq 1} \) 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 \( Y_1, \ldots, Y_n \),
\[
\Pr\left\{ \sum_{j=1}^{n} s_jY_j > K \ \middle| \ Y_1, \ldots, Y_n \right\} \leq \exp\left( -K^2 \left/ \left( 2 \sum_{j=1}^{n} Y_j^2 \right) \right. \right).
\]

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.
Introduction: Donsker’s Theorem, Metric Entropy, and Inequalities

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 $X_1, \ldots, X_n$ 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 $0 < t < 1 - \mu$,

$$
\Pr \left\{ \overline{X} - \mu \geq t \right\} \leq \left\{ \left( \frac{\mu}{\mu + t} \right)^{\mu + t} \left( \frac{1 - \mu}{1 - \mu - t} \right)^{1 - \mu - t} \right\}^n
$$

\leq e^{-n^2 g(\mu)} \leq e^{-2nt^2},

where

$$
g(\mu) := (1 - 2\mu)^{-1} \log((1 - \mu)/\mu) \quad \text{for } 0 < \mu < 1/2, \text{ or}
$$

$$
:= 1/(2\mu(1 - \mu)) \quad \text{for } 1/2 \leq \mu \leq 1.
$$

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.

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 $e^{vx} \leq 1 - x + xe^v$. Taking expectations gives $E \exp(vX_j) \leq 1 - \mu_j + \mu_j e^v$, where $\mu_j := E X_j$. (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

$$
\Pr \left\{ \overline{X} - \mu \geq t \right\} = \Pr(S_n - ES_n \geq nt)
$$

\leq E \exp(v(S_n - ES_n - nt))

= e^{-nv(t+\mu) + \sum_{j=1}^n E \exp(vX_j)}

\leq e^{-nv(t+\mu) + \sum_{j=1}^n (1 - \mu_j + \mu_j e^v)}.
$$

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 $t_1, \ldots, t_n$, the geometric mean is less than or equal to the arithmetic mean, in other words

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

Applying the lemma gives

\[
\Pr \{ \overline{X} - \mu > t \} \leq e^{-nv(t+\mu)} \left( \frac{1}{n} \sum_{j=1}^{n} 1 - \mu_j + \mu e^v \right)^n
\]

\[
\leq e^{-nv(t+\mu)} (1 - \mu + \mu e^v)^n.
\]

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 \( \Pr(\overline{X} - \mu \geq t) \leq \exp(-nt^2G(t, \mu)) \), 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

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

In \( \frac{\partial G(t, \mu)}{\partial t} \), the terms not containing logarithms cancel, giving

\[
t^2 \frac{\partial G(t, \mu)}{\partial t} = \left[ 1 - \frac{2}{t} (1 - \mu) \right] \log\left( 1 - \frac{t}{1 - \mu} \right) - \left[ 1 - \frac{2}{t} (\mu + t) \right] \log\left( 1 - \frac{t}{\mu + t} \right)
\]

\[ = H\left( \frac{t}{1 - \mu} \right) - H\left( \frac{t}{\mu + t} \right). \]

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 \( \mu \), has a minimum with respect to \( t > 0 \) at \( t = 1 - 2\mu \), giving \( g(\mu) \) for that case as stated. Or if \( \mu \geq 1/2 \), then \( G(t, \mu) \) is increasing in \( t > 0 \), with \( \lim_{t \to 0} G(t, \mu) = g(\mu) \) 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 \( \mu \geq 1/2 \), \( g \) 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 \( \mu \), 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

\[
B(k; n, p) := \sum_{0 \leq j \leq k} \binom{n}{j} p^j q^{n-j}, \quad 0 \leq q := 1 - p \leq 1,
\]

\[
E(k; n, p) := \sum_{k \leq j \leq n} \binom{n}{j} p^j q^{n-j}.
\]

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

\[
E(k; n, p) \leq \left( \frac{np}{k} \right)^k \left( \frac{nq}{n-k} \right)^{n-k} \quad \text{if } k \geq np.
\]

\[
B(k; n, p) \leq \exp \left( - (np - k)^2 / (2npq) \right) \quad \text{if } k \leq np \leq n/2.
\]

**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 \( x := nq/(n-k) \leq e^{x-1} \), it follows that

\[
E(k; n, p) \leq (np/k)^k e^{k-np} \quad \text{if } k \geq np.
\]

The next inequality is for the special value \( p = 1/2 \).

