Institute of Mathematics, EPFL MATH524

Chapter 5: Introduction to empirical processes

Lecturer: Rajita Chandak *Spring 2025*

So far we have focused on specific non-parametric estimators of functions of i.i.d. data. We have evaluated these estimators against measures like the mean-squared error, bias and integrated mean squared error. While these estimators were natural in the parametric setting for point estimates, they don't always make sense when evaluating estimators for functions. Consider for example, the CDF F and our empirical CDF estimator given by

$$F_n(x) = \frac{1}{n}\sum_i \mathbf{1}(X_i \leq x).$$

We know that $F_n(x)$ is binomially distributed with mean $F(x)$ and variance $F(x)(1-F(x))/n$. We have already seen that by LLN, the estimator is consistent and by the CLT it is asymptotically normal with mean 0 and variance $F(x)(1-F(x))$. However, these results only capture the properties of the estimator at a given point x. These results tell us nothing about how we expect the estimator to perform over the entire domain, or what the worst case behavior of the estimator may look like when F can be any possible CDF. To address such questions, instead of thinking of the ECDF as a point estimate for an infinite sequence of x, we need to think of the estimator as a random function.

Lets consider the induced random probability measure from the i.i.d. sample $\mathbb{P}_n$ as an integral operator on $L^1(\mathbb{P})$ where

$$\mathbb{P}_n(f) = \frac{1}{n}\sum_i f(X_i).$$

The population analogue of which is the limiting measure,

$$\mathbb{P}(f) = \mathbb{E}[f(X)].$$

We can then study the ECDF (an similar functionals) as a stochastic process, $(\mathbb{P}_n f - \mathbb{P}f : f \in \mathcal{F})$ for $\mathcal{F} \subseteq L^1(\mathbb{P})$.

1 A new statistic

Lets start with looking at the result of LLN again. We see that for the ECDF, LLN implies

$$F_n(x) \xrightarrow{a.s.} F(x) \quad \textbf{for every } x.$$

When we want to understand *functional behavior*, we want to make conclusions of the form

$$\|F_n - F\|_\infty \to ?$$

We can then use the test statistic $\|F_n - F\|_\infty = \sup_x |F_n(x) - F(x)|$, and attempt to understand its distribution under the null hypothesis. This statistic is also known as the Kolmogorov-Smirnov statistic.

2 Uniformity of the ECDF

The following theorem establishes the behavior of the Kolmogorov-Smirnov statistic.

Theorem 1 (Glivenko-Cantelli)
Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$. *Then,* $\|F_n - F\|_\infty \overset{a.s.}{\to} 0$.

Proof of Theorem 1. By SLLN, $F_n(x) \overset{a.s.}{\to} F(x)$ and $F_n(x-) \overset{a.s.}{\to} F(x-)$ for every x. For a fixed $\varepsilon > 0$, there exists a partition $-\infty = t_0 < t_1 < \ldots < t_k = \infty$ such that $F(t_i-) - F(t_{i-1}) < \varepsilon$ for every i (when F jumps by more than ε, the pont is taken to be one of the points in the partition). Then, for $t_{i-1} < x < t_i$,

$$F_n(x) - F(x) \leq F_n(t_i-) - F(t_i-) + \varepsilon,$$
$$F_n(x) - F(x) \geq F_n(t_{i-1}) - F(t_{i-1}) - \varepsilon.$$

We then see that

$$\|F_n - F\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \leq \max_{i \in \{1,\ldots,k\}} |F_n(t_k) - F(t_k)| + \varepsilon.$$

Since $\max_{i \in \{1,\ldots,k\}} |F_n(t_k) - F(t_k)| \to 0$ almost surely by SLLN, we can choose k appropriately such that $\|F_n - F\|_\infty \leq \varepsilon$, giving us almost sure convergence.. ■

We can use this result to define a new class of measurable functions: the *Glivenko-Cantelli* class.

Definition 1 ($\mathbb{P}$-GC)**.** *A class of measurable functions* $\mathcal{F}$ *is* $\mathbb{P}$*-GC if*

$$\|\mathbb{P}_n(f) - \mathbb{P}(f)\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbb{P}_n(f) - \mathbb{P}(f)| \overset{a.s.}{\to} 0.$$

Any finite class of integrable functions can be shown to be $\mathbb{P}$-GC. But it should be straightforward to see that the class of *all* square-integrable is not $\mathbb{P}$-GC.

While this notion of uniform (almost sure over the whole space) convergence is already a very powerful feature of the ECDF, we can make the following stronger statement about the empirical CDF.

Theorem 2 (Dvoretzky-Kiefer-Wolfowitz inequality)
Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$. *Then, for every* $\varepsilon > 0$,

$$\mathbb{P}(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| > \varepsilon) \leq 2e^{-2n\varepsilon^2}$$

This result allows for the direct construction of non-parametric confidence intervals: Let $\varepsilon_n^2 = \log(2/\alpha)/(2n)$, $L(x) = \max\{\hat{F}(x) - \varepsilon_n, 0\}$ and $U(x) = \max\{\hat{F}(x) + \varepsilon_n, 1\}$. Then,

$$\mathbb{P}(\forall x : L(x) \leq F(x) \leq U(x)) \geq 1 - \alpha.$$

This is a fully non-parametric confidence band for the ECDF.

Corollary 1 (Uniform Glivenko-Cantelli)
Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$. Let $\mathcal{F}$ be a class of functions that is uniformly bounded (i.e., $\|F\|_\infty \leq b, \forall F \in \mathcal{F}$). Then, for every $\varepsilon > 0$,

$$\sup_{F \in \mathcal{F}} \mathbb{P}(\sup_{m \geq n} \sup_{x \in \mathbb{R}} |\hat{F}_m(x) - F(x)| > \varepsilon) \to 0$$

as $n \to \infty$.

