Chapter 4: Complexity of classes of functions and sets

Empirical Processes (MATH-522)

Myrto Limnios

18th of March, 2025

This chapter focuses on measuring the size of uncountable classes of functions, or of sets, defined as their *complexity*. The intuition being that, on the one hand, the bigger the size, the likelier we can approximate the *true* model/measure. On the other hand, if the class is extremely large, then one can yield a NP-hard problem. It is important to keep in mind that an empirical process will concentrate with high probability, if the complexity of its index class can be upperbounded. We will detail three important tools, discuss their relations to uniform asymptotic convergences, and permanence properties.

Reminder 0.1 (Union bound.)**.** *We will (massively) use in the proofs the so-called* **union bound** *that we recall below. Consider a countable set of measurable events $A_1, \dots, A_N$, with $N \in \mathbb{N}^*$ then*

$$\mathbb{P}\left(\bigcup_{i=1}^{N} A_i\right) \leq \sum_{i=1}^{N} \mathbb{P}(A_i) \leq N \max_{i \leq N} \mathbb{P}(A_i) \ .$$

Apply this to the maximum of a set.

Motivation. If the class is supposed of finite cardinality, or countable, then the union bound can be used directly. Let $\mathcal{H} = \{h_1, \dots, h_N\}$ composed of a finite number of measurable functions $N \in \mathbb{N}^*$. Then, the uniform deviation of the empirical measure can be upperbounded as follows (we suppose everything is well-defined). We want to control the probability for any $\delta > 0$,

$$\mathbb{P}\left\{\|P_n - P\|_{\mathcal{H}} \geq \delta\right\} \ .$$

Notice that, because $|\mathcal{H}| = N < \infty$, $\|P_n - P\|_{\mathcal{H}} = \sup_{h \in \mathcal{H}} |P_n h_i - P h_i| = \max_{i \leq N} |P_n h_i - P h_i|$, thus, by the union bound

$$\mathbb{P}\left\{\|P_n - P\|_{\mathcal{H}} \geq \delta\right\} \leq \sum_{i=1}^{N} \mathbb{P}\left\{|P_n h_i - P h_i| \geq \delta\right\} \ .$$

We want to apply Hoeffding's inequality. Suppose that there exists a constant $C > 0$, such that $\sup_{x \in \mathcal{X}} |h_i(x)| \leq C$, for all $i \leq N$, and $Ph = 0$ for simplicity, for all $h \in \mathcal{H}$.

$$\mathbb{P}\left\{\|P_n\|_{\mathcal{H}} \geq \delta\right\} \leq 2N e^{-2n\delta^2/C^2} \ .$$

Notations. In the following sections, we will define a series of results, for which we place ourselves in a general setting. We denote by $\{Z(t)\}_{t \in T}$ a process, where T is a generic set endowed with a (pseudo)metric d. We suppose for technical reasons the process $Z(t)$ to have bounded sample functions[1].

[1] Versions with bounded sample paths a.s. is sufficient.

1 Entropic Measures of Classes of Functions

We start by defining a purely deterministic way of controlling the size of a semimetric space, and thus of a stochastic process.

Consider $(\Omega, \mathcal{A}, \mathbb{P})$ to be an arbitrary probability space and let $X : \Omega \to \mathcal{X}$ be an arbitrary random map of distribution P. Let (T, d) be a semimetric space. We think of subset of (T, d) mainly as subset of the $L_q(P)$-space, with $q \in \mathbb{N}^*$, defined by

$$L_q(P) = \{h : \mathcal{X} \to \mathbb{R}, \quad P|h|^q < \infty\}$$

endowed with the associated norm

$$\|h\|_q = (P|h|^q)^{1/q} \ .$$

The constant ε will always be in $(0, \infty)$ in this chapter.

This section can be extended to general semimetric spaces.

1.1 Covering and Bracketing Numbers

Covering Numbers.

