The Glivenko-Cantelli Theorem and Introduction to VC Dimension

Daniel Raban

Contents

1 Introduction

In this expository note, we discuss the classical Glivenko-Cantelli theorem and use it to motivate the idea of VC dimension. We prove some properties of VC dimension and relate it to other notions of size, such as cardinality and covering/packing numbers. Finally, we prove general Glivenko-Cantelli type results using the VC dimension machinery, providing effective rates for convergence.

2 The Glivenko-Cantelli Theorem

2.1 Motivation

In the most general statistical setup, we have some unknown probability distribution μ, and we try to estimate properties of it by taking independent, identically distributed samples $X_1, X_2, \ldots, X_n$ with distribution μ (written $X_i \sim \mu$). The general hope is that as we take enough samples ($n \to \infty$), we can recover properties of the distribution μ. For example, the Strong Law of Large Numbers says that, with probability 1, the sample average of our data converges to the true average of the distribution μ:

$$\mathbb{P}_\mu\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}_\mu[X_1]\right) = 1.$$

This theorem holds for any functions $f(X_i)$ of our data, as well, provided that $\mathbb{E}[|f(X_1)|] < \infty$.

Intuitively, the hope and general assumption of statistics is that it is possible to estimate any property of the distribution μ consistently. A distribution μ on $\mathbb{R}$ specifies a **cumulative distribution function (CDF)**

$$F_\mu(t) := \mathbb{P}_\mu(X \leq t).$$

The CDF uniquely specifies the distribution it arises from, as well. We have $\mu((a, b]) = F_\mu(b) - F_\mu(a)$, which determines the value of μ by the Carathéodory outer measure construction (see chapter 1 of [Fol13]). So if we can consistently estimate F_μ with the data $X_1, \ldots, X_n$ as $n \to \infty$, we should be able to estimate any property of the distribution μ. The Glivenko-Cantelli theorem says that this estimation of the entire distribution is indeed possible.

2.2 Statement and proof of the Glivenko-Cantelli theorem

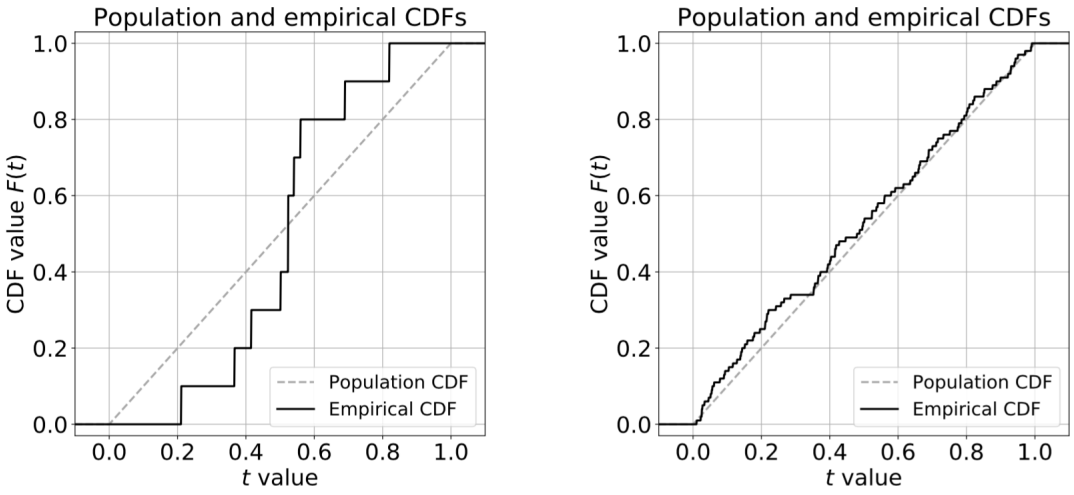
First, we mention what our estimator of F_μ should be.

Definition 2.1. The **sample** or **empirical CDF** is

$$F_{\mu_n}(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,t]}(X_i).$$

This is a nondecreasing, right continuous function which jumps up by $1/n$ whenever we hit any of the values X_i.