The proof of the corollary follows from a union bound and applying the DKW inequality.

We have used GC to extend LLN to a *uniform* notion. Can we do something similar to extend the CLT? We start by defining the empirical process $G_n = \sqrt{n}(F_n - F)$ and the associated covariance function $\Sigma = \mathbb{C}\text{ov}[G_n(x_i), G_n(x_j)] = \mathbb{E}[F(x_i \wedge x_j)] - \mathbb{E}[F(x_i)]\mathbb{E}[F(x_j)]$. Then, the Gaussian process indexed by F is given by $G_F \sim \mathcal{N}(0, \Sigma)$. We consider the *Skorohod space* $D[-\infty, \infty]$, defined as follows.

Definition 2 (Skorohod space)**.** *Let (M, d) be a metric space and $E \subseteq \mathbb{R}$. Functions $f : E \to M$ such that for every $t \in E$,*

- *the left limit $f(t_-) = \lim_{s \to t_-} f(s)$ exists*
- *the right limit $f(t_+) = \lim_{s \to t_-} f(s)$ exists and is equal to $f(t)$*

(i.e., f is right-continuous with left limits), are known as cadlag functions. The collection of these functions makes up the Skorohod space, typically denoted as $D(M)$.

The Skorohod space equipped with the uniform norm wherein the limiting process G_F is referred to as a *Brownian bridge* (a conditional Gaussian process). Now, we can state Donsker's theorem.

Theorem 3 (Donsker)
Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$. Then, $\sqrt{n}(F_n - F)$ converges in the space $D[-\infty, \infty]$ to a random element G_F, whose distribution is the Brownian bridge with mean zero and covariance function Σ.

This theorem in combination with additional methods for handling random functions can be used to yield strong results on the functional approximation of the ECDF. Here we briefly discuss the concept of *strong approximations*. Consider a probability space with i.i.d $\{X_i\} \sim F$ and a sequence of Brownian bridges $G_{F,n}$ such that, almost surely,

$$\limsup_{n \to \infty} \frac{\sqrt{n}}{(\log n)^2} \| \sqrt{n}(F_n - F) - G_{F,n} \|_\infty < \infty,$$

Then, by Donsker's theorem we have distributional convergence. Furthermore, a bound of the type

$$\mathbb{P}\left(\|\sqrt{n}(F_n - F) - G_{F,n}\|_\infty > \frac{a\log n + x}{\sqrt{n}}\right) \leq be^{-cx}$$

for some fixed constants a, b, c and every $x > 0$ can be established. This is known as the Komlos-Major-Tusnady (KMT) approximation (or Hungarian embedding).

Now, analogous to the GC function class, we can define the Donsker class. Consider the empirical process evaluated at f, given by $G_n(f) = \sqrt{n}(\mathbb{P}_n(f) - \mathbb{P}(f))$. By the multivariate CLT, for any finite set of functions f_i with $\mathbb{P}(f_i^2) < \infty$,

$$(G_n(f_1), \ldots, G_n(f_k)) \rightsquigarrow (G_{\mathbb{P}}(f_1), \ldots, G_{\mathbb{P}}(f_k))$$

where the covariance of the multivariate normal on the right-hand side is given by $\Sigma_{fg} = \mathbb{P}(fg) - \mathbb{P}(f) - \mathbb{P}(g)$. Then, by the Donsker theorem, this result can be made uniform in the class of functions.

Definition 3 (Donsker class)**.** *A class of measurable functions $\mathcal{F}$ is $\mathbb{P}$-Donsker if the sequence $\{G_n(f) : f \in \mathcal{F}\}$ converges in distribution to a limiting process in the $L^\infty(\mathcal{F})$ space. The limit process is given by a Gaussian process $G_{\mathbb{P}}$ with zero mean and covariance $\Sigma_{fg} = \mathbb{P}(fg) - \mathbb{P}(f)\mathbb{P}(g)$.*

The Donsker class includes the requirement that all sample paths $f \mapsto G_n(f)$ are uniformly bounded for all n and all realizations of $\{X_i\}$. This is automatically satisfied if, for example, $\mathcal{F}$ has a finite and integrable *envelope function*: $|f(x)| < F_e(x) < \infty$ for every x and f. Note that this definition does not require F_e to be uniformly bounded.

Once again, we see that any finite class of square-integrable functions is $\mathbb{P}$-Donsker. However, infinite classes of square-integrable functions will not always be $\mathbb{P}$-Donsker.

So far, we have focused our definitions and results to the class of functions that are associated with the ECDF. More generally, we would like to establish results of the type

$$\sup_{f\in\mathcal{F}_n}\left|\frac{1}{n}\sum f(X_i) - \mathbb{E}[f(X)]\right| \xrightarrow{n\to\infty} 0 \qquad \text{a.s.}$$

wherein $\mathcal{F}_n$ is a general function class. Recall that by the SLLN we can establish the following under the assumption that for some f such that $\mathbb{E}[|f(X)|] < \infty$,

$$\lim_{n\to\infty}\left|\frac{1}{n}\sum f(X_i) - \mathbb{E}[f(X)]\right| = 0 \qquad \text{a.s.}$$

The following inequality will allow us to extend this statement to the class $\mathcal{F}_n$, provided that $f : \mathbb{R}^d \to [0, B]$ (all functions are bounded).

$$\mathbb{P}\left(\left|\frac{1}{n}\sum f(X_i) - \mathbb{E}[f(X)]\right| > \varepsilon\right) \leq 2\exp\left(-2\frac{n\varepsilon^2}{B^2}\right).$$

This inequality is widely referred to as Hoeffding's inequality. Which, with the union bound, can be extended to the whole class as

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}_n}\left|\frac{1}{n}\sum f(X_i)-\mathbb{E}[f(X)]\right|>\varepsilon\right)\leq 2|\mathcal{F}_n|\exp\left(-2\frac{n\varepsilon^2}{B^2}\right).$$

Of course, for the statement to be non-vacuous, we need $\mathcal{F}_n$ to be a finite class.