Definition 1.1 (ε-cover)**.** A ε-cover of the set T w.r.t. d is a set $\{t_1, \dots, t_N\} \subset T$, such that for any $t \in T$, there exists a $m \in \{1, \dots, N\}$, such that $d(t, t_m) \le \varepsilon$.

Definition 1.2 (Covering number)**.** Let $\varepsilon > 0$ be fixed. The ε-covering number of (T, d) is the cardinality of the smallest ε-cover of the set T.
We consider in particular the ε-cover in terms of the balls. The ε-covering number of (T, d) is the smallest number of open balls of radius at most ε needed to cover T, defined by

$$N(\varepsilon, T, d) = \min\{N \in \mathbb{N}^*, \ \exists (t_i)_{i \le N} \ T \subset \cup_{i \le N} B_d(t_i, \varepsilon)\} \ ,$$

where the balls are defined by $B_d(t_i, \varepsilon) := \{t \in T, \ d(t, t_i) < \varepsilon\}$. The entropy of (T, d) is defined by

$$H(\varepsilon, T, d) = \log N(\varepsilon, T, d) \ .$$

The function $\varepsilon \mapsto H(\varepsilon, T, d)$ is called the metric entropy.

Remark 1.1. *This definition provides a control of a set T in terms of the set of ε-balls covering the set T. We think of metric entropy when T is totally bounded, i.e., when the covering number is finite for all $\varepsilon > 0$.*

Remark 1.2. *The covering number is a decreasing function of the radius ε: for all $\varepsilon \le \varepsilon'$, it holds that $N(\varepsilon, T, d) \ge N(\varepsilon', T, d)$. Notice that $N(\varepsilon, T, d) \to \infty$, for $\varepsilon \to 0$. Our interest is when this is at a logarithm rate. In particular, if $\lim_{\varepsilon \to 0} \log N(\varepsilon, T, d) \log(1/\varepsilon)$ exists, then it is the metric dimension.*

Example 1.3 (Covering numbers of cubes)**.** *Consider $T = [-1, 1]$, equipped with the semimetric $d(t, t') = |t - t'|$. If we divide $[-1, , 1]$ into $N = \lfloor 1/\varepsilon \rfloor = 1$ sub-intervals, centers at the points $t_i = -1 + 2(i-1)\varepsilon$, for all $i \le N$, and of length at most 2ε, then it is an ε-cover by construction: for any point $t \in [-1, 1]$, there exists some $m \le N$, such that $d(t, t_m) \le \varepsilon$. Thus*

$$N(\varepsilon, [-1, 1], d) \le \frac{1}{\varepsilon} + 1 \ .$$

NB: It can be generalized to d-dimentional cube, with bound $(\frac{1}{\varepsilon} + 1)^d$.

We state and prove a very useful result.

Lemma 1.3. *Let $T = \{h : [0,1] \to [0,1],\ h$ is 1-Lipschitz$\}$. Then, for some constant A, it holds that*

$$\log N(\varepsilon, T, \|\cdot\|_\infty) \le \frac{A}{\varepsilon}\ ,$$

for all $\varepsilon > 0$.

Proof. Suppose that $h(0) = 0$ for all functions $h \in \mathcal{H}$ for simplicity. Notice first that, if $\varepsilon \ge 1$, then only one ball is necessary to cover T, thus $N(\varepsilon, T, \|\cdot\|_\infty) = 1$.
We now consider $\varepsilon \in (0,1)$. We construct an ε-cover of T w.r.t. the uniform norm $\|\cdot\|_\infty$ that has cardinality upperbounded by $e^{A/\varepsilon}$ for some constant $A > 0$.

We construct am ε-grid of the y-axis, and a $\varepsilon(/L)$-grid of the x-axis.

