

SDS 384 11: Theoretical Statistics

Lecture 14: Uniform Law of Large Numbers- Covering number

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin

Definitions

- Recall that a metric space (\mathcal{T}, ρ) consists of a nonempty set \mathcal{T} and a mapping $\rho : \mathcal{T} \times \mathcal{T} \rightarrow \mathbb{R}$ that satisfies:
 - Non-negative: $\rho(\theta, \theta') \geq 0$ for all (θ, θ') with equality iff $\theta = \theta'$.
 - Symmetric: $\rho(\theta, \theta') = \rho(\theta', \theta)$ for all pairs (θ', θ) , and
 - Triangle ineq holds: $\rho(\theta, \theta') + \rho(\theta', \theta'') \geq \rho(\theta, \theta'')$
- Examples:
 - $\mathcal{T} = \mathbb{R}^d$, $\rho(\theta, \theta') = \|\theta - \theta'\|_2$
 - $\mathcal{T} = \{0, 1\}^d$ with $\rho(\theta, \theta') = \frac{1}{d} \sum_i 1(\theta_i \neq \theta'_i)$

Covering numbers

Definition

A δ cover of a set \mathcal{T} w.r.t to a metric ρ is a set $\{\theta^1, \dots, \theta^N\}$ such that for every $\theta \in \mathcal{T}$, $\exists i \in [N]$, s.t. $\rho(\theta, \theta^i) \leq \delta$. The δ covering number $N(\delta; \mathcal{T}, \rho)$ is the cardinality of the smallest δ cover.

- We will consider metric spaces which are totally bounded, i.e. $N(\delta; \mathcal{T}, \rho) < \infty$ for all $\delta > 0$.
- The covering number is non-increasing in δ , i.e. $N(\delta) \geq N(\delta')$ for all $\delta < \delta'$
- We are interested in something called Metric entropy, which is the logarithm of the covering number.

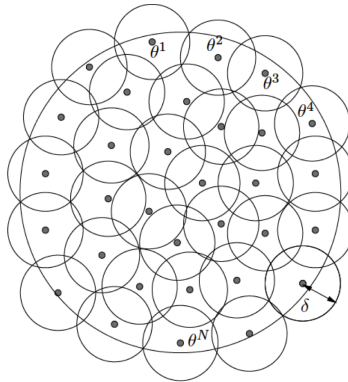


Figure 1: [courtesy: Martin Wainwright's book]

- A δ covering can be thought of as a union of balls with radius δ .

Covering number of a unit cube

Example

Consider the interval $[-1, 1]$ with $\rho(\theta, \theta') = |\theta - \theta'|$. We have $N(\delta; [-1, 1], |\cdot|) \leq \frac{1}{\delta} + 1$

- Divide the interval into L sub-intervals centered at $\theta^i := -1 + (2i - 1)\delta$ for $i \in [L]$ and each of length at most 2δ .
- By construction this is a δ covering.
- So $L \leq 1 + 1/\delta$

Covering the binary hypercube

Example

Consider a d dimensional binary hypercube $\mathcal{T} = \{0,1\}^d$ with the Hamming metric defined before.

$$\frac{\log N(\delta; \mathcal{T}, \rho)}{\log 2} \leq \lceil d(1 - \delta) \rceil$$

- Let $S = \{1, 2, \dots, \lceil \delta d \rceil\}$
- Consider the set of binary vectors $\mathcal{S}(\delta) := \{\theta \in \mathcal{T} : \theta_j = 0\}$.
- By construction, for every binary vector $\theta' \in \mathcal{T}$, we can find a vector $\theta \in \mathcal{S}(\delta)$ such that $\rho(\theta, \theta') \leq \delta$
- $N(\delta; \mathcal{T}, \rho) \leq |\mathcal{S}(\delta)| = 2^{\lceil d(1-\delta) \rceil}$

Lower bound on Covering number of the binary hypercube

- Let $\delta \in (0, 1/2)$
- If $\{\theta^1, \dots, \theta^N\}$ is a δ covering, then the (unrescaled) Hamming balls of radius $s = \delta d$ around each θ^ℓ must contain all 2^d vectors.
- Let $s = \lfloor \delta d \rfloor$
- For each θ^i there are exactly $\sum_{j=0}^d \binom{d}{j}$ vectors within δd distance.
- So $N \sum_{j=0}^d \binom{d}{j} \geq 2^d$

Lower bound on Covering number of the binary hypercube

- Let $\delta \in (0, 1/2)$
- So $N \sum_{j=0}^s \binom{d}{j} \geq 2^d$
- Now take a Binomial $(d, 1/2)$ random variable X .
- $P(X \leq \delta d) = \sum_{j=0}^s \binom{d}{j} / 2^d$
- So $N \geq \frac{1}{P(X \leq \delta d)}$
- Using the Hoeffding bound gives: $N \geq \exp(\frac{d}{2}(1/2 - \delta)^2)$
- Using the refined version in your homework gives:
 $N \geq \exp(dKL(\delta||1/2))$

Definition

An δ -packing of \mathcal{T} w.r.t a metric ρ is a set $\{\theta^1, \dots, \theta^M\}$ such that $\rho(\theta^i, \theta^j) > \delta$ for every distinct pair $i, j \in [M]$. The δ packing number $M(\delta; \mathcal{T}, \rho)$ is the cardinality of the largest δ packing.