We will now develop notions to define the size of classes to help further identify other definitions of "finite" function classes.

3 Bracketing numbers

We will first introduce the idea of measuring function class sizes measured by the *bracketing number*. Start by defining the bracketing entropy for the $L_r(\mathbb{P})$ norm as

$$\|f\|_{\mathbb{P},r}=(\mathbb{P}(|f|^r))^{1/r}.$$

Given two functions l and u, the bracket $[l,u]$ is the collection of all functions f such that $l\leq f\leq u$. An ε-bracket in $L_r(\mathbb{P})$ is the bracket such that $\mathbb{P}((u-l)^r)<\varepsilon$. Then, the bracketing number $N_{[]}(\varepsilon,\mathcal{F},L_r(\mathbb{P}))$ is the minimum number of ε-brackets needed to cover $\mathcal{F}$. By definition, u and l need to have finite $L_r(\mathbb{P})$ norm but they do not have to belong to $\mathcal{F}$.

Now we can use this measure of class size to identify $\mathbb{P}$-GC and $\mathbb{P}$-Donsker classes:

Theorem 4 (GC identification)
Every class $\mathcal{F}$ of measurable functions with a finite L_1-bracketing number ($N_{[]}(\varepsilon,\mathcal{F},L_1(\mathbb{P}))<\infty$) for every $\varepsilon>0$ is $\mathbb{P}$-GC.

We would like to work with the bracketing number. Unfortunately in many cases, $N_{[]}(\varepsilon,\mathcal{F},L_r(\mathbb{P}))\to\infty$ as $\varepsilon\to 0$. For Donsker classes, fortunately, a sufficient condition is that the class size (bracketing number) does not grow too fast. The rate of growth can be understood with the L_2-bracketing integral:

$$J_{[]}(\delta,\mathcal{F},L_2(\mathbb{P}))=\int_0^\delta\sqrt{\log N_{[]}(\varepsilon,\mathcal{F},L_2(\mathbb{P}))}\mathrm{d}\varepsilon.$$

Theorem 5 (Donsker class identification)
Every class $\mathcal{F}$ of measurable functions with a finite bracketing integral ($J_{[]}(1,\mathcal{F},L_2(\mathbb{P}))<\infty$) is $\mathbb{P}$-Donsker.

Note that $J_{[]}$ is a decreasing function of ε. Since $\int_0^1\varepsilon^{-r}\mathrm{d}\varepsilon$ converges for $r<1$ and diverges for $r\geq 2$, the finite integral condition translates to the log-bracketing number growing slower than $(1/\varepsilon)^2$.

Lets now look at some simple examples of how the bracketing number for different classes can be computed and then used to identify GC and Donsker classes.

Example 1 (Distribution function)**.** *Let $\mathcal{F} = \{f : f(x) = \mathbf{1}(x), x \in \mathbb{R}\}$. Then, the process $G_n(f)$ is the empirical process we are already familiar with. We can directly identify from the theorems that the class is both Glivenko-Cantelli and Donsker. Lets start with the bracketing number. Consider the brackets $[\mathbf{1}(x_{i-1}), \mathbf{1}(x_i)]$ for a set of ordered points $-\infty = x_0 < x_1 < \cdots < x_k = \infty$ such that $F(x_i) - F(x_{i-1}) < \varepsilon$ for each i. Then, by construction each bracket has size ε in L_1 and k can be chosen carefully to be smaller than $2/\varepsilon$. Now, since $\mathbb{P}(f^2) \leq \mathbb{P}(f)$ and $0 \leq f \leq 1$ for all $f \in F$, the L_2 brackets are no larger than $\sqrt{\varepsilon}$. Thus, $N_{[]}(\sqrt{\varepsilon}, \mathcal{F}, L_2) \leq (2/\varepsilon)$. Of course, the bracketing integral is also bounded by this argument.*

Example 2 (Parametric class)**.** *Consider $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ be a collection of measurable functions and let $\Theta \subset \mathbb{R}^d$ be bounded. Suppose there exists a measurable function m s.t.*

$$|f_{\theta_1} - f_{\theta_2}| \leq m(x)\|\theta_1 - \theta_2\|$$

for all θ_1, θ_2. If $\mathbb{P}(|m|^r) < \infty$, then, there exists a constant K such that

$$N_{[]}(\varepsilon\|m\|_{\mathbb{P},r}, \mathcal{F}, L_2) \leq K\left(\frac{diam(\Theta)}{\varepsilon}\right)^d$$

for every $0 < \varepsilon < diam(\Theta)$. This gives a bounded bracketing integral, and so the class $\mathcal{F}$ is Donsker.

4 Covering numbers

An alternative to computing the bracketing number for identification of GC or Donsker classes is to instead compute the L_p-covering number, $N(\varepsilon, \mathcal{F}, L_p)$, which is the minimal number of L_p balls of radius ε needed to cover $\mathcal{F}$. Usually, we take $p = 2$.

