Appendix C

Packing, Covering, Bracketing and Entropy Numbers

Covering and bracketing numbers provide ways to measure the complexity of a metric space, or a set of functions. In this appendix we give their definitions, and a number of examples.

C.1 Definitions

A subset S of a semimetric space (T, d) is said to be *ϵ-dispersed* if $d(s, s') \geq \epsilon$ for all $s, s' \in S$ with $s \neq s'$. The maximum cardinality of an ϵ-dispersed subset of T is known as the *ϵ-packing number* of T and is denoted by $D(\epsilon, T, d)$.

A set S is called an *ϵ-net* for T if for every $t \in T$, there exists $s \in S$ such that $d(s, t) < \epsilon$, or equivalently T is covered by the collection of balls of radius ϵ around the points in S. The minimal cardinality of an ϵ-net is known as the *ϵ-covering number* of T and is denoted by $N(\epsilon, T, d)$. If T is itself a subset of a larger semimetric space, then one may allow the points of the ϵ-net not to belong to T. In general this results in a smaller covering number, but as every ball of radius ϵ around a point not in T is contained in a ball of radius 2ϵ around a point of T (unless it does not intersect T) covering numbers at 2ϵ that are restricted to centers in T will not be bigger.

Because a maximal ϵ-dispersed set is necessarily an ϵ-net, and $\epsilon/2$-balls centered at points in an ϵ-dispersed set are disjoint, we have the inequalities, for every $\epsilon > 0$,

$$N(\epsilon, T, d) \leq D(\epsilon, T, d) \leq N(\epsilon/2, T, d). \tag{C.1}$$

This means that packing and covering numbers can be used interchangeably if their order of magnitude and not exact constants are important.

For two given functions $u, l: \mathfrak{X} \to \mathbb{R}$ on a set $\mathfrak{X}$ with $l \leq u$, the *bracket* $[l, u]$ is defined as the set of all functions $f: \mathfrak{X} \to \mathbb{R}$ such that $l \leq f \leq u$ everywhere. For a semimetric d that is compatible with pointwise partial ordering in the sense that $d(l, u) = \sup\{d(f, g): f, g \in [l, u]\}$, the bracket is said to be of size ϵ if $d(l, u) < \epsilon$. For a given set T of functions the minimal number of ϵ-brackets needed to cover T is called the *ϵ-bracketing number* of T, and is denoted by $N_{[]}(\epsilon, T, d)$. In this definition the boundary functions l, u of the brackets are restricted to a given function space that contains T, and the bracketing numbers may depend on this space.

Because a bracket of size ϵ is contained in the ball of radius ϵ around its lower bracket, it follows that, for every $\epsilon > 0$,[1]

[1] Typically this can be slightly improved, with 2ϵ rather than ϵ in the right side, by considering balls around the mid function $(l + u)/2$, but this requires a stronger condition on the metric.

$$N(\epsilon, T, d) \leq N_{[]}(\epsilon, T, d). \tag{C.2}$$

In general, there is no inequality in the reverse direction, except when d is the uniform distance, in which case the left side coincides with the expression on the right for ϵ replaced by 2ϵ.

The logarithms of the packing (or covering) and bracketing numbers are called the (metric) *entropy* and the *bracketing entropy*, respectively.

All these numbers grow to infinity as ϵ decays to zero, unless T is a finite set, in which case the both packing and covering numbers are bounded by the cardinality of the sets. The rate of growth qualitatively measures the size of the space, and is fundamental in empirical process theory and Bayesian nonparametrics.

In some applications the lower bounds of the brackets are not needed. It is therefore useful to define *upper bracketing numbers* $N_{]}(\epsilon, T, d)$ as the minimal number of functions $u_1, \ldots, u_m$ such that for every $p \in T$, there exist a function u_i such that both $p \leq u_i$ and $d(u_i, p) < \epsilon$. The upper bracketing numbers are clearly smaller than the corresponding bracketing numbers.