Example 2.1. If μ is the uniform distribution on $[0,1]$, $F_\mu(t) = t$ for $t \in [0,1]$. Here is a comparison of what the empirical CDF F_{μ_n} may look like compared to F_μ for $n = 10$ and $n = 100$ samples.[1]



The next order of business is to specify what type of functional convergence we will be referring to when we discuss consistency. Amazingly, the convergence is *uniform*, which is essentially the most powerful form of convergence one can hope for.

Theorem 2.1 (Glivenko-Cantelli)**.** *Let μ be a distribution on $\mathbb{R}$, let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} \mu$, and let $F_{\mu_n} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(-\infty,t]}(X_i)$. Then, as $n \to \infty$,*

$$\sup_{t\in\mathbb{R}} |F_{\mu_n}(t) - F_\mu(t)| \xrightarrow{a.s.} 0.$$

Here is an elementary proof of the theorem. First, we prove a deterministic lemma.

Lemma 2.1. *Let F_n and F be uniformly bounded, nondecreasing, right-continuous functions. If*

(a) $F_n(t) \to F(t)$ for each rational t,

(b) $F_n(t) \to F(t)$ for each atom of F (points where $F(t) \neq \lim_{s\uparrow t} F(s)$),

[1]Picture taken from chapter 4 of [Wai19].

then $\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow{n\to\infty} 0$.

Here is the proof of the lemma in the case where F has no atoms; the general case is the same, but constructing the ε-net in the proof is a bit more notationally messy, so we omit it.

Proof. Let $\varepsilon > 0$. There exists a finite net $T = \{t_1 < \cdots < t_\ell\}$ of rationals and atoms of F such that for any $t \in \mathbb{R}$, there is a $t' \in T$ with $t' > t$ and $F(t') - F(t) < \varepsilon$. In the case where there are no atoms of F, we can just set $t_j = \sup\{x \in \mathbb{R} : F(x) \le j/N\}$ for $1 \le j \le N$.

Now, for any $t \in \mathbb{R}$, letting $t_j \in T$ be the smallest element of T which is $\ge t$, we have

$$\begin{aligned} |F_n(t) - F(t)| &\le \begin{cases} F_n(t_j) - F(t_{j-1}) & \text{if } F_n(t) \ge F(t) \\ F(t_j) - F_n(t_{j-1}) & \text{if } F_n(t) < F(t) \end{cases} \\ &\le \begin{cases} |F_n(t_j) - F(t_j)| + |F(t_j) - F(t_{j-1})| & \text{if } F_n(t) \ge F(t) \\ |F(t_j) - F(t_{j-1})| + |F(t_{j-1}) - F_n(t_{j-1})| & \text{if } F_n(t) < F(t) \end{cases} \end{aligned}$$

Picking n large enough such that $|F_n(t') - F(t')| < \varepsilon$ for all $t' \in T$,

$$< 2\varepsilon.$$

So, for large enough n, $\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| < 2\varepsilon$. Letting $\varepsilon \downarrow 0$ completes the proof. □

Now, here is the proof of the Glivenko-Cantelli theorem.

Proof. For a fixed t, the Strong Law of Large Numbers says that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,t]}(X_i) \xrightarrow{a.s.} \mathbb{E}_\mu[\mathbb{1}_{(-\infty,t]}(X_i)] = \mathbb{P}_\mu(X_1 \le t).$$

That is, $F_{\mu_n}(t) \to F_\mu(t)$ with probability 1. So if we let $S = \mathbb{Q} \cup \{\text{atoms of } F_\mu\}$, the convergence holds with probability 1 for each of the countably many points in S, so

$$\mathbb{P}_\mu(F_{\mu_n}(t) \to F_\mu(t) \quad \forall t \in S) = 1.$$

Thus, by the lemma,

$$\mathbb{P}_\mu\left(\sup_{t\in\mathbb{R}} |F_{\mu_n}(t) - F_\mu(t)| \xrightarrow{n\to\infty} 0\right) = 1.$$ □

There are two important things to note about this proof:

- The proof does not give any information about the rate of convergence of F_{μ_n} to F_μ.
- The proof relies heavily on the linear ordering of $\mathbb{R}$, so it is hard to generalize to, say, $\mathbb{R}^n$.

The concept of **VC dimension** will fix both of these issues.

