2 Size/complexity of a function class

Let $\mathcal{F}$ be a class of measurable real-valued functions defined on $\mathcal{X}$. Whether a given class of function $\mathcal{F}$ is "Glivenko-Cantelli" or "Donsker" depends on the *size* (or *complexity*) of the class. A finite class of square integrable functions is always Donsker, while at the other extreme the class of all square integrable, uniformly bounded functions is almost never Donsker.

2.1 Covering numbers

A relatively simple way to measure the size of any set is to use *covering numbers*. Let (Θ, d) be an arbitrary semi-metric space[23]; we will assume that $\Theta \subset \Xi$ and that $d(\cdot, \cdot)$ is defined on the space Ξ. Let $\varepsilon > 0$.

Definition 2.1 (ε-cover)**.** *A ε-cover of the set Θ with respect to the semi-metric d is a set $\{\theta_1, \ldots, \theta_N\} \subset \Xi$[24] such that for any $\theta \in \Theta$, there exists some $v \in \{1, \ldots, N\}$ with $d(\theta, \theta_v) \leq \varepsilon$.*

Definition 2.2 (Covering number)**.** *The ε-covering number of Θ is*

$$N(\varepsilon, \Theta, d) := \inf\{N \in \mathbb{N} : \exists \text{ a } \varepsilon\text{-cover } \theta_1, \ldots, \theta_N \text{ of } \Theta\}.$$

Equivalently, the ε-covering number $N(\varepsilon, \Theta, d)$ is the minimal number of balls $B(x; \varepsilon) := \{y \in \Theta : d(x, y) \leq \varepsilon\}$ of radius ε needed to cover the set Θ.

Definition 2.3 (Metric entropy)**.** *The metric entropy of the set Θ with respect to the semi-metric d is the logarithm of its covering number:* $\log N(\varepsilon, \Theta, d)$.

Note that a semi-metric space (Θ, d) is said to be *totally bounded* if the ε-covering number is finite for every $\varepsilon > 0$. We can define a related measure of size that relates to the number of disjoint balls of radius $\varepsilon > 0$ that can be placed into the set Θ.

Definition 2.4 (ε-packing)**.** *A ε-packing of the set Θ with respect to the semi-metric d is a set $\{\theta_1, \ldots, \theta_D\} \subseteq \Theta$ such that for all distinct $v, v' \in \{1, \ldots, D\}$, we have $d(\theta_v, \theta_{v'}) > \varepsilon$.*

Definition 2.5 (Packing number)**.** *The ε-packing number of Θ is*

$$D(\varepsilon, \Theta, d) := \sup\{D \in \mathbb{N} : \exists \text{ a } \varepsilon\text{-packing } \theta_1, \ldots, \theta_D \text{ of } \Theta\}.$$

Equivalently, call a collection of points ε-separated if the distance between each pair of points is larger than ε. Thus, the packing number $D(\varepsilon, \Theta, d)$ is the maximum number of ε-separated points in Θ.

[23] By a semi-metric space (Θ, d) we mean, for any $\theta_1, \theta_2, \theta_3 \in \Theta$, we have: (i) $d(\theta_1, \theta_2) = 0 \Rightarrow \theta_1 = \theta_2$; (ii) $d(\theta_1, \theta_2) = d(\theta_2, \theta_1)$; and (iii) $d(\theta_1, \theta_3) \leq d(\theta_1, \theta_2) + d(\theta_2, \theta_3)$.

[24] The elements $\{\theta_1, \ldots, \theta_N\} \subset \Xi$ need not belong to Θ themselves.

A minimal ϵ-cover and or maximal ϵ-packing do not have to be finite. In the proofs of the following results, we do not separate out the case when they are infinite (in which case there is nothing show).

Lemma 2.6. *Show that*

$$D(2\varepsilon, \Theta, d) \leq N(\varepsilon, \Theta, d) \leq D(\varepsilon, \Theta, d), \qquad \text{for every } \varepsilon > 0.$$

Thus, packing and covering numbers have the same scaling in the radius ε.

Proof. Let us first show the second inequality. Suppose $E = \{\theta_1, \ldots, \theta_D\} \subseteq \Theta$ is a maximal packing. Then for every $\theta \in \Theta \setminus E$, there exists $1 \leq i \leq D$ such that $d(\theta, \theta_i) \leq \varepsilon$ (for if this does not hold for θ then we can construct a bigger packing set with $\theta_{D+1} = \theta$). Hence E is automatically an ε-covering. Since $N(\varepsilon, \Theta, d)$ is the minimal size of all possible coverings, we have $D(\varepsilon, \Theta, d) \geq N(\varepsilon, \Theta, d)$.

