Printed: 2 March 2024 at 14:43

Chapter 10

Chaining

Chaining::Chaining

SECTION *10.1 introduces the idea of a maximal inequalities. It explains why it often suffices to consider finite, or countable, collections of random variables.*

SECTION *10.2 summarizes some useful facts about orlicz norms.*

SECTION *10.3 introduces chaining, a strategy for approximation by means of a sequence of finite sets linked together to decompose a process into a sum of increments.*

SECTION *10.4 defines covering and packing numbers, which are often used in the construction of chaining approximations.*

SECTION *10.5 gives four examples of chaining with packing numbers, each based on control of increments by an orlicz norm.*

SECTION *10.6 shows how chaining can be used to prove uniform continuity of sample paths, at least when the inequalities are derived from nested chaining frameworks.*

*SECTION *10.7 presents an example to point out some some subtle differences between the approach described in Section 10.3 and the way chaining was employed in earlier literature.*

10.1 Maximal inequalities

Chaining::S:intro

This Chapter introduces a powerful method for deriving various probabilistic bounds for stochastic processes, $X = \{X_t : t \in T\}$, where the index set T is equipped with a metric d that gives some control over the increments of the process. For example, we might have

\E@ norm.incr <1>
$$\|X_s - X_t\| \le K_0 d(s,t)$$

for some norm $\|\cdot\|$ and some constant K_0, or we might have a tail bound

\E@ tail.incr <2>
$$\mathbb{P}\{|X_s - X_t| > K_1 d(s,t) r\} \le \beta(r) \qquad \text{for all } s,t \in T, \text{ and } r \ge 0.$$

for some decreasing function β with $\lim_{r\to\infty} \beta(r) = 0$ and some other constant K_1.

To simplify the exposition, in this Chapter I'll always take the norm in <1> to be $\|\cdot\|_\Psi$ for some orlicz function Ψ, with β derived from Ψ. To avoid having to fuss about some special cases I'll also assume that Ψ is strictly increasing on $\mathbb{R}^+$, with no flat spot. That is, $\Psi(x) = 0$ only at $x = 0$. For your convenience the relevant properties from Chapter 5 are summarized in Section 10.2.

Remark. It is not essential that $d(s,t) = 0$ should imply $s = t$. That is, d might be a semi-metric, in which case, it suffices to have $X_s = X_t$ almost surely if $d(s,t) = 0$. Inequality <2> ensures that this property holds. Similarly, $\|\cdot\|$ could be a semi-norm provided $\|W\| = 0$ if and only if $W = 0$ almost surely, as happens with the usual $\mathcal{L}^p$ and orlicz "norms". I leave it to the fastidious reader to modify the arguments to accommodate semi-metrics and semi-norms.

The central task in many probabilistic and statistical problems is to find good upper bounds for quantities such as $\sup_{t\in T}|X_t|$ or the ***oscillation***,

\E@ osc.def <3>
$$\operatorname{osc}(\delta, X, T) := \sup\{|X_s - X_t| : s,t \in T,\, d(s,t) < \delta\}.$$

For example, we might seek bounds on $\|\sup_{t\in T}|X_t|\,\|$ or $\|\operatorname{osc}(\delta, X, T)\,\|$, or on their tail probability analogs

$$\mathbb{P}\{\sup_{t\in T}|X_t| > \eta\} \qquad \text{or} \qquad \mathbb{P}\{\operatorname{osc}(\delta, X, T) > \eta\}.$$

Such bounds are often referred to as ***maximal inequalities*** because they involve maxima (or suprema) over potentially large sets of random variables.

Maximal inequalities play an important role in the theory of stochastic processes, empirical process theory, and statistical asymptotic theory. Bounds for the oscillation are essential for the construction of processes with continuous sample paths (Section 10.6) and for the study of convergence in distribution of sequences of stochastic processes (Chapter 14). In particular, oscillation control is the key ingredient in the proofs of donsker theorems for empirical processes. In the literature on the asymptotic theory for estimators defined by optimization over random processes, oscillation bounds (or something similar) have played a major role under various names, such as "stochastic asymptotic equicontinuity" or "uniform tightness".

For uncountable index sets T we encounter complications regarding the integrability of quantities like $\sup_{t\in T}|X_t|$, because uncountable unions need not preserve measurability. Some authors, such as Ledoux and Talagrand (1991, page 298) and Talagrand (2021, page 13), got around this difficulty by interpreting the inequalities as statements about arbitrarily large finite subsets of T: they interpreted

$$\mathbb{P}\sup_{t\in T} X_t \quad \text{to mean} \quad \sup\{\mathbb{P}\sup_{t\in F} X_t : \text{finite } F \subset T\}.$$

With such a convention the real challenge becomes: find bounds for finite F that do not grow unhelpfully large as the cardinality (size) of F increases.

This convention is (essentially) equivalent to replacing $\sup_{t\in T} X_t$ by the essential supremum of $\{X_t : t \in T\}$. See Section 9.2 for details.

Another solution to the problem when T is an uncountable, but separable, metric space involves the replacement of $\{X_t : t \in T\}$ by a new ***version*** of the process: a new stochastic process $\{\widetilde{X}_t : t \in T\}$ on the same probability space, for which the set $\mathcal{N}_t := \{\omega : X_t(\omega) \neq \widetilde{X}_t(\omega)\}$ is $\mathbb{P}$-negligible for each fixed t. Such a modification is particulary unobjectionable if the X_t's are really only defined up to an almost sure equivalence. Doob (1953, Section II.2) showed that the version can always be chosen in such a way that there exists a countable, dense, ***approximating subset*** T_∞ of T and a single $\mathbb{P}$-negligible set $\mathcal{N}$ such that:

> for each ω in $\mathcal{N}^c$ and each $t \in T$,
> there is a sequence $\{s_n\}$ (depending on ω) in T_∞
> for which $\widetilde{X}(s_n, \omega) \to \widetilde{X}(t, \omega)$.

For the technical details see Section 9.4, where I use the term ***doob-separable*** to avoid confusion with separability (meaning existence of a countable, dense subset) of a metric space.

If X is doob-separable then $\sup_{t\in T} |\widetilde{X}(t,\omega)| = \sup_{t\in T_\infty} |\widetilde{X}(t,\omega)|$ for each ω in $\mathcal{N}^c$. Similar equalities holds for the oscillation and other quantities involving suprema and infima over T. Effectively all probability calculations can be carried out using only $\{\widetilde{X}_t : t \in T_\infty\}$, a process with a countable index set.

There is a significant simplification if the doob-separable process is ***continuous in probability***, that is, for each t in T,

\E@ cts.in.prob <4> $$X_{s_n} \to X_t \qquad \text{in probability as } s_n \to t.$$

In that case every countable, dense subset of T is an approximating subset. Property <4> follows from either of <1> or <2>. Again see Section 9.4 for details (and fine print) regarding doob-separable stochastic processes.

In view of the previous few paragraphs, you will be quite safe if you prefer to ignore all that stuff about doob-separable processes and just focus on how to handle $\{X_t : t \in T_\infty\}$ for some countable, dense subset T_∞ of T. Equivalently, you could just consider very simple processes, $\{X_t : t \in F\}$ with F a finite subset of T, as long as the bounds do not blow up when the size of F is sent off to infinity.

10.2 A reminder about orlicz norms

Chaining::S:Orlicz

If an orlicz function Ψ has no flat spot then $g := \log(1+\Psi)$ is continuous and strictly increasing with $g(0) = 0$, and $\Psi = e^g - 1$. The inverse functions Ψ^{-1} and g^{-1} are well defined, continuous, and strictly increasing.

Assumption <1> with $\|\cdot\| = \|\cdot\|_\Psi$ becomes

$$\|X_s - X_t\|_\Psi := \inf\{c > 0 : \mathbb{P}\Psi\left(|X_s - X_t|/c\right) \le 1\} \le K_0 d(s,t).$$

If $d(s,t) > 0$, the standardized increment $Y_{s,t} := |X_s - X_t|/K_0 d(s,t)$ has $\|Y_{s,t}\|_\Psi \le 1$, so that $\mathbb{P}\Psi(Y_{s,t}) \le 1$ and

$$2 \ge 1 + \mathbb{P}\Psi(Y_{s,t}) = \mathbb{P}\exp(g(Y_{s,t})).$$

It then follows that

\E@ Orlicz.tails <5> $$\mathbb{P}\{|X_s - X_t| \ge K_0 d(s,t) r\} \le \mathbb{P}\exp\left(g(Y_{s,t}) - g(r)\right) \le 2e^{-g(r)}.$$

That is, the $\beta(r)$ in <2> can be taken to equal $2e^{-g(r)}$.

For the special case where X is a centered gaussian process it is natural to define the metric by $d(s,t) = \|X_s - X_t\|_2$ and use the orlicz function corresponding to $g(x) = x^2$. For those choices (see Section 5.6) the constant K_0 equals the awkward $\sqrt{8/3}$. Even with the choice $g(x) = x^2/2$, which would produce the 'subgaussian tail bound' $2\exp(-r^2/2)$, the constant K_0 would be $\sqrt{2}$, only slightly less awkward. I'll often ignore such nuisances, perhaps with the suggestion that we could just work with X/K_0.