When more than one semimetric is relevant, the corresponding packing and covering numbers may be related. If d_1 and d_2 are semimetrics such that $d_1(t, t') \leq H(d_2(t, t'))$ for all $t, t' \in T$, for some strictly increasing continuous function $H: [0, \infty) \to [0, \infty)$ with $H(0) = 0$, then $N(\epsilon, T, d_1) \leq N(H^{-1}(\epsilon), T, d_2)$. Similar conclusions hold for packing and bracketing numbers. In particular, if $\mathcal{F}$ is a set of probability densities, then $N(\epsilon, \mathcal{F}, d_H) \leq N(\epsilon^2, \mathcal{F}, \|\cdot\|_1)$.

C.2 Examples

In this section we present examples of bounds for covering and bracketing numbers, first for subsets of Euclidean spaces and next for the most important spaces.

Proposition C.1 (Unit simplex) *For the norm $\|x\|_1 = \sum_i |x_i|$ on the m-dimensional unit simplex $\mathbb{S}_m$, for $0 < \epsilon \leq 1$,*

$$D(\epsilon, \mathbb{S}_m, \|\cdot\|_1) \leq \left(\frac{5}{\epsilon}\right)^{m-1}.$$

Proof For $x \in \mathbb{S}_m$, let x^* denote the vector of its first $m-1$ coordinates. Then x^* belongs to the set $\mathbb{D}_{m-1} := \{(y_1, \ldots, y_{m-1}): y_i \geq 0, \sum_{i=1}^{m-1} y_i \leq 1\}$. The correspondence $x \mapsto x^*$ is one-to-one and $\|x_1 - x_2\|_1 \leq 2\|x_1^* - x_2^*\|_1$. Let $x_1, \ldots, x_N \in \mathbb{S}_m$ such that $\|x_i - x_j\|_1 > \epsilon$, for $i, j = 1, \ldots, N$, $i \neq j$. Then $\|x_i^* - x_j^*\|_1 > \epsilon/2$, for $i, j = 1, \ldots, N$, $i \neq j$, and so that the ℓ_1-balls in $\mathbb{R}^{m-1}$ of radius $\epsilon/4$ centered at x_i^*, are disjoint. The union of these balls is clearly contained in the set

$$\left\{(y_1, \ldots, y_{m-1}): \sum_{i=1}^{m-1} |y_i| \leq (1 + \epsilon/4)\right\}. \tag{C.3}$$

Let V_{m-1} be the volume of the unit ℓ_1-ball in $\mathbb{R}^{m-1}$. Then the volume of the set in (C.3) is $(1 + \epsilon/4)^{m-1} V_{m-1} \leq (5/4)^{m-1} V_{m-1}$, while the volume of each $\epsilon/4$-ball around a point x_i^* is $(\epsilon/4)^{m-1} V_{m-1}$. A comparison of volumes gives the desired bound. □

Proposition C.2 (Euclidean space) *For* $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$ *and* $p \geq 1$, *for any* M *and* $0 < \epsilon < M$,

$$D(\epsilon, \{x \in \mathbb{R}^m : \|x\|_p \leq M\}, \|\cdot\|_p) \leq \left(\frac{3M}{\epsilon}\right)^m.$$

The proof of Proposition C.2 is similar to that of Proposition C.1. The lemma, shows, in particular, that the *local entropy*

$$\log D(\epsilon, \{x \in \mathbb{R}^m : \|x\|_p \leq k\epsilon\}, \|\cdot\|_p)$$

is bounded by $m \log(3k)$, independently of ϵ, for any fixed k. Thus this quantity behaves essentially as the dimension of the space.

Recall the variation Δp of a function p over a given partition, defined in (B.12).

Corollary C.3 *Given a measurable partition* $\{\mathfrak{X}_1, \ldots, \mathfrak{X}_m\}$ *of a probability space* $(\mathfrak{X}, \mathscr{X}, \nu)$, *let* $\mathcal{P}_1$ *and* $\mathcal{P}_2$ *be the classes of all probability densities* p *with* $\Delta p < \epsilon$ *and* $\Delta \log p < \epsilon$, *respectively. Then* $N(2\epsilon, \mathcal{P}_1, \|\cdot\|_1) \leq (5/\epsilon)^m$ *and* $N(\epsilon + \sqrt{2\epsilon}, \mathcal{P}_2, \|\cdot\|_1) \leq (5/\epsilon)^m$.