Define the uniform grid of $[0,1]$ as follows: $0 = x_0 < x_1 < \ldots < x_N = 1$, with $x_k = k\varepsilon$, for all $k \le N-1$, and $N = \lfloor 1/\varepsilon \rfloor$, where we recall that $\lfloor \cdot \rfloor$ is the floor function defined by: for any $x \in \mathbb{R}$, $\lfloor x \rfloor = \max\{n \in \mathbb{Z},\ n \le x\}$.

We define the intervals by $I_k = (x_{k-1}, x_k]$, and $I_0 = [x_0, x_1]$.

For all $h \in \mathcal{H}$, consider the function $\bar{h} : [0,1] \to \mathbb{R}$ by

$$\bar{h}(x) = \sum_{k \le N} \varepsilon \lfloor \frac{h(x_k)}{\varepsilon} \rfloor 1_{I_k}(x)\ .$$

Notice first that for all $k \le N$, $h(x_k) - \varepsilon \lfloor \frac{h(x_k)}{\varepsilon} \rfloor = \varepsilon (h(x_k) - \lfloor \frac{h(x_k)}{\varepsilon} \rfloor) \le \varepsilon$ by definition of the floor function (approximation error).

Then, the function $\bar{h}$ is piecewise constant on the intervals I_k and takes values of the form of $i\varepsilon$, with $i \in \{0, \ldots, \lfloor 1/\varepsilon \rfloor\}$.

For any $k \le N$, for all $x \in I_k$ we have

$$|h(x) - \bar{h}(x)| \le |h(x) - h(x_k)| + |h(x_k) - \bar{h}(x)| \le 2\varepsilon\ ,$$

because h is 1-Lipschitz and the approximation error. Thus $\|h - \bar{h}\| \le 2\varepsilon$.

Now, we need to count how many distinct functions $\bar{h}$ when considering all h-s in $\mathcal{H}$.

For $\bar{h}(x_1)$, there are N choices, as the origin has been fixed. Then for all $\bar{h}$,

$$|\bar{h}(x_k) - \bar{h}(x_{k-1})| \le |h(x_k) - \bar{h}(x_k)| + |h(x_{k-1}) - h(x_k)| + |h(x_{k-1}) - \bar{h}(x_{k-1})| \le 3\varepsilon$$

there are 7 choices for the next value $\bar{h}(x_k)$. Thus, the set $\{\bar{h},\ h \in \mathcal{H}\}$ is a 2ε-cover if $\mathcal{H}$ an is composed of $N 7^{\lfloor 1/\varepsilon \rfloor}$ distinct points. Thus $\log N(\varepsilon, T, \|\cdot\|_\infty) \le \log N + \lfloor 1/\varepsilon \rfloor \log 7$, that concludes the proof.

□

Bracketing Numbers. Now we consider only T to be subset of $L_q(P)$ space.

We define a different approach to measure the size of a functional space, namely through constructing pointwise functions bounding the elements of the set T.

Definition 1.4 (Bracketing number)**.** Let two functions s and t, the bracket $[l,u]$ is the set of all function h such that $l \le h \le u$. Let $\varepsilon > 0$ be fixed. The ε-bracketing number of $(T, \|\cdot\|_q)$ is the smallest number of ε-brackets needed to cover T, defined by

$$N_B(\varepsilon, T, \|\cdot\|_q) = \min\{N \in \mathbb{N}^*,\ \exists (l_i, u_i)_{i \le N} \in T,\ \|l_i - u_i\|_q \le \varepsilon,\ T \subset \cup_{i \le N} [l_i, u_i]\}\ .$$