For some calculations it is convenient to have a graceful way to bound products of the form $\Psi(x)\Psi(y)$. The easiest approach is to assume that Ψ belongs to $\mathcal{Y}_{\exp}$, as described in Section 5.3. That is, there exists a constant C_g such that

\E@ g.bnd <6> $$g(x) + g(y) \le g\left(C_g \max(x,y)\right) \qquad \text{for all } x, y \in \mathbb{R}^+.$$

Equivalently,

\E@ ig.bnd <7> $$g^{-1}(a+b) \le C_g \max\left(g^{-1}(a), g^{-1}(b)\right) \qquad \text{for all } a, b \text{ in } \mathbb{R}^+.$$

These inequalities for g imply analogous inequalities for Ψ:

\E@ Psi.bnd <8> $$\Psi(x)\Psi(y) = \left(e^{g(x)} - 1\right)\left(e^{g(y)} - 1\right) \le e^{g(x)+g(y)} - 1 \le \Psi\left(C_g \max(x,y)\right)$$

for all x, y in $\mathbb{R}^+$ and

\E@ yyexp.inv <9> $$\Psi^{-1}(ab) \le C_g \max\left(\Psi^{-1}(a), \Psi^{-1}(b)\right) \qquad \text{for all } a, b \text{ in } \mathbb{R}^+.$$

For $1 \le \alpha < \infty$, each of the orlicz functions $\Psi_\alpha := e^{g_\alpha} - 1$, with $g_\alpha(x) = x^\alpha$, belongs to $\mathcal{Y}_{\exp}$. Moreover, g_α is convex and $g_\alpha(x) + g_\alpha(y) \le g_\alpha(x+y)$ for all $x, y \in \mathbb{R}^+$, so that <6> actually holds with $C_{g_\alpha} = 2$.

The bounds from Section 5.4 for handling maxima of finitely many variables will also be particularly useful in the present Chapter. Here is a slight restatement of the main result:

Chaining::Orlicz.maximal <10> **Theorem.** *Let $W_1, \ldots, W_N$ be random variables, **not necessarily independent**, but all defined on the same probability space. Define $M := \max_{i \le N} |W_i|$. Suppose $\max_{i \le N} \|W_i\|_\Psi \le 1$ for some orlicz function Ψ. Then:*

(i) $\mathbb{P}M \le \Psi^{-1}(N)$.

(ii) $\mathbb{P}_B M \le \Psi^{-1}(N/\mathbb{P}B)$ where $\mathbb{P}_B$ denotes conditional expectation given an event B with $\mathbb{P}B > 0$.

(iii) If $\Psi \in \mathcal{Y}_{\exp}$ then $\|M\|_\Psi \le 2C\Psi^{-1}(N)$ for some constant C.

10.3 What is chaining?

Chaining::S:chains

The workhorse of the modern approach to approximating stochastic processes is called ***chaining***, which seems to have an undeserved reputation as a complicated and difficult idea. I suspect the reputation comes from the earlier literature where it was mostly applied to bound tail probabilities. As you will soon see, it is not difficult to bound tails by sums involving arbitrarily chosen small constants; but it can be messy to transform such sums into easily interpreted expressions. Many earlier papers gave little insight into how those constants were chosen. I suspect the choices involved some trial and error. (Of course I intend no criticism of those earlier authors. Indeed, I count myself amongst those who engaged in much trial and error.) In this Chapter I hope to dispel some of the mysteries of chaining by departing from the historical order in which chaining ideas first emerged, proceeding instead from simpler to more subtle constructions.

Remark. A wise probabilist once told me that he couldn't understand all the mystery surrounding chaining, for 'it is really not much more than the triangle inequality'. It is indeed (mostly) just the triangle inequality, applied many times. The cleverness comes from the choices of when to invoke the inequality and how to collect terms into manageable chunks.

The following Example introduces the key to the chaining idea's success.

Chaining::BM.bnd <11>

Example. Consider the problem of bounding $\mathbb{P}\sup_t |B_t|$ for a brownian motion process $\{B_t : 0 \le t \le 1\}$. Quite sharp inequalities have been derived using special properties of the process, such as linearity of the index set and independence of the increments across disjoint intervals. Less precise bounds, to show that the expected supremum is finite, can be derived using only the fact that $B_t - B_s$ has a subgaussian distribution with scale parameter equal to $\sqrt{|t-s|}$.

The main tool is a simple maximal inequality: if $X_1, \ldots, X_N$ are random variables with $X_i \in \text{SUBG}(\tau_i^2)$ then

\E@ max.subg <12>

$$\mathbb{P}\max_i X_i \le \tau\sqrt{2\log N} \qquad \text{where } \tau := \max_i \tau_i.$$

To derive this bound, argue for $\lambda > 0$, that

$$\begin{aligned}\exp\left(\lambda \mathbb{P}\max_i X_i\right) &\le \mathbb{P}\exp\left(\lambda \max_i X_i\right) \qquad \text{by the jensen inequality} \\ &= \mathbb{P}\max_i e^{\lambda X_i} \le \sum\nolimits_i \mathbb{P}e^{\lambda X_i} \\ &\le \sum\nolimits_i e^{\lambda^2\tau_i^2/2} \le N e^{\lambda^2\tau^2/2}.\end{aligned}$$

Take logs, divide through by λ, then choose $\lambda = \tau^{-1}\sqrt{2\log N}$ to minimize.

It is natural to approximate the supremum by a maximum taken over a large, finite subset such as

$$T_i := \{j\delta_i : j = 0, 1, \ldots, 2^i\} \qquad \text{where } \delta_i := 2^{-i}.$$

Then we can let i tend to infinity, with the hope that the limit will be finite.

To emphasize the fact that the bound will only depend on information about the increments of the process I'll write $B_t - B_0$ instead of B_t, using the fact that $B_0 = 0$. Inequality <12> with $\tau^2 = 1$ gives

$$\mathbb{P} \max_{t\in T_i}(B_t - B_0) \le \sqrt{2\log(1+2^i)}\,,$$

which blows up as $i \to \infty$. You might protest that $\tau^2 = 1$ is a gross overestimate when $j\delta_i$ is close to zero. However, half of the members of T_i have a τ_j^2 of at least $1/2$, which is enough to lead to failure.

For the next part of the discussion it will be cleaner to work with two-sided bounds. If one of the X_i's in <12> is identically zero, as will always be the case in what follows, then

\E@ max.subg2 <13>
$$\mathbb{P} \max_i |X_i| \le \mathbb{P}\left(\max_i X_i + \max_i(-X_i)\right) \le 2\tau\sqrt{2\log N}\,.$$

Remark. Another way to derive a two-sided bound like <13> is: for the orlicz function $\Psi(x) = \exp(x^2) - 1$ we have $\|X_i\|_\Psi \le C_0\tau_i$ for some constant C_0. Theorem <10>(i) then gives

$$\mathbb{P} \max_i |X_i| \le C_0\tau\Psi^{-1}(N) = C_0\tau\sqrt{\log(1+N)}\,.$$

It is indeed wasteful to apply the subgaussian bound directly to the $B_t - B_0$ increments. It would be better to relate each $B_{j\delta_i}$ to a value $B_{\ell_{i-1}(j\delta_i)}$, where ℓ_{i-1} rounds down to the nearest multiple of δ_{i-1}. Equivalently, ℓ_{i-1} maps T_i into the set $T_{i-1} := \{j\delta_{i-1} : 0 \le j \le 2^{i-1}\}$ with $|t - \ell_{i-1}(t)| \le \delta_i$. The triangle inequality,

$$|B_t - B_0| \le |B_t - B_{\ell_{i-1}(t)}| + |B_{\ell_{i-1}(t)} - B_0| \le \max_{t\in T_i} |B_t - B_{\ell(t)}| + \max_{s\in T_{i-1}} |B_s - B_0|,$$

then leads us to the bound

$$\begin{aligned}\mathbb{P} \max_{t\in T_i} |B_t - B_0| &\le \mathbb{P} \max_{t\in T_i} |B_t - B_{\ell_{i-1}(t)}| + \mathbb{P} \max_{s\in T_{i-1}} |B_s - B_0| \\ &\le \sqrt{\delta_i}\sqrt{2\log(1+2^i)} + \mathbb{P} \max_{s\in T_{i-1}} |B_s - B_0|.\end{aligned}$$

By matching each t in T_i with an s in T_{i-1} with $|t - s|$ significantly smaller than $|t - 0|$ we have have solved one explosive problem (for T_i) at the cost of creating an analogous problem with a slightly smaller index set T_{i-1}.

So how do we handle the T_{i-1} problem? Of course, the answer is: use the same idea to approximate T_{i-1} by an smaller T_{i-2}. And so on. Eventually we get a bound

$$\mathbb{P} \max_{t\in T_i} |B_t - B_0| \le \sum_{j=1}^{i} \sqrt{2^{-j}}\sqrt{2\log(1+2^j)}\,,$$

which does not blow up as $i \to \infty$.

An appeal to monotone convergence then bounds $\mathbb{P}\sup_{t\in T_\infty} |B_t - B_0|$ for a countable, dense subset T_∞ of $[0,1]$. Continuity of sample paths, or even just doob-separabilty, then takes care of the whole interval $[0,1]$. □

Remark. You might experiment with different choices of T_i in the previous Example. For example, for which positive α would the T_i consisting of integer multiples of $i^{-\alpha}$ lead to a finite bound? Historically speaking, the 2^{-i} grid seems to have been the favorite choice. It has the advantage that there is a simple upper bound, δ_i, for $|t - \ell(t)|$. Something similar, based on packing numbers, will be used in Section 10.5. In Chapters 11 and 12 you will learn that a different strategy is sometimes better.

Keeping the message from Example <11> in mind, let us now return to the general problem of obtaining maximal inequalities for a process $\{X_t : t \in T\}$ indexed by a metric space (T, d) with a metric that provides some probabilistic control over the increments of the process, as <1> or <2>. As a typical example, suppose we wish to bound $\mathbb{P}\{\sup_{t\in T} |X_t| > \eta\}$, for a countably infinite or finite T. We might try the union bound,

$$\mathbb{P}\{\sup_{t\in T} |X_t| > \eta\} \le \sum\nolimits_{t\in T} \mathbb{P}\{|X_t| > \eta\},$$

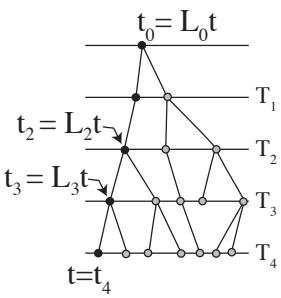
Unfortunately, the sum might be larger than 1, especially if T is big.

Instead we could creep up on T through a sequence of finite subsets, $T_0, T_1, \ldots$ such that, for each fixed s in T,

\E@ denser <14>
$$d(s, T_m) := \min\{d(s,t) : t \in T_m\} \to 0 \qquad \text{as } m \to \infty.$$

It simplifies some notation if we insist that $T_0 = \{t_0\}$, a singleton set.