Proof For a given probability density p let p^* be its projection on the set of probability densities that are constant on each partitioning set, as in (B.11). By Lemma B.10 $\|p - p^*\|_1 \leq \Delta p \leq \epsilon$, for every $p \in \mathcal{P}_1$, and $\|p - p^*\|_1^2 \leq 2K(p; p^*) < 2\epsilon$, for every $p \in \mathcal{P}_2$. Furthermore, for any pair of discretizations, $\|p^* - q^*\|_1 = \sum_{j=1}^m |P(\mathfrak{X}_j) - Q(\mathfrak{X}_j)|$, so that the ϵ-covering number of the set of discretizations is bounded by $(5/\epsilon)^m$, by Proposition C.1. □

The preceding bounds show that entropy numbers of sets in Euclidean spaces grow logarithmically. For infinite-dimensional spaces the growth is much faster, as is illustrated by the following examples.

Definition C.4 (Hölder space) The *Hölder space* $\mathfrak{C}^\alpha(\mathfrak{X})$ is the class of all functions $f : \mathfrak{X} \to \mathbb{R}$ with domain a bounded, convex subset $\mathfrak{X} \subset \mathbb{R}^m$ such that $\|f\|_{\mathfrak{C}^\alpha} < \infty$, where the *Hölder norm* of order α is defined as

$$\|f\|_{\mathfrak{C}^\alpha} = \max_{k:|k|\leq \underline{\alpha}} \sup_{x \in \mathcal{X}} |D^k f(x)| + \max_{k:|k|=\underline{\alpha}} \sup_{x,y \in \mathcal{X}: x \neq y} \frac{|D^k f(x) - D^k f(y)|}{\|x - y\|^{\alpha - \underline{\alpha}}}.$$

Here $\underline{\alpha}$ is the biggest integer strictly smaller than α, and for a vector $k = (k_1, \ldots, k_m)$ of integers with sum $|k|$, D^k is the differential operator

$$D^k = \frac{\partial^{|k|}}{\partial x_1^{k_1} \cdots \partial x_m^{k_m}}.$$

Proposition C.5 (Hölder space) *There exists a constant* K *depending only on the dimension* m *and smoothness* α *such that, with* $\mathfrak{X}^* = \{y : \|y - x\| \leq 1\}$,

$$\log N(\epsilon, \{f \in \mathfrak{C}^\alpha : \|f\|_{\mathfrak{C}^\alpha} \leq M\}, \|\cdot\|_\infty) \leq K \operatorname{vol}(\mathfrak{X}^*) \left(\frac{M}{\epsilon}\right)^{m/\alpha}.$$

Definition C.6 (Sobolev space) The *Sobolev space* $\mathfrak{W}^\alpha(\mathfrak{X})$ of order $\alpha \in \mathbb{N}$ for an interval $\mathfrak{X} \subset \mathbb{R}$ is the class of all functions $f \in \mathbb{L}_2(\mathfrak{X})$ that possess an absolutely continuous $(\alpha - 1)$th derivative whose (weak) derivative $f^{(\alpha)}$ is contained in $\mathbb{L}_2(\mathfrak{X})$; the space is equipped with the norm $\|f\|_{2,2,\alpha} = \|f\|_2 + \|f^{(\alpha)}\|_2$. The Sobolev space $\mathfrak{W}^\alpha(\mathbb{R}^m)$ of order $\alpha > 0$ is the class of all functions $f \in \mathbb{L}_2(\mathbb{R}^m)$ with Fourier transform $\hat{f}$ satisfying $v_\alpha(f)^2 := \int |\lambda|^{2\alpha} |\hat{f}(\lambda)|^2 \, d\lambda < \infty$; the space is equipped with the norm $\|f\|_{2,2,\alpha} = \|f\|_2 + v_\alpha(f)$.