Definition 4 (ε-covering)**.** *Let $\varepsilon > 0$ and $\mathcal{F} = \{f_i : \mathbb{R}^d \to \mathbb{R}\}$ be a set of functions. The finite collection of functions $\{f_j\}_{j=1}^N$ such that for each $f \in \mathcal{F}$ there is a $j = j(f) \in \{1, \ldots, N\}$ such that*

$$\|f - f_j\|_\infty := \sup_x |f(x) - f_j(x)| < \varepsilon,$$

is called an ε-cover of $\mathcal{F}$ with respect to the L_∞ norm.

Definition 5 (ε-covering number)**.** *Let $\varepsilon > 0$ and $\mathcal{F} = \{f_i : \mathbb{R}^d \to \mathbb{R}\}$ be a set of functions. We use $N(\varepsilon, \mathcal{F}, L_2)$ to be the size of the smallest ε cover of $\mathcal{F}$. $N(\varepsilon, \mathcal{F}, L_2) = \infty$ if there does not exist any finite cover.*

Lemma 1
Let $\mathcal{F} = \{f : \mathbb{R}^d \to [0, B]\}$ and $\varepsilon > 0$. Then,

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}_n}\left|\frac{1}{n}\sum f(X_i) - \mathbb{E}[f(X)]\right| > \varepsilon\right) \leq 8\mathbb{E}[N(\varepsilon/8, \mathcal{F}, X_1^n)]\exp\left(-\frac{n\varepsilon^2}{128B^2}\right).$$

where the X_1^n implies we use the empirical measure with respect to the observed data.

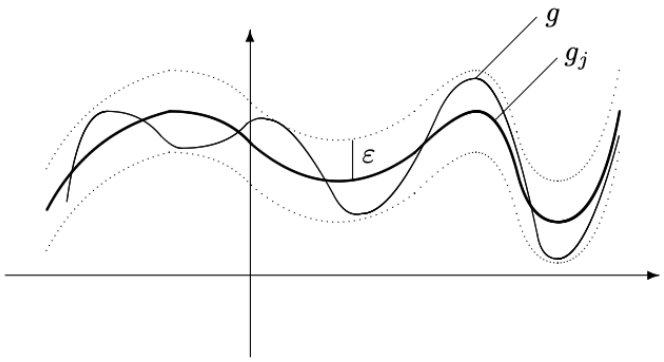

Figure 1: Sup norm distance between function g and member g_j of cover is less than ε (taken from Györfi et al. (2006))

As it turns out, all the results with bracketing numbers can be replaced by the uniform covering number $\sup_{\mathbb{P}} N(\varepsilon\|F\|_r, \mathcal{F}, L_r(\mathbb{P}))$ where the supremum is over all probability measures $\mathbb{P}$ for which $\mathcal{F}$ is not identically zero. The uniform entropy integral is then defined as

$$J(\delta, \mathcal{F}, L_2) = \int_0^\delta \sqrt{\log \sup_{\mathbb{P}} N(\varepsilon\|F\|_2, \mathcal{F}, L_2)} \mathrm{d}\varepsilon$$

Then, we can state the alternative theorems:

Theorem 6 (GC identification)
For the class $\mathcal{F}$ of measurable functions with a finite covering number ($\sup_{\mathbb{P}} N(\varepsilon\|F\|_1, \mathcal{F}, L_1(\mathbb{P})) < \infty$) for every $\varepsilon > 0$. If $\mathbb{P}(F) < \infty$, then $\mathcal{F}$ is $\mathbb{P}$-GC.

Theorem 7 (Donsker class identification)
Every class $\mathcal{F}$ of measurable functions with a finite covering integral ($J(1, \mathcal{F}, L_2) < \infty$) and $\mathbb{P}(F^2) < \infty$ is $\mathbb{P}$-Donsker.

5 Packing numbers

We now introduce one final notion of class size:

Definition 6 (L_p packing numbers)**.** *Let $\varepsilon > 0$ and $\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R}\}$, $1 \leq p < \infty$ and ν is a probability measure on $\mathbb{R}^d$. For every finite collection $f_1, \ldots, f_N \in \mathcal{F}$ with*

$$\|f_j - f_k\|_{L_p(\nu)} \geq \varepsilon$$

for all $1 \leq j < k \leq N$ is called an ε-packing of $\mathcal{G}$ for the L^p norm. Let $\mathcal{M}(\varepsilon, \mathcal{F}, L_p(\nu))$ be the largest possible ε-packing and set it equal to infinity if there exists a packing for every N.

Now we can establish the following relationship between covering and packing numbers.

Lemma 2
$\mathcal{F} = \{f \in \mathbb{R}^d\}$ and ν is a probability measure with $p \geq 1$, $\varepsilon > 0$. Then,

$$\mathcal{M}(2\varepsilon, \mathcal{F}, L_p(\nu)) \leq N(\varepsilon, \mathcal{F}, L_p(\nu)) \leq \mathcal{M}(\varepsilon, \mathcal{F}, L_p(\nu)).$$

6 VC Dimension

Lets now turn back to the covering number. It turns out that there is a special class of functions for which the covering numbers can be easily computed: the *Vapnik-Červonenkis* (VC) class.

Definition 7 (Shattering coefficient)**.** *Let $\mathcal{A}$ be a class of subsets. For some $x_1, \ldots, x_n \in \mathbb{R}^d$, define*

$$s(\mathcal{A}, \{x_1, \ldots, x_n\}) = |\{A \cap \{x_1, \ldots, x_n\} : A \in \mathcal{A}\}|,$$

*as the number of different subsets that can be generated by intersecting elements of $\mathcal{A}$ with the collection of points. $\mathcal{A}$ is said to **shatter** $\{x_1, \ldots, x_n\}$ if each of the 2^n subsets can be picked out by some element $A \in \mathcal{A}$. The **shatter coefficient** of $\mathcal{A}$ is given by*

$$S(\mathcal{A}, n) = \max_{\{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d} s(\mathcal{A}, \{x_1, \ldots, x_n\}).$$

i.e., the shatter coefficient is the maximal number of different subsets of n points that can be picked up by sets from $\mathcal{A}$.