The entropy of (T, d) is defined by

$$H_B(\varepsilon, T, \|\cdot\|_q) = \log N_B(\varepsilon, T, \|\cdot\|_q)\ .$$

Example 1.4 (Distribution function)**.** *Let $X_1, \ldots, X_n$ i.i.d. sample drawn from P, and $\varepsilon < 1$. Suppose that $T = \{1(-\infty, x](\cdot),\ x \in \mathbb{R}\}$, and consider a grid points of $\bar{\mathbb{R}}$: $-infty = x_0 < x_1 < \ldots < x_N = \infty$ that generate the brackets of the form $[1(-\infty, x_{i-1}], 1(-\infty, x_i]]$. These brackets have $L_1(P)$ size ε using that $F(x_i-) - F(x_i) < \varepsilon$. And the total number of N can be chosen smaller than $2/\varepsilon$. Notice that, because we can bound the L_2-norm by the L_1 in this case, the size of the $L_2(P)$-brackets are bounded by $\sqrt{\varepsilon}$. Thus $N_B(\sqrt{\varepsilon}, T, \|\cdot\|_2) \leq 2/\varepsilon$ hence they are of polynomial order $1/\varepsilon^2$.*

Remark 1.5.

1. *The centers are not necessarily required to be in T, but they need to have finite norm.*
2. *Consider $T \subset L_q(P)$, then for all $\varepsilon > 0$,*

$$N(\varepsilon, T, \|\cdot\|_q) \leq N_B(2\varepsilon, T, \|\cdot\|_q) \ .$$

If h is in a 2ε-bracket, then $\|l-u\|_q \leq 2\varepsilon$, and $l \leq h \leq u$. Thus $l - (l+u)/2 \leq h - (l+u)/2 \leq u - (l+u)/2$ iff. $(l-u)/2 \leq h - (l+u)/2 \leq (u-l)/2$ iff. $|h - (l+u)/2| \leq (u-l)/2$, thus $\|h - (l+u)/2\|_q \leq 2\varepsilon/2 = \varepsilon$.

3. *Notice that in general, converse inequalities do not hold.*
4. *In conclusion, N_B is always bigger than N. The main argument for using in some applications bracketing numbers is related to the pointwise control of any function h: $l(x) \leq h(x) \leq u(x)$ for any $x \in \mathcal{X}$, that is different from the integrated L_q-norm.*

1.2 Uniform Asymptotic Laws

In order to relate complexity measures with Glivenko-Cantelli classes, we first need to exhibit a function that upperbounds pointwise all functions in T defined as *envelope function* $x \mapsto H(x)$, such that $|h(x)| \leq H(x)$.

The key observation being that, for any q-norm, if $|h(x)| \leq H(x)$ then $\|h\|_q \leq \|H\|_q$. (i.e. from pointwise to norm comparison)

Notice also that the minimal envelope function is given by $x \mapsto \sup_{h \in \mathcal{H}} |h(x)|$.

Definition 1.5. The uniform entropy number of T relative to $L_q(P)$ is defined by

$$\sup_P \log N(\varepsilon\|H\|_q, T, \|\cdot\|_q) \ ,$$

where the sup is taken over all probability measures such that $(PH^q)^{1/q} = \|H\|_q < \infty$.

We are ready now to state two important theorems.

Theorem 1.6. *Let $T \subset L_1(P)$. Then, T is P-Glivenko-Cantelli if the associate bracketing number is finite, i.e., $N_B(\varepsilon, T, \|\cdot\|_1) < \infty$, for all $\varepsilon > 0$.*

Proof. Exercise 4 week 1! □

Theorem 1.7. *Let $T \subset L_1(P)$ composed of P-measurable functions. If T has P-integrable envelope function H, such that the entropy measure of the subset $T_K = \{h1\{H \leq K\},\ h \in T\}$ w.r.t. the empirical measure $\log N(\varepsilon, T_K, L_1(P_n))$ is $o_P(n)$ for all $\varepsilon, K > 0$, then $\|P_n - P\|_T \to 0$ a.s. and in mean.*

In particular T is P-Glivenko-Cantelli.

Remark 1.6.

