# Maximal inequalities

Paul Delatte
delatte@usc.edu
University of Southern California

Last updated: 31 December, 2022

In many problems, we want to control the behavior (e.g. boundedness or continuity of the sample path) of stochastic processes $(X_t)_{t\in T}$ indexed by infinite sets $T$ (e.g., a continuous-time real-valued process which is indexed by some subset of the real line, an empirical process which is indexed by some set of functions), and consequently that of families $((X_t^{(i)})_{t\in T})_{i\in I}$ of stochastic processes $(X_t^{(i)})_{t\in T}$ indexed by infinite sets $T$ (e.g. a sequence of empirical processes). This generally translates into finding maximal inequalities, that is, bounds on $\sup_{t\in T} X_t$ for some stochastic process $(X_t)_{t\in T}$. Important examples of maximal inequalities for real-valued processes include: (a) bounds on $\sup_{t\in T} |X_t|$ yielding boundedness of the sample path; (b) bounds on the modulus of continuity $w(\delta) := \sup\{|X_s - X_t| : s,t \in T,\ d(s,t) < \delta\}$ yielding uniform continuity of the sample path. For instance:

(a) control over $\sup_{t\in T} |X_t^{(n)}|$ where $((X_t^{(n)})_{t\in T})_{n\in\mathbb{N}} = ((\mathbb{P}_n - \mathbb{P})f)_{f\in\mathcal{F}})_{n\in\mathbb{N}}$ yields uniform law of large numbers through symmetrization;

(b) control over $\sup\{|X_s - X_t| : s,t \in T,\ d(s,t) < \delta\}$ where $((X_t^{(n)})_{t\in T})_{n\in\mathbb{N}} = ((\sqrt{n}(\mathbb{P}_n - \mathbb{P})f)_{f\in\mathcal{F}})_{n\in\mathbb{N}}$ yields uniform limit central theorems through symmetrization and the asymptotic equicontinuity characterization of weak convergence.

It turns out that maximal inequalities for sub-Gaussian stochastic processes indexed by infinite sets (from which more general maximal inequalities can be derived) depend intimately on the regularity of the index sets. This forces us to develop finer ways to measure the size or complexity of infinite sets than the one provided by cardinality. Several solutions exist: we develop a metric one, a combinatorial one, and a probabilistic one and see how they can be used to generate maximal inequalities.

## 1 Covering, packing, and metric entropy

One of the ideas developed by the Russian school (to measure the size of infinite sets) is to approximate any infinite set by a minimal finite subset such that every point in the infinite set is close (to a given degree) to a point in the finite subset.

Recall that a metric on a set $T$ is a function $d\colon T\times T \to \mathbb{R}$ such that for all $x,y,z \in T$, $d(x,y) \geq 0$ with equality if and only $x = y$, $d(x,y) = d(y,x)$, and $d(x,y) + d(y,z) \leq d(x,z)$. A **pseudo-metric** is a metric without the requirement that $d(x,y) = 0$ implies $x = y$. (In particular, any metric space is a pseudo-metric space). Contrarily to points in a metric space, points in a pseudo-metric space need not be distinguishable. (It is easy

to see that by quotienting a pseudo-metric space by the equivalence relation induced by the vanishing of the pseudo-metric, we obtain a metric space. The metric identification preserves the induced topologies.)

**Definition 1** ($\varepsilon$-Cover)**.** Let $(T, d)$ be a pseudo-metric space, $S \subseteq T$, and $\varepsilon > 0$. A subset $A \subseteq S$ is said to be a $\boldsymbol{\varepsilon}$**-cover** of $S$ if for all $x \in S$, there exists $y \in A$ such that $d(x, y) \leq \varepsilon$.

Equivalently, $A \subseteq S$ is an $\varepsilon$-cover of $S$ if and only if $S \in \bigcup_{x \in A} \bar{B}(x, \varepsilon)$ (that is, $S$ can be covered by closed balls of radius $\varepsilon$ centered at points in $A$), hence the terminology. An $\varepsilon$-cover(ing) is also called an $\varepsilon$-net, an internal $\varepsilon$-cover, or a proper $\varepsilon$-cover. If we weaken the requirement that $A \subseteq S$ to $A \subseteq T$, then the resulting object is said to be an **external** $\boldsymbol{\varepsilon}$**-cover**. Some authors reverse the terminology and call $\varepsilon$-cover what we define as external $\varepsilon$-cover.

**Definition 2** ($\varepsilon$-Covering Number)**.** Let $(T, d)$ be a pseudo-metric space, $S \subseteq T$, and $\varepsilon > 0$. The smallest cardinality of any $\varepsilon$-cover of $S$ is said to be the $\boldsymbol{\varepsilon}$**-covering number** of $S$ and is denoted $N(S, d, \varepsilon)$. That is,

$$N(S, d, \varepsilon) = \inf\{|A| : A \text{ is an } \varepsilon\text{-cover of } S\}.$$

The covering number is thus the minimal number of closed balls centered at points in $S$ of radius $\varepsilon$ needed to cover $S$. It is an intuitive measure of the size or complexity of a set. It can be proved that the covering number is decreasing in $\varepsilon$ and in many cases diverges as $\varepsilon \to 0+$. By definition, a pseudo-metric space $(T, d)$ is totally bounded (or precompact in French) if and only if for all $\varepsilon > 0$, $N(T, d, \varepsilon) < +\infty$.