It should be clear from the definition that $s(\mathcal{A}, \{x_1, \ldots, x_n\}) \leq 2^n$ and $S(\mathcal{A}, n) \leq 2^n$. Furthermore, if $S(\mathcal{A}, n) < 2^n$, then by construction $s(\mathcal{A}, \{x_1, \ldots, x_n\}) < 2^n$ for all $x_1, \ldots, x_n$. Importantly, if $s(\mathcal{A}, \{x_1, \ldots, x_n\}) < 2^n$ then there exists a subset of $\{x_1, \ldots, x_n\}$ that cannot be picked out by any set in $\mathcal{A}$.

The VC dimension, $V(\mathcal{A})$, is then defined as the smallest n for which no set of size n can be shattered by $\mathcal{A}$ or, (essentially) equivalently, sometimes, as the largest n for which the set can be shattered. Note that these definitions are equivalent up to being off by exactly 1.

Definition 8 (VC dimension)**.** *$\mathcal{A} \neq \emptyset$ is a class of subsets. The VC dimension is defined to be*

$$V(\mathcal{A}) = \sup\{n : S(\mathcal{A}, n) = 2^n\} = \inf\{n : S(\mathcal{A}, n) < 2^n\} - 1.$$

The collection C is called a VC class if $V(C)$ is finite. We can define VC classes of functions in a similar manner. $\mathcal{F}$ is a VC class of functions if all $\{(x, t) : f(x) < t\}$ for all $f \in \mathcal{F}$ forms a VC class of sets in $\mathcal{X} \times \mathbb{R}$. There is a one-to-one correspondence in the definition of VC classes for sets and functions: A collection of sets C is a VC class if and only if the collection of $\mathbf{1}(C), C \in C$ is a VC class of functions.

Example 3. *The class $\{(-\infty, b], b \in \mathbb{R}\}$ cannot shatter any two distinct points in $\mathbb{R}$. Thus, the VC dimension is 1.*

While the class $\{(a, b], a, b \in \mathbb{R}\}$ can shatter any two point set in $\mathbb{R}$. However, this class cannot shatter any combination of 3 points on $\mathbb{R}$. Thus, the VC dimension of this class is 2.

The next lemma presents the surprising fact that if the shattering coefficient is strictly less than 2^n it can be bounded by a polynomial of degree $V(\mathcal{A})$.

Lemma 3 (Sauer-Shelah)
Let $\mathcal{A}$ be a collection of sets in $\mathbb{R}^d$ with VC dimension $V(\mathcal{A})$. Then,

$$S(\mathcal{A}, n) \leq \sum_{i=0}^{V(\mathcal{A})} \binom{n}{i}$$

Theorem 8 (Bounded shattering coefficient)
Let $\mathcal{A}$ be a collection of sets in $\mathbb{R}^d$ with VC dimension $V(\mathcal{A}) < \infty$. Then,

$$S(\mathcal{A}, n) \leq (n+1)^{V(\mathcal{A})}$$

and for all $n \geq V(\mathcal{A})$,

$$S(\mathcal{A}, n) \leq \left(\frac{en}{V(\mathcal{A})}\right)^{V(\mathcal{A})}$$

Proof of Theorem 8. By Lemma 3 and the binomial theorem,

$$S(\mathcal{A}, n) \leq \sum_{i=0}^{V(\mathcal{A})} \binom{n}{i} = \sum_{i=0}^{V(\mathcal{A})} \frac{n!}{(n-i)!i!} \leq \sum_{i=0}^{V(\mathcal{A})} \binom{V(\mathcal{A})}{i} n^i = (n+1)^{V(\mathcal{A})}.$$

If $V(\mathcal{A})/n \leq 1$, then by Lemma 3 and the binomial theorem, again,

$$\begin{aligned}\left(\frac{V(\mathcal{A})}{n}\right)^{V(\mathcal{A})} S(\mathcal{A}, n) \leq \left(\frac{V(\mathcal{A})}{n}\right)^{V(\mathcal{A})} \sum_{i=0}^{V(\mathcal{A})} \binom{n}{i} &\leq \sum_{i=0}^{V(\mathcal{A})} \left(\frac{V(\mathcal{A})}{n}\right)^i \binom{n}{i} \\ &\leq \sum_{i=0}^{n} \left(\frac{V(\mathcal{A})}{n}\right)^i \binom{n}{i} \\ &= \left(1 + \frac{V(\mathcal{A})}{n}\right)^n \leq e^{V(\mathcal{A})}\end{aligned}$$

■

The following lemma relates the VC class to the covering number (and therefore implicitly also to the packing number).

Lemma 4
There exists a universal constant K such that for any VC class $\mathcal{F}$, $0 < \varepsilon < 1$ and $r \geq 1$

$$\sup_{\mathbb{P}} N(\varepsilon \|F_e\|_r, \mathcal{F}, L_r) \leq KVC(\mathcal{F})(16e)^{VC(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{r(VC(\mathcal{F})-1)}.$$

This statement tells us that VC classes have covering numbers that are bounded by some polynomial in ε^{-1}, and are thus considered to be relatively small. This bound also shows that VC classes satisfy the conditions for $\mathbb{P}$-GC and $\mathbb{P}$-Donsker.

