1

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 $F(x) := P((−\infty, x])$. 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 $\delta_x(A) = 1_A(x)$. For any given $X_1, \ldots, X_n$, $P_n$ is a probability measure called the empirical measure. Let $F_n$ be the distribution function of $P_n$. Then $F_n$ is called the empirical distribution function.

Let $U$ be the $U[0, 1]$ distribution function $U(x) = \min(1, \max(0, x))$ for all $x$, so that $U(x) = x$ for $0 \leq x \leq 1$, $U(x) = 0$ for $x < 0$ and $U(x) = 1$ for $x > 1$. To relate $F$ and $U$ we have the following.

**Proposition 1.1** For any distribution function $F$ on $\mathbb{R}$:

(a) For any $y$ with $0 < y < 1$, $F^−(y) := \inf\{x : F(x) \geq y\}$ is well-defined and finite.

(b) For any real $x$ and any $y$ with $0 < y < 1$ we have $F(x) \geq y$ if and only if $x \geq F^−(y)$.

(c) If $V$ is a random variable having $U[0, 1]$ distribution, then $F^−(V)$ has distribution function $F$.

**Proof.** For (a), recall that $F$ is nondecreasing, $F(x) \to 0$ as $x \to -\infty$, and $F(x) \to +\infty$ as $x \to +\infty$. So the set $\{x : F(x) \geq y\}$ is nonempty and bounded below, and has a finite infimum.

For (b), $F(x) \geq y$ implies $x \geq F^−(y)$ by definition of $F^−(y)$. Conversely, as $F$ is continuous from the right, $F(F^−(y)) \geq y$, and as $F$ is nondecreasing, $x \geq F^−(y)$ implies $F(x) \geq y$.

For (c), and any $x$, we have by (b)

$$\Pr(F^−(V) \leq x) = \Pr(V \leq F(x)) = U(F(x)) = F(x)$$
Recall that for any function \( f \) defined on the range of a function \( g \), the composition \( f \circ g \) is defined by \( (f \circ g)(x) := f(g(x)) \). We can then relate empirical distribution functions \( F_n \) for any distribution function \( F \) to those \( U_n \) for \( U \), as follows.

**Proposition 1.2** For any distribution function \( F \), and empirical distribution functions \( F_n \) for \( F \) and \( U_n \) for \( U \), \( U_n \circ F \) have all the properties of \( F_n \).

Proof. Let \( V_1, \ldots, V_n \) be i.i.d. \( U \), so that \( U_n(t) := \frac{1}{n} \sum_{j=1}^{n} 1_{V_j \leq t} \) for \( 0 \leq t \leq 1 \). Thus for any \( x \), by Proposition 1.1(b) and (c),

\[
U_n(F(x)) = \frac{1}{n} \sum_{j=1}^{n} 1_{V_j \leq F(x)}
\]

\[
= \frac{1}{n} \sum_{j=1}^{n} 1_{F^{-1}(V_j) \leq x}
\]

\[
= \frac{1}{n} \sum_{j=1}^{n} 1_{X_j \leq x}
\]

where \( X_j := F^{-1}(V_j) \) are i.i.d. \( (F) \). Thus \( U_n(F(x)) \) has all properties of \( F_n(x) \). \( \square \)

The developments to be described in this book began (in 1933) with the Glivenko–Cantelli theorem, a uniform law of large numbers. Probability distribution functions can converge pointwise but not uniformly: for example, as \( n \to \infty \), \( 1_{[-1/n, \infty)}(x) \to 1_{[0, +\infty)}(x) \) for all \( x \) but not uniformly.

**Theorem 1.3 (Glivenko–Cantelli)** For any distribution function \( F \), almost surely, \( \sup_x |(F_n - F)(x)| \to 0 \) as \( n \to \infty \).