The chaining method works by breaking the process on T_m into a contribution from t_0 plus a sum of increments across each T_i to T_{i-1} pair. To achieve such a decomposition, for each i in $\mathbb{N}$ we need a map $\ell_{i-1} : T_i \to T_{i-1}$, thereby arranging the T_i's into a tree. The tree then defines maps $L_i : T_m \to T_i$ for $i \le m$ by following paths towards t_0: for each t in T_m,



$$L_i t := \text{the point of } T_i \text{ on the tree path from } t \text{ to } t_0.$$

More formally, if $t \in T_m$ and $i < m$ then $L_i t := \ell_i \circ \ell_{i+1} \circ \cdots \circ \ell_{m-1}(t)$.

Remark. To cut down on parenthetic clutter I'll abbreviate $\ell_i(t)$ to $\ell_i t$ and $L_i(t)$ to $L_i t$ whenever there is no ambiguity. Parentheses seem advisable for things like $\ell_{i-1}(j\delta_i)$.

The edges of the tree correspond to the increments of the process,

\E@ X.increments <15>
$$X_t - X_{t_0} = \sum\nolimits_{i=1}^{m} X(L_i t) - X(L_{i-1} t) \qquad \text{if } t \in T_m,$$

with a corresponding bound derived from the triangle inequality,

$$|X_t - X_{t_0}| \le \sum\nolimits_{i=1}^{m} |X(L_i t) - X(L_{i-1} t)| \qquad \text{if } t \in T_m.$$

Remark. Here the $|\cdot|$ stands for the absolute value of a real-valued random variable. The same argument works if X takes values in some more exotic normed vector spaces (such as $\mathbb{R}^N$) with $|\cdot|$ denoting the corresponding norm.

Clearly we must lose something by using such an inequality. (See Problem [2] for an example.) Against that cost we receive a payoff from the simple fact that many different t's in T_m can share the same increment:

$$X(L_i t) - X(L_{i-1}t) = X(s) - X(\ell_{i-1}s) \qquad \text{if } L_i t = s.$$

We need to control that increment only once, not separately for each path. Inequalities <1> and <2> are tailor-made for that purpose. This sharing of increments is the main reason for the effectiveness of chaining arguments.

If you were suspecting that the cleverness of the chaining method lies mostly in the choice of the T_i's then you would, in general, be correct. However, for the most important special case based on packing numbers (discussed in the next two Sections), the choices are fairly straightforward. Moreover, for that special case the T_i's can be chosen as nested, $T_0 = \{t_0\} \subset T_1 \subset T_2 \subset \ldots$. The countable set $T_\infty := \cup_{j\in\mathbb{N}_0} T_j$ is dense in T. It becomes a natural candidate for the approximating subset of a doob-separable version of a process satisfying <1> or <2>. The passage from a bound over T_m to a bound over T_∞ then involves little more than an appeal to the monotone convergence theorem.

For future reference, I'll give a name to the approximation scheme described in the previous few pages.

Chaining::framework <16> **Definition.** *A **chaining framework** $\{(T_i, \ell_i) : i \in \mathbb{N}_0\}$ on a metric space (T, d) consists of:*

- *a sequence $\{T_i : i \in \mathbb{N}_0\}$ of finite subsets of T with T_0 a singleton set*
- *maps $\ell_{i-1} : T_i \to T_{i-1}$ for each $i \in \mathbb{N}$*

for which

(i) $d(s, T_m) := \min\{d(s,t) : t \in T_m\} \to 0$ *as $m \to \infty$ for each fixed s in T.*

(ii) The $\{\ell_i\}$ define a tree on $T_\infty := \cup_{j\in\mathbb{N}_0} T_j$ rooted at t_0, with edges $(s, \ell_{i-1}s)$ for $s \in T_i$. The paths along the tree toward the root define maps $L_i : T_\infty \to T_i$ such that $\ell_{i-1} \circ L_i = L_{i-1}$ for each i.

*(iii) The framework is said to be **nested** if $T_{i-1} \subset T_i$ for each i. In that case it is required that $\ell_{i-1}t = t$ when $t \in T_{i-1}$.* □

Remark. It turns out that for some purposes (such as the construction in Section 11.6) it is useful to allow frameworks $\{(T_i, \ell_i) : i = 0, \ldots, m\}$ for a fixed finite m. Clearly the assumption that $d(s, T_m) \to 0$ should be ignored in such a case. Rather than complicate the current definition—which is adequate for the present Chapter—with further subcases (such as dense frameworks and finite frameworks) I'll leave it to you to ignore assertion involving $m \to \infty$ when there are only finitely many T_i's.

Sometimes—particularly when the ultimate aim is control of the oscillation of a process—it pays to stop following the links of the chain (= the edges in the tree) before they reach the T_0 level. For $0 \le k < m$ we then have

$$|X(t) - X(L_k t)| \le \sum_{i=k+1}^{m} |X(L_i t) - X(L_{i-1} t)| \qquad \text{for } t \in T_m,$$

with a corresponding inequality that provides a uniform approximation for $\{X_t : t \in T_m\}$ via the simpler process $\{X_t : t \in T_k\}$:

$$\begin{aligned} \mathcal{M}_{k,m} &:= \max_{t \in T_m} |X(t) - X(L_k t)| \\ &\le \max_{t \in T_m} \sum_{i=k+1}^{m} |X(L_i t) - X(L_{i-1} t)| \qquad \text{\E@ max.sum } \langle 17 \rangle \\ &\le \sum_{i=k+1}^{m} \max_{s \in T_i} |X(s) - X(\ell_{i-1} s)|. \qquad \text{\E@ sum.max } \langle 18 \rangle \end{aligned}$$

The early literature typically worked with the upper bound <18>, with the ith summand a maximum of at most $|T_i|$ random variables, which was usually controlled by simple tail bounds. For example, if <2> holds then

$$\begin{aligned} \mathbb{P}\{\max_{s \in T_i} |X(s) - X(\ell_{i-1} s)| > \eta_i\} &\le \sum_{s \in T_i} \mathbb{P}\{|X(s) - X(\ell_{i-1} s)| > \eta_i\} \\ &\le |T_i| \max_{s \in T_i} \beta\left(\eta_i / (K_1 d(s, \ell_{i-1} s)\right). \qquad \text{\E@ tail.max } \langle 19 \rangle \end{aligned}$$

Actually, I should be more careful regarding the possibility that $\ell_{i-1} s$ might be equal to s, particularly if the framework is nested.

In the later literature, bounds based on orlicz norms (of the type described in Theorem <10>) became popular. Those bounds involve maxima of random variables with orlicz norms at most 1, which suggests a preliminary standardization: for s in T_i,

$$\Delta_i(s) := \begin{cases} \dfrac{|X(s) - X(\ell_{i-1} s)|}{\rho_i(s)} & \text{if } \rho_i(s) := \|X(s) - X(\ell_{i-1} s)\|_\Psi \ne 0 \\ 0 & \text{otherwise} \end{cases}. \qquad \text{\E@ Del.def } \langle 20 \rangle$$

With $\mathbb{M}_i := \max_{s \in T_i} \Delta_i(s)$, inequality <17> then gives

$$\begin{aligned} \mathcal{M}_{k,m} &\le \max_{t \in T_m} \sum_{i=k+1}^{m} \rho_i(L_i t) \Delta_i(L_i t) \qquad \text{\E@ max.Deli } \langle 21 \rangle \\ &\le \max_{t \in T_m} \sum_{i=k+1}^{m} \|X(L_i t) - X(L_{i-1} t)\|_\Psi \, \mathbb{M}_i \qquad \text{\E@ max.MMi } \langle 22 \rangle \\ &\le \max_{t \in T_m} \sum_{i=k+1}^{m} K_0 d(L_i t, L_{i-1} t) \mathbb{M}_i \qquad \text{by } \langle 1 \rangle. \qquad \text{\E@ diMM } \langle 23 \rangle \end{aligned}$$

As you will learn in Section 10.4, by design the packing method starts from given constants $\{\delta_i\}$ to construct T_i's for which $\max_{s \in T_i} d(s, \ell_{i-1} s) \le \delta_{i-1}$. In that case, the bound <23> is at most $\sum_{i=k+1}^{m} K_0 \delta_{i-1} \mathbb{M}_i$. From there a host of probabilistic bounds are easy to derive. For example, by Theorem <10>(i) we have

$$\mathbb{P}\mathcal{M}_{k,m} \le \sum_{i=k+1}^{m} K_0 \delta_{i-1} \mathbb{P}\mathbb{M}_i \le \sum_{i=k+1}^{m} K_0 \delta_{i-1} \Psi^{-1}(|T_i|). \qquad \text{\E@ pack.bnd1 } \langle 24 \rangle$$

Once we have a suitable chaining framework in place and some control over the increments of the process it then takes little further effort to control $\mathcal{M}_{k,m}$ in various ways. See Section 10.5.

An inequality like <24> is particularly useful if we can bound the size of the finite set T_i by some function of δ_i (and the sum on the right-hand side stays bounded as m increases). Such bounds have been available for quite some time. For example, Kolmogorov and Tikhomirov (1959) had obtained results for many kinds of function spaces; and in empirical process theory (see Chapters 15 and 16) packing bounds were later obtained by combinatorial arguments. All in all, the results derivable from an inequality like <24> justify the position of traditional chaining arguments in the toolbox of every user of probability theory.

Nevertheless, largely through the work of Fernique and Talagrand and their collaborators, it became known that some more delicate problems lie beyond the reach of traditional chaining methods. That difficulty led to a searching re-examination of the chaining idea, which resulted in modifications that created new tools capable of handling some of the delicate problems. For the start of that story see Chapter 11, then move on to Chapter 12 for a discussion of how the new ideas play out for gaussian processes.

10.4 Covering and packing numbers

Chaining::S:packing

This Section describes two classical ways of quantifying the 'size' of a metric space (T, d).