The following result on the VC dimension of vector spaces is very useful in practice.

Theorem 9
Let $\mathcal{G}$ be an r-dimensional vector space of real-valued functions on $\mathbb{R}^d$ (i.e., $dim(\mathcal{G}) = r < d$) and let

$$\mathcal{A} = \{\{x : g(x) \geq 0\} : g \in \mathcal{G}\}.$$

Then,

$$V(\mathcal{A}) \leq r.$$

Proof of Theorem 9. In order to prove this theorem, note that it is sufficient to show that no set of size $r + 1$ can be shattered. WLOG lets select $\{x_1, \ldots, x_{r+1}\} \in \mathbb{R}^{r+1}$ (all distinct). Define the map $M : \mathcal{G} \to \mathbb{R}^{r+1}$ as

$$M(g) = (g(x_1), \ldots g(x_{r+1}))^{\mathrm{T}}.$$

$M(\mathcal{G})$ is then a linear subspace of $\mathbb{R}^{r+1}$ and the dimensionality of $M(\mathcal{G})$ must be no larger than that of $\mathcal{G}$, i.e., $\dim(M(\mathcal{G})) \leq r$. Thus, there is a non-zero vector $\gamma \in \mathbb{R}^{r+1}$ such that $\gamma^{\mathrm{T}} M(g) = 0$ for all $g \in \mathcal{G}$. i.e., γ is orthogonal to $M(\mathcal{G})$. We can assume that at least one of the γ_i is negative (otherwise replace γ with $-\gamma$). This means that

$$\sum_{i:\gamma_i \geq 0} \gamma_i g(x_i) = \sum_{i:\gamma_i < 0} -\gamma_i g(x_i)$$

for all $g \in \mathcal{G}$. Suppose there is a $g \in \mathcal{G}$ that picks out precisely the corresponding x_i for which $\gamma_i > 0$. In this case, the LHS above would be non-negative ($\gamma_i \geq 0$ and $g(x_i) \geq 0$ implies $\gamma_i g(x_i) \geq 0$). But then the RHS would be negative and so we have a contradiction. Thus, $\{x_i, \ldots, x_{n+1}\}$ cannot be shattered. ■

7 Uniform Law of Large Numbers

We have already seen a version of this uniform law of of large numbers for the empirical CDF in the Glivenko-Cantelli theorem. Ideally we would like to generalize this uniform bound to other types of functions. Our work with covering numbers and VC dimension will help us formalize precisely this idea.

Theorem 10 (Uniform Law of Large Numbers (ULLN))
Suppose $X_1, \ldots, X_n$ are i.i.d. on some measurable space $\{\mathcal{X}, \mathcal{F}\}$. Let $\mathcal{G}$ be a class of measurable functions on $\mathcal{X}$. Then,

$$\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_i g(X_i) - \mathbb{E}[g(X)] \right| \xrightarrow{a.s.} 0$$

as $n \to \infty$.

Theorem 1 is a special case of this by taking $\mathcal{G} = \{\mathbf{1}(\cdot \le x), x \in \mathbb{R}\}$. In general, to establish the ULLN for a class of functions, the *size* of $\mathcal{G}$ must be controlled in some way.

We will prove the following version of ULLN:

Theorem 11 (ULLN for VC bounded functions)
Let $\mathcal{G} : \mathbb{R}^d \to \mathbb{R}$ be a class of measurable functions and let

$$G : \mathbb{R}^d \to \mathbb{R}, \qquad G(x) := \sup_{g \in \mathcal{G}} |g(x)|, x \in R^d$$

be an envelope of $\mathcal{G}$. Assume $\mathbb{E}[G] < \infty$ and for

$$\mathcal{G}^+ = \{\{(x,t) \in \mathbb{R}^d \times \mathbb{R}; t \le g(x)\}; g \in \mathcal{G}\},$$

with $VC(\mathcal{G}^+) < \infty$ ($\mathcal{G}^+$ is called the subgraph of $\mathcal{G}$). Then,

$$\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_i g(X_i) - \mathbb{E}[g(X)] \right| \xrightarrow{a.s.} 0$$

as $n \to \infty$.

Before we prove this result we need the following bound.

Theorem 12
Let $\mathcal{G} = \{g : \mathbb{R}^d \to [0, B]\}$ with $VC(\mathcal{G}^+) \ge 2$ and let $p \ge 1$. Let ν be a probability measure on $\mathbb{R}^d$ and let $0 < \varepsilon < B/4$. Then,

$$\mathcal{M}(\varepsilon, \mathcal{G}, L_p(\nu)) \le 3 \left(\frac{2eB^p}{\varepsilon^p} \log \frac{2eB^p}{\varepsilon^p} \right)^{VC(\mathcal{G}^+)}$$

We will use this additional result, without proof, to prove Theorem 11.