We naturally define the external covering number as $N_{ext}(S, d, \varepsilon) := \inf\{|A| : A \text{ is an external } \varepsilon\text{-cover of } S\}$, that is, the minimal number of closed balls centered at points in $T$ of radius $\varepsilon$ needed to cover $S$. By definition, $N_{ext}(S, d, \varepsilon) \leq N(S, d, \varepsilon)$, but one can prove more generally (E.4.2.9. in Vershynin HDP p.83) that

$$N_{ext}(S, d, \varepsilon) \leq N(S, d, \varepsilon) \leq N_{ext}(S, d, \varepsilon/2).$$

**Definition 3** (Metric Entropy)**.** Let $(T, d)$ be a pseudo-metric space, $S \subseteq T$, and $\varepsilon > 0$. Then $\ln N(S, d, \varepsilon)$ is said to be the ($\boldsymbol{\varepsilon}$-)**metric entropy** of $S$.

The metric entropy is alternatively called the Kolmogorov entropy. (Sometimes it is more convenient to take the logarithm in the definition in base 2. In this case, the metric entropy can be shown to be equivalent to the number of bits needed to specify any point in $S$ up to a given error – P.4.3.1. in Vershynin HDP p.86).

A dual approach to covering is to look at the maximal number points in $S$ that stand apart from each other by at least $\varepsilon$. This leads to the notion of packing numbers. The advantage of this approach is that there is only an "internal" packing number.

**Definition 4** ($\varepsilon$-Packing)**.** Let $(T, d)$ be a pseudo-metric space, $S \subseteq T$, and $\varepsilon > 0$. A subset $A \subseteq S$ is said to be an $\boldsymbol{\varepsilon}$**-packing** of $S$ if $x, y \in A$, $x \neq y$ implies that $d(x, y) > \varepsilon$.

Equivalently, $A \subseteq S$ is an $\varepsilon$-packing of $S$ if and only if $\bar{B}(x, \varepsilon) \cap \bar{B}(x', \varepsilon) = \emptyset$ for all $x \neq x' \in A$ (that is, every distinct pair in $A$ is at least separated by $\varepsilon$). An $\varepsilon$-packing is hence also called an $\varepsilon$-separated set.

**Definition 5** ($\varepsilon$-Packing Number)**.** Let $(T, d)$ be a pseudo-metric space, $S \subseteq T$, and $\varepsilon > 0$. The largest cardinality of any $\varepsilon$-packing of $S$ is said to be the **$\varepsilon$-packing number** of $S$ and is denoted $\text{pack}(S, d, \varepsilon)$. That is,

$$\text{pack}(S, d, \varepsilon) = \sup\{|A| : A \text{ is an } \varepsilon\text{-packing of } S\}.$$

In other words, the $\varepsilon$-packing number of $S$ is the maximal number points in $S$ that can be made to stand apart from each other by at least $\varepsilon$. The dual flavor of the definitions suggests that covering and packing can be used interchangeably (up to constants) as measures of the size of infinite sets; this is indeed the case as proved in next results.

**Lemma 6.** *Let $(T, d)$ be a pseudo-metric space and $S \subseteq T$. If $A \subseteq S$ is a maximal $\varepsilon$-packing of $S$, then $A$ is an $\varepsilon$-cover of $S$.*

*Proof.* L.4.2.6. in Vershynin HDP p.82. □

**Proposition 7** (**Duality of Packing and Covering**)**.** *Let $(T, d)$ be a pseudo-metric space, $S \subseteq T$, and $\varepsilon > 0$. Then*

$$N(S, d, \varepsilon) \leq \text{pack}(S, d, \varepsilon) \leq N(S, d, \varepsilon/2).$$

*Proof.* L.4.2.8. in Vershynin HDP p.83. □

Last lemma also suggests a simple algorithm to exhibit an $\varepsilon$-cover. Start with $A = \emptyset$ and add a point to $A$ at a distance at least $\varepsilon$ from all other points until it is not possible anymore. If the space is totally bounded, then the algorithm terminates.

We now introduce a few classical examples of covering and packing computations: the first relates the standard notion of volume in $\mathbb{R}^d$ (that is, the Lebesgue measure, which can be used as a measure of the size of sets) to that of covering number (since flat sets have volume zero but non-zero covering numbers, there is no strict equivalence); the second provides bounds on the covering numbers of some function spaces (illustrating as an aside the curse of dimensionality for rich enough sets). For any set, $A \subseteq \mathbb{R}^d$, we denote $\text{vol}(A) = \lambda(A)$ where $\lambda$ is the Lebesgue measure on $\mathbb{R}^d$.