The simplest strategy for chaining is to minimize the size (cardinality) of the approximating subsets T_i for a chaining framework subject to a given upper bound on $\sup_{t\in T} \min_{s\in T_i} d(t, s)$. That idea translates into a statement about ***covering numbers***.

Remark. The logarithm of the covering number is sometimes called the ***metric entropy***. See Dudley (1973, page 70) for the origin of that name.

Chaining::covering <25>

Definition. *For a subset S of T write $\text{COVER}_T(\delta, S, d)$ for the δ-**covering number**, the smallest number of closed δ-balls whose union contains S. That is, the covering number is the smallest N for which there exist points $t_1, \ldots, t_N$ in T with $\min_{i\le N} d(t, t_i) \le \delta$ for each t in S. The set of centers $\{t_i\}$ is called a δ-**net** for S.* □

Remark. Notice a small subtlety related to the subscript T in the definition. If we regard S as a metric space in its own right, not just as a subset of T, then the covering numbers might be larger because the centers t_i would (implicitly) be forced to lie in S. It is an easy exercise (select a point of S from each covering ball that actually intersects S) to show that $\text{COVER}_S(2\delta, S, d) \le \text{COVER}_T(\delta, S, d)$. The extra factor of 2 would usually be of little consequence.

Some metric spaces (such as the whole real line under its usual metric) cannot be covered by a finite set of balls of a fixed radius. A metric space T for which $\text{COVER}_T(\delta, T, d) < \infty$ for every $\delta > 0$ is said to be ***totally bounded***, a concept very close to compactness: a metric space is compact if and only if it is both complete and totally bounded (Dudley, 2003, Section 2.3).

I prefer to work with the ***packing number*** $\text{PACK}(\delta, S, d)$, defined as the largest N for which there exist points $t_1, \ldots, t_N$ in S that are ***δ-separated***, that is, for which $d(t_i, t_j) > \delta$ if $i \neq j$. Notice the lack of a subscript T; the packing numbers are an intrinsic property of S, and do not depend on T except through the metric it defines on S.

Chaining::cover.pack <26>

Lemma. *For each $\delta > 0$,*

$$\text{COVER}_S(\delta, S, d) \leq \text{PACK}(\delta, S, d) \leq \text{COVER}_T(\delta/2, S, d).$$

Proof. Suppose $\text{PACK}(\delta, S, d) = N$ and the points of $F = \{t_1, \ldots, t_N\}$ are δ-separated. If $t \in S \backslash F$ then the set $F \cup \{t\}$ cannot be δ-separated, which implies $\min_i d(t, t_i) \leq \delta$. The closed balls $B[t_i, \delta]$ for $1 \leq i \leq N$ cover S.

For the second inequality, observe that no closed ball of radius $\delta/2$ can contain two points more than δ apart. In particular, each of the points t_i must lie in a different ball for a $\delta/2$ covering of S. □

The Lemma suggests that it is largely a matter of taste whether one works with covering or packing numbers, provided an occasional extra factor of 2 causes no significant concerns.

Henceforth I'll call a (finite) subset F of S a ***δ-packing set*** if it is δ-separated and maximal, in the sense that it is a not a proper subset of any other δ-separated set. As in the proof of the Lemma, for such an F the maximality implies $\sup_{s \in S} d(s, F) \leq \delta$. Also I'll write $\text{PACK}(\epsilon, S)$, instead of $\text{PACK}(\epsilon, S, d)$, when there is no ambiguity regarding the metric d.

Chaining::rr.norm <27>

Example. Let $\|\cdot\|$ denote some norm on $\mathbb{R}^n$, such as an ℓ^p norm, $|x|_p = \left(\sum_{i \leq n} |x_i|^p\right)^{1/p}$ for $p \geq 1$. As usual, write $B[t, r] := \{x \in \mathbb{R}^n : \|x - t\| \leq r\}$, the closed ball centered at t with radius r. Let $\mathfrak{m}_n$ denote Lebesgue's measure on $\mathcal{B}(\mathbb{R}^n)$.

The covering/packing numbers for such norms share a common geometric bound, a property derived from the fact that

$$\mathfrak{m}_n B[t, r] := \{x \in \mathbb{R}^n : \|x - t\| \leq r\} = r^n \Lambda_n \qquad \text{where } \Lambda_n := \mathfrak{m}_n B[0, 1].$$

Let $\{x_1, \ldots, x_N\}$ be any ϵr-separated set of points in B_r. The closed balls $B[x_i, \epsilon r/2]$, of radius $\epsilon r/2$ centered at the x_i, are disjoint and their union lies within $B_{r + \epsilon r/2}$. Thus

$$N \leq \frac{(r + \epsilon r/2)^n \Lambda_n}{(\epsilon r/2)^n \Lambda_n} = \left(\frac{2 + \epsilon}{\epsilon}\right)^n \leq (3/\epsilon)^n \qquad \text{if } 0 < \epsilon \leq 1.$$

That is, $\text{PACK}(\epsilon r, B_r, d) \leq (3/\epsilon)^n$ for $0 < \epsilon \leq 1$, where d denotes the metric corresponding to $\|\cdot\|$. □

Another reason for preferring packing to covering is that it easily generates nested chaining frameworks, as shown by the following construction.

Chaining::framePack <28>

Lemma. *Suppose (T, d) is a totally bounded metric space with diameter D. Let $\delta_0 = D > \delta_1 > \delta_2 \ldots$ be a sequence of positive numbers that decreases to zero. Then there exist a nested chaining framework for which:*

(i) Each T_i is a δ_i-packing set, so that $|T_i| \le \text{PACK}(\delta_i, T)$ for each i.

(ii) If $t \in T_i$ then $d(t, \ell_{i-1}t) \le \delta_{i-1}$ and $\ell_{i-1}t = t$ if $t \in T_{i-1}$. Consequently, the countable set $T_\infty := \cup_i T_i$ is dense in T. □

Proof. By total boundedness, D and each of the numbers $\text{PACK}(\delta_i, T)$ is finite for each i.

We have $\text{PACK}(D, T) = 1$ for the trivial reason that no pair of points in T is greater than D apart. Define $T_0 = \{t_0\}$ for an arbitrarily chosen point t_0 in T. Build T_0 up to a δ_1-separated set: start with $T_1 = T_0$ then loop,

$$\begin{aligned} &\text{while } R := \{t \in T : d(t, T_1) > \delta_1\} \neq \emptyset \\ &\qquad T_1 \leftarrow T_1 \cup \{t\} \text{ for some } t \text{ in } R \qquad \# \text{ a new } \delta_1\text{-separated set} \end{aligned}$$

This procedure must exit with $R = \emptyset$ after a finite number of steps because $\text{PACK}(\delta_1, T)$ is finite. By construction $d(t, T_1) := \min\{d(t, s) : s \in T_1\} \le \delta_1$ for all t in $T \backslash T_1$. Let $\ell_1 t$ be a point of T_1 for which $d(t, \ell_1 t) = d(t, T_1)$, with some suitable tie-breaking rule such as: choose the point that was added earlier to T_1. And, of course, choose $\ell_1 t = t$ for $t \in T_1$.

To construct T_2, start with $T_2 = T_1$ then repeat the loop with T_1 replaced by T_2 and δ_1 by δ_2. And so on.

Assertions (i) and (ii) are easy consequences of the way the iterative construction was carried out. □

For future reference, I'll call the chaining framework described by the Lemma as a $\{\delta_i : i \in \mathbb{N}_0\}$-***packing framework***, or $\{\delta_i\}$-packing framework, for short.

10.5 Chaining with packing numbers

Chaining::S:chainPack

This Section presents several ways to obtain probabilistic bounds involving $\sup_{t \in T} |X_t|$, where $\{X_t : t \in T\}$ is a doob-separable stochastic process indexed by a totally bounded metric space (T, d). The real work involves only calculations with $\sup_{t \in T_\infty} |X_t|$, where T_∞ is a countable, dense subset of T. If you prefer to ignore all that stuff about doob-separable processes you could just focus on T_∞. As you will see, most of that real work involves only $\{X_t : t \in F\}$, for finite subsets F of T, followed by a passage to a limit.

From now on I'll omit the annoying constant K_0 from <1> and just assume

\E@ incr.norm <29>

$$\|X_s - X_t\|_\Psi \le d(s, t) \qquad \text{for all } s, t \in T \ .$$

If you find this simplification troubling, you could work with X/K_0.

The inequalities in the rest of the Section usually involve various constants, all denoted by C (or something similar), which depend only on Ψ (and the abolished K_0). Here I am following the lead of Talagrand (2021, page 24), to avoid having to keep track of how a host of different constants are interrelated. I leave for you the instructive, if somewhat tedious, exercise of determining how the constants are interrelated.

The increment bound <29> ensures that the X process is continuous in probability: $X_s \to X_t$ in probability if $d(s,t) \to 0$. Thus it suffices to consider the behavior of X on any countable, dense subset of T.

Consider any $\{\delta_i\}$-packing framework, constructed using the method described in Lemma <28> with $\delta_i = D/2^i$, where $D := \operatorname{diam}(T)$. That is, we have an increasing sequence of finite subsets $T_0 = \{t_0\} \subset T_1 \subset \dots$ with $N_i := |T_i| \le \text{PACK}(\delta_i, T)$ for each i. The ℓ_{i-1} maps each t in T_i to its closest point in T_{i-1}. In particular, $d(s, \ell_{i-1}s) \le \delta_{i-1}$ for each s in $S_i := T_i \backslash T_{i-1}$. The set $T_\infty = \cup_{i \in \mathbb{N}_0} T_i$ is countable and dense in T.