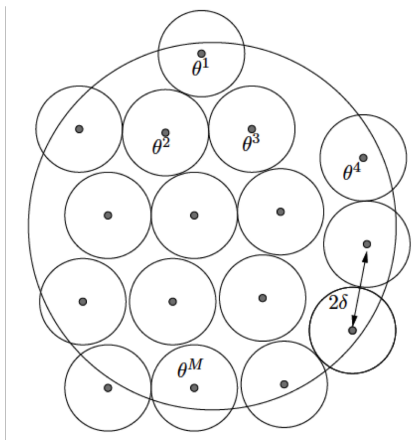


Figure 2: [courtesy: Martin Wainwright's book]

- A δ covering can be thought of as a union of balls with radius δ such that no two balls touch.

Relationship between packing and covering numbers

Theorem

For all $\delta > 0$,

$$M(2\delta; \mathcal{T}, \rho) \leq N(\delta; \mathcal{T}, \rho) \leq M(\delta; \mathcal{T}, \rho)$$

- This is saying that packing and covering numbers exhibit the same scaling behavior as $\delta \rightarrow 0$.

- For each element of $N(\delta; \mathcal{T}; \rho)$, there can be only one or less element of $M(2\delta; \mathcal{T}; \rho)$. Because otherwise, two elements of the 2δ packing will be within 2δ of each other via triangle inequality.
- Consider a δ packing of \mathcal{T} . Since it is maximal, there are no more points in \mathcal{T} which can be added without falling within δ distance of one of the elements. Hence this is also an epsilon cover. Hence the last inequality.

Covering and Packing numbers-example

Theorem

Let ρ be the Euclidean norm on \mathbb{R}^d . Let $B_1(0)$ be the unit ball centered at the origin (WLOG).

$$\frac{1}{\epsilon^d} \leq N(\epsilon, B_1, \rho) \leq (1 + 2/\epsilon)^d$$

- Consider an ϵ cover $\{\theta^1, \dots, \theta^N\}$. Now,

$$B_1 \subseteq \bigcup_{i=1}^N B_\epsilon(\theta^i)$$

$$\text{vol}(B_1) \leq N \text{vol}(B_\epsilon(\theta^i)) = N \epsilon^d \text{vol}(B_1)$$

$$N \geq 1/\epsilon^d$$

Proof-upper bound

- Consider a ϵ packing $\{\theta^1, \dots, \theta^M\}$
- This is an union of disjoint balls of radius $\epsilon/2$

$$\bigcup_i B_{\epsilon/2}(\theta^i) \subseteq B_{1+\epsilon/2}$$

$$M \text{vol}(B_{\epsilon/2}(\theta^i)) \leq (1 + \epsilon/2) \text{vol}(B_{1+\epsilon/2})$$

$$M(\epsilon/2)^d \text{vol}(B_1) \leq (1 + \epsilon/2)^d \text{vol}(B_1)$$

$$M \leq (1 + 2/\epsilon)^d$$

Example-smoothly parametrized problems

- Consider the following function class parametrized by $\theta \in \Theta$.

$$\mathcal{F} := \{f_\theta(\cdot) : \theta \in \Theta\}$$

- Let $\|\cdot\|_\Theta$ be the norm for θ and $\|\cdot\|_{\mathcal{F}}$ be the norm for \mathcal{F} .
- Say $\|f_\theta(\cdot) - f_{\theta'}(\cdot)\|_{\mathcal{F}} \leq L\|\theta - \theta'\|_\Theta$
- Then $N(\epsilon; \mathcal{F}, \|\cdot\|_{\mathcal{F}}) \leq N(\epsilon/L; \Theta, \|\cdot\|_\Theta)$

Example-smoothly parametrized problems

- A Lipschitz parametrization allows us to go from cover of the Θ space to cover of the f_θ space with a loss of L .
- If \mathcal{F} is parametrized by a compact set of d parameters then $N(\epsilon, \mathcal{F}) = O(1/\epsilon^d)$

Example-Lipschitz functions on the unit interval

Example

$$\mathcal{F}_L = \{g : [0, 1] \rightarrow \mathbb{R} \mid g(0) = 0, |g(x) - g(y)| \leq L|x - y|, \forall x, y \in [0, 1]\}$$

Metric entropy scales as $\log N(\delta; \mathcal{F}_L) \asymp L/\delta$ for small enough $\delta > 0$.

- Its sufficient to consider a sufficiently large packing of \mathcal{F}_L
- For a given ϵ define $M = \lfloor \frac{1}{\epsilon} \rfloor$
- Let $x_i = (i - 1)\epsilon$ for $i = 1, \dots, M + 1$
-

$$\phi(x) := \begin{cases} 0 & x < 0 \\ x & x \in [0, 1] \\ 1 & x > 1 \end{cases} \quad (1)$$

- Define $f(\beta) = \sum_{i=1} \beta_i L_{\epsilon} \phi\left(\frac{y - x_i}{\epsilon}\right)$ for $\beta \in \{-1, 1\}^M$