We next prove the first inequality by contradiction. Suppose that there exists a 2ε-packing $\{\theta_1, \ldots, \theta_D\}$ and an ε-covering $\{x_1, \ldots, x_N\}$ such that $D \geq N + 1$. Then by pigeonhole, we must have θ_i and θ_j belonging to the same ε-ball $B(x_k, \varepsilon)$ for some $i \neq j$ and k. This means that the distance between θ_i and θ_j cannot be more than the diameter of the ball, i.e., $d(\theta_i, \theta_j) \leq 2\varepsilon$, which leads to a contradiction since $d(\theta_i, \theta_j) > 2\epsilon$ for a 2ε-packing. Hence the size of any 2ε-packing is less or equal to the size of any ε-covering. □

Remark 2.1. *As shown in the preceding lemma, covering and packing numbers are closely related, and we can use both in the following. Clearly, they become bigger as $\varepsilon \to 0$.*

Let $\|\cdot\|$ denote any norm on $\mathbb{R}^d$. The following result gives the (order of) covering number for any bounded set in $\mathbb{R}^d$.

Lemma 2.7. *For a bounded subset $\Theta \subset \mathbb{R}^d$ there exist constants $c < C$ depending on Θ (and $\|\cdot\|$) only such that, for $\epsilon \in (0, 1)$,*

$$c\left(\frac{1}{\epsilon}\right)^d \leq N(\epsilon, \Theta, \|\cdot\|) \leq C\left(\frac{1}{\epsilon}\right)^d.$$

Proof. If $\theta_1, \ldots, \theta_D$ are ϵ-separated points in Θ, then the balls of radius $\epsilon/2$ around the θ_i's are disjoint, and their union is contained in $\Theta' := \{\theta \in \mathbb{R}^d : \|\theta - \Theta\| \leq \epsilon/2\}$. Thus, the sum $Dv_d(\epsilon/2)^d$ of the volumes of these balls, where v_d is the volume of the unit ball, is bounded by $\mathrm{Vol}(\Theta')$, the volume of Θ'. This gives the upper bound of the lemma, as

$$N(\epsilon, \Theta, \|\cdot\|) \leq D(\epsilon, \Theta, \|\cdot\|) \leq \frac{2^d \,\mathrm{Vol}(\Theta')}{v_d}\left(\frac{1}{\epsilon}\right)^d.$$

Let $\theta_1, \ldots, \theta_N$ be an ϵ-cover of Θ, i.e., the union of the balls of radius ϵ around them covers Θ. Thus the volume of Θ is bounded above by the sum of the volumes of the N

balls, i.e., by $Nv_d\epsilon^d$. This yields the lower bound of the lemma, as

$$N(\epsilon, \Theta, \|\cdot\|) \geq \frac{\text{Vol}(\Theta)}{v_d}\left(\frac{1}{\epsilon}\right)^d.$$

□

The following result gives an upper bound (which also happens to be optimal) on the entropy numbers of the class of Lipschitz functions[25].

Lemma 2.8. *Let $\mathcal{F} := \{f : [0,1] \to [0,1] \mid f \text{ is 1-Lipschitz}\}$. Then for some constant A, we have*

$$\log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq A\frac{1}{\epsilon}, \qquad \text{for all } \epsilon > 0.$$

Proof. If $\epsilon > 1$, there is nothing to prove as then $N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) = 1$ (take the function $f_0 \equiv 0$ and observe that for any $f \in \mathcal{F}$, $\|f - f_0\|_\infty \leq 1 < \epsilon$).

Let $0 < \epsilon < 1$. We will explicitly exhibit an ϵ-cover of $\mathcal{F}$ (under $\|\cdot\|_\infty$-metric) with cardinality less than $\exp(A/\epsilon)$, for some $A > 0$. This will complete the proof as $N(\epsilon, \mathcal{F}, \|\cdot\|_\infty)$ will then be automatically less than $\exp(A/\epsilon)$.