In each case the starting point is inequality derived from <23> (with K_0 equal to 1): for $0 \le k < m$,

\E@ pack.bnd2 <30>
$$\mathcal{M}_{k,m} := \max_{t \in T_m} |X(t) - X(L_k t)| \le \sum_{i=k+1}^{m} \delta_{i-1} \mathbb{M}_i.$$

As m increases to ∞ the set T_m expands up to T_∞, which implies that $\mathcal{M}_{k,m}$ increases to $\mathcal{M}_{k,\infty} := \sup_{t \in T_\infty} |X(t) - X(L_k t)|$. Probabilistic bounds for $\mathcal{M}_{k,\infty}$ can usually be obtained by taking limits from bounds for $\mathcal{M}_{k,m}$.

The following four subsections present four types of chaining argument, to illustrate the different ways that packing numbers lead to different sorts of probability bound. For the first subsection Ψ can be any orlicz function; for the other three Ψ is assumed to belong to $\mathcal{Y}_{\exp}$, which makes it easier to simplify some products. The fourth subsection attacks the complication mentioned at the start of Section 10.3, the mystery of how to choose the sequences involved in bounding tail probabilities using a union bound. The other three subsections use the maximal inequalities from Theorem <10>, which I'll refer to as (orlicz.i), (orlicz.ii), and (orlicz.iii).

10.5.1 Expected values

Chaining::expected.chain

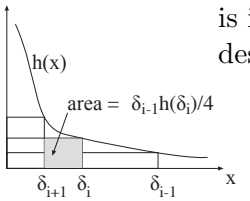
Take expected values of both sides of <30>, invoking (orlicz.i) for $\mathbb{P}\mathbb{M}_i$ to get

$$\mathbb{P}\mathcal{M}_{k,m} \le \sum_{i=k+1}^{m} \delta_{i-1} \Psi^{-1}(N_i) \le \sum_{i=k+1}^{\infty} \delta_{i-1} \Psi^{-1}(\text{PACK}(\delta_i, T)).$$

Invoke Monotone Convergence as m increases to deduce that

$$\mathbb{P}\mathcal{M}_{k,\infty} := \mathbb{P} \sup_{t \in T_\infty} |X(t) - X(L_k t)| \le \sum_{i=k+1}^{\infty} \delta_{i-1} \Psi^{-1}(\text{PACK}(\delta_i, T)).$$

It is traditional to exploit the geometric rate of decrease of the δ_i's to bound such a sum by an integral. If the function $h(x) = \Psi^{-1}(\text{PACK}(x, T))$



is integrable on the interval $(0, \operatorname{diam}(T)]$ and $\delta_i = D/2^i$ then the method described in Section 2.7, gives

$$\mathbb{P}\mathcal{M}_{k,\infty} \le 4J(\delta_{k+1}) := 4\int_0^{\delta_{k+1}} \Psi^{-1}\left(\text{PACK}(x,T)\right)\,dx.$$

Remark. For simplicity, the picture represents h as smoothly decreasing even though it is a decreasing step function. However, in practice, the step function $\text{PACK}(\,\cdot\,,T)$ is usually not known exactly but, rather, is bounded from above by a smooth function.

10.5.2 Orlicz norms

Chaining::Orlicz.chain

Suppose $\Psi \in \mathcal{Y}_{\exp}$. This time take Ψ-norms of both sides of <30>, invoking (orlicz.iii) for $\|M_i\|_\Psi$ to get

$$\|\mathcal{M}_{k,m}\|_\Psi \le C_0 \sum_{i=k+1}^{m} \delta_{i-1}\Psi^{-1}(N_i).$$

Then take a limit, as in Section 10.5.1, to deduce that

$$\|\mathcal{M}_{k,\infty}\|_\Psi \le C_1 \int_0^{\delta_{k+1}} \Psi^{-1}\left(\text{PACK}(x,T)\right)\,dx$$

for a constant C_1 (depending only on Ψ and K_0).

Remark. For processes with subgaussian increments the inequality $\|W\|_p \le C_p \|W\|_{\Psi_2}$, for $p \ge 1$, leads to upper bounds for $\|\mathcal{M}_{k,\infty}\|_p$ similar to those derived by Pollard (1989, Section 3). Those bounds became the main technical tool for the cube-root asymptotic theory developed by Kim and Pollard (1990).

10.5.3 Conditional expected values

haining::conditional.chain

The bound based on (orlicz.ii) is the most surprising. It uses a delightful trick (apparently due to Fernique, 1983) that I learned from Ledoux and Talagrand (1991, Section 11.1). For a fixed event B with $\mathbb{P}B > 0$, the argument parallels the one for expected values until we get to

$$\mathbb{P}_B\mathcal{M}_{k,\infty} \le \sum_{i=k+1}^{\infty} \delta_{i-1}\Psi^{-1}(N_i/\mathbb{P}B) \qquad \text{with } N_i \le \text{PACK}(\delta_i, T).$$

To disentangle the N_i from the $\mathbb{P}B$ use <9> to split the $\Psi^{-1}(N_i/\mathbb{P}B)$:

$$\begin{aligned}\mathbb{P}_B\mathcal{M}_{k,\infty} &\le C_0 \sum_{i=k+1}^{\infty} \delta_{i-1}\Psi^{-1}(N_i) + C_0\Psi^{-1}(1/\mathbb{P}B)\sum_{i=k+1}^{\infty}\delta_{i-1}\\ &\le 4CJ(\delta_{k+1}) + 4C\delta_{k+1}\Psi^{-1}(1/\mathbb{P}B),\end{aligned}$$

with $J(r) = \int_0^r \Psi^{-1}(\text{PACK})x, T)\,dx$, as in Section 10.5.1.

Now comes the very clever part. Choose

$$B = \{\mathcal{M}_{k,\infty} \ge 4CJ(\delta_{k+1}) + 4C\delta_{k+1}r\} \qquad \text{with } r > 0,$$

so that $\mathbb{P}_B\mathcal{M}_{k,\infty} \geq 4CJ(\delta_{k+1})+4C\delta_{k+1}r$. It then follows that $r \leq \Psi^{-1}(1/\mathbb{P}B)$, which rearranges to

\E@ tail-by-condit <31>
$$\mathbb{P}\{\mathcal{M}_{k,\infty} \geq 4CJ(\delta_{k+1}) + 4C\delta_{k+1}r\} \leq 1/\Psi(r).$$

Remark. For the subgaussian case we have $1/\Psi_2(r) \leq 2\exp(-r^2/2)$ if $r \geq \sqrt{\log 2} \approx 0.83$. Thus <31> is a subgaussian tail bound for deviations above $4CJ(\delta_{k+1})$, which is an upper bound for $\mathbb{P}\mathcal{M}_{k,\infty}$.

10.5.4 Tail probabilities

Chaining::pack.tail

Finally I come to the chaining based directly on tail probabilities, the method mentioned (with an oblique warning about the difficulties of interpreting bounds derived by trial and error) at the start of Section 10.3.

Inequality <30> again provides the starting point. Suppose $\{\eta_i\}$ is a sequence of strictly positive numbers. Define $G_i := \{\mathbb{M}_i \leq \eta_i\}$. On the set $\cap_{i=k+1}^m G_i$ we have

$$|X(t) - X(L_k t)| \leq \sum\nolimits_{i=k+1}^m \delta_{i-1}\eta_i \qquad \text{for } t \in T_m;$$

and a union bound based on inequality <5> gives $\mathbb{P}G_i^c \leq 2N_i e^{-g(\eta_i)}$. Thus

\E@ tail1 <32>
$$\mathbb{P}\{\mathcal{M}_{k,m} > \sum\nolimits_{i=k+1}^m \delta_{i-1}\eta_i\} \leq \sum\nolimits_{i=k+1}^m \mathbb{P}G_i^c \leq 2\sum\nolimits_{i=k+1}^m N_i e^{-g(\eta_i)}.$$

So much for the easy bit.

Now we have to figure out a good choice for the η_i's. If I hadn't fixed the choice $\delta_i = \operatorname{diam}(T)/2^i$ then we would be facing something like an optimization over a pair of sequences.

Consider first the special case where $\Psi_1(x) := e^x - 1$ and $g(x) = x$. Inequality <32> is not of much use unless $\sum_{i\geq k+1} \exp(\log N_i - \eta_i)$ is small. That suggests that η_i should be big enough to kill off the $\log N_i$, leaving a remainder γ_i for which $\sum_{i\geq k+1} e^{-\gamma_i}$ is small. If we replaced $\log(N_i)$ by the slightly larger $\log(1 + N_i) = \Psi_1^{-1}(N_i)$ then we would have a contribution $\sum_{i=k+1}^m \delta_{i-1}\Psi_1^{-1}(N_i)$ to the sum inside the left side of <32>, an expression that Section 10.5.1 identified as an upper bound for $\mathbb{P}\mathcal{M}_{k,m}$. It seems that we might be headed for a bound like inequality <31> for deviations beyond the expected value. That suggests we choose $\eta_i = \Psi_1^{-1}(N_i) + \gamma_i + r$, with γ_i large enough to make both $\sum_i e^{-\gamma_i}$ and $\sum_{i\geq k+1} \delta_{i-1}\gamma_i$ not too large. For example, the choice $\gamma_i = \log(1/\delta_i)$ would result in sums not much larger than their first terms. The r would contribute a $2\delta_k r$ to the left side and $2e^{-r}$ to the right side of the tail bound. For universal constants c_1 and c_2 we would have

$$\mathbb{P}\{\mathcal{M}_{k,m} \geq \sum\nolimits_{i=k+1}^m \delta_{i-1}\Psi_1^{-1}(N_i) + c_1\delta_k \log(1/\delta_k) + 2\delta_k r\} \leq c_2 e^{-r},$$

which looks a lot like <31>. In fact, if $\text{PACK}(\delta, T)$ were increasing at least like $1/\delta$, which usually happens in nontrivial cases, the contribution of $\{\gamma_i\}$ to $\sum_i \delta_{i-1}\eta_i$ could be absorbed into the $\Psi_1^{-1}(N_i)$ contribution.