The preceding definition is restricted to functions of "integral" smoothness $\alpha \in \mathbb{N}$ with domain an interval in the real line or functions of general smoothness $\alpha > 0$ with domain a full Euclidean space. The relation between the two cases is that for $\alpha \in \mathbb{N}$ the function $\lambda \mapsto (i\lambda)^\alpha \hat{f}(\lambda)$ is the Fourier transform of the (weak) αth derivative $f^{(\alpha)}$ of a function $f: \mathbb{R} \to \mathbb{R}$ and hence $v_\alpha(f) = \|f^{(\alpha)}\|_2$. The Fourier transform is not entirely natural for functions not defined on the full Euclidecan space, which makes definitions of Sobolev spaces for general domains and smoothness a technical matter. One possibility suggests itself by the fact that in both cases the Sobolev space is known to be equivalent to the Besov space $\mathfrak{B}^\alpha_{2,2}(\mathfrak{X})$, defined in Definition E.8. Hence we may define a Sobolev space $\mathfrak{W}^\alpha(\mathfrak{X})$ in general as the corresponding Besov space. This identification suggested the notation $\|\cdot\|_{2,2,\alpha}$ for the norm. It is known that functions defined on a sufficiently regular domain (such as an interval) that belong to a given Besov space extend to a function on the full Euclidean domain of the same Besov norm. Thus it is also reasonable to define the Sobolev space $\mathfrak{W}^\alpha(\mathfrak{X})$ on a domain $\mathfrak{X} \subset \mathbb{R}^m$ in general as the set of functions $f: \mathfrak{X} \to \mathbb{R}$ that possess an extension belonging to $\mathfrak{W}^\alpha(\mathbb{R}^d)$, equipped with the norm of a "minimal" extension. Another possibility arise for *periodic* functions, which we briefly indicate below.

The Sobolev spaces of functions on a compact domain (identified with the corresponding Besov space) are compact in $\mathbb{L}_r(\mathbb{R})$ if the smoothness level is high enough: $\alpha > m(1/2 - 1/r)_+$. The entropy is not bigger than the entropy of the smaller Hölder spaces and hence the following proposition generalizes and improves Proposition C.5.

Proposition C.7 (Sobolev space) *For $\alpha > m(1/2 - 1/r)_+$ and $r \in (0, \infty]$ there exists a constant K that depends only on α and r such that*

$$\log N(\epsilon, \{f \in \mathfrak{W}^\alpha[0,1]^m : \|f\|_{2,2,\alpha} \le M\}, \mathbb{L}_r([0,1]^m) \le K \left(\frac{M}{\epsilon}\right)^{m/\alpha}.$$

Proposition C.8 (Monotone functions) *The collection $\mathcal{F}$ of monotone functions $f: \mathfrak{X} \to [-M, M]$ on an interval $\mathfrak{X} \subset \mathbb{R}$ satisfies, for $\|\cdot\|_{r,Q}$ the $\mathbb{L}_r(Q)$ norm relative to a probability measure Q, any $r \ge 1$ and a constant K that depends on r only,*

$$\log N_{[\,]}(\epsilon, \mathcal{F}, \|\cdot\|_{r,Q}) \le K \frac{M}{\epsilon}.$$

Proposition C.9 (Analytic functions) *The class $\mathfrak{A}_A[0,1]^m$ of all functions $f: [0,1]^m \to \mathbb{R}$ that can be extended to an analytic function on the set $G = \{z \in \mathbb{C}^m : \|z - [0,1]^m\|_\infty < A\}$ with $\sup_{z \in G} |f(z)| \le 1$ satisfies, for $\epsilon < 1/2$ and a constant c that depends on m only,*

$$\log N(\epsilon, \mathfrak{A}_A[0,1]^m, \|\cdot\|_\infty) \le c\left(\frac{1}{A}\right)^m \left(\log \frac{1}{\epsilon}\right)^{1+m}.$$

A function $f \in \mathbb{L}_2[0, 2\pi]$ may be identified with its sequence of Fourier coefficients $(f_j) \in \ell_2$. The αth derivative of f has Fourier coefficients $((ij)^\alpha f_j)$. This suggests to think of the ℓ_2-norm of the sequence $(j^\alpha f_j)$ as a *Sobolev norm* of order α. The following proposition gives the entropy of the corresponding Sobolev sequence space relative to the ℓ_2-norm.