Let us define a ϵ-grid of the interval [0,1], i.e., $0 = a_0 < a_1 < \ldots < a_N = 1$ where $a_k := k\epsilon$, for $k = 1, \ldots, N-1$; here $N \leq \lfloor 1/\epsilon \rfloor + 1$ (where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x). Let $B_1 := [a_0, a_1]$ and $B_k := (a_{k-1}, a_k]$, $k = 2, \ldots, N$. For each $f \in \mathcal{F}$ define $\tilde{f} : [0,1] \to \mathbb{R}$ as

$$\tilde{f}(x) = \sum_{i=1}^{N} \epsilon \left\lfloor \frac{f(a_k)}{\epsilon} \right\rfloor \mathbf{1}_{B_k}(x). \tag{7}$$

Thus, $\tilde{f}$ is constant on the interval B_k and can only take values of the form $i\epsilon$, for $i = 0, \ldots, \lfloor 1/\epsilon \rfloor$. Observe that for $x \in B_k$ (for some $k \in \{1, \ldots, N\}$) we have

$$|f(x) - \tilde{f}(x)| \leq |f(x) - f(a_k)| + |f(a_k) - \tilde{f}(a_k)| \leq 2\epsilon,$$

where the first ϵ comes from the fact that f is 1-Lipschitz, and the second appears because of the approximation error in (7)[26]. Thus, $\|f - \tilde{f}\|_\infty \leq 2\epsilon$.

New, let us count the number of distinct $\tilde{f}$'s obtained as f varies over $\mathcal{F}$. There are at most $\lfloor 1/\epsilon \rfloor + 1$ choices for $\tilde{f}(a_1)$. Further, note that for any $\tilde{f}$ (and any $k = 2, \ldots, N$),

$$|\tilde{f}(a_k) - \tilde{f}(a_{k-1})| \leq |\tilde{f}(a_k) - f(a_k)| + |f(a_k) - f(a_{k-1})| + |f(a_{k-1}) - \tilde{f}(a_{k-1})| \leq 3\epsilon.$$

Therefore once a choice is made for $\tilde{f}(a_{k-1})$ there are at most 7 choices left for the next value of $\tilde{f}(a_k)$, $k = 2, \ldots, N$.

[25] Note that $f : \mathcal{X} \to \mathbb{R}$ is L-Lipschitz if $|f(x) - f(y)| \leq L\|x - y\|$ for all $x, y \in \mathcal{X}$.

[26] Note that, for $x \in B_k$, $\tilde{f}(x) = \tilde{f}(a_k) = \epsilon \left\lfloor \frac{f(a_k)}{\epsilon} \right\rfloor \leq f(a_k)$, and $f(a_k) - \tilde{f}(a_k) = \epsilon \left(\frac{f(a_k)}{\epsilon} - \left\lfloor \frac{f(a_k)}{\epsilon} \right\rfloor \right) \leq \epsilon$.

Now consider the collection $\{\tilde{f} : f \in \mathcal{F}\}$. We see that this collection is a 2ϵ-cover of $\mathcal{F}$ and the number of distinct functions in this collection is upper bounded by

$$\left(\left\lfloor \frac{1}{\epsilon} \right\rfloor + 1\right) 7^{\lfloor 1/\epsilon \rfloor}.$$

Thus, $N(2\epsilon, \mathcal{F}, \|\cdot\|_\infty)$ is bounded by the right-side of the above display, which completes the proof the result. □

Thus, the set of Lipschitz functions is much "larger" than a bounded set in $\mathbb{R}^d$, since its metric entropy grows as $1/\epsilon$ as $\epsilon \to 0$, as compared to $\log(1/\epsilon)$ (cf. Lemma 2.7).

Exercise (HW1): For $L > 0$, let $\mathcal{F}_L := \{f : [0,1] \to \mathbb{R} \mid f \text{ is } L\text{-Lipschitz}\}$. Show that, for $\epsilon > 0$, $N(\epsilon, \mathcal{F}_L, \|\cdot\|_\infty) \geq a\frac{L}{\epsilon}$, for some constant $a > 0$. Then, using Lemma 2.8 show that $N(\epsilon, \mathcal{F}_L, \|\cdot\|_\infty) \asymp \frac{L}{\epsilon}$, for $\epsilon > 0$ sufficiently small.

2.2 Bracketing numbers

Let $(\mathcal{F}, \|\cdot\|)$ be a subset of a normed space of real functions $f : \mathcal{X} \to \mathbb{R}$ on some set $\mathcal{X}$. We are mostly thinking of $L_r(Q)$-spaces for probability measures Q. We shall write $N(\varepsilon, \mathcal{F}, L_r(Q))$ for covering numbers relative to the $L_r(Q)$-norm $\|f\|_{Q,r} = \left(\int |f|^r dQ\right)^{1/r}$.

Definition 2.9 (ε-bracket)**.** *Given two functions $l(\cdot)$ and $u(\cdot)$, the bracket $[l, u]$ is the set of all functions $f \in \mathcal{F}$ with $l(x) \leq f(x) \leq u(x)$, for all $x \in \mathcal{X}$. An ε-bracket is a bracket $[l, u]$ with $\|l - u\| < \varepsilon$.*

Definition 2.10 (Bracketing numbers)**.** *The bracketing number $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ε-brackets needed to cover $\mathcal{F}$.*

Definition 2.11 (Entropy with bracketing)**.** *The entropy with bracketing is the logarithm of the bracketing number.*

In the definition of the bracketing number, the upper and lower bounds u and l of the brackets need not belong to $\mathcal{F}$ themselves but are assumed to have finite norms.