Proof. By Proposition 1.2, and since \( U \circ F \equiv F \), it suffices to prove this for the \( U[0, 1] \) distribution \( U \). Given \( \varepsilon > 0 \), take a positive integer \( k \) such that \( 1/k < \varepsilon/2 \). For each \( j = 0, 1, \ldots, k \), \( U_n(j/k) \to j/k \) as \( n \to \infty \) with probability 1 by the ordinary strong law of large numbers. Take \( n_0 = n_0(\omega) \) such that for all \( n \geq n_0 \) and all \( j = 0, 1, \ldots, k \), \( |U_n(j/k) - j/k| < \varepsilon/2 \). For \( t \) outside \( [0, 1] \) we have \( U_n(t) \equiv U(t) = 0 \) or 1. For each \( t \in [0, 1] \) there is at least one \( j = 1, \ldots, k \) such that \( (j - 1)/k \leq t \leq j/k \). Then for \( n \geq n_0 \),

\[
(j - 1)/k - \varepsilon/2 < U_n((j - 1)/k) \leq U_n(t) \leq U_n(j/k) < (j + 1)/k + \varepsilon/2.
\]

It follows that \( |U_n(t) - t| < \varepsilon \), and since \( t \) was arbitrary, the theorem follows. \( \square \)
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 $F(t)(1 - F(t))$.

In what follows (as mentioned in the Note after the Preface), “RAP” will mean the author’s book Real Analysis and Probability.

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 $N(0, C_T)$ with mean 0 and covariance $C_T(s, t) = F(s)(1 - F(t))$ 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 $N(0, C_S)$, where $C = C_U$ for the uniform distribution function $U(t) = t$, $0 \leq t \leq 1$. 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 (1933b) showed that when $F$ is continuous, the supremum sup, $\alpha_n(t)$ and the supremum of absolute value, $\sup |\alpha_n(t)|$, converge in distribution to the laws of the same functionals of $y_T$. Then, these functionals of $y_T$ 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.

To work with the Brownian bridge process it will help to relate it to the well-known Brownian motion process $x_t$, defined for $t \geq 0$, also called the Wiener process. This process is such that for any any finite set $T \subset [0, +\infty)$, the joint distribution of $\{x_t\}_{t \in F}$ is $N(0, C)$ where $C(s, t) = \min(s, t)$. This process has independent increments, namely, for any $0 = t_0 < t_1 < \cdots < t_k$, the increments $x_{t_j} - x_{t_{j-1}}$ for $j = 1, \ldots, k$ are jointly independent, with $x_t - x_s$ having distribution $N(0, t - s)$ for $0 \leq s < t$. Recall that for jointly Gaussian (normal) random variables, joint independence, pairwise independence, and having covariances equal to 0 are equivalent. Having independent increments with the given distributions clearly implies that $E(x_s x_t) = \min(s, t)$ and so is equivalent to the definition of Brownian motion with that covariance.

Brownian motion can be taken to be sample continuous, i.e. such that $t \mapsto x_t(\omega)$ is continuous in $t$ for all (or almost all) $\omega$. This theorem, proved by Norbert Wiener in the 1920s, is Theorem 12.1.5 in RAP; a proof will be indicated here.
4

Donsker’s Theorem and Inequalities

If \( Z \) has \( N(0,1) \) distribution, then for any \( c > 0 \), \( \Pr(Z \geq c) \leq \exp(-c^2/2) \)
(RAP, Lemma 12.1.6(b)). Thus if \( X \) has \( N(0, \sigma^2) \) distribution for some \( \sigma > 0 \),
then \( \Pr(X \geq c) = \Pr(X/\sigma > c/\sigma) \leq \exp(-c^2/(2\sigma^2)) \). It follows that for any
\( n = 1, 2, \ldots \) and any \( j = 1, 2, \ldots \),

\[
\Pr\left( \left| x_{j/2^n} - x_{(j-1)/2^n} \right| \geq \frac{1}{n^2} \right) \leq 2 \exp\left(-\frac{n^4}{2n^4}\right).
\]

It follows that for any integer \( K > 0 \), the probability of any of the above
events occurring for \( j = 1, \ldots, 2^n K \) is at most \( 2^{n+1} K \exp(-2n^2/(2n^4)) \),
which approaches 0 very fast as \( n \to \infty \), because of the dominant factor \(-2n^2\)
in the exponent. Also, the series \( \sum_n 1/n^2 \) converges. It follows by the Borel–Cantelli
Lemma (RAP, Theorem 8.3.4) that with probability 1, for all \( t \in [0, K] \),
for a sequence of dyadic rationals \( t_n \to t \) given by the binary expansion of \( t, x_{t_n} \)
will converge to some limit \( X_t \), which equals \( x_t \) almost surely. Specifically,
for \( t < K \), let \( t_n = (j-1)/2^n \) for the unique \( j \leq 2^n K \) such that \((j-1)/2^n \leq t < j/2^n\).
Then \( t_{n+1} = t_n = 2j/2^{n+1} \) or \( t_{n+1} = (j-1)/2^{n+1} \), so that \( t_{n+1} \) and \( t_n \) are either equal or are adjacent
dyadic rationals with denominator \( 2^{n+1} \), and the above bounds apply to the differences \( x_{t_{n+1}} - x_{t_n} \).