1. *The condition $\log N(\varepsilon, T_K, L_1(P_n)) = o_P(n)$ is equivalent to $(1/n) \log N(\varepsilon, T_K, L_1(P_n)) \to 0$ in probability when $n \to \infty$.*

2. *The statement of the Theorem based on entropy and its proof are more complicated than the one based on bracketing numbers. However, it provides a sufficient and necessary condition that can be more easily checked, as we will see in the next section in particular.*

In fact, similar conditions can be showed to characterize Donsker classes yielding uniform central limit theorems. The first is based on bracketing numbers, while the second on requires some finite integral of the uniform entropy number. We will not prove the latter, as it requires techniques that will be studied at length in the next chapters.

Theorem 1.8. *Let $T \subset L_2(P)$ composed of measurable functions. Then, T is P-Donsker if*

$$\int_0^\infty \sqrt{\log N_B(\varepsilon, T, \|\cdot\|_2)}\, \mathrm{d}\varepsilon < \infty \ .$$

Theorem 1.9. *Let $T \subset L_1(P)$. Then, T is P-Donsker if it is P-measurable, with P-integrable envelope function H, if its uniform covering number is finite, i.e.,*

$$\int_0^1 \sup_Q \sqrt{\log N(\varepsilon\|H\|_1, T, L_1(Q))}\, \mathrm{d}\varepsilon < \infty \ .$$

Now that the necessary conditions for weak uniform convergence to hold true, we need practical conditions to check. The following section exhibits classes of functions and sets with combinatorial properties fulfill the necessary conditions above.

2 A special case: Vapnik-Chervonenkis classes

This section will show that particular classes, known as Vapnik-Chervonenkis (VC) classes of sets, provide polynomial upperbound of the uniform covering number in $1/\varepsilon$.

2.1 Definitions

Example 2.1 (Binary class of functions.)**.** *We start with a simple class $\mathcal{H}$ of measurable functions taking values in $\{0,1\}$. For any n-sample $x^n = \{x_1, \dots, x_n\}$, we want to measure the complexity of the class of functions $\mathcal{H}$ valued at x^n. The complexity is related to the size of the set valued at x^n, that is related to its cardinality.*

Because the functions take binary values, then $\mathcal{H}(x^n)$ contains at most 2^n elements. If there exists a **finite** *sample size $n < \infty$ such that $\mathcal{H}$ describes all possible values, i.e., $|\mathcal{H}(x^n)| = 2^n$, then $\mathcal{H}$ is said to be a* VC*-class.*

We now define the general notions underlying Example 2.1.

Definition 2.1. Let $\mathcal{C}$ be a class of measurable sets of the sample space $\mathcal{X}$. We say that $\mathcal{C}$ picks out a subset A of the sample $x^n = \{x_1, \dots, x_n\}$ if it can be described by a subset of $C \subset \mathcal{C}$. That is, if there exists a $C \subset \mathcal{C}$ such that

$$\{x_1, \dots, x_n\} \cap C = A \ .$$

We define the shattering coefficient by

$$m_n(\mathcal{C}) = \sup_{|A|=n,\ A\subset\mathcal{X}} |\{A \cap C,\ C \in \mathcal{C}\}| \ .$$

The VC-dimension of $\mathcal{C}$ is defined by

$$\mathcal{V} = \sup_{n\in\mathbb{N}}\{m_n(\mathcal{C}) = 2^n\} \ ,$$

i.e. such that $\mathcal{C}$ shatters the sample of size $\mathcal{V}$. We say that $\mathcal{C}$ is a Vapnik-Chervonenkis class (VC-class) if its VC-dimension is finite, i.e., $\mathcal{V} < \infty$.

The VC-dimension of a function class is related to the largest n (sample size), for which there is some collection of points x^n, that is shattered by $\mathcal{C}$, i.e., such that for any function of $\mathcal{C}$, the function valued at this collection of points x^n can describe all possible 2^n values.

Example 2.2 (Half-spaces in $\mathbb{R}^d$)**.** *Show that the class of half-spaces ($\{(-\infty, x],\ x \in \mathbb{R}\}$) is a* VC*-class of sets of* VC*-dimension equal to* 1*.*

Proof for $d = 1$: For any single point x_1, both subsets $\{x_1\}$ and the empty set can be picked out by that class.