3 VC Dimension

To generalize the Glivenko-Cantelli theorem to more settings, let's rephrase the theorem in the following way:

Theorem 3.1 (Glivenko-Cantelli)**.** *Let* $\mathcal{I} = \{(-\infty, a] : a \in \mathbb{R}\}$*. Then, as* $n \to \infty$*,*

$$\sup_{S \in \mathcal{I}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_S(X_i) - \mathbb{P}_\mu(S) \right| \xrightarrow{a.s.} 0.$$

We want to be able to look at sets S other than intervals. Here is an example of a class of measurable sets where this won't work, however.

Example 3.1. Let $\mathcal{F}$ be the collection of finite subsets of $\mathbb{R}$, and let μ be some distribution with no atoms (e.g. the uniform distribution on $[0, 1]$). Then, if we let $S_n = \{X_1, \ldots, X_n\}$, we get

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{S_n}(X_i) - \mathbb{P}_\mu(S_n) \right| = |1 - 0| = 1$$

for each n. So in this situation, we have

$$\sup_{S \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_S(X_i) - \mathbb{P}_\mu(S) \right| = 1$$

for every n, and the convergence to 0 does not occur.

This example shows that some collections of sets might be too large for Glivenko-Cantelli results to hold. If we include even more sets, we get the following extreme example.

Example 3.2. Let $\mathcal{F}$ be the Borel σ-field on $\mathbb{R}$. Then the Glivenko-Cantelli result is asking for

$$\|\mu_n - \mu\|_{\text{TV}} \xrightarrow{a.s.} 0,$$

where $\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ is the empirical distribution measure.

The solution to this issue, introduced by Vapnik and Chervonenkis, is to provide a combinatorial "dimension" for the class $\mathcal{F}$ of sets. This concept will help us determine what kinds of sets enjoy Glivenko-Cantelli type results and how fast the convergence occurs.

Definition 3.1. Let $\mathcal{F}$ be a collection of subsets of some space Ω. We say that $\mathcal{F}$ **shatters** a set T if for every $U \subseteq T$, there is an $S \in \mathcal{F}$ with $S \cap T = U$. The **VC dimension** $\text{vc}(\mathcal{F})$ is the maximum cardinality of a set that shatters $\mathcal{F}$.

The VC dimension is the largest number of points where $\mathcal{F}$ can distinguish all possible subsets of the points.

Example 3.3. Let $\mathcal{I} = \{(-\infty, a] : a \in \mathbb{R}\}$, as before. If $T = \{0\}$, then $(-\infty, -1] \cap T = \varnothing$ and $(-\infty, 1] \cap T = T$, so $\mathcal{I}$ shatters T.

On the other hand, for any two point set $T = \{x, y\}$ with $x < y$, $\mathcal{I}$ cannot pick out the set $\{y\} \subseteq T$. So $\mathcal{I}$ cannot shatter any set with at least two points, and we get $\text{vc}(\mathcal{I}) = 1$.

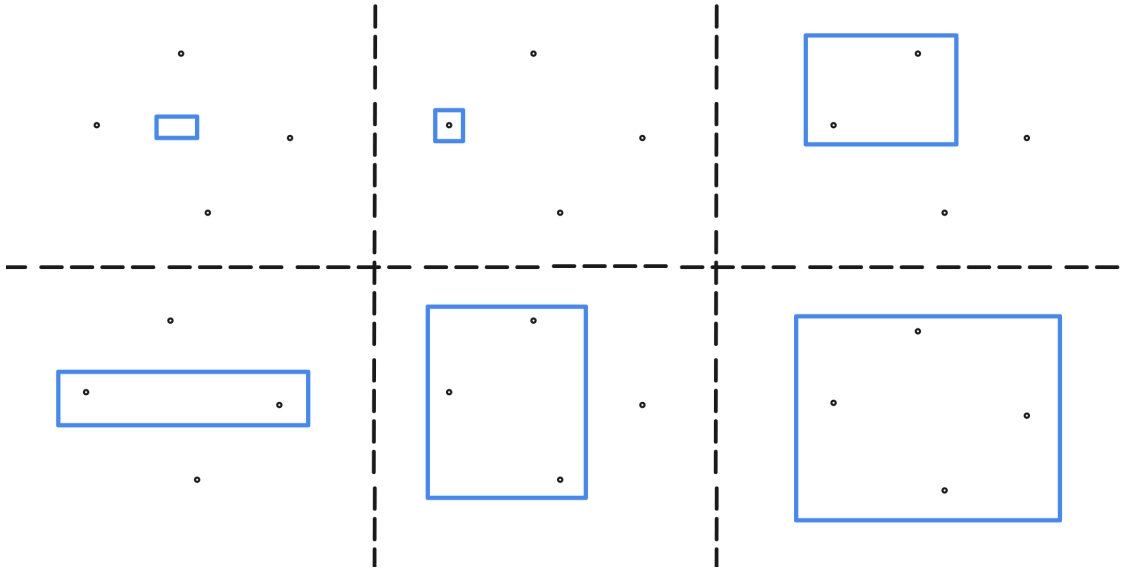
Example 3.4. Let $\mathcal{F} = \{(a, b) : a, b \in \mathbb{R}\}$ be the collection of finite length open intervals. $\mathcal{F}$ can shatter $T = \{0, 1\}$:

$$(-1, 0) \cap T = \varnothing, \quad (-1/2, 1/2) \cap T = \{0\}, \quad (1/2, 3/2) \cap T = \{1\}, \quad (-1, 2) \cap T = T.$$

However, $\mathcal{F}$ cannot shatter any set containing at least three points. If $x < y < z$, then $\mathcal{F}$ cannot pick out $\{x, z\}$. So $\text{vc}(\mathcal{F}) = 2$.

Example 3.5. Let $\mathcal{F}$ be the set of (axis parallel) rectangles in $\mathbb{R}^2$. Then $\text{vc}(\mathcal{F}) = 4$.

Here is how $\mathcal{F}$ can shatter a four point set:



The example from before which did not satisfy the Glivenko-Cantelli theorem has infinite VC-dimension:

Example 3.6. Let $\mathcal{F}$ be the set of finite subsets of $\mathbb{R}$. Then $\mathcal{F}$ shatters any finite set T because for any $J \subseteq T$, $J \in \mathcal{F}$, and $J \cap T = J$. So $\text{vc}(\mathcal{F})$ is infinite.

4 Uses of VC Dimension

In the first half of this section, we prove some relationships between VC dimensions and other notions of size. In the second half, we use these relationships to provide a general proof of the Glivenko-Cantelli theorem for a number of classes $\mathcal{F}$.

4.1 Relating VC dimension to other notions of size

We begin this section by discussing the case where $\mathcal{F}$ is finite. If $\mathcal{F}$ is a finite collection of subsets of $\Omega = \{x_1, \ldots, x_n\}$, then there is an interesting relationship between $\text{vc}(\mathcal{F})$ and $|\mathcal{F}|$. We first have $|\mathcal{F}| \geq 2^{\text{vc}(\mathcal{F})}$ because the right hand side is the size of the collection of elements of $\mathcal{F}$ intersected with A, where A is any maximal shattered set.

We can actually prove an upper bound, as well.

Lemma 4.1 (Pajor)**.** *Let $\mathcal{F} \subseteq \Omega = \{x_1, \ldots, x_n\}$, and denote* $\mathrm{SH}(\mathcal{F}) = \{A \subseteq \Omega :$ *A is shattered by $\mathcal{F}\}$ (which includes $\varnothing$). Then*

$$|\mathcal{F}| \leq |\,\mathrm{SH}(\mathcal{F})|.$$

Proof. Proceed by induction on $|\Omega|$. When $|\Omega| = 1$, we are done because the right side includes the empty set. Now assume the lemma holds for $|\Omega| = n$. For $|\Omega| = n+1$, write $\Omega = \Omega_0 \cup \{x_0\}$, where $|\Omega_0| = n$. We can split $\mathcal{F}$ into

$$\mathcal{F}_+ = \{S \in \mathcal{F} : x_0 \in S\}, \qquad \mathcal{F}_- = \{S \in \mathcal{F} : x_0 \notin S\}.$$

By the inductive hypothesis,

$$|\mathcal{F}| = |\mathcal{F}_+| + |\mathcal{F}_-| \leq |\,\mathrm{SH}(\mathcal{F}_+)| + |\,\mathrm{SH}(\mathcal{F}_-)|.$$