**Proposition 8.** *Let $\|\cdot\|$ and $\|\cdot\|'$ be two norms on $\mathbb{R}^d$ and $B$ and $B'$ the corresponding unit balls. Then*

$$\left(\frac{1}{\varepsilon}\right)^d \frac{\text{vol}(B)}{\text{vol}(B')} \leq N(B_{\|\cdot\|}, \|\cdot\|', \varepsilon) \leq \frac{\text{vol}(\frac{2}{\varepsilon}B + B')}{\text{vol}(B')}.$$

*Proof.* L.5.7. in Wainwright HDS p.125. □

**Corollary 9** (Covering Number of Unit Ball in $\mathbb{R}^d$ in Own Norm)**.** *Let $\|\cdot\|$ be any norm on $\mathbb{R}^d$ and $B$ the corresponding unit ball. Then*

$$\left(\frac{1}{\varepsilon}\right)^d \leq N(B_{\|\cdot\|}, \|\cdot\|, \varepsilon) \leq \left(1 + \frac{2}{\varepsilon}\right)^d.$$

*Proof.* Last proposition with $\|\cdot\|' = \|\cdot\|$ (see E.5.8. in Wainwright HDS p.126.). □

Note that this covers also $\|\cdot\| = \|\cdot\|_\infty$ for which the unit ball is $B = [-1, 1]^d$. This result more generally implies that for fixed $d$, as $\varepsilon \to 0+$, the metric entropy of the unit ball $\ln N(B_{\|\cdot\|}, \|\cdot\|, \varepsilon)$ scales as $d \ln(1/\varepsilon)$.

**Proposition 10** (Covering Number of Lipschitz Function Space)**.** *Let* $\mathscr{F} = \mathrm{Lip}_K([0,1]^d)$ *for* $K > 0$ *and endow the space with the supremum norm* $\|\cdot\|_\infty$*. Then*

$$2^{(1/2\varepsilon)^d} \leq N(\mathscr{F}, \|\cdot\|_\infty, \varepsilon) \leq \left(\frac{4K}{\varepsilon} + 1\right) 2^{(1/\varepsilon)^d}$$

*Proof.* (240) in A. N. Kolmogorov, V. M. Tikhomirov "$\varepsilon$-entropy and $\varepsilon$-capacity of sets in functional spaces" (1959) in Selected Works of Kolmogorov (1993) p.163. See also E.5.10. in Wainwright HDS p.127. □

## 2 Vapnik–Chervonenkis theory

If $\mathcal{C}$ is a collection of subsets of a set $T$ and $A$ a set, let us denote $\mathcal{C} \cap A = A \cap \mathcal{C} = \{A \cap C : C \in \mathcal{C}\}$. In what follows, we should abbreviate Vapnik–Chervonenkis as VC.

**Definition 11** (Shattering)**.** Let $T$ be any set and $\mathcal{C}$ a collection of subsets of $T$. A set $A \subseteq T$ is said to be **shattered** by $\mathcal{C}$ if $\mathcal{C} \cap A = 2^A$.

In other words, $A$ is shattered by $\mathcal{C}$ if and only if for every $B \subseteq A$, there exists $C \in \mathcal{C}$ such that $A \cap C = B$, that is, we can recover each subset of $A$ by intersecting $A$ with elements of $C$. Hence, the bigger $\mathcal{C}$, the more likely it can shatter $A$: in particular, if $\mathcal{C}_1 \subseteq \mathcal{C}_2$ and $\mathcal{C}_1$ shatter $A$, then $\mathcal{C}_2$ shatter $A$. If $A$ is finite, then $A$ is shattered by $\mathcal{C}$ if and only if $|\mathcal{C} \cap A| = 2^{|A|}$ (note that we always have $|\mathcal{C} \cap A| \leq 2^{|A|}$).

**Definition 12** (VC Dimension)**.** Let $T$ be any set and $\mathcal{C}$ a collection of subsets of $T$. The quantity

$$\mathrm{vc}(\mathcal{C}) = \sup\{|A| : A \subseteq T \text{ is shattered by } \mathcal{C}\}$$

is said to be the **VC dimension** of $\mathcal{C}$.

If $\mathrm{vc}(\mathcal{C}) < +\infty$, then we say that $\mathcal{C}$ is a **VC class** for $T$.

The VC dimension of $\mathcal{C}$ is thus the cardinality of the largest set of points in $T$ for which we can recover all possible subsets of these points by intersecting the whole set of points with elements of $\mathcal{C}$ (and $+\infty$ if it is always possible). The bigger $\mathcal{C}$, the larger the set of points in $T$ that $\mathcal{C}$ can shatter, and so the bigger the VC dimension of $\mathcal{C}$: hence $\mathrm{vc}(\mathcal{C})$ provides an intuitive measure of the size or complexity of $\mathcal{C}$.

The VC dimension can be equivalently defined as the largest integer $n$ such that $|\mathcal{C} \cap \{t_1, \ldots, t_n\}| = 2^n$ for some points $t_1, \ldots, t_n \in T$ (and $+\infty$ if no such largest integer exists). This suggests a related measure of the complexity of $\mathcal{C}$.

**Definition 13** (Shattering Coefficient)**.** Let $T$ be any set and $\mathcal{C}$ a collection of subsets of $T$. Then the integers $S_{\mathcal{C}}(n)$ defined for all $n \in \mathbb{N}$ by

$$S_{\mathcal{C}}(n) = \begin{cases} 0 & \text{if } |T| < n, \\ \max_{t_1,\ldots,t_n \in T} |\mathcal{C} \cap \{t_1, \ldots, t_n\}| & \text{otherwise,} \end{cases}$$

are said to be the **shattering coefficients** or the **growth function** of $\mathcal{C}$.

