Chapter 3 Part I: Covering number, packing number and metric entropy

Di SU

June 2024

1 Introduction

This Section describes two classical ways of quantifying the 'size' of a metric space (T, d). Usually a metric space is too large to deal with, and one can 'decompose' the space into many smaller spaces that are smaller and easier to study. The 'size' of these smaller spaces are hence important and is used in the chaining strategy.

2 Covering number and packing number

Let $(\mathbb{T}, \rho)$ be a metric space which consists of a non-empty set $\mathbb{T}$ and a metric ρ.

Definition 2.1 (δ-cover)**.** A δ-cover of a set $\mathbb{T}$ with respect to a metric ρ is a set $\left\{\theta^1, \ldots, \theta^N\right\} \subset \mathbb{T}$ such that for each $\theta \in \mathbb{T}$, there exists some $i \in \{1, \ldots, N\}$ such that $\rho\left(\theta, \theta^i\right) \leq \delta$.

Definition 2.2 (Covering number)**.** The δ-covering number $N(\delta; \mathbb{T}, \rho)$ is the cardinality of the smallest δ-cover.

Lemma 2.1. *Given a function class $\mathcal{H}$ with pesudo metrics μ and μ' such that*

$$\mu(h, h') \leq c\mu'(h, h'), \forall h, h^{'} \in \mathcal{H},$$

then

$$N(\varepsilon; \mathcal{H}, \mu) \leq N(\varepsilon/c; \mathcal{H}, \mu').$$

Moreover, for any $h_0 \in \mathcal{H}$ and $c > 0$, then

$$N(\varepsilon; c\mathcal{H} + h_0, \mu) = N(\varepsilon; c\mathcal{H}, \mu) = N(\varepsilon/c; \mathcal{H}, \mu).$$

Definition 2.3 (δ-packing)**.** A δ-packing of a set $\mathbb{T}$ with respect to a metric ρ is a set $\left\{\theta^1, \ldots, \theta^M\right\} \subset \mathbb{T}$ such that $\rho\left(\theta^i, \theta^j\right) > \delta$ for all distinct $i, j \in \{1, 2, \ldots, M\}$.

Definition 2.4 (Packing number)**.** The δ-packing number $M(\delta; \mathbb{T}, \rho)$ is the cardinality of the largest δ-packing.

The following lemma gives a relation between covering numbers and packing numbers, and shows that working with either measure of 'size' is fine.

Lemma 2.2 (Covering-packing duality)**.** *For all $\delta > 0$, the packing and covering numbers are related as follows:*

$$M(2\delta; \mathbb{T}, \rho) \overset{(a)}{\leq} N(\delta; \mathbb{T}, \rho) \overset{(b)}{\leq} M(\delta; \mathbb{T}, \rho).$$

Proof. Suppose $M(\delta; \mathbb{T}, \rho) = N$ and the points of $F := \left\{\theta^1, \ldots, \theta^N\right\}$ are δ-separated. If $\theta \in \mathbb{T} \backslash F$ then the set $F \cup \{\theta\}$ cannot be δ-separated, which implies $\min_i \rho\left(\theta, \theta^i\right) \leq \delta$. The closed balls $B\left[\theta^i, \delta\right]$ for $1 \leq i \leq N$ cover $\mathbb{T}$.
For the second inequality, observe that no closed ball of radius δ can contain two points more than 2δ apart. In particular, each of the points θ^i must lie in a different ball for a δ covering of $\mathbb{T}$. □

The above proof basically follows from the geometry.