If we consider two distinct points $x_1 < x_2$, then it is impossible to find an interval $(-\infty, x]$ containing x_2 but not x_1. Thus $\mathcal{V} = 1$.

Example 2.3 (Intervals in $\mathbb{R}^d$)**.** *Show that the class of half-spaces ($\{(y, x],\ x > y \in \mathbb{R}\}$) is a* VC*-class of sets of* VC*-dimension equal to $d + 1$.*

Proof for $d = 1$: For any single point x_1, both subsets $\{x_1\}$ and the empty set can be picked out by that class.

If we consider two distinct points $x_1 < x_2$, then it is possible to find an interval $(y, x]$ containing x_2 but not x_1. Now consider three distinct points $x_1 < x_2 < x_3$, it is impossible to shatter the extremes x_1, x_3, without containing x_2. Thus $\mathcal{V} = 2$.

Exercise, extend to $d \geq 2$.

Although the intuition of the previous example is quite straightforward, other classes of functions might require some strong results, such as the following ones.

Lemma 2.2 (Sauer's lemma)**.** *Let $\mathcal{C}$ be a* VC*-class of finite dimension $\mathcal{V}$. Then, for any $n \geq \mathcal{V}$*

$$m_n(\mathcal{C}) \leq \sum_{i=1}^{\mathcal{V}} \binom{n}{i} .$$

In particular $m_n(\mathcal{C}) \leq (n+1)^{\mathcal{V}}$.

Lemma 2.3 (Vector spaces)**.** *Suppose $\mathcal{H}$ to be a class of measurable functions $h : \mathcal{X} \to \mathbb{R}$ of finite dimension. Then, $\mathcal{H}$ forms a* VC*-class with* VC*-dimension bounded by* $\dim(\mathcal{H}) + 2$*.*

Notice the relation we have with the Example 2.1 by defining the class $\mathcal{H} = \{1_C,\ C \in \mathcal{C}\}$. More generally, we can define VC-classes of functions as follows.

Definition 2.4. Consider a class $\mathcal{H}$ of measurable functions $h : \mathcal{X} \to \mathbb{R}$. The *subgraph* of any function $h \in \mathcal{H}$, is defined by $\{(x,t),\ t < h(x)\}$, for any $t \in \mathbb{R}$. Then, we say that $\mathcal{H}$ is VC-class, if the class of subgraphs on $\mathcal{X} \times \mathbb{R}$ is VC-class of sets.

In applications such as in statistical learning, VC-classes of functions are fundamental classes (see Exercise session).

We state basic operations for VC-classes of measurable functions such that the VC-type is preserved. Similar properties can be formulated in terms of classes of sets.

Proposition 2.5 (VC-permanence properties)**.** *Let $\mathcal{H}$, $\mathcal{G}$ be two* VC*-classes of measurable functions. Let $\Phi : \mathbb{R} \to \mathbb{R}$ monotone and $\Psi : \mathcal{X}' \to \mathcal{X}$ be fixed. The following classes are* VC*-subgraphs:*

- $-\mathcal{H}$
- $\mathcal{H} \vee \mathcal{G} = \{h \vee g,\ h \in \mathcal{H},\ g \in \mathcal{G}\}$, $\mathcal{H} \wedge \mathcal{G}$

- $\mathcal{H} \circ \Psi = \{h(\Psi),\ h \in \mathcal{H}\}$
- $\Phi \circ \mathcal{H} = \{\Phi(h),\ h \in \mathcal{H}\}$

Example 2.4. *Show that the class of translates $\{h(x-t), t \in \mathbb{R}\}$, with h being fixed, has* VC*-dimension equal to* 2*.*

2.2 VC-dimension and Covering Numbers

A very useful property is that the uniform entropy number can be upperbounded by an exponentially decreasing function of the VC-dimension. That is to say that if a class of sets is VC, then we can apply uniform convergence theorems.