The coefficients are well defined since $|\mathcal{C} \cap \{t_1, \ldots, t_n\}| \leq 2^n$ for all $n \in \mathbb{N}$. (The coefficients are mostly of interest when $T$ is infinite). If $S_{\mathcal{C}}(n) = 2^n$, then it means that there is a at least one subset of $T$ of cardinality $n$ which can be shattered by $\mathcal{C}$. If

$S_{\mathcal{C}}(m) < 2^m$ for some $m \in \mathbb{N}$, then $S_{\mathcal{C}}(n) < 2^n$ for all $n \geq m$. As said, the vc dimension can then be rewritten as

$$\mathrm{vc}(\mathcal{C}) = \sup\{n : n \in \mathbb{N} \text{ s.t. } S_{\mathcal{C}}(n) = 2^n\}.$$

The shattering coefficients of $\mathcal{C}$ provide another intuitive measure of the complexity or size of $\mathcal{C}$: the bigger $\mathcal{C}$, the slowest the decrease of $S_{\mathcal{C}}(n)$ in $n$ (as bigger sets can be shattered by $\mathcal{C}$). If $\mathcal{C}$ is not a VC class, then $|\mathcal{C} \cap \{t_1, \ldots, t_n\}| = 2^n$ for all $n$. But even if $\mathcal{C}$ is a VC class with $\mathrm{vc}(\mathcal{C}) = m$, it could be possible that $2^n > S_{\mathcal{C}}(n) \geq 2^n - 1$ for $n > m$. The Sauer–Shelah lemma shows that it is not possible: if $\mathcal{C}$ is a VC class, then $S_{\mathcal{C}}(n)$ grows at most polynomially in $n$.

**Theorem 14 (Sauer–Shelah Lemma).** *Let $\mathcal{C}$ be VC class for a set $T$. Then for all $n \geq 1$ and all $t_1, \ldots, t_n \in T$,*

$$S_{\mathcal{C}}(n) \leq \sum_{k=0}^{\mathrm{vc}(\mathcal{C})} \binom{n}{k}.$$

*Proof.* T.13.2. in Lugosi PTPR p.216 or T.8.3.16. in Vershynin HDP p.211. □

**Corollary 15.** *Let $\mathcal{C}$ be VC class in a set $T$. Then for all $n \geq 1$,*

$$S_{\mathcal{C}}(n) \leq (n+1)^{\mathrm{vc}(\mathcal{C})}$$

*and, if $n \geq \mathrm{vc}(\mathcal{C})$,*

$$S_{\mathcal{C}}(n) \leq \left(\frac{en}{\mathrm{vc}(\mathcal{C})}\right)^{\mathrm{vc}(\mathcal{C})}$$

*Proof.* Smart bounding and binomial formula. See C.III-3.14. in Moulines p.191. □

Let us consider a few VC dimension and shattering coefficients computations (see S.13.2. in Lugosi PTPR p.219 for proofs) which also illustrate the Sauer–Shelah lemma.

**Example 16.** (from S.13.2. in Lugosi PTPR p.219)

(i) $T$ is finite with $|T| = m$, then $\mathrm{vc}(2^T) = m$ and $S_{2^T}(n) = 2^n$ if $n \leq m$ and $S_{2^T}(n) = 0$ otherwise;

(ii) $T = \mathbb{R}$ and $\mathcal{C} = \{(-\infty, a] : a \in \mathbb{R}\}$, then $\mathrm{vc}(\mathcal{C}) = 1$ and $S_{\mathcal{C}}(n) = n + 1$;

(iii) $T = \mathbb{R}$ and $\mathcal{C} = \{[a, b] : a, b \in T, a \leq b\}$, then $\mathrm{vc}(\mathcal{C}) = 2$ and $S_{\mathcal{C}}(n) = (n(n+1)/2) + 1$;

(iv) $T = \mathbb{R}^d$ and $\mathcal{C} = \{\{t \in \mathbb{R}^d : a^T t \geq b\} : a \in \mathbb{R}^d, b \in \mathbb{R}\}$, then $\mathrm{vc}(\mathcal{C}) = d + 1$ and $S_{\mathcal{C}}(n) = 2\sum_{i=0}^{d} \binom{n-1}{i} \leq 2(n-1)^d + 2$.

(v) $T = \mathbb{R}^2$ and $\mathcal{C} = \{C \subseteq \mathbb{R}^2 : C \text{ is compact and convex}\}$, then $\mathrm{vc}(\mathcal{C}) = +\infty$.

**Theorem 17 (Dudley (1978)).** *There exists a constant $K \in (0, +\infty)$ such that for any class $\mathscr{F}$ of measurable Booleans functions over $\mathbb{R}^d$, any probability measure $\mu$ on $\mathbb{R}^d$, and any $\varepsilon \in (0, 1)$,*

$$N(\mathscr{F}, \|\cdot\|_{L^2(\mu)}, \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^{K\,\mathrm{vc}(\mathscr{F})}.$$

*Proof.* T.8.3.18. in Vershynin HDP p.206. □

# 3 Rademacher complexity

Metric entropy provides a measure of the size (or richness) of (infinite) sets based on metric considerations only. Alternatively, it is possible to obtain measures of the size of (infinite) sets based on probabilistic considerations. One idea is to consider the maximum correlation between elements of a set and noise from some distribution. The bigger the set, the higher the probability to find an element that correlates well with the noise. Both metric entropy and VC theory can then be used to bound these probabilistic measures of complexity.