Example 2.1. *(from Pollard's) Let $\|\cdot\|$ denote some norm on $\mathbb{R}^n$, such as an ℓ^p norm, $|x|_p = \left(\sum_{i \leq n} |x_i|^p\right)^{1/p}$ for $p \geq 1$. As usual, write $B[t, r] := \{x \in \mathbb{R}^n : \|x - t\| \leq r\}$, the closed ball centered at t with radius r. Let $\mathfrak{m}_n$ denote Lebesgue's measure on $\mathcal{B}\left(\mathbb{R}^n\right)$.*
The covering/packing numbers for such norms share a common geometric bound, a property derived from the fact that

$$\mathfrak{m}_n B[t, r] := \{x \in \mathbb{R}^n : \|x - t\| \leq r\} = r^n \Lambda_n \quad \text{where } \Lambda_n := \mathfrak{m}_n B[0, 1].$$

Let $\{x_1, \ldots, x_N\}$ be any ϵr-separated set of points in B_r. The closed balls $B\left[x_i, \epsilon r/2\right]$, of radius $\epsilon r/2$ centered at the x_i, are disjoint and their union lies within $B_{r+\epsilon r/2}$. Thus

$$N \leq \frac{(r + \epsilon r/2)^n \Lambda_n}{(\epsilon r/2)^n \Lambda_n} = \left(\frac{2 + \epsilon}{\epsilon}\right)^n \leq (3/\epsilon)^n \quad \text{if } 0 < \epsilon \leq 1.$$

That is, $\operatorname{PACK}(\epsilon r, B_r, d) \leq (3/\epsilon)^n$ *for $0 < \epsilon \leq 1$, where d denotes the metric corresponding to $\|\cdot\|$.*

Example 2.2 (Covering of the binary hypercube)**.** *Consider the binary hypercube $\mathbb{H}^d := \{0, 1\}^d$ equipped with the rescaled Hamming metric $\rho_H(\theta, \widetilde{\theta}) := \frac{1}{d} \sum_{j=1}^d \mathbb{I}\left[\theta_j \neq \theta'_j\right]$. First, let us upper bound its δ covering number. Let $S = \{1, 2, \ldots, \lceil(1-\delta)d\rceil\}$, where $\lceil(1-\delta)d\rceil$ denotes the smallest*

integer larger than or equal to $(1-\delta)d$. Consider the set of binary vectors

$$\mathbb{T}(\delta) := \left\{\theta \in \mathbb{H}^d \mid \theta_j = 0 \quad \text{for all } j \notin S\right\}.$$

By construction, for any binary vector $\widetilde{\theta} \in \mathbb{H}^d$, we can find a vector $\theta \in \mathbb{T}(\delta)$ such that $\rho_H(\theta, \widetilde{\theta}) \leq \delta$. (Indeed, we can match $\widetilde{\theta}$ exactly on all entries $j \in S$, and, in the worst case, disagree on all the remaining $\lfloor \delta d \rfloor$ positions.) Since $\mathbb{T}(\delta)$ contains $2^{\lceil (1-\delta)d \rceil}$ vectors, we conclude that

$$\frac{\log N_H\left(\delta; \mathbb{H}^d\right)}{\log 2} \leq \lceil d(1-\delta) \rceil.$$

Let us lower bound its δ-covering number, where $\delta \in \left(0, \frac{1}{2}\right)$. If $\left\{\theta^1, \ldots, \theta^N\right\}$ is a δ-covering, then the (unrescaled) Hamming balls of radius $s = \delta d$ around each θ^ℓ must contain all 2^d vectors in the binary hypercube. Let $s = \lfloor \delta d \rfloor$ denote the largest integer less than or equal to δd. For each θ^ℓ, there are exactly $\sum_{j=0}^{s} \binom{d}{j}$ binary vectors lying within distance δd from it, and hence we must have $N\left\{\sum_{j=0}^{s} \binom{d}{j}\right\} \geq 2^d$. Now let $X_i \in \{0,1\}$ be i.i.d. Bernoulli variables with parameter $1/2$. Rearranging the previous inequality, we have

$$\frac{1}{N} \leq \sum_{j=0}^{s} \binom{d}{j} 2^{-d} = \mathbb{P}\left[\sum_{i=1}^{d} X_i \leq \delta d\right] \overset{\text{(i)}}{\leq} e^{-2d\left(\frac{1}{2}-\delta\right)^2},$$

where inequality (i) follows by applying Hoeffding's bound to the sum of d i.i.d. Bernoulli variables. Following some algebra, we obtain the lower bound

$$\log N_H\left(\delta; \mathbb{H}^d\right) \geq 2d\left(\frac{1}{2} - \delta\right)^2, \qquad \text{valid for } \delta \in \left(0, \frac{1}{2}\right)$$