Proof of Theorem 11. For some $L > 0$,

$$\mathcal{G}_L = \{g\mathbf{1}(G \le L) : g \in \mathcal{G}\}.$$

For any $g \in \mathcal{G}$,

$$\begin{aligned}
&\left| \frac{1}{n} \sum_i g(X_i) - \mathbb{E}[g(X)] \right| \\
&\le \left| \frac{1}{n} \sum_i g(X_i) - \frac{1}{n} \sum_i g(X_i)\mathbf{1}(G(X_i) \le L) \right| + \left| \frac{1}{n} \sum_i g(X_i)\mathbf{1}(G(X_i) \le L) - \mathbb{E}[g(X)\mathbf{1}(G(X) \le L)] \right| \\
&\qquad + |\mathbb{E}[g(X)\mathbf{1}(G(X) \le L)] - \mathbb{E}[g(X)]| \\
&\le \left| \frac{1}{n} \sum_i g(X_i)\mathbf{1}(G(X_i) \le L) - \mathbb{E}[g(X)\mathbf{1}(G(X) \le L)] \right| + \frac{1}{n} \sum_i G(X_i)\mathbf{1}(G(X_i) > L) + \mathbb{E}[G(X)\mathbf{1}(G(X) > L)].
\end{aligned}$$

This implies that

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_i g(X_i)-\mathbb{E}[g(X)]\right|$$
$$\leq \sup_{g\in\mathcal{G}_L}\left|\frac{1}{n}\sum_i g(X_i)-\mathbb{E}[g(X)]\right|+\frac{1}{n}\sum_i G(X_i)\mathbf{1}(G(X_i)>L)+\mathbb{E}[G(X)\mathbf{1}(G(X)>L)].$$

Using the fact that $\mathbb{E}[G(X)]<\infty$ and the SLLN,

$$\frac{1}{n}\sum_i G(X_i)\mathbf{1}(G(X_i)>L)\xrightarrow{a.s.}\mathbb{E}[G(X)\mathbf{1}(G(X)>L)].$$

Furthermore, by definition of G,

$$\mathbb{E}[G(X)\mathbf{1}(G(X)>L)]\to 0 \qquad \text{as } L\to\infty.$$

Then, we only need to show

$$\sup_{g\in\mathcal{G}_L}\left|\frac{1}{n}\sum_i g(X_i)-\mathbb{E}[g(X)]\right|\xrightarrow{a.s.}0$$

in order to complete the proof. Note that the functions in $\mathcal{G}_L$ are bounded in absolute value by L by definition. Now, for any $\varepsilon>0$, we can apply Lemmas 1 and 2 and Theorem 12 to get the following:

$$\begin{aligned}\mathbb{P}\left(\sup_{g\in\mathcal{G}_L}\left|\frac{1}{n}\sum_i g(X_i)-\mathbb{E}[g(X)]\right|>\varepsilon\right)&\leq 8\mathbb{E}[N(\varepsilon/8,\mathcal{G}_L,X_1^n)]\exp\left(-\frac{n\varepsilon^2}{128(2L)^2}\right)\\&\leq 8\mathbb{E}[\mathcal{M}(\varepsilon/8,\mathcal{G}_L,X_1^n)]\exp\left(-\frac{n\varepsilon^2}{512L^2}\right)\\&\leq 24\left(\frac{32eL}{\varepsilon}\log\frac{48eL}{\varepsilon}\right)^{VC(\mathcal{G}_L^+)}\exp\left(-\frac{n\varepsilon^2}{512L^2}\right)\end{aligned}$$

If a collection of points are shattered by $\mathcal{G}_L^+$ then they are also shattered by $\mathcal{G}^+$. This means that $VC(\mathcal{G}_L^+)\leq VC(\mathcal{G}^+)$ which implies

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}_L}\left|\frac{1}{n}\sum_i g(X_i)-\mathbb{E}[g(X)]\right|>\varepsilon\right)\leq 24\left(\frac{32eL}{\varepsilon}\log\frac{48eL}{\varepsilon}\right)^{VC(\mathcal{G}^+)}\exp\left(-\frac{n\varepsilon^2}{512L^2}\right).$$

The RHS of this inequality can be shown to be summable for each $\varepsilon>0$, which combined with the Borel-Cantelli lemma yields the desired result and concludes the proof.

■

8 Kolmogorov–Smirnov testing

We can now return to the Kolmogorov-Smirnov test (widely referred to as a *goodness of fit* test). The hypothesis is formulated as follows:

$$H_0 : X \sim F \qquad\qquad H_1 : X \nsim F.$$

The idea now is to use $\|F_n - F\|_\infty = \sup_x |F_n(x) - F(x)|$ as the test statistic for testing the hypothesis. We have already seen that GC and Donsker convergence results depend on bracketing numbers in L_1 and L_2, respectively. We have also established that the limiting distribution of the empirical CDF is a Brownian bridge. It can be seen that under the null hypothesis that the i.i.d. sample is generated by F,

$$\sqrt{n}\|F_n - F\| = \sqrt{n} \sup_x |F_n(x) - F(x)| \xrightarrow{n\to\infty} \sup_x |G_F(x)|,$$

where G_F is the Brownian bridge. It turns out that if F is continuous under the null, the statistic converges to the Kolmogorov distribution ($\sup_{x\in[0,1]} |G(x)|$), which in itself does not depend on F.

Then, using properties of the Kolmogorov distribution, the critical region for any level α can be determined. For example, at the 5% level, the null is rejected if

$$\sup_x |F_n(x) - F(x)| > \frac{1.358}{\sqrt{n}}.$$

9 Applications

We will now look at two specific applications of the theory developed in this chapter. In particular, we will take two fairly general class of estimators and use empirical process theory to prove uniform consistency.

9.1 M-estimation

We will first look at empirical risk minimization, or what may be familiar as M-estimation. This estimation procedure is concerned with estimators of the form

$$\hat{\theta}_n = \operatorname*{argmax}_{\theta\in\Theta} \mathbb{P}_n(m_\theta) = \operatorname*{argmax}_{\theta\in\Theta} \frac{1}{n}\sum_i m_\theta(X_i)$$

where $X_i \sim F$ and take values in some well-defined space $\mathcal{X}$. $m_\theta : \mathcal{X} \to \mathbb{R}$ is typically a real-valued loss function parametrized by $\theta \in \Theta$. $\hat{\theta}_n$ is called the M-estimator. The mapping

$$\theta \mapsto -\mathbb{P}_n(m_\theta) = -\frac{1}{n}\sum_i m_\theta(X_i)$$

can be understood as the empirical risk, and so $\hat{\theta}_n$ is the *empirical risk minimizer*. The MLE, L_2 regression estimator, Median and Mode estimators are all examples of M-estimators.