Lemma 2.6. *Let P be a probability measure on $(\mathcal{X}, \mathcal{A})$. Let $\mathcal{H}$ be a class of measurable functions, such that it has a square-integrable measurable envelope function $H(t)$, i.e., $PH^2 < \infty$*

$$N(\varepsilon \|H\|_2, \mathcal{H}, \|\cdot\|_2) \leq C \left(\frac{1}{\varepsilon}\right)^{\mathcal{V}} ,$$

where the norm here is relative to $L_2(P)$, and $C > 0$ is a universal constant.

Lemma 2.6 can be extended to subsets of $L_q(P)$, with $q \geq 1$, as soon as the envelope function has $L_q(P)$-norm strictly positive. However, the constant C will not be universal anymore, and depends on the VC-dimension. See Theorem 2.6.7. in (van der Vaart and Wellner, 1996).

Remark 2.5. *Lemma 2.6 and generalizations of it are fundamental in empirical process theory, insofar as they provide guarantees for stochastic processes indexed by infinite classes satisfying the conditions for their uniform asymptotic behaviors (e.g. Glivenko-Cantelli's and Donsker's types of theorems). In statistical applications, under some conditions, it can be shown that plug-in estimator are uniformly consistent.*

Remark 2.6. *This* VC*-classes are more conservative as Lemma 2.6. holds true for the covering number and not their entropy.*

3 Random measures of complexity: Rademacher and Gaussian processes

In the last sections, we saw how to describe the size of the set T in terms of deterministic concepts, w.r.t. probability measures. We present, here, a modern approach to describe the structure of the set T: the intuition is to deduce information about T through the knowledge of the data that we have. Precisely, consider $T \subset \mathbb{R}^d$, we define the following two complexity measures.

Example 3.1 (Binary classification)**.** *Consider the binary classification problem, i.e., based on an i.i.d. sample, we want to find the best classifier $h : \mathcal{X} \to \{-1, +1\}$ that assigns to each observation the true label $Y_i \in \{-1, +1\}$. Then, minimizing the classification error, $\sum_{i=1}^n 1\{h(X_i) \neq Y_i\}$, is equivalent to minimizing $\sum_{i=1}^n Y_i h(X_i)$ w.r.t. $h \in \mathcal{H}$.*

A classic risk measure in learning theory is known as the Rademacher empirical complexity taking the form of $\sup_{h \in \mathcal{H}} \sum_{i=1}^n \varepsilon_i h(X_i)$, where the ε_i-s are i.i.d. ± 1 r.v.s. It quantifies the worst statistical error we could commit based on the function class $\mathcal{H}$.

3.1 Definitions and Examples

Definition 3.1 (Rademacher complexity)**.** Let ε be a Rademacher r.v., such that $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2$. The *Rademacher canonical process* indexed by $T \subset \mathbb{R}^d$ is defined by

$$R_t = \langle \varepsilon, t \rangle = \sum_{i=1}^d \varepsilon_i t_i \ .$$

The *Rademacher complexity* of T is given by $\mathcal{R}(T) = \mathbb{E}[\sup_{t\in T} R_t]$.

We can similarly define standard Gaussian complexity.

Definition 3.2 (Gaussian complexity)**.** Let η be a standard Gaussian r.v.. The *Gaussian canonical process* indexed by $T \subset \mathbb{R}^d$ is defined by

$$G_t = \langle \eta, t\rangle = \sum_{i=1}^{d} \eta_i t_i \ .$$

The *Gaussian complexity* of T is given by $\mathcal{G}(T) = \mathbb{E}[\sup_{t\in T} G_t]$.

These canonical complexities and why they play a key role in the control of the complexity of sets as shown below.