3 Metric Entropy

Remark 3.1. *The logarithm of the covering number is sometimes called the metric entropy. See Dudley (1973, page 70) for the origin of that name.*

Theorem 3.1 (Upper bound)**.** *(One-step discretization bound) Let $D := \sup_{\theta, \tilde{\theta} \in \mathcal{T}} \rho_X(\theta, \tilde{\theta})$ denote the diameter of $\mathcal{T}$, and let $N_X(\delta; \mathbb{T})$ denote the δ-covering number of $\mathbb{T}$ in the ρ_X-metric. Let $\{X_\theta, \theta \in \mathbb{T}\}$ be a zero-mean sub-Gaussian process with respect to the metric ρ_X. Then for any $\delta \in [0, D]$ such that $N_X(\delta; \mathbb{T}) \geq 10$, we have*

$$\mathbb{E}\left[\sup_{\theta, \bar{\theta} \in \mathbb{T}} (X_\theta - X_{\bar{\theta}})\right] \leq 2\mathbb{E}\left[\sup_{\substack{\gamma, \gamma' \in \mathbb{T} \\ \rho_x(\gamma, \gamma') \leq \delta}} (X_\gamma - X_{\gamma'})\right] + 4\sqrt{D^2 \log N_X(\delta; \mathbb{T})}.$$

The above result involves the approximation error and the estimation error (the one involving the metric entropy). As $\delta \to 0^+$, the approximation error shrinks to 0 whereas the estimation error grows. In practice we can choose an optimal δ w.r.t. the two terms.

Proof. For a given $\delta \geq 0$ and associated covering number $N = N_X(\delta; \mathbb{T})$, let $\{\theta^1, \ldots, \theta^N\}$ be a δ-cover of $\mathbb{T}$. For any $\theta \in \mathbb{T}$, we can find some θ^i such that $\rho_X(\theta, \theta^i) \leq \delta$, and hence

$$X_\theta - X_{\theta^1} = (X_\theta - X_{\theta^i}) + (X_{\theta^i} - X_{\theta^1})$$

$$\leq \sup_{\gamma, \gamma' \in \mathbb{T}; \rho_X(\gamma, \gamma') \leq \delta} (X_\gamma - X_{\gamma'}) + \max_{i=1,2,\ldots,N} |X_{\theta^i} - X_{\theta^1}|.$$

Given some other arbitrary $\tilde{\theta} \in \mathbb{T}$, the same upper bound holds for $X_{\theta^1} - X_{\tilde{\theta}}$, so that adding together the bounds, we obtain

$$\sup_{\theta, \tilde{\theta} \in \mathbb{T}} (X_\theta - X_{\tilde{\theta}}) \leq 2 \sup_{\gamma, \gamma' \in \mathcal{T}, \rho_X(\gamma, \gamma') \leq \delta} (X_\gamma - X_{\gamma'}) + 2 \max_{i=1,\ldots,N} |X_{\theta^i} - X_{\theta^1}|. \tag{5.34}$$

Now by assumption, for each $i = 1, 2, \ldots, N$, the random variable $X_{\theta^i} - X_{\theta^1}$ is zero-mean and sub-Gaussian with parameter at most $\rho_X(\theta^i, \theta^1) \leq D$. Consequently, by the behavior of sub-Gaussian maxima (see Exercise 2.12(c)), we are guaranteed that

$$\mathbb{E}\left[\max_{i=1,\ldots,N} |X_{\theta^i} - X_{\theta^1}|\right] \leq 2\sqrt{D^2 \log N},$$

which yields the claim.