The process \( X_t \) is sample-continuous and is itself a Brownian motion, as
desired. From here on, a “Brownian motion” will always mean a sample-
continuous one.

Here is a reflection principle for Brownian motion (RAP, 12.3.1). A proof
will be sketched.

**Theorem 1.4** Let \( \{x_t\}_{t \geq 0} \) be a Brownian motion, \( b > 0 \) and \( c > 0 \). Then

\[
\Pr(\sup\{x_t : t \leq b\} \geq c) = 2 \Pr(x_b \geq c) = 2 \Pr(N(0,b)([c, +\infty))).
\]

**Sketch of proof:** If \( \sup\{x_t : t \leq b\} \geq c \), then by sample continuity there is a least
time \( \tau \) with \( 0 < \tau \leq b \) such that \( x_{\tau} = c \). The probability that \( \tau = b \) is 0, so we
can assume that \( \tau < b \) if it exists. Starting at time \( \tau, x_b \) is equally likely to be \( > c \)
or \( < c \). [This holds by an extension of the independent increment property or the
strong Markov property (RAP, Section 12.2); or via approximation by suitably
normalized simple symmetric random walks and the reflection principle for
them.] Thus

\[
\Pr(x_b \geq c) = \frac{1}{2} \Pr(\sup\{x_t : t \leq b\} \geq c),
\]

which gives the conclusion. \( \square \)

One way to write the Brownian bridge process \( y_t \) in terms of Brownian
motion is \( y_t = x_t - tx_1, \ 0 \leq t \leq 1 \). It is easily checked that this a Gaussian
process (\( y_t \) for \( t \) in any finite subset of \([0,1]\) have a normal joint distribution,
with zero means) and that the covariance \( E_{y_t,y_s} = s(1-t) \) for \( 0 \leq s \leq t \leq 1 \),
fitting the definition of Brownian bridge. It follows that the Brownian bridge
process, on \([0, 1]\), is also sample continuous, i.e., we can and will take it such that \(t \mapsto y_t(\omega)\) is continuous for almost all \(\omega\).

Another relation is that \(y_t\) is \(x_t\) for \(0 \leq t \leq 1\) conditioned on \(x_1 = 0\) in a suitable sense, namely, it has the limit of the distributions of \(\{x_t\}_{0 \leq t \leq 1}\) given \(|x_1| < \varepsilon\) as \(\varepsilon \downarrow 0\) (RAP, Proposition 12.3.2). A proof of this will also be sketched here. Suppose we are given a Brownian bridge \(\{y_t\}_{0 \leq t \leq 1}\). Let \(Z\) be a \(N(0, 1)\) random variable independent of the \(y_t\) process. Define \(\xi_t = y_t + tZ\) for \(0 \leq t \leq 1\). Then \(\xi_t\) is a Gaussian stochastic process with mean 0 and covariance given, for \(0 \leq s \leq t \leq 1\), by

\[
E \xi_s \xi_t = s(1 - t),
\]

so \(\xi_t\) for \(0 \leq t \leq 1\) has the distribution of Brownian motion restricted to \([0, 1]\). The conditional distribution of \(\xi_t\) given \(|\xi_1| < \varepsilon\), in other words \(|Z| < \varepsilon\), is that of \(y_t + tZ\) given \(|Z| < \varepsilon\), and since \(Z\) is independent of \(\{y_t\}_{0 \leq t \leq 1}\), this conditional distribution clearly converges to that of \(\{y_t\}\) as \(\varepsilon \downarrow 0\), as claimed.