It now suffices to show that $|\,\mathrm{SH}(\mathcal{F})| \geq |\,\mathrm{SH}(\mathcal{F}_+)| + |\,\mathrm{SH}(\mathcal{F}_-)|$. First, if A is shattered by one of $\mathcal{F}_+, \mathcal{F}_-$, then it is shattered by $\mathcal{F}$. And if A is shattered by both $\mathcal{F}_+, \mathcal{F}_-$, then $A \cup \{x_0\}$ is shattered by $\mathcal{F}$ but not by either of $\mathcal{F}_+, \mathcal{F}_-$. This proves the desired inequality. □

One way of stating the condition that $\mathcal{F}$ shatters $\{x_1, \ldots, x_n\}$ is that

$$|\{S \cap \{x_1, \ldots, x_n\} : S \in \mathcal{F}\}| = 2^n.$$

Using Pajor's theorem, we can get a bound on the number of pieces we get if we try to shatter a large set using $\mathcal{F}$.

Lemma 4.2 (Sauer-Shelah)**.** *Let $x_1, \ldots, x_n \in \Omega$, and let $\mathcal{F}$ be a class of subsets of Ω. Then*

$$|\{S \cap \{x_1, \ldots, x_n\} : S \in \mathcal{F}\}| \leq \sum_{k=0}^{\mathrm{vc}(\mathcal{F})} \binom{n}{k} \leq \left(\frac{en}{\mathrm{vc}(\mathcal{F})}\right)^{\mathrm{vc}(\mathcal{F})}.$$

Proof. Let the collection in the left hand side be $\mathcal{G}$. By Pajor's lemma, we have

$$|\mathcal{G}| \leq |\{A \subseteq \{x_1, \ldots, x_n\} : A \text{ is shattered by } \mathcal{G}\}|.$$

If A is shattered by $\mathcal{G}$, then it is shattered by $\mathcal{F}$, so the cardinality of any such A is bounded: $|A| \leq \mathrm{vc}(\mathcal{F})$. So we get

$$|\mathcal{G}| \leq |\{A \subseteq \{x_1, \ldots, x_n\} : |A| \leq \mathrm{vc}(\mathcal{F})| = \sum_{k=0}^{\mathrm{vc}(\mathcal{F})} \binom{n}{k},$$

proving the first inequality.

The second inequality follows from the following computation involving the binomial theorem: If $d \leq n$, then

$$\left(\frac{d}{n}\right)^d \sum_{k=0}^{d} \binom{n}{k} \leq \sum_{k=0}^{d} \binom{n}{k} \left(\frac{d}{n}\right)^k = \left(1 + \frac{d}{n}\right)^n \leq e^d. \qquad \square$$

The following theorem is our destination for this section. It relates the VC dimension to another notion of size, the covering number.

Theorem 4.1 (Dudley)**.** *Let μ be a distribution on Ω, and let $\mathcal{F}$ be a collection of subsets of Ω. There is a universal constant K such that*

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

for all $\varepsilon < 1$.

Here, the metric on $\mathcal{F}$ is $\rho(A, B) := \|\mathbb{1}_A - \mathbb{1}_B\|_{L^2(\mu)}$.

Remark 4.1. This bound is independent of the distribution μ, so we could take the supremum of the left hand side over all probability distributions μ.

Also compare this bound on the covering number to the covering number of the unit ball in $\mathbb{R}^d$: $(1/\varepsilon)^d$. This gives a bit more justification of why we think of $\operatorname{vc}(\mathcal{F})$ as a measure of dimension.

The idea is as follows. Covering and packing numbers are equivalent, so we can just focus on packing numbers. By the Strong Law of Large Numbers, $\|\mathbb{1}_A - \mathbb{1}_B\|_{L^2(\mu)}$ can be approximated by $\|\mathbb{1}_A - \mathbb{1}_B\|_{L^2(\mu_r)}$ for sufficiently large r. Since μ_r is discrete, we have a packing in $L^2(\mu_r)$, then all of the sets belonging to the packing must have distinct intersections with $\{X_1, \ldots, X_r\}$ (as long as $\varepsilon < 1/r$). Thus, we can upper bound the packing number by counting the number of these intersections, which will be accomplished via Sauer-Shelah.

The following lemma makes precise what value of r we can use.