(Note that for clarity we abuse notation by subscripting the expectation with respect to the pushforward measure.)

**Definition 18** (Rademacher Complexity)**.** Let $T \subseteq \mathbb{R}^n$ be a set. The **Rademacher complexity** of $T$ is defined as

$$\mathrm{Rad}(T) = \frac{1}{n}\mathbb{E}_{\sigma}\left(\sup_{t\in T}\sum_{i=1}^{n}\sigma_i t_i\right),$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)$ is a vector of i.i.d. Rademacher variables.

The Rademacher complexity is also known as the Rademacher average or mean. It is also of interest to define an alternative version with absolute value, namely

$$\overline{\mathrm{Rad}}(T) = \frac{1}{n}\mathbb{E}_{\sigma}\left(\sup_{t\in T}\left|\sum_{i=1}^{n}\sigma_i t_i\right|\right).$$

Then we have (see SLT Raginsky p.68)

$$\mathrm{Rad}(T) \leq \overline{\mathrm{Rad}}(T) = \mathrm{Rad}(T \cup -T).$$

**Definition 19** (Rademacher Complexity of a Function Class)**.** Let $(E, \mathcal{E}, \mathbb{P})$ be a probability space and $(X_1, \ldots, X_n)$ i.i.d. random variables drawn from $\mathbb{P}$. Let $\mathscr{F} \subseteq \mathbb{R}^E$ be a class of real-valued functions $f\colon E \to \mathbb{R}$. Then the **Rademacher complexity** of $\mathscr{F}$ with respect to $\mathbb{P}$ for sample size $n$ is defined as

$$\mathrm{Rad}_{\mathbb{P},n}(\mathscr{F}) = \mathbb{E}_{\mathbb{P}^n}\left(\frac{1}{n}\mathbb{E}_{\sigma}\left(\sup_{f\in\mathscr{F}}\sum_{i=1}^{n}\sigma_i f(X_i)\right)\right),$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)$ is a vector of i.i.d. Rademacher variables.

The quantity

$$\mathrm{Rad}_{x_{1:n}}(\mathscr{F}) = \frac{1}{n}\mathbb{E}_{\sigma}\left(\sup_{f\in\mathscr{F}}\sum_{i=1}^{n}\sigma_i f(x_i)\right)$$

for some $x_{1:n} = (x_1, \ldots, x_n) \in E^n$ is called the **empirical Rademacher complexity** of $\mathscr{F}$ given $x_{1:n}$. Then

$$\mathrm{Rad}_{x_{1:n}}(\mathscr{F}) = \mathrm{Rad}(\mathscr{F} \circ x_{1:n}),$$

where $F \circ x_{1:n} = \{(f(x_1), \ldots, f(x_n)) : f \in \mathscr{F}\}$, and

$$\mathrm{Rad}_{\mathbb{P},n}(\mathscr{F}) = \mathbb{E}_{\mathbb{P}^n}(\mathrm{Rad}_{X_{1:n}}(\mathscr{F})) = \mathbb{E}_{\mathbb{P}^n}(\mathrm{Rad}(\mathscr{F} \circ X_{1:n})),$$

where $X_{1:n}$ is a random vector drawn from $\mathbb{P}^n$.

We similarly define absolute versions of these notions, namely

$$\overline{\mathrm{Rad}}_{x_{1:n}}(\mathscr{F}) = \overline{\mathrm{Rad}}(\mathscr{F} \circ x_{1:n}),$$

and

$$\overline{\mathrm{Rad}}_{\mathbb{P},n}(\mathscr{F}) = \mathbb{E}_{\mathbb{P}^n}(\overline{\mathrm{Rad}}(\mathscr{F} \circ X_{1:n})).$$

If we substitute $g_i$ for $\sigma_i$ where $g_i \sim N(0,1)$ in all the definitions above, then we get **Gaussian complexity** counterparts of the Rademacher complexities so defined. We denote them by substituting Gau for Rad in all the notations above. Gaussian and Rademacher complexities are equivalent up to logarithmic factors.

**Lemma 20.** *For any set $T \subseteq \mathbb{R}^n$,*

$$\sqrt{\frac{2}{\pi}}\mathrm{Rad}(T) \leq \mathrm{Gau}(T) \leq 2\sqrt{\ln(n)}\mathrm{Rad}(T).$$

*Proof.* E.5.5. in Wainwright HDS p.155. □

**Lemma 21.** *Let $S, T \subseteq \mathbb{R}^n$, $a \in \mathbb{R}$, $b \in \mathbb{R}^n$. For every $i = 1, \ldots, n$, let $f_i \in \mathrm{Lip}_K(\mathbb{R})$ where $K > 0$. Then:*

*(i)* $\mathrm{Rad}(aT + b) = |a|\mathrm{Rad}(T)$*;*

*(ii)* $\mathrm{Rad}(S + T) = \mathrm{Rad}(S) + \mathrm{Rad}(T)$*;*

*(iii)* $\mathrm{Rad}(\mathrm{conv}(T)) = \mathrm{Rad}(T)$ *where* conv *denotes the convex hull;*

*(iv)* $\mathrm{Rad}(f \circ T) \leq K\mathrm{Rad}(T)$ *where* $f \circ T = \{(f_1(t_1), \ldots, f_n(t_n)) : (t_1, \ldots, t_n) \in T\}$.