The lessons I draw from the special case are: the main role of η_i is to kill off the N_i from the union bound; and the little bit extra needed to keep the right-hand side under control can be tacked on to (or tacked off from?) the N_i. Here is one way to apply these lessons for a more general g: choose η_i so that

$$g(\eta_i) = \log(1 + N_i/\delta_i) + g(r).$$

Take g^{-1} of both sides, invoking inequality <7>:

$$\eta_i \le C_g^{-1}\left(\log(1 + N_i/\delta_i)\right) + C_g r = C_g \Psi^{-1}(N_i/\delta_i) + C_g r$$

and

$$\begin{aligned}\mathbb{P}\{\mathcal{M}_{k,m} \ge C_g \textstyle\sum_{i=k+1}^{m} \delta_{i-1}\Psi^{-1}(N_i/\delta_i) + 2C_g\delta_k r\} \\ \le 2\textstyle\sum_{i=k+1}^{m} \dfrac{N_i}{1 + N_i/\delta_i} e^{-g(r)} \le 2\delta_k e^{-g(r)}.\end{aligned}$$

Again we have a bound that involves deviations beyond a multiple of $\mathbb{P}\mathcal{M}_{k,m}$ with a tail that decreases like the tail for a single random variable with Ψ-norm at most 1. (And if $g(x) = x^\alpha$ with $\alpha \ge 1$ we would have $C_g = 1$ and $x_0 = 0$.)

You might ask why I went to so much trouble with the tail probabilities just to get an inequality comparable with that obtainable using conditional expectations, as in Section 10.5.3. The reason is that the first three chaining methods depend on the existence of a deterministic upper bound, δ_{i-1}, for the link lengths $\max_{s\in S_i} d(s, \ell_{i-1}s)$; it was essential that $\mathbb{P}$, $\mathbb{P}_B$, and $\|\cdot\|_\Psi$ could be moved inside the sum in inequality <30> to act directly on the $\mathbb{M}_i$. Chaining with tail probabilities would also work without the uniform bound on the link lengths, which will be essential for Talagrand's modification of the chaining idea.

10.6 Oscillation and continuity of sample paths

Chaining::S:oscF

As in Section 10.1, suppose $X = \{X_t : t \in T\}$ is a stochastic process whose index set is a totally bounded metric space (T, d). Remember that a sample path X_ω is d-uniformly continuous if

$$\begin{aligned}\textsc{osc}(\delta, X_\omega, T) &:= \sup\{|X(\omega, s) - X(\omega, t)| : s, t \in T \text{ and } d(s, t) < \delta\} \\ &\to 0 \qquad \text{as } \delta \to 0.\end{aligned}$$

If the process is doob-separable (or it is replaced by a doob-separable version) then we can replace T by any countable, dense subset T_∞ of T.

To control the oscillation over T_∞ it is enough to show that to each $\epsilon > 0$ there exists a $\delta > 0$ such that $\mathbb{P}\textsc{osc}(\delta, X, T_\infty) \le \epsilon$ Indeed, it is then just a matter of finding the $\{\delta_j\}$ correponding to $\epsilon_j = 2^{-j}$ then deducing from $\sum_j \mathbb{P}\textsc{osc}(\delta_j, X, T_\infty) < \infty$ that $\sum_j \textsc{osc}(\delta_j, X_\omega, T_\infty) < \infty$ for almost all ω. If

you wanted a version with all (not just almost all) sample paths uniformly continuous then you could redefine X_ω to be the zero function on a suitable negligible set.

If $\{F_j\}$ is a sequence of finite subsets of T that increase to the countable, dense T_∞ then $\text{osc}(\delta, X_\omega, F_j)$ increases to $\text{osc}(\delta, X_\omega, T_\infty)$ as $j \to \infty$. Thus it suffices to show that for each $\epsilon > 0$ there exists a $\delta > 0$ such that

\E@ osc.F <33> $$\mathbb{P}\text{osc}(\delta, X, F) \le \epsilon \qquad \text{for every finite subset } F \text{ of } T.$$

That is, everything comes down to proving an inequality that holds uniformly over all finite subsets of T.

Remark. Almost all of the stochastic process maximal inequalities that I know about seem to have this form: some sort of limit of inequalities involving finitely many random variables, $\{X_t : t \in F\}$ where the size of the index set F does not enter the bounds.

Instead of hard-coding in assumptions about packing numbers and chaining frameworks I think it preferable to start from a more primitive requirement, which captures the idea that the behavior of the process is controlled by an approximation property for finite subsets of T.

Chaining::PP.osc <34> **Theorem.** *Let $\{X_t : t \in T\}$ be a stochastic process indexed by a metric space (T, d) for which:*

(i) $\|X_s - X_t\|_\Psi \le K_0 d(s,t)$ *for all s, t in T, for some orlicz norm.*

(ii) For each $\epsilon > 0$ there exists a finite $N(\epsilon)$ such that: to each finite subset F of T there corresponds an $F_0 \subset T$ with $|F_0| \le N(\epsilon)$ and a map $\gamma : F \to F_0$ such that

$$\mathbb{P}\mathcal{M}_F \le \epsilon \qquad \text{where } \mathcal{M}_F := \max_{s \in F} |X(s) - X(\gamma s)|.$$

Then there is a version of the process with d-uniformly continuous sample paths.

Remark. Problem [3] shows how a typical chaining argument leads to approximation property (ii).

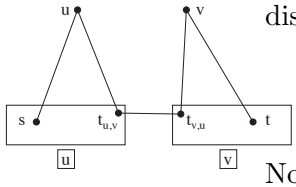
Proof. Fix an $\epsilon > 0$ and a finite F. Choose F_0 as in (ii).

For each u in F_0 define $\boxed{u} := \gamma^{-1}(u) = \{s \in F : \gamma s = u\}$. The sets $\{\boxed{u} : u \in F_0\}$ partition F into at most $N := N(\epsilon)$ equivalence classes. For each distinct pair u, v in F_0 choose points $t_{u,v} \in \boxed{u}$ and $t_{v,u} \in \boxed{v}$ such that

$$d(\boxed{u}, \boxed{v}) := \min(d(s,t) : s \in \boxed{u}, t \in \boxed{v}\} = d(t_{u,v}, t_{v,u}).$$

Define $\mathbb{M}_\square := \max_{u \ne v} |X(t_{u,v}) - X(t_{v,u})|/d(t_{u,v}, t_{v,u})$, a maximum of at most $n_\epsilon := \binom{N}{2}$ standardized increments. By inequality <5>, we have

$$\mathbb{P}\mathbb{M}_\square \le \Psi^{-1}(n_\epsilon).$$



Consider a pair $s, t \in T_m$ with $d(s,t) < \delta$. If s, t belong to the same $\boxed{\text{u}}$ then $|X_s - X_t| \le |X_s - X_u| + |X_u - X_t| \le 2\mathcal{M}_F$. If $s \in \boxed{\text{u}}$ and $t \in \boxed{\text{v}}$ for distinct u and v then

$$\begin{aligned}|X(s) - X(t)| &\le |X(s) - X(t_{u,v})| + |X(t_{u,v}) - X(t_{v,u})| + |X(t_{v,u}) - X(t)| \\ &\le 2\mathcal{M}_F + d(t_{u,v}, t_{v,u})\mathbb{M}_\square + 2\mathcal{M}_F.\end{aligned}$$

Note that $d(t_{u,v}, t_{v,u}) = d(\boxed{\text{u}}, \boxed{\text{v}}) < \delta$ if $s \in \boxed{\text{u}}$ and $t \in \boxed{\text{v}}$ and $d(s,t) < \delta$. Take a maximum over all pairs $\{s,t\}$ with $d(s,t) < \delta$ to deduce that

\E@ oscF.bnd <35>

$$\mathbb{P}\text{OSC}(\delta, X, F) \le 4\mathbb{P}\mathcal{M}_F + \delta\mathbb{P}\mathbb{M}_\square \le 4\epsilon + \delta\Psi^{-1}(n_\epsilon),$$

which is $\le 5\epsilon$ if δ is chosen small enough. □

Remark. If analogous inequalities hold for a whole sequence of processes $\{X_n(t,\omega) : t \in T\}$ with the same δ and ϵ for each n (or, sometimes, just for all n large enough) then we have an example of ***stochastic equicontinuity***. In its tail probability version, this condition asserts: for each $\eta > 0$ and $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\limsup_n \mathbb{P}\{\text{OSC}(\delta, X_n, T) > \eta\} < \epsilon.$$

Such a property plays a central role in the theory of convergence in distribution of random processes. See Chapter 14 for details.

*10.7 An example of a classical chaining argument

Chaining::S:classical

The chaining arguments used to derive maximal inequalities in well-known texts, such as Billingsley (1968, Sections 12, 13), usually involved the index set $[0,1]$. They focussed on processes with sample paths in the space $C[0,1]$ of continuous real functions or the space $D[0,1]$ of right-continuous functions with left limits at each point of $(0,1]$, the famous ***cadlag*** functions. It was natural to exploit the ordering of the index set, as in the notion of a limit from the left or from the right. However, particularly for the theory of empirical processes, the reliance on ordering properties leads into a chaining cul-de-sac when one seeks generalizations to more complicated index sets.

Remark. Unfortunately, the empirical distribution function F_n for a distribution such as the Unif$[0,1]$ is not measurable with respect to the sigma-field generated by the uniform metric on $D[0,1]$. However, F_n is measurable with respect to various sigma-fields defined by Skorohod (1956). He actually defined five different ways for a sequence functions x_n in $D[0,1]$ to converge to a limit function x. Each involved different ways to handle behavior near an index point t at which x has a discontinuity.

The difficulties involved in efforts to extend Skorohod's approach to more general index sets greatly complicated the early attempts to develop a general empirical process theory. It took many years for this measurability obstacle to be removed from the general theory. See Chapter 14 for some of the details.

Even more importantly from my point of view, the special features of $[0,1]$ allowed for some clever tricks that, to me, made chaining seem more mysterious than it really was. To illustrate this point I'll reinterpret the presentation by Parthasarathy (1967, page 216) of a technique apparently first introduced by Kolmogorov (see Notes). The subject is a stochastic process $\{X_t : 0 \le t \le 1\}$ for which

\E@ incr.moment <36>
$$\mathbb{P}|X_s - X_t|^\alpha \le d(s,t)^{1+r} \qquad \text{for all } s,t \in [0,1],$$

where $d(s,t) = |s-t|_2$ (the usual Eucliden distance) and both α and r are strictly positive constants. The aim is to show that there is a version of the process with continuous sample paths.