Kolmogorov evaluated the distributions of \(\sup_0 \leq t \leq 1 y_t\) and \(\sup_0 \leq t \leq 1 |y_t|\) explicitly. For the first (1-sided) supremum this follows from a reflection principle (RAP, Proposition 12.3.3) for \(y_t\) whose proof will be sketched:

**Theorem 1.5** For a Brownian bridge \(\{y_t\}_{0 \leq t \leq 1}\) and any \(c > 0\),

\[
\Pr(\sup_{0 \leq t \leq 1} y_t > c) = \exp(-2c^2).
\]

**Sketch of proof:** The probability is, for a Brownian motion \(x_t\),

\[
\lim_{\varepsilon \downarrow 0} \Pr\left(\sup_{0 \leq t \leq 1} x_t > c \mid |x_1| < \varepsilon\right)
= \lim_{\varepsilon \downarrow 0} \Pr\left(\sup_{0 \leq t \leq 1} x_t > c \text{ and } |x_1| < \varepsilon\right) / \Pr(|x_1| < \varepsilon)
= \lim_{\varepsilon \downarrow 0} \Pr\left(\sup_{0 \leq t \leq 1} x_t > c \text{ and } |x_1 - 2c| < \varepsilon\right) / \Pr(|x_1| < \varepsilon)
\]

where the last equality is by reflection. For \(\varepsilon\) small enough, \(\varepsilon < c\), and then the last quotient becomes simply \(\Pr(|x_1 - 2c| < \varepsilon) / \Pr(|x_1| < \varepsilon)\). Letting \(\phi\) be the standard normal density function, the quotient is asymptotic as \(\varepsilon \downarrow 0\) to \(\phi(2c) \cdot 2\varepsilon / (\phi(0) \cdot 2\varepsilon) = \exp(-2c^2)\) as stated. \(\square\)

The distribution of \(\sup_{0 \leq t \leq 1} |y_t|\) is given by a series (RAP, Proposition 12.3.4) as follows:

**Theorem 1.6** For any \(c > 0\), and a Brownian bridge \(y_t\),

\[
\Pr\left(\sup_{0 \leq t \leq 1} |y_t| > c\right) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2c^2).
\]
1 Donsker’s Theorem and Inequalities

The proof is by iterated reflections: for example, a Brownian path which before time 1 reaches $+c$, then later $-c$, then returns to (near) 0 at time 1, corresponds to a path which reaches $c$, then $3c$, then (near) $4c$, and so on.

Doob (1949) asked whether the convergence in distribution of empirical processes to the Brownian bridges held for more general functionals (other than the supremum and that of absolute value). 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, first in this chapter on speed of convergence in the classical case. In the rest of the book, the collection of half-lines $(-\infty, x]$, $x \in \mathbb{R}$, will be 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 of Donsker’s theorem with a statement on rate of convergence, to be proved in Section 1.4. Sections 1.2 on metric entropy and 1.3 on inequalities provide concepts and facts to be used in the rest of the book.

1.1 Empirical Processes: The Classical Case

In this section, a form of Donsker’s theorem with rates of convergence will be stated for the $U[0, 1]$ distribution with distribution function $U$ and empirical distribution functions $U_n$. This would imply a corresponding limit theorem for a general distribution function $F$ via Proposition 1.2. Let $\alpha_n := n^{-1/2}(U_n - U)$ 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).

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

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

Here is a formulation of Donsker’s theorem.

**Theorem 1.7** 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 $X_i$;
1.2 Metric Entropy and Capacity

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

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

The theorem just stated is a consequence of the following facts giving rates of convergence. Komlós, Major, and Tusnády (1975) stated a sharp rate of convergence in Donsker’s theorem, namely, that on some probability space there exist $X_i$ i.i.d. $U[0, 1]$ and Brownian bridges $Y_n$ such that

$$P\left(\sup_{0 \leq t \leq 1} |(\alpha_n - Y_n)(t)| > \frac{x + c \log n}{\sqrt{n}}\right) < Ke^{-\lambda x}$$

(1.1)

for all $n = 1, 2, \ldots$ and $x > 0$, where $c$, $K$, and $\lambda$ are positive absolute constants. If we take $x = a \log n$ for some $a > 0$, so that the numerator of the fraction remains of the order $O(\log n)$, the right side becomes $Kn^{-\lambda a}$, decreasing as $n \to \infty$ as any desired negative power of $n$.

More specifically, Bretagnolle and Massart (1989) proved the following:

**Theorem 1.8 (Bretagnolle and Massart)** The approximation (1.1) of empirical processes by Brownian bridges holds with $c = 12$, $K = 2$, and $\lambda = 1/6$ for $n \geq 2$.

Bretagnolle and Massart’s theorem is proved in Section 1.4.

1.2 Metric Entropy and Capacity

The notions in this section will be applied in later chapters, first, to Gaussian processes, then later in adapted forms, metric entropy with inclusion for sets, or with bracketing for functions, as applied to empirical processes in later chapters.