*Proof.* S.26.1.1. in Shai&Shai UML p.329 or T.1.16. in Wolf TUM MFSL p.38. □

**Theorem 22** (**Massart's Lemma (2000)**)**.** *Let $T$ be a finite subset of $\mathbb{R}^n$ and $r = \sup_{t \in T} \|t\|_2$. Then*

$$\mathrm{Rad}(T) \leq \frac{r}{n}\sqrt{2\ln|T|}.$$

*Proof.* L.5.2. in Massart (2000) "Some applications of concentration inequalities to statistics" p.300. □

**Corollary 23.** *Let $(E, \mathcal{E}, \mathbb{P})$ be a probability space. Let $K > 0$ and $S \subseteq \mathbb{R}$ a finite set of real numbers such that $|s| \leq K$ for all $s \in S$. Let $\mathscr{F} \subseteq S^E$. Then*

$$\mathrm{Rad}_{\mathbb{P},n}(\mathscr{F}) \leq K\sqrt{\frac{2\ln S_{\mathscr{F}}(n)}{n}}.$$

*Proof.* C.1.18. in Wolf TUM MFSL p.39 or C.3.8. in Mohri FML p.35. □

Let $\mathscr{F} \subseteq \mathbb{R}^E$ and $x \in E^n$. Define the pseudonorms $\|\cdot\|_{p,x}$ on the linear span of $\mathscr{F}$ by

$$\|f\|_{p,x} = \left(\frac{1}{n}\sum_{i=1}^{n}|f(x_i)|^p\right)^{1/p}$$

for $p \in [1, +\infty)$, and

$$\|f\|_{\infty,x} = \max_{i \in \{1,\ldots,n\}}|f(x_i)|.$$

(They are nothing more than the $L^p$ (pseudo)norms for the probability space with measure the uniform distribution over the $x_i$.) These pseudonorms naturally induce pseudometrics by taking $(f, g) \mapsto \|f - g\|_{p,x}$. By Jensen's inequality, $\|f\|_{p,x} \leq \|f\|_{p,x}$ for $p \leq q$. Then one gets $N(\mathscr{F}, \|\cdot\|_{p,x}, \varepsilon) \leq N(\mathscr{F}, \|\cdot\|_{q,x}, \varepsilon)$ and $\operatorname{pack}(\mathscr{F}, \|\cdot\|_{p,x}, \varepsilon) \leq \operatorname{pack}(\mathscr{F}, \|\cdot\|_{q,x}, \varepsilon)$ for $p \leq q$.

**Proposition 24** (**One-Step Discretization Bound**). *Let $(x_1, \ldots, x_n) \in E^n$ and $\mathscr{F} \subseteq \mathbb{R}^E$. If $\sup_{f \in \mathscr{F}} \|f\|_{2,x} \leq K$, then*

$$\operatorname{Rad}_{x_{1:n}}(\mathscr{F}) \leq \inf_{\varepsilon > 0} \left( \varepsilon + K \sqrt{\frac{2}{n} \ln N(\mathscr{F}, \|\cdot\|_{1,x}, \varepsilon)} \right).$$

*Proof.* P.5.2. in Rebeschini AFoL notes (2021) L.5. □

**Theorem 25** (**Dudley's Entropy Integral Bound**). *Let $(x_1, \ldots, x_n) \in E^n$ and $\mathscr{F} \subseteq \mathbb{R}^E$. If $\sup_{f \in \mathscr{F}} \|f\|_{2,x} \leq K$, then*

$$\operatorname{Rad}_{x_{1:n}}(\mathscr{F}) \leq \inf_{\varepsilon \in [0, K/2]} \left( 4\varepsilon + \frac{12}{\sqrt{n}} \int_{\varepsilon}^{K/2} \sqrt{\ln N(\mathscr{F}, \|\cdot\|_{2,x}, \tau)}\, d\tau \right).$$

*Proof.* P.5.3. in Rebeschini AFoL notes (2021) L.5 (also E.5.24. in Wainwright HDS p.142 and T.1.19. in Wolf TUM MFSL p.39, but both losing something in the integral upper limit). See also C.13.2. in BLM CI p.365 for Dudley's entropy integral. □

**Corollary 26.** *Let $(x_1, \ldots, x_n) \in E^n$ and $\mathscr{F} \subset \{0, 1\}^E$. Then*

$$\operatorname{Rad}_{x_{1:n}}(\mathscr{F}) \leq 19 \sqrt{\frac{\operatorname{vc}(\mathscr{F})}{n}}.$$

*Proof.* T.5.6. in Rebeschini AFoL notes (2021) L.5 or C.1.25. in Wolf TUM MFSL p.48 (with $K = 1$ but control of the upper limit in the integral yielding 19 instead of 31). □

# 4 Chaining and sub-Gaussian maximal inequalities

**Definition 27** (Sub-Gaussian Process)**.** A stochastic process $(X_t)_{t \in T}$ is said to be a **sub-Gaussian process** with respect to a pseudodistance $\rho$ on $T$ if $\mathbb{E}(X_t) = 0$ and

$$\mathbb{E}\left( e^{\lambda (X_t - X_s)} \right) \leq e^{\lambda^2 \rho(t,s)^2 / 2}$$

for all $s, t \in T$ and all $\lambda \in \mathbb{R}$.