The moment assumption <36> implies a tail bound,

\E@ incr.bnd <37>
$$\mathbb{P}\{|X_s - X_t| \ge \eta\} \le d(s,t)^{1+r}/\eta^\alpha \qquad \text{for } \eta > 0.$$

Note that neither <36> nor <37> fits neatly into either of my schemes <1> or <2> for how a metric should control increments. Nevertheless, chaining arguments very similar to those in Section 10.5 will provide the necessary maximal inequalities.

The dyadic rationals provide a very natural chaining framework for $[0,1]$, namely $T_0 = \{0\}$ and $T_i = \{j/2^i : j = 0, 1, \ldots, 2^i\}$, with ℓ_{i-1} the map that rounds down to an integer multiple of $(1/2)^{i-1}$ when $t \in S_i = T_i \backslash T_{i-1}$. Thus $d(t, \ell_{i-1}t) \le \delta_i := (1/2)^i$ for each t in T_i. Instead of working with the scaled maxima $\mathbb{M}_i$, as in Section 10.5, this time use the unscaled random variables $M_i := \max_{s \in S_i} |X(s) - X(\ell_{i-1}s)|$, so that

$$\mathcal{M}_{k,m} := \max_{t \in T_m} |X(t) - X(L_k t)| \le \sum_{i=k+1}^m M_i.$$

Inequality <37> gives

$$\mathbb{P}\{M_i \ge \eta_i\} \le \sum_{s \in S_i} \mathbb{P}\{|X_s - X_{\ell_{i-1}s}| \ge \eta_i\} \le 2^i \delta_i^{1+r}/\eta_i^\alpha = \delta_i^r/\eta_i^\alpha.$$

Notice that 2^i is effectively the covering number for a δ_i-approximation. (My calculation differs a little from Parthsarathy's at this point: he was working with a continuous approximation obtained by linear interpolation between the values $\{X_s : s \in T_i\}$.) The 1 in the $1 + r$ exponent has been used to cancel out the size of the set T_i, a slightly subtle idea. Even more subtle is the choice $\eta_i = \delta_i^\theta$ where $0 < \theta < r/\alpha$ so that an analog of my inequality <32> gives

$$\mathbb{P}\{\mathcal{M}_{k,m} > \sum_{i=k+1}^m C\delta_i^\theta\} \le C \sum_{i=k+1}^m \delta_i^{r-\theta\alpha}.$$

The rest of the argument could then follow the method described in Section 10.6.

10.8 Problems

Chaining::S:Problems

Chaining::P:indep.incr

[1] Suppose $Y_1, \ldots, Y_n$ are independent random variables. For a fixed $y > 0$ define $M = \mathbb{P}\{\max_{i\le n} Y_i \ge y\}$ and $S = \sum_{i\le n} \mathbb{P}\{Y_i \ge y\}$. Show that

$$S \ge M \ge 1 - e^{-S} = S + o(S)$$

The union bound can be quite sharp when dealing with tails of independent random variables.

Chaining::P:chain.cost

[2] This problem provides some insight into the cost of chaining, using the heuristics from Section 10.5.4. Suppose $\{B_t : 0 \le t \le 1\}$ is a standard Brownian motion process. The increment $B_1 - B_0$ has a $N(0,1)$ distribution. The usual subgaussian tail bound gives $\mathbb{P}\{|B_1 - B_0| > r\} \le 2\exp(-r^2/2)$ for $r \in \mathbb{R}^+$. What sort of bound can we get using

$$|B(t_m) - B(t_0)| \le \sum_{i=1}^m |B(t_i) - B(t_{i-1})| \qquad \text{where } t_i = 2^{-i}?$$

(i) The natural metric is $d(s,t) = \|B_s - B_t\|_2 = \sqrt{|s-t|}$. Show that the increment $B(t_i) - B(t_{i-1})$ has a $N(0, \delta_{i-1}^2)$, where $\delta_i = \sqrt{(1/2)^{i+1}}$.

(ii) Define $\Delta_i := |B(t_i) - B(t_{i-1})|/\delta_{i-1}$. Show that $\mathbb{P}\{\Delta_i > \eta_i\} \le 2\exp(-\eta_i^2/2)$ for $\eta_i \in \mathbb{R}^+$.

(iii) The heuristic from Section 10.5.4 suggests we choose η_i so that $\eta_i^2/2 = r^2/2 + \gamma_i$, with $\{\gamma_i\}$ increasing rapidly enough to make $\sum_{i\in\mathbb{N}} \exp(-\gamma_i) < \infty$. Show that this choice leads to

$$\mathbb{P}\{|B_1 - B_0| > \sum\nolimits_{i\in\mathbb{N}} \delta_{i-1}(r + \sqrt{2\gamma_i})\} \le 2\sum\nolimits_{i\in\mathbb{N}} e^{-\gamma_i} e^{-r^2/2}.$$

Experiment with some γ_i values. With $\gamma_i = \log(i\log(2+i)^2)$ I got

$$\sum\nolimits_{i\in\mathbb{N}} \delta_{i-1} \approx 2.4 \qquad \sum\nolimits_{i\in\mathbb{N}} \delta_{i-1}\sqrt{2\gamma_i} \approx 4.2 \qquad \sum\nolimits_{i\in\mathbb{N}} 2e^{-\gamma_i} \approx 3.7.$$

Chaining::P:FapproxPP

[3] As in Section 10.3, suppose $\{X_t : t \in T\}$ is a stochastic process and $\Psi = e^g - 1$ is an orlicz function for which:

(a) $\|X_s - X_t\|_\Psi \le K_0 d(s,t)$ for all $s, t \in T$.

(b) $\{(T_i, \ell_i) : i \in \mathbb{N}_0\}$ is a chaining framework, not necessarily nested, on T for which to each $\epsilon > 0$ there exists a finite $k(\epsilon)$ such that $\mathbb{P}\mathcal{M}_{k,m} \le \epsilon$ if $m > k \ge k(\epsilon)$.

Show that there is a finite $N(\epsilon)$ for which: to each finite subset F of T there exists an $F_0 \subset T$ with $|F_0| \le N(\epsilon)$ and a map $\gamma : F \to F_0$ such that

$$\mathbb{P}\max_{s\in F} |X(s) - X(\gamma s)| \le \epsilon.$$

Argue as follows. Consider any fixed, finite F.

(i) Deduce from Definition <16>(i) that $\delta_{m,F} := \max\{d(s,T_m) : s \in F\} \to 0$ as $m \to \infty$. Choose $\pi_m : F \to T_m$ so that $d(s,\pi_m s) = \delta_{m,F}$. Define $\gamma := L_k \circ \pi_m$

(ii) For $m > k \geq k(\epsilon/2)$ show that

$$\begin{aligned}\mathbb{P}\max_{s\in F}|X(s)-X(\gamma s)| &\leq \mathbb{P}\max_{s\in F}|X(s)-X(\pi_m s)| + \mathbb{P}\mathcal{M}_{k,m}\\ &\leq K_0\delta_{m,F}\Psi^{-1}(|F|)+\epsilon/2,\end{aligned}$$

which is smaller than ϵ if m is large enough.

Chaining::P:Hellinger [4] The following statistical example is based loosely on results from Ibragimov and Has'minskii (1981, Section 1.5). It involves a set of probability measures $\{P_\theta : 0 \leq \theta \leq 1\}$ all defined on the same $(\mathcal{X},\mathcal{F})$, with densities $p(x,\theta) = dP_\theta/d\mu$ for some dominating measure μ. For simplicity assume $\theta \mapsto p(x,\theta)$ is continuous. Define $\mathbb{P}_{\theta,n} = P_\theta^n$, the probability measure on the product sigma-field on $\mathcal{X}^n$ under which the coordinate maps $x_1,\ldots,x_n$ represent independent observations from P_θ. As a regularity condition assume existence of positive constants C_i for which

\E@ euc.hell <38>
$$C_1|\theta-\theta'|_2 \leq h(P_\theta,P_{\theta'}) \leq C_2|\theta-\theta'|_2 \qquad \text{for all } \theta,\theta' \in [0,1],$$

where $h(\theta,\theta') = \left(\mu|\sqrt{p(x,\theta)}-\sqrt{p(x,\theta')}|\right)^{1/2}$ denotes the Hellinger distance between P_θ and $P_{\theta'}$. See Pollard (2001, Problem 4.18) for useful properties of Hellinger distance, such as $h^2(P^n,Q^n) \leq nh^2(P,Q)$.

The maximum likelihood estimator $\widehat{\theta}_n$ is defined to maximize the likelihood function

$$\mathcal{L}_n(\theta) = \mathcal{L}_n(\theta,x_1,\ldots,x_n) := \prod_{i\leq n} p(x_i,\theta).$$

Show that there exists constants K_1 and K_2 for which

$$\mathbb{P}_{\theta_0,n}\{\sqrt{n}|\widehat{\theta}_n-\theta_0| \geq y\} \leq K_1\exp(-K_2y^2) \qquad \text{all } y \geq 0, \text{ all } \theta_0 \in [0,1], \text{ all } n.$$

Argue as follows. (If issues of uniqueness worry you, feel free to generalize.) To simplify notation, for a fixed θ_0 abbreviate $\mathbb{P}_{\theta_0,n}$ to $\mathbb{P}$, write $\widehat{t}_n$ for $\sqrt{n}(\widehat{\theta}_n-\theta_0)$, and define

$$\xi(z,t) = \sqrt{p(z,\theta_0+t/\sqrt{n})/p(z,\theta_0)}\,.$$

(Add some indicator functions if you worry about $p(z,\theta_0) = 0$.)