□

Theorem 3.2 (Upper bound)**.** *(Dudley's entropy integral) Let X_θ be a zero-mean stochastic process that is sub-Gaussian with respect to a pseudo-metric d on the indexing set T. Then*

$$\mathbb{E} \sup_\theta X_\theta \leq 8\sqrt{2} \int_0^\infty \sqrt{\log N(\epsilon, T, d)}, d\epsilon.$$

Proof.

$$E(\sup_\theta X_\theta) = E \sup_\theta (X_\theta - X_{\theta'}) \leq E \sup_{\theta, \theta'} (X_\theta - X_{\theta'}),$$

and choosing $\hat{\theta} \in \hat{T}$ (a minimal ε-cover) with $d(\hat{\theta}, \theta) \leq \varepsilon$ (and similarly for θ'), we have

$$\begin{aligned} X_\theta - X_{\theta'} =& X_\theta - X_{\hat{\theta}} + X_{\hat{\theta}} - X_{\hat{\theta}'} + X_{\hat{\theta}'} - X_{\theta'} \\ \leq& 2 \sup_{d(\theta, \hat{\theta}) \leq \varepsilon} (X_\theta - X_{\hat{\theta}}) + 2 \sup_{\hat{\theta}, \hat{\theta}' \in \hat{T}} (X_{\hat{\theta}} - X_{\hat{\theta}'}). \end{aligned}$$

We first define $\hat{T}_k = \hat{T}$, and think of it as a $(2^{-k}D)$-cover of $\hat{T}$, where $k = \lfloor \log_2(D/\epsilon) \rfloor$ ensures that $2^{-k}D \leq \epsilon$. Then we define $\hat{T}_{i-1} =$ a minimal $(2^{-(i-1)}D)$-cover of $\hat{T}_i$, for i going from $k-1$ down to 0. Notice that $\hat{T}_0$ is a minimal D-cover of $\hat{T}_1$, so $|\hat{T}_0| = 1$.
Pick $\hat{\theta}_k = \hat{\theta}$, and then pick $\hat{\theta}_{i-1} \in \hat{T}_{i-1}$ as the best approximation of $\hat{\theta}_i$. We can write $\hat{\theta}_{i-1} = f_{i-1}(\hat{\theta}_i)$, where $f_{i-1} : \hat{T}_i \to \hat{T}_{i-1}$ is the best approximation operator. Then we can write

$$X_{\hat{\theta}} = X_{\hat{\theta}_k} = X_{\hat{\theta}_0} + \sum_{i=1}^k \left(X_{\hat{\theta}_i} - X_{\hat{\theta}_{i-1}}\right)$$

and, using the same notation for $\hat{\theta}'$, we have

$$\begin{aligned} X_{\hat{\theta}} - X_{\hat{\theta}'} &= X_{\hat{\theta}_k} - X_{\hat{\theta}'_k} \\ &= \sum_{i=1}^{k} \left(X_{\hat{\theta}_i} - X_{\hat{\theta}_{i-1}} \right) - \sum_{i=1}^{k} \left(X_{\hat{\theta}'_i} - X_{\hat{\theta}'_{i-1}} \right). \end{aligned}$$

Thus,

$$\mathbf{E} \sup_{\hat{\theta},\hat{\theta}' \in \hat{T}} X_{\hat{\theta}} - X_{\hat{\theta}'} \leq 2 \sum_{i=1}^{k} \mathbf{E} \sup_{\hat{\theta}_i \in \hat{T}_i} \left(X_{\hat{\theta}_i} - X_{f_{i-1}(\hat{\theta}_i)} \right).$$

Since $d(\hat{\theta}_i, \hat{\theta}_{i-1}) \leq 2^{-(i-1)} D$, the Finite Lemma shows that

$$\mathbf{E} \sup_{\hat{\theta}_i \in \hat{T}_i} \left(X_{\hat{\theta}_i} - X_{f_{i-1}(\hat{\theta}_i)} \right) \leq 2^{-(i-1)} D \sqrt{2 \log |\hat{T}_i|} \leq 2^{-(i-1)} D \sqrt{2 \log N(2^{-i} D, T)}.$$

Since $\log N(2^{-i} D) \leq \log N(u)$ for $u \leq 2^{-i} D$, we can approximate the area of the rectangle from $(2^{-(i+1)} D, 0)$ to $(2^{-i} D, \sqrt{2 \log N(2^{-i} D)})$ by the integral under $\sqrt{2 \log N(u)}$ for u in that interval (which has length $2^{-(i+1)} D$)):

$$\begin{aligned} 2^{-(i-1)} D \sqrt{2 \log N(2^{-i} D)} &= 4 \times 2^{-(i+1)} D \sqrt{2 \log N(2^{-i} D)} \\ &\leq 4 \int_{2^{-(i+1)} D}^{2^{-i} D} \sqrt{2 \log N(u, T)} du. \end{aligned}$$