Equivalently, $(X_t)_{t \in T}$ is sub-Gaussian if and only if for all $s, t \in T$, $X_t - X_s$ is sub-Gaussian with proxy variance $\rho(t, s)^2$. The pseudo-metric $\rho$ is called the canonical pseudo-metric. For instance, if the variables $X_t$ are Gaussian, then $(X_t)_{t \in T}$ is sub-Gaussian by taking $\rho(s, t) = \sqrt{\operatorname{Var}(X_t - X_s)}$.

Using the characterization of sub-Gaussian variables in terms of tail probability, the sub-Gaussian property of a process can be interpreted as a Lipschitz property in probability: indeed, up to constants, the sub-Gaussian property is equivalent to

$$\mathbb{P}(|X_t - X_s| \geq x \rho(t, s)) \leq C e^{-x^2 / C}.$$

The processes $(\sum_{i=1}^n \sigma_i t_i)_{t\in T}$ where $T \subseteq \mathbb{R}^n$ in the definition of Rademacher and Gaussian complexities are also sub-Gaussian with respect to the distance induced by the 2-norm on $\mathbb{R}^n$ (since $\mathrm{Var}(\sigma^T t - \sigma^T s) \leq \|t-s\|_2^2\sigma^2$, with $\sigma^2 = 1$ for Rademacher). For fixed $x = (x_1, \ldots, x_n)$, the processes $(n^{-1/2}\sum_{i=1}^n \sigma_i f(x_i))_{f\in\mathscr{F}}$ where $\mathscr{F} \subseteq \mathbb{R}^E$ in the definition of the empirical complexities are also sub-Gaussian with respect to the pseudo-distance induced by the $\|\cdot\|_{2,x}$ pseudo-norm. Hence the results of the previous section are only particular cases (up to constants for which we do not optimize) of the general bound for sub-Gaussian processes we now develop.

**Definition 28** (Separable Process)**.** Let $(T,\rho)$ be a pseudo-metric space. A stochastic process $(X_t)_{t\in T}$ on $(\Omega,\mathcal{F},P)$ is said to be **separable** if there exists a countable dense subset $S \subseteq T$ and $\Omega_0 \subseteq \Omega$ with $P(\Omega_0) = 1$ such that for all $t \in T$ and all $\omega \in \Omega_0$, there exists a sequence $(s_n)_{n\in\mathbb{N}}$ in $S$ for which $s_n \to t$ and $X_{s_n}(\omega) \to X_t(\omega)$.

A separable process $(X_t)_{t\in T}$ is thus such that each sample path $t \mapsto X_t(\omega)$ is controlled by its behavior on a fixed countable subset of $T$. The definition itself imposes the space $(T,\rho)$ to be separable (as $T$ is supposed to be the sequential closure of $S$). Many common processes are separable: for instance, if $(T,\rho)$ is separable and $(X_t)_{t\in T}$ has a.s. continuous sample path, then $(X_t)_{t\in T}$ is separable. However, not all: for instance, $X_t(\omega) = \mathbb{1}_{\{t=\omega\}}(\omega,t)$ with $T = \Omega = [0,1]$ is not separable. Separability also ensures the measurability of $\sup_{t\in T} X_t$ since under separability $\sup_{t\in T} X_t = \sup_{s\in S} X_s$ a.s. (see section on "measurability woes").

**Theorem 29** (**Sub-Gaussian Dudley's Entropy Integral Bound**)**.** *Let $(T,\rho)$ be a pseudo-metric space and $(X_t)_{t\in T}$ a separable sub-Gaussian process with respect to $\rho$. Let $D = \sup_{s,t}\rho(s,t)$. Then for all $\varepsilon \in [0,D]$,*

$$\mathbb{E}\left(\sup_{s,t\in T}(X_t - X_s)\right) \leq 2\mathbb{E}\left(\sup_{s',t'\in T:\,\rho(s',t')\leq\varepsilon}(X_{t'} - X_{s'})\right) + 16\int_{\delta/4}^{D}\sqrt{N(T,\rho,\tau)}\,d\tau.$$

*Proof.* If $\sup_{s,t}\rho(s,t) = +\infty$, then the inequality is trivially true (since in this case $N(T,\rho,\varepsilon) = +\infty$ for all $\varepsilon$). For $\sup_{s,t}\rho(s,t) < +\infty$, chaining argument: see T.5.22. in Wainwright HDS p.140. □

# 5 Symmetrization and empirical processes

The importance of the Rademacher complexity is that it appears naturally when trying to obtain maximal inequalities for empirical processes (inequalities which are central to derive uniform convergence results). The connection comes from what is known as **symmetrization**. The idea is that for any $f \in \mathscr{F}$, the random variable $\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)$ (where $\sigma_1,\ldots,\sigma_n$ are i.i.d. Rademacher, independent of $X_1,\ldots,X_n$) is simply a symmetrized (or randomized) version of the random variable $(\mathbb{P}_n - \mathbb{P})f := \frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}_{\mathbb{P}}(f(X_i)))$. (Note that both variables have mean zero.) As one may expect, it is possible to use one to control the other. In particular, the following result shows that $\mathbb{E}_{\mathbb{P}^n}(\sup_{f\in\mathscr{F}}(\mathbb{P}_n - \mathbb{P})f)$ (that is, the average worst-case deviation of the mean from the sample average) is intimately related to the Rademacher complexity $\mathrm{Rad}_{\mathbb{P},n}(\mathscr{F})$. Given the interpretation of the Rademacher complexity as a measure of the size of $\mathscr{F}$, this result exemplifies the idea that the behavior of stochastic processes indexed by infinite sets is directly related to the size of the index sets.