Example 2.12. *(Distribution function). When $\mathcal{F}$ is equal to the collection of all indicator functions of the form $f_t(\cdot) = \mathbf{1}_{(-\infty,t]}(\cdot)$, with t ranging over $\mathbb{R}$, then the empirical process $\mathbb{G}_n(f_t)$ is the classical empirical process $\sqrt{n}(\mathbb{F}_n(t) - F(t))$ (here $X_1, \ldots, X_n$ are i.i.d. P with cd.f. F).*

Consider brackets of the form $[\mathbf{1}_{(-\infty,t_{i-1}]}, \mathbf{1}_{(-\infty,t_i)}]$ for a grid points $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ with the property $F(t_i-) - F(t_{i-1}) < \varepsilon$ for each $i = 1, \ldots, k$; here we assume that $\varepsilon < 1$. These brackets have $L_1(P)$-size ε. Their total number k can be chosen smaller than $2/\varepsilon$. Since $Pf^2 \leq Pf$ for every $0 \leq f \leq 1$, the $L_2(P)$-size of the brackets is bounded by $\sqrt{\varepsilon}$. Thus $N_{[]}(\sqrt{\varepsilon}, \mathcal{F}, L_2(P)) \leq 2/\varepsilon$, whence the bracketing numbers are of the polynomial order $1/\varepsilon^2$.

Exercise (HW1): Show that $N(\varepsilon, \mathcal{F}, \|\cdot\|) \le N_{[]}(2\varepsilon, \mathcal{F}, \|\cdot\|)$, for every $\varepsilon > 0$.

In general, there is no converse inequality. Thus, apart from the constant 1/2, bracketing numbers are bigger than covering numbers. The advantage of a bracket is that it gives pointwise control over a function: $l(x) \le f(x) \le u(x)$, for every $x \in \mathcal{X}$. In comparison an $L_r(P)$-ball gives integrated, but not pointwise control.

Definition 2.13 (Envelope function)**.** *An envelope function of a class $\mathcal{F}$ is any function $x \mapsto F(x)$ such that $|f(x)| \le F(x)$, for every $x \in \mathcal{X}$ and $f \in \mathcal{F}$. The minimal envelope function is $x \mapsto \sup_{f\in\mathcal{F}} |f(x)|$.*

Consider a class of functions $\{m_\theta : \theta \in \Theta\}$ indexed by a parameter θ in an arbitrary index set Θ with a metric d. Suppose that the dependence on θ is Lipschitz in the sense that

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \le d(\theta_1, \theta_2)F(x),$$

for some function $F : \mathcal{X} \to \mathbb{R}$, for every $\theta_1, \theta_2 \in \Theta$, and every $x \in \mathcal{X}$. The bracketing numbers of this class are bounded by the covering numbers of Θ as shown below.

Lemma 2.14. *Let $\mathcal{F} = \{m_\theta : \theta \in \Theta\}$ be a class of functions satisfying the preceding display for every θ_1 and θ_2 and some fixed function F. Then, for any norm $\|\cdot\|$,*

$$N_{[]}(2\epsilon\|F\|, \mathcal{F}, \|\cdot\|) \le N(\epsilon, \Theta, d).$$

Proof. Let $\theta_1, \ldots, \theta_p$ be an ϵ-cover of Θ (under the metric d). Then the brackets $[m_{\theta_i} - \epsilon F, m_{\theta_i} + \epsilon F]$, $i = 1, \ldots, p$, cover $\mathcal{F}$. The brackets are of size $2\epsilon\|F\|$. □

Exercise (HW1): Let $\mathcal{F}$ and $\mathcal{G}$ be classes of measurable function. Then for any probability measure Q and any $1 \le r \le \infty$,

(i) $N_{[]}(2\epsilon, \mathcal{F} + \mathcal{G}, L_r(Q)) \le N_{[]}(\epsilon, \mathcal{F}, L_r(Q))\, N_{[]}(\epsilon, \mathcal{G}, L_r(Q))$;

(ii) provided $\mathcal{F}$ and $\mathcal{G}$ are bounded by 1,

$$N_{[]}(2\epsilon, \mathcal{F} \cdot \mathcal{G}, L_r(Q)) \le N_{[]}(\epsilon, \mathcal{F}, L_r(Q))\, N_{[]}(\epsilon, \mathcal{G}, L_r(Q)).$$