1.3.12 Proposition If \( k \leq n/2 \) then \( 2^n B(k; n, 1/2) \leq (ne/k)^k \).

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

A form of Stirling’s formula with error bounds is:
1.3 Inequalities

1.3.13 Theorem For \( n = 1, 2, \ldots, e^{1/(12n+1)} \leq n!(e/n)^n(2\pi n)^{-1/2} \leq e^{1/12n} \).

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 \( m \geq 1 \),

\[ E(z - m)^+ \geq m^{1/2}/8. \]

Proof We have \( E(z - m)^+ = \sum_{k>m} e^{-m}m^k(k - m)/k! \). Let \( j := \lfloor m \rfloor \), meaning \( j \) is the greatest integer with \( j \leq m \). 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),

\[
E(z - m)^+ \geq e^{-m}m^{j+1}(e/j)!/(2\pi j)^{-1/2}e^{-1/(12j)}
\]

\[
\geq (m^{j+1}/j^{j+1/2})e^{-13/12}(2\pi)^{-1/2} \geq m^{1/2}/8.
\]

In the following two facts, let \( X_1, X_2, \ldots, X_n \) 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 \( j = 1, \ldots, n \).

1.3.15 Ottaviani’s inequality If for some \( \alpha > 0 \) and \( c \) with \( 0 < c < 1 \), we have \( \Pr(\|S_j - S_j\| > \alpha) \leq c \) for all \( j = 1, \ldots, n \), then

\[ P\{ \max_{j \leq n} \|S_j\| \geq 2\alpha \} \leq P(\|S_n\| \geq \alpha)/(1 - c). \]

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 \( (\Omega, \mathcal{P}) \) and a countable set \( Y \), let \( X_1, X_2, \ldots \) be stochastic processes defined on \( \Omega \) indexed by \( Y \), in other words for each \( j \) and \( y \in Y \), \( X_j(y)(\cdot) \) is a random variable on \( \Omega \). 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 $M > 0$,

$$P \left( \max_{1 \leq j \leq n} \|S_j\|_Y > M \right) \leq 2P \left( \|S_n\|_Y > M \right).$$

**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 $M_k(\omega) := \max_{j \leq k} \|S_j\|_Y$. Let $C_k$ be the disjoint events $\{M_{k-1} \leq M < M_k\}$, $k = 1, 2, \cdots$, where we set $M_0 := 0$. Then for $1 \leq m \leq n$, $2\|S_m\|_Y \leq \|S_n\|_Y + \|2S_m - S_n\|_Y$. 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 $m < j \leq n$ preserves probabilities, by symmetry and independence. Then $S_n$ is interchanged with $2S_m - S_n$, while $X_j$ are preserved for $j \leq m$. So $P(C_m \cap \{\|S_n\|_Y > M\}) = P \left( C_m \cap \{\|2S_m - S_n\|_Y > M\} \right) \geq P(C_m)/2$, and $P(M_n > M) = \sum_{m=1}^n P(C_m) \leq 2P(\|S_n\|_Y > M)$.

---

**Problems**

1. Find the covariance matrix on $\{0, 1/4, 1/2, 3/4, 1\}$ of

   (a) the Brownian bridge process $y_t$;

   (b) $U_d - U$. *Hint:* Recall that $n^{1/2}(U_n - U)$ has the same covariances as $y_t$.

2. Let $0 < t < u < 1$. Let $\alpha_n$ 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 \(\mu\) be a law on \(S \times S\) and let \(\delta > 0\) satisfy
\[
\mu(\{(x, y) : d(x, y) > 2\delta\}) \leq 3\delta.
\]
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 \(\rho\) 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 \(\alpha\) on \(S \times S \times S\) such that \(\alpha \circ \pi_{12}^{-1} = \mu, \alpha \circ \pi_3^{-1} = Q,\) and
\[
\alpha(\{(x, y, z) : d(x, z) > 3\delta\}) \leq 4\delta.
\]

**Hint:** Use Strassen’s theorem, which implies that for some law \(\nu\) 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 \(\gamma\) 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 \(\epsilon > 0\), evaluate \(D(\epsilon, I, d), N(\epsilon, I, d),\) and \(N(\epsilon, I, I, d)\). **Hint:** The ceiling function \([x]\) is defined as the least integer \(\geq x\). Answers can be written in terms of \([\cdot]\).

6. For a Poisson variable \(X\) with parameter \(\lambda > 0\), that is, \(P(X = k) = e^{-\lambda}\lambda^k/k!\) for \(k = 0, 1, 2, \ldots\), evaluate the moment-generating function \(E e^{tX}\) 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 \(\alpha_n\) 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.


**Notes to Section 1.2.** Apparently the first publication on \( \varepsilon \)-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 Tihomirov (1959, Section 1, Theorem IV).

Lorentz (1966) proposed the name “metric entropy” rather than “\( \varepsilon \)-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, “\( \varepsilon \)-entropy” has been used for different quantities. Posner, Rodemich, and Rumsey (1967, 1969) define an \( \varepsilon, \delta \) 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 \( \varepsilon \) and one set of probability at most \( \delta \). Also, Posner et al. define \( \varepsilon \)-entropy as the infimum of entropies \( - \sum P(U_j) \log(P(U_j)) \) where the \( U_j \) have diameters at most \( \varepsilon \). 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.


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