Example 3.2 (Euclidean ball)**.** *Define the Euclidean ball by* $T = \{t \in \mathbb{R}^d,\ \|t\|_2 \le 1\}$*. First, by Cauchy-Schwarz inequality we have*

$$\mathcal{R}(T) = \mathbb{E}\left[\sup_{\|t\|_2\le 1} \langle \varepsilon, t\rangle\right] = \mathbb{E}\left[(\sum_{i=1}^{d} \varepsilon_i^2)^{1/2} \sup_{\|t\|_2\le 1} (\sum_{i=1}^{d} t_i^2)^{1/2}\right] = \mathbb{E}\left[(\sum_{i=1}^{d} \varepsilon_i^2)^{1/2}\right] = \sqrt{d}\ . \tag{1}$$

For the Gaussian complexity, we similarly obtain

$$\mathcal{G}(T) = \mathbb{E}\left[\sup_{\|t\|_2\le 1} \langle \eta, t\rangle\right] = \mathbb{E}\left[(\sum_{i=1}^{d} \eta_i^2)^{1/2} \sup_{\|t\|_2\le 1} (\sum_{i=1}^{d} t_i^2)^{1/2}\right] = \mathbb{E}\left[(\sum_{i=1}^{d} \eta_i^2)^{1/2}\right] \tag{2}$$

Then, the square-root function being concave, Jensen's inequality applies

$$\mathcal{G}(T) \le \sqrt{\mathbb{E}\left[\sum_{i=1}^{d} \eta_i^2\right]} = \sqrt{d}\ . \tag{3}$$

Recall that the ℓ_2-norm of a standard d-dimentional Gaussian variable equals to d.

Example 3.3 (ℓ_1-balls)**.** *Define the ℓ_1-ball by* $T = \{t \in \mathbb{R}^d,\ \|t\|_1 \le 1\}$*. Prove that* $\mathcal{R}(T) = 1$ *and* $\mathcal{G}(T) \le C\sqrt{\log d}$*.* **Exercise**

Remark 3.4. *Notice that Rademacher and Gaussian complexities are comparable, and we will give an explicit constant factor in the next Proposition.*

3.2 Properties

Proposition 3.3 (Comparison $\mathcal{R}$ and $\mathcal{G}$)**.** *Consider T to be composed of elements valued in $\mathbb{R}^d$. Let η be a standard Gaussian r.v. and ε be a Rademacher r.v.*
Then,

$$\sqrt{\frac{2}{\pi}}\,\mathcal{R}(T) \le \mathcal{G}(T) \le 2\sqrt{\log d}\,\mathcal{R}(T)\ .$$

Proof. Board. □

Remark 3.5. *Proposition 3.3 shows that, although Rademacher and Gaussian complexities are comparable, the latter can be substantially larger than the Rademacher one. In fact, Rademacher complexity will be mainly used in the problems studied in this class (and also in the literature).*

Another property of both complexities that turns out to be very useful in practice, especially in Empirical Risk Minimization problems, is that of contraction. Consider a fixed function $\Phi = (\Phi_i,\ i = 1, \dots, d)$, centered , i.e., $\Phi_i(0) = 0$, and L-Lipschitz. Suppose we are interested in the image of the complexities by this map defined by

$$\mathcal{R}(\Phi(T)) = \mathbb{E}\left[\sup_{t \in T} \sum_{i=1}^{d} \Phi_i(t_i)\varepsilon_i\right] \ ,$$

and

$$\mathcal{G}(\Phi(T)) = \mathbb{E}\left[\sup_{t \in T} \sum_{i=1}^{d} \Phi_i(t_i)\eta_i\right] \ .$$

Proposition 3.4 (Contraction property)**.** *Consider T to be composed of elements valued in $\mathbb{R}^d$, and a centered L-Lipschitz function $\Phi = (\Phi_1, \dots, \Phi_d)$, i.e., $\Phi_i(0) = 0$. The following assertions hold true.*

1. $\mathcal{R}(\Phi(T)) \leq 2L\mathcal{R}(T)$
2. $\mathcal{G}(\Phi(T)) \leq L\mathcal{G}(T)$

Proof. Exercise. □