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.
1 Donsker’s Theorem and Inequalities

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 \(A\).

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 \(F\).

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

**Theorem 1.9** 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 \(\overline{B}(x, \varepsilon) := \{y : d(x, y) \leq \varepsilon\}\) 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 \(N(\varepsilon, A, d)\). \(N(\varepsilon, A, S, d)\) depends not only on \(A\) but also 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 \(N(\varepsilon, B, B, d) = 1 < N(\varepsilon, A, A, d)\). 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 \( \varepsilon^{-r} \). 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. Bernštein and P. L. Chebyshev:

**Theorem 1.10** For any real random variable \( X \) and \( t \in \mathbb{R} \),

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

**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 \( X_1, \ldots, X_n \), let \( S_n := X_1 + \cdots + X_n \).

**Theorem 1.11 (Bernštein’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 \( \sigma^2_j := \text{var}(X_j) \) and \( \tau^2_n := \text{var}(S_n) = \sigma^2_1 + \cdots + \sigma^2_n \). Then for any \( K > 0 \),

\[
\Pr(|S_n| \geq Kn^{1/2}) \leq 2 \cdot \exp(-nK^2/(2\tau^2_n + 2Mn^{1/2}K/3)). \tag{1.2}
\]

**Proof.** We can assume that \( \tau^2_n > 0 \) since otherwise \( S_n = 0 \) a.s. and the inequality holds. For any \( u \geq 0 \) and \( j = 1, \ldots, n \),

\[
E \exp(uX_j) = 1 + u^2 \sigma^2_j F_j/2 \leq \exp(\sigma^2_j F_j u^2/2) \tag{1.3}
\]

where \( F_j := 2\sigma^{-2}_j \sum_{r=2}^{\infty} u^{r-2} E|X'_j/r|! \), or \( F_j = 0 \) if \( \sigma^2_j = 0 \). For \( r \geq 2 \), \( |X'_j| \leq X^2_j M^{r-2} \) 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 \( u := v/(\tau^2_n + Mu/3) \), so that \( v = \tau^2_n u/(1 - Mu/3) \).

Then \( 0 < u < 3/M \). Thus, multiplying the factors on the right side of (1.3) by...
1 Donsker’s Theorem and Inequalities

independence, we have

$$E \exp(u S_n) \leq \exp\left(\tau^2_n u^2 / 2 (1 - Mu / 3)\right) = \exp(\tau n u^2 / 2).$$

So by Theorem 1.10, \( \Pr\{S_n \geq v\} \leq e^{-uv/2} \) and

$$e^{-uv/2} = \exp\left(-v^2 / (2\tau^2_n + 2Mv / 3)\right) = \exp\left(-nK^2 / (2\tau^2_n + 2MKn^{1/2} / 3)\right). \quad \square$$

Here are some remarks on Bernšteín’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 \( 2 \exp(-K^2/(2\sigma^2)) \), 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 \( X_j \), replacing them by truncated ones, say replacing \( X_j \) by \( f_{M_n}(X_j) \) where \( f_M(x) = x 1_{|x| \leq M} \). In that case the probability \( \Pr(|X_j| > M_n \text{ for some } j \leq n) \) needs to be small enough so that the inequality with this additional probability added to the bound is still useful.

Next, let \( s_1, s_2, \ldots \) be i.i.d. 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 \( \{s_i Y_j\}_{j \geq 1} \) 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 \( Y_1, \ldots, Y_n \),

$$\Pr(\sum_{j=1}^{n} s_i Y_j > K | Y_1, \ldots, Y_n) \leq \exp(-K^2/(2\sum_{j=1}^{n} Y_j^2))$$

**Proposition 1.12 (Hoeffding)** For any \( t \geq 0 \), and real \( a_j \) not all 0,

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

**Proof.** Since \( 1/(2n)! \leq 2^{-n} / n! \) for \( n = 0, 1, \ldots \), we have \( \cosh x \equiv (e^x + e^{-x}) / 2 \leq \exp(x^2 / 2) \) for all \( x \). Applying Theorem 1.10, the probability on the left is bounded above by \( \inf_u \exp(-ut + \sum_{j=1}^{n} a_j^2 u^2 / 2) \), which by calculus is attained at \( u = t / \sum_{j=1}^{n} a_j^2 \), and the result follows. \( \square \)