Lemma 4.3 (Probabilistic extraction)**.** *Let $S_1, \ldots, S_m$ be subsets of Ω such that $\|\mathbb{1}_{S_i} - \mathbb{1}_{S_j}\|_{L^2(\mu)} > \varepsilon$ for all $i \neq j$. Then there exist $r \leq c\varepsilon^{-4} \log m$ points $x_1, \ldots, x_r \in \Omega$ such that*

$$\|\mathbb{1}_{S_i} - \mathbb{1}_{S_j}\|_{L^2(\mu^x)} > \varepsilon/2$$

for all $i \neq j$. Here, $\mu^x := \frac{1}{r}\sum_{k=1}^{r} \delta_{x_k}$ is the empirical distribution for these points, and c is a universal constant.

Proof. Let $X_1, \dots, X_r \overset{\text{iid}}{\sim} \mu$, and let μ_r be the empirical measure. Then

$$\mathbb{P}\left(\|\mathbb{1}_{S_i} - \mathbb{1}_{S_j}\|^2_{L^2(\mu_r)} \le \frac{\varepsilon^2}{4} \right) \le \mathbb{P}\left(\|\mathbb{1}_{S_i} - \mathbb{1}_{S_j}\|^2_{L^2(\mu_r)} \le \|\mathbb{1}_{S_i} - \mathbb{1}_{S_j}\|^2_{L^2(\mu)} - \frac{3\varepsilon^2}{4} \right)$$

$\|\mathbb{1}_{S_i} - \mathbb{1}_{S_j}\|^2_{L^2(\mu)}$ is the expectation of $\|\mathbb{1}_{S_i} - \mathbb{1}_{S_j}\|^2_{L^2(\mu_r)}$, so using the Azuma-Hoeffding inequality,

$$\le e^{-r\varepsilon^4/15}.$$

Using a union bound over all $i \neq j$, we get

$$\mathbb{P}\Big(\|\mathbb{1}_{S_i} - \mathbb{1}_{S_j}\|_{L^2(\mu_r)} > \frac{\varepsilon}{2} \quad \forall i \neq j \Big) \ge 1 - m^2 e^{-r\varepsilon^4/15}.$$

For $r > 30\varepsilon^{-4} \log m$, this is > 0, so there exist some points which work. □

Now let's prove Dudley's theorem.

Proof. Let $S_1, \dots, S_m$ be a maximal ε-packing of $(\mathcal{F}, \|\cdot\|_{L^2(\mu)})$. By the lemma, we can pick $r \le c\varepsilon^{-4} \log m$ points $x_1, \dots, x_r$ such that $S_1, \dots, S_m$ is an $\varepsilon/2$-packing of $(\mathcal{F}, \|\cdot\|_{L^2(\mu^x)})$. We can bound

$$m \le |\{S \cap \{x_1, \dots, x_r\} : S \in \mathcal{F}\}|$$

Using the Sauer-Shelah lemma,

$$\begin{aligned}
&\le \left(\frac{er}{\operatorname{vc}(\mathcal{F})} \right)^{\operatorname{vc}(\mathcal{F})} \\
&\le \left(\frac{ec \log m}{\operatorname{vc}(\mathcal{F})\varepsilon^4} \right)^{\operatorname{vc}(\mathcal{F})} \\
&= 2^{\operatorname{vc}(\mathcal{F})} \left(\frac{\log m}{2 \operatorname{vc}(\mathcal{F})} \right)^{\operatorname{vc}(\mathcal{F})} \left(\frac{(ec)^{1/4}}{\varepsilon} \right)^{4 \operatorname{vc}(\mathcal{F})}
\end{aligned}$$

Using the bound $\alpha \log m \le m^\alpha$ with $\alpha = 1/(2 \operatorname{vc}(\mathcal{F}))$,

$$\le m^{1/2} \left(\frac{(2ec)^{1/4}}{\varepsilon} \right)^{4 \operatorname{vc}(\mathcal{F})}.$$

So we get

$$m \le \left(\frac{(2ec)^{1/4}}{\varepsilon} \right)^{8 \operatorname{vc}(\mathcal{F})},$$

which provides a bound on the covering number. □

4.2 Uniform Glivenko-Cantelli rates

We can now give effective rates for the Glivenko-Cantelli theorem for any classes of sets with finite VC dimension. Here is a symmetrization lemma we will not prove.