The general approach for proving this uniform consistency is done in two steps:

1. First, the rate of convergence is estimated by controlling the uniform deviation given by $\sup |\mathbb{P}(m_\theta) - \mathbb{P}_n(m_\theta)|$, then
2. Show that $m_{\hat{\theta}_n}$ is close to m_{θ_0}.

We will follow this structure here.

Lets assume that Θ is a metric space equipped with the metric function d. We will use our empirical process theory to prove consistency of the M-estimator. That is, we will show that

$$d(\hat{\theta}_n, \theta_0) \xrightarrow{p} 0 \qquad \text{where} \qquad \theta_0 = \operatorname*{argmax}_{\Theta} \mathbb{P}(m_\theta).$$

We introduce the following notation to simplify our derivation.

$$M_n(\theta) = \mathbb{P}_n(m_\theta) \qquad \text{and} \qquad M(\theta) = \mathbb{P}(m_\theta), \qquad \forall \theta \in \Theta.$$

We further assume that the class of M-estimators $\mathcal{F} = \{m_\theta : \theta \in \Theta\}$ is $\mathbb{P}$-GC and that θ_0 is a well-separated maximizer, i.e.,, for every $\delta > 0$,

$$M(\theta_0) > \sup_{\theta : d(\theta, \theta_0) \geq \delta} M(\theta).$$

Now, fix some $\delta > 0$ and let

$$\psi(\delta) = M(\theta_0) - \sup_{\theta : d(\theta, \theta_0) \geq \delta} M(\theta)$$

Then, observe that

$$\begin{aligned}
\{d(\hat{\theta}_n, \theta_0) \geq \delta\} &\Rightarrow M(\hat{\theta}_n) \leq \sup_{\theta : d(\theta, \theta_0) \geq \delta} M(\theta) \\
&\Leftrightarrow M(\hat{\theta}_n) - M(\theta_0) \leq -\psi(\delta) \\
&\Rightarrow M(\hat{\theta}_n) - M(\theta_0) + (M_n(\theta_0) - M_n(\hat{\theta}_n)) \leq -\psi(\delta) \\
&\Rightarrow 2 \sup_\theta |M_n(\theta) - M(\theta)| \geq \psi(\delta)
\end{aligned}$$

Therefore,

$$\mathbb{P}(d(\hat{\theta}_n, \theta_0) \geq \delta) \leq \mathbb{P}(\sup_\theta |M_n(\theta) - M(\theta)| \geq \psi(\delta)/2) \to 0$$

by the fact that $\mathcal{F}$ is Glivenko-Cantelli.

Of course, we still need to verify this assumption that $\mathcal{F}$ is GC. This can be done by bounding the L_1 bracketing number. An example of this will be worked through in the exercises.

9.2 Kernel density estimation

Lets now look at our working example of the kernel density estimator. Recall our setup: $\{X_1, \ldots, X_n\} \overset{iid}{\sim} F$ on $\mathbb{R}$. Suppose F corresponds to the distribution with a continuous density f and $\|f\|_\infty < \infty$. Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function. Then, recall that the KDE is defined as

$$\hat{f}_{n,h}(x) = \frac{1}{nh} \sum_i K\left(\frac{X_i - x}{h}\right) = \mathbb{P}_n \left(K_h X - x)\right)$$

Now, if we choose $h_n \to 0$ at the right rate, we can show uniform consistency of the KDE. Starting with the decomposition

$$\hat{f}_{n,h}(x) - f(x) = \hat{f}_{n,h}(x) - f_h(x) + f_h(x) - f(x)$$

where

$$f_h(x) = \mathbb{P}\left(K_h(X - x)\right) = \int K(u) f(x - hu) \mathrm{d}u$$

is the smoothed version of $\hat{f}_{n,h}$. Proving the convergence of $f_h(x) - f(x)$ only requires smoothness conditions. For example, if f is uniformly continuous, then

$$\sup_{h \le b_n} \sup_x |f_h(x) - f(x)| \to 0$$

for any sequence $b_n \to 0$. We are only left with $\hat{f}_{nh}(x) - f_h(x)$, which can be re-written as

$$(\mathbb{P}_n - \mathbb{P})(K_h(X - x)).$$

Think of the KDE as a process indexed by the evaluation point x and h, instead of a point estimator. Then, we can form the class of functions

$$\mathcal{F} = \{y \mapsto K\left((y - x)/h\right) : x \in \mathbb{R}, h > 0\}.$$

If the kernel is right-continuous with bounded variation, the covering number can be bounded

$$N(\varepsilon \|K\|_\infty, \mathcal{F}, L_2) \le (A/\varepsilon)^V,$$

for some constants $A > e^2$, $V \ge 2$ (depending on the precise form of the kernel). Then, by the Glivenko-Cantelli theorem, it follows directly that

$$\sup_{h>0,x} |(\mathbb{P}_n - \mathbb{P})(K((X - x)/h)| \overset{a.s.}{\to} 0.$$

Note that there is a h^{-1} term in the Kernel function that is not accounted for in $\mathcal{F}$. To get the right rate of convergence (especially in higher dimensions where the correction h^{-d} is larger), maximal inequalities (like of the form in Hoeffding's inequality) and strong approximation tools will need to be used.

References

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge University Press.