**Lemma 30** (**Symmetrization Inequalities**). *Let $X_1, \ldots, X_n$ be independent random elements where $X_i = (X_{i,t})_{t \in T}$ for some index set $T$ and some real random variables $X_{i,t}$. Assume that $E(X_{i,t}) = 0$ for all $i = 1, \ldots, n$ and all $t \in T$. Let $\sigma_1, \ldots, \sigma_n$ be independent Rademacher variables independent of $X_1, \ldots, X_n$. Then*

$$\frac{1}{2}\mathbb{E}\left(\sup_{t\in T}\left|\sum_{i=1}^{n}\sigma_i X_{i,t}\right|\right) \leq \mathbb{E}\left(\sup_{t\in T}\left|\sum_{i=1}^{n} X_{i,t}\right|\right) \leq 2\mathbb{E}\left(\sup_{t\in T}\left|\sum_{i=1}^{n}\sigma_i X_{i,t}\right|\right),$$

*and*

$$\mathbb{E}\left(\sup_{t\in T}\sum_{i=1}^{n} X_{i,t}\right) \leq 2\mathbb{E}\left(\sup_{t\in T}\sum_{i=1}^{n}\sigma_i X_{i,t}\right).$$

*Proof.* L.11.4. in BLM CI p.322. The idea is to introduce a "ghost sample" $Y_1, \ldots, Y_n$ which is an independent copy of $X_1, \ldots, X_n$ and to note that the random elements $X_i - Y_i$ are symmetric and distributed as $\sigma_i(X_i - Y_i)$. □

**Corollary 31** (**Rademacher Symmetrization Bound**). *Let $X_1, \ldots, X_n$ be i.i.d. random variables in $E$ with distribution $\mathbb{P}$. Let $\mathscr{F} \subseteq \mathbb{R}^E$ be a class of integrable functions (with respect to $\mathbb{P}$) and $\mathscr{F}^* = \{f - \mathbb{E}\, f : f \in \mathscr{F}\}$. Then*

$$\frac{1}{2}\overline{\mathrm{Rad}}_{\mathbb{P},n}(\mathscr{F}^*) \leq \mathbb{E}_{\mathbb{P}^n}\left(\sup_{f\in\mathscr{F}}\left|(\mathbb{P}_n - \mathbb{P})f\right|\right) \leq 2\overline{\mathrm{Rad}}_{\mathbb{P},n}(\mathscr{F}),$$

*and*

$$\mathbb{E}_{\mathbb{P}^n}\left(\sup_{f\in\mathscr{F}}\left((\mathbb{P}_n - \mathbb{P})f\right)\right) \leq 2\mathrm{Rad}_{\mathbb{P},n}(\mathscr{F}).$$

*Proof.* Application of last result with $(X_{i,t})_{t\in T} = (f(X_i))_{f\in\mathscr{F}}$ and rough bounding for non-centered variables in the right hand-side inequalities. First inequalities: P.4.11. in Wainwright HDS p.107. Second inequality: L.7.4. in van Handel APC550 p.201. □

It is possible to derive probabilistic (instead of moment) versions of these results.

In any case, bounding the Rademacher complexity of $\mathscr{F}$ then directly yields maximal inequalities for the processes $((\mathbb{P}_n - \mathbb{P})f)_{f\in\mathscr{F}}$.

# References


Boucheron, S., G. Lugosi, and P. Massart (2013): *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Devroye, L., L. Györfi, and G. Lugosi (2013): *A probabilistic theory of pattern recognition*, vol. 31. Springer.

Fort, G., M. Lerasle, and E. Moulines (2020): "Statistique et apprentissage," *Notes du cours MAP433 at Polytechnique*.

Hajek, B., and M. Raginsky (2021): "Statistical learning theory," *Lecture notes for ECE543 at UIUC*.

Massart, P. (2000): "Some applications of concentration inequalities to statistics," in *Annales de la Faculté des sciences de Toulouse: Mathématiques*, vol. 9, pp. 245–303.

Mohri, M., A. Rostamizadeh, and A. Talwalkar (2018): *Foundations of machine learning*. MIT Press.

Rebeschini, P. (2021): "Algorithmic foundations of learning," *Lecture notes at Oxford*.

Shalev-Shwartz, S., and S. Ben-David (2014): *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Shiryayev, A. N. (1993): *Selected Works of A. N. Kolmogorov*, vol. 3. Springer.

van Handel, R. (2016): "Probability in high dimension," *Lecture notes for ACP550 at Princeton*.

Vershynin, R. (2018): *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.

Wainwright, M. J. (2019): *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

Wolf, M. M. (2022): "Mathematical foundations of supervised learning," *Lecture notes at TUM*.