Lemma 4.4 (Symmetrization and chaining)**.**

$$\mathbb{E}\left[\sup_{S\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_S(X_i)-\mathbb{P}_\mu(S)\right|\right]\lesssim\frac{1}{\sqrt{n}}\,\mathbb{E}\left[\int_0^1\sqrt{\log N(\mathcal{F},\|\cdot\|_{L^2(\mu_n)},\varepsilon)}\,d\varepsilon\right],$$

where $\mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ *is the empirical distribution of the data* $X_1,\dots,X_n$*.*

Proof. See chapter 7 of [vH14]. □

Theorem 4.2 (Uniform Glivenko-Cantelli rates)**.** *There is a universal constant* L *such that for any distribution* μ *on* Ω *and a collection* $\mathcal{F}$ *of subsets of* Ω*,*

$$\mathbb{E}\left[\sup_{S\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_S(X_i)-\mathbb{P}_\mu(S)\right|\right]\le L\sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}}.$$

Remark 4.2. This provides L^1-type Glivenko-Cantelli results that hold regardless of the distribution μ chosen, as long as the VC dimension of $\mathcal{F}$ is finite. It also provides an explicit rate of convergence of the error, $1/\sqrt{n}$, which only scales with the VC dimension as a constant factor.

If $\mathcal{F}$ is just a single set S, then the bound gives us the usual rate for the Central Limit Theorem. So for the convergence to hold uniformly over $\mathcal{F}$, we only pay the price of a constant factor: the VC dimension.

Proof. Using symmetrization and then the previous theorem,

$$\begin{aligned}\mathbb{E}\left[\sup_{S\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_S(X_i)-\mathbb{P}_\mu(S)\right|\right]&\lesssim\frac{1}{\sqrt{n}}\,\mathbb{E}\left[\int_0^1\sqrt{\log N(\mathcal{F},\|\cdot\|_{L^2(\mu_n)},\varepsilon)}\,d\varepsilon\right]\\&\le\sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}}\cdot\sqrt{K}\int_0^1\sqrt{\log(K/\varepsilon)}\,d\varepsilon.\end{aligned}$$

□

Although this is an L^1 convergence result, we can use it to obtain an almost sure convergence result, more in the vein of the original Glivenko-Cantelli theorem.

Corollary 4.1 (Glivenko-Cantelli for finite VC classes)**.** *Let* μ *be a distribution on* Ω*, let* $X_1, X_2, \dots \overset{\text{iid}}{\sim} \mu$*, and let* $\mathcal{F}$ *be a collection of subsets of* Ω *with* $\mathrm{vc}(\mathcal{F}) < \infty$*. Then, as* $n\to\infty$*,*

$$\sup_{S\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_S(X_i)-\mathbb{P}_\mu(S)\right|\xrightarrow{a.s.}0.$$

Proof. For large n,

$$\begin{aligned}
&\mathbb{P}_\mu\left(\sup_{S\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_S(X_i)-\mathbb{P}_\mu(S)\right|>\varepsilon\right)\\
&\quad\le\mathbb{P}_\mu\left(\sup_{S\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_S(X_i)-\mathbb{P}_\mu(S)\right|-\mathbb{E}\left[\sup_{S\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_S(X_i)-\mathbb{P}_\mu(S)\right|\right]>\frac{\varepsilon}{2}\right)
\end{aligned}$$

By the bounded differences inequality,

$$\le\exp(-c\varepsilon^2 n),$$

where c is a universal constant. By the Borel-Cantelli lemma,

$$\mathbb{P}_\mu\left(\sup_{S\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_S(X_i)-\mathbb{P}_\mu(S)\right|>\varepsilon\quad\text{for infinitely many } n\right)=0.$$

This holds for all $\varepsilon>0$, so

$$\sup_{S\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_S(X_i)-\mathbb{P}_\mu(S)\right|\xrightarrow{a.s.}0.$$ □

Remark 4.3. It can be shown that if $\mathrm{vc}(\mathcal{F})=\infty$, then the Glivenko-Cantelli theorem fails.

References

[Fol13] G.B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[vH14] Ramon van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.