We have

$$\begin{aligned} \mathbb{E} \sup_{\theta} X_\theta &\leq 2\mathbb{E} \sup_{d(\theta,\tilde{\theta}) \leq \epsilon} (X_\theta - X_{\tilde{\theta}}) + 2 \sum_{i=1}^{k} \mathbb{E} \sup_{\tilde{\theta}_i \in T_i} \left(X_{\tilde{\theta}_i} - X_{f_{i-1}(\tilde{\theta}_i)} \right) \\ &\leq 2\mathbb{E} \sup_{d(\theta,\tilde{\theta}) \leq \epsilon} (X_\theta - X_{\tilde{\theta}}) + 2 \sum_{i=1}^{k} 2^{-(i-1)} D \sqrt{2 \log N(2^{-i} D, T)} \\ &\leq 2\mathbb{E} \sup_{d(\theta,\tilde{\theta}) \leq \epsilon} (X_\theta - X_{\tilde{\theta}}) + 8\sqrt{2} \int_{2^{-(k+1)} D}^{D/2} \sqrt{\log N(u, T)} du. \end{aligned}$$

When $\epsilon \to 0$, the first term goes to zero and (since $k = \lfloor \log_2(D/\epsilon) \rfloor$), the second term approaches the integral from 0 to $D/2$, which gives the result. □

Theorem 3.3 (Lower bound)**.** *(Sudakov minoration) Let $\{X_\theta, \theta \in \mathbb{T}\}$ be a zero-mean Gaussian process defined on the non-empty set $\mathbb{T}$. Then*

$$\mathbb{E} \left[\sup_{\theta \in \mathbb{T}} X_\theta \right] \geq \sup_{\delta > 0} \frac{\delta}{2} \sqrt{\log M_X(\delta; \mathbb{T})},$$

where $M_X(\delta; \mathbb{T})$ is the δ-packing number of $\mathbb{T}$ in the metric $\rho_X(\theta, \tilde{\theta}) := \sqrt{\mathbb{E}[(X_\theta - X_{\tilde{\theta}})^2]}$.

Proof. For any $\delta > 0$, let $\{\theta^1, \ldots, \theta^M\}$ be a δ-packing of $\mathbb{T}$, and consider the sequence $\{Y_i\}_{i=1}^M$ with elements $Y_i := X_{\theta^i}$. Note that by construction, we have the lower bound

$$\mathbb{E}[(Y_i - Y_j)^2] = \rho_X^2(\theta^i, \theta^j) > \delta^2 \quad \text{for all } i \neq j.$$

Now let us define an i.i.d. sequence of Gaussian random variables $Z_i \sim N(0, \delta^2/2)$ for $i = 1, \ldots, M$. Since $\mathbb{E}[(Z_i - Z_j)^2] = \delta^2$ for all $i \neq j$, the pair of random vectors Y and Z satisfy the Sudakov-Fernique condition (5.59), so that we are guaranteed that

$$\mathbb{E}\left[\sup_{\theta \in \mathbb{T}} X_\theta\right] \geq \mathbb{E}\left[\max_{i=1,\ldots,M} Y_i\right] \geq \mathbb{E}\left[\max_{i=1,\ldots,M} Z_i\right].$$

Since the variables $\{Z_i\}_{i=1}^M$ are zero-mean Gaussian and i.i.d., we can apply standard results on i.i.d. Gaussian maxima (viz. Exercise 2.11) to obtain the lower bound $\mathbb{E}[\max_{i=1,\ldots,M} Z_i] \geq \frac{\delta}{2}\sqrt{\log M}$, thereby completing the proof. □

4 Reference

https://www.stat.berkeley.edu/ bartlett/courses/2013spring-stat210b/notes/14notes.pdf
Wainwright book
CUHK STAT6050 Lecture 5