Proposition C.10 (Sobolev sequence) *For $\|\theta\|_2 = (\sum_{i=1}^\infty \theta_i^2)^{1/2}$ the norm of ℓ_2 and $\alpha > 0$, for all $\epsilon > 0$,*

$$\log D\Big(\epsilon, \{\theta \in \ell_2 : \sum_{i=1}^\infty i^{2\alpha}\theta_i^2 \le B^2\}, \|\cdot\|_2\Big) \le \log(4(2e)^{2\alpha})\Big(\frac{3B}{\epsilon}\Big)^{1/\alpha}.$$

C.3 Historical Notes

Metric entropy was introduced by Kolmogorov and Tihomirov (1961) and subsequently developed by many authors. The results mentioned here are only a few examples from the literature. Proofs can be found at many places, including Edmunds and Triebel (1996), Dudley (1984, 2014), van der Vaart and Wellner (1996, 2017) and Giné and Nickl (2015). For instance, for Propositions C.5, see van der Vaart and Wellner (1996), pages 155–156; for Proposition C.8 see pages 159–162 in the same reference; for Proposition C.7, see van der Vaart and Wellner (2017) or Edmunds and Triebel (1996), page 105; for Proposition C.9, see Kolmogorov and Tihomirov (1961) or van der Vaart and Wellner (2017); for Proposition C.10 see e.g. Belitser and Ghosal (2003). These references also give many other results and a more extensive bibliography.

Problems

C.1 (Wu and Ghosal 2010) Consider a class $\mathcal{F}_{a,\Sigma}$ of mixtures of multivariate normal density given by $p(x) = p_{F,\Sigma} = \int \phi_d(x; \theta, \Sigma)\, dF(\theta)$, where Σ is a fixed $d \times d$ nonsingular matrix and $F(\|\theta\|_\infty \le a) = 1$. Show that

$$\log N(2\epsilon, \mathcal{F}_{a,\Sigma}, \|\cdot\|_1) \le \left(\sqrt{\frac{8d}{\pi \det(\Sigma)}}\frac{a}{\epsilon} + 1\right)[1 + \log(1 + \epsilon^{-1})].$$

Now let $\mathcal{F}^M_{a,h,\epsilon}$ stand for the class of all normal mixtures $p_{F,\Sigma}$, where $F(\|\theta\|_\infty > a) \le \epsilon$, $h \le \mathrm{eig}_1(\Sigma) \le \mathrm{eig}_d(\Sigma) \le M$, and $a > \sqrt{d}M\epsilon^{-1/2}$. Then

$$\log N(4\epsilon, \mathcal{F}^M_{a,h,\epsilon}, \|\cdot\|_1) \le \left(\sqrt{\frac{8d}{\pi}}\frac{2a}{h^{d/2}\epsilon} + 1\right)[1 + \log(1 + \epsilon^{-1})].$$

C.2 (Separability of Hölder spaces) Show that the functions $f_b : [0, 1] \to \mathbb{R}$ given by $f_b(x) = |x - b|^\alpha$, for $b \in (0, 1)$ and $\alpha \in (0, 1]$, satisfy $\sup_{x \ne y} |(f_b - f_c)(x) - (f_b - f_c)(y)|/|x - y|^\alpha \ge 2\alpha|b - c|^{\alpha-1}$. Conclude that $\mathfrak{C}^\alpha[0, 1]$ is not separable under its norm. [Hint: for $b \le c$ consider the increase of $f_b - f_c$ over the interval $[b, c]$.]