(i) Show that $\widehat{t}_n$ maximizes the process

$$Z_n(t) := \sqrt{\mathcal{L}_n(\theta_0+t/\sqrt{n})/\mathcal{L}_n(\theta_0)} = \prod_{i\leq n}\xi(x_i,t).$$

(ii) Show that $\mathbb{P}|Z_n(t_1)-Z_n(t_2)|^2 \leq C_2^2|t_1-t_2|^2$ and $\mathbb{P}Z_n(t) \leq \exp(-\frac{1}{2}C_1^2t^2)$.

(iii) Use the fact that $Z_n(\widehat{t}_n) \geq Z_n(0) = 1$ to show that

$$\mathbb{P}\{|\widehat{t}_n| \geq y\} \leq \mathbb{P}\{\sup_{|t|\geq y} Z_n(t) \geq 1\} \leq \mathbb{P} \sup_{|t|\geq y} Z_n(t).$$

(iv) Split the set $\{t : |t| \geq y \text{ and } \theta_0 + t/\sqrt{n} \in [0,1]\}$ into a union of intervals like $J = [y_k, y_k + 1]$ where $y_k = y + k$ for an integer k. For a suitable δ, let S_J be a δ-packing set. Use the chaining bound from Section 10.5.1 to show that

$$\mathbb{P} \sup_{t\in J} Z_n(t) \leq (1/\delta)e^{-1/2 C_1^2 y_k^2} + C_2\sqrt{8}\delta^{1/2}.$$

Minimize over δ.

(v) Sum over intervals like J to get a suitable bound for $\mathbb{P} \sup_{|t|\geq y} Z_n(t)$.

10.9 Notes

Chaining::S:Notes

Credit for the idea of chaining as a method of successive approximations seems to belong to Kolmogorov, at least for the case of a one-dimensional index set. For example, the start of the paper of Chentsov (1956) commented that "In 1934 A. N. Kolmogorov proved [the result described in my Section 10.7] ... A generalization of this theorem is the following proposition which was suggested to the author by A. N. Kolmogorov". He added the footnote: "This theorem was first published in a paper by E. E. Slutskii" (=Slutsky, 1937), a paper that I have not seen. See Billingsley (1968, Section 12) for a small generalization—with credit to Kolmogorov, via Slutsky, and Chentsov—and a chaining proof.

The Billingsley book was my main source of information when I first learned about stochastic processes with well behaved sample paths. It prepared me for the study of general empirical processes, which I first encountered in the ground-breaking paper of Dudley (1978). To grapple with the host of ideas in that paper I needed to study the earlier work by Dudley (1967, 1973) on gaussian processes, which became my textbooks for general chaining methods. Only later did I stumble on the French work expounded so convincingly by Ledoux and Talagrand (1991).

Dudley (1973) used chaining with covering numbers and exponential tail bounds to establish various probabilistic bounds for gaussian processes. Dudley (1978) adapted the methods using Bernstein's inequality for the increments and metric entropy plus inclusion assumptions (now called bracketing—see Chapter 19) to extend the gaussian techniques to empirical processes indexed by collections of sets. He also derived bounds for processes indexed by VC classes of sets (see Chapter 15) via symmetrization (see Chapter 13) arguments.

Dudley's 1978 paper became the basis for his famous St. Flour lectures (Dudley, 1984), which were widely circulated in note form. Those notes turned into the 1999 first edition of the definitive Dudley (2014) text. See Dudley (1973, Section 1) and Dudley (2014, Section 1.2 and Notes) for more about packing and covering.

Dudley (2016) took pains to give V.N. Sudakov credit for an earlier use of covering numbers to bound the expected supremum of a Gaussian process. (A scholar less scrupulous than Dudley might have been contented to accept Sudakov's generous comment that "The idea of using ϵ-entropy characteristics here is due independently to several authors.")

Pisier (1983) is usually credited for realizing that the entropy methods used for Gaussian processes could also be extended to nongaussian processes with orlicz norm control of the increments. However, as Pisier (page 127) remarked:

> For the proof of this theorem, we follow essentially [10]; I have included a slight improvement over [10] which was kindly pointed out to me by X. Fernique. Moreover, I should mention that N. Kôno [6] proved a result which is very close to the above; at the time of [10], I was not aware of Kôno's paper [6].

Here [10] = Pisier (1980) and [6] = Kôno (1980). The earlier paper [10] included extensive discussion of other precursors for the idea. See also the Notes to Section 2.8 of Dudley (2014).

Using essentially the method in my Section 10.5.1, Pisier (1983) proved existence of continuous versions of stochastic processes $\{X_t : t \in T\}$ indexed by a (semi-)metric space for which $\|X_s - X_t\|_\Psi \le d(s,t)$. Under an integral condition on covering numbers for T he derived a bound Pisier (1980), for the case $\Psi(x) = x^p$, following comments (Pisier, 1983, page 124) regarding that possibility by Fernique. In his July 1981 Saint-Flour lectures, Fernique (1983) gave his own proofs of Pisier's result and generalizations.

I learned the idea behind the proof of Theorem <34> from Ledoux and Talagrand (1991, page 306).

Using methods like those in Section 10.5, Nolan and Pollard (1988) proved a functional central limit for the U-statistic analog of the empirical process. Kim and Pollard (1990) and Pollard (1990) proved limit theorems for a variety of statistical estimators using second moment control for suprema of empirical processes. See also Pollard (1985) for one way to use a form of oscillation bound (under the name *stochastic differentiability*) to establish central limit theorems for M-estimators. Pakes and Pollard (1989, Lemma 2.17) used a property more easily recognized as oscillation around a fixed index point.

My analysis in Problem [4] is based on arguments of Ibragimov and Has'minskii (1981, Section 1.5), with the chaining bound replacing their method for deriving maximal inequalities. The analysis could be extended to unbounded subsets of $\mathbb{R}$ by similar adaptations of their arguments for unbounded sets.

References

Billingsley68book

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

Chentsov1956TPA Chentsov, N. N. (1956). Weak convergence of stochastic processes whose trajectories have no discontinuities of the second kind and the "heuristic" approach to the Kolmogorov-Smirnov tests. *Theory of Probability and Its Applications 1*(1), 140–144.

Doob53book Doob, J. L. (1953). *Stochastic Processes*. New York: Wiley.

Dudley1967JFnAl Dudley, R. M. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis 1*(3), 290–330.

Dudley73gauss Dudley, R. M. (1973). Sample functions of the Gaussian process. *Annals of Probability 1*, 66–103.

Dudley78clt Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability 6*, 899–929.

Dudley84StFlour Dudley, R. M. (1984). A course on empirical processes. *Springer Lecture Notes in Mathematics 1097*, 1–142. École d'Été de Probabilités de St-Flour XII , 1982.

Dudley2003RAP Dudley, R. M. (2003). *Real Analysis and Probability* (2nd ed.), Volume 74 of *Cambridge studies in advanced mathematics*. Cambridge University Press.

Dudley2014UCLT Dudley, R. M. (2014). *Uniform Central Limit Theorems* (2nd ed.), Volume 142 of *Cambridge studies in advanced mathematics*. Cambridge University Press. (First edition, 1999).

Dudley2016sudakov Dudley, R. M. (2016). V. N. Sudakov's work on expected suprema of Gaussian processes. In *High Dimensional Probability VII*, pp. 37–43. Springer.

Fernique83StFlour Fernique, X. (1983). Regularité de fonctions aléatoires non gaussiennes. *Springer Lecture Notes in Mathematics 976*, 1–74. École d'Été de Probabilités de St-Flour XI, 1981.

bragimov:Hasminskii:81book Ibragimov, I. A. and R. Z. Has'minskii (1981). *Statistical Estimation: Asymptotic Theory*. New York: Springer. (English translation from 1979 Russian edition).

KimPollard90cuberoot Kim, J. and D. Pollard (1990). Cube root asymptotics. *Annals of Statistics 18*, 191–219.

KolmogorovTikhomirov1959 Kolmogorov, A. N. and V. M. Tikhomirov (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Mat. Nauk 14*(2 (86)), 3–86. Review by G. G. Lorentz at MathSciNet MR0112032. Included as paper 7 in volume 3 of Selected Works of A. N. Kolmogorov.

Kono1980JMKY Kôno, N. (1980). Sample path properties of stochastic processes. *J. Math. Kyoto Univ. 20*(2), 295–313.

LedouxTalagrand91book Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. New York: Springer.

NolanPollard88Uproc2 Nolan, D. and D. Pollard (1988). Functional limit theorems for U-processes. *Annals of Probability 16*, 1291–1298.

PakesPollard89simulation Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica 57*, 1027–1058.

Parthasarathy67book Parthasarathy, K. R. (1967). *Probability Measures on Metric Spaces*. New York: Academic.

Pisier7980 Pisier, G. (1980). Conditions d'entropie assurant la continuité de certains processus et applications à l'analyse harmonique. In *Séminaire d'analyse fonctionnelle, 1979-80*, pp. 1–41. École Polytechnique Palaiseau. Available from http://archive.numdam.org/.

Pisier83metricEntropy Pisier, G. (1983). Some applications of the metric entropy condition to harmonic analysis. *Springer Lecture Notes in Mathematics 995*, 123–154. (A collection of lecture notes from the 1980-81 Special Year in Analysis at the University of Connecticut Department of Mathematics).

Pollard85NewWays Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory 1*, 295–314.

Pollard89StatSci Pollard, D. (1989). Asymptotics via empirical processes (with discussion). *Statistical Science 4*, 341–366.

Pollard90Iowa Pollard, D. (1990). *Empirical Processes: Theory and Applications*, Volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Hayward, CA: Institute of Mathematical Statistics.

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

Skorohod56metrics Skorohod, A. V. (1956). Limit theorems for stochastic processes. *Theory of Probability and Its Applications 1*, 261–290.

Slutsky1937qualche Slutsky, E. (1937). Qualche proposizione relativa alla teoria delle funzioni aleatorie. *Giorn. Ist. Ital. Attuari 8*, 183–199.

Talagrand2021MMbook Talagrand, M. (2021). *Upper and Lower Bounds for Stochastic Processes: Decomposition Theorems* (Second ed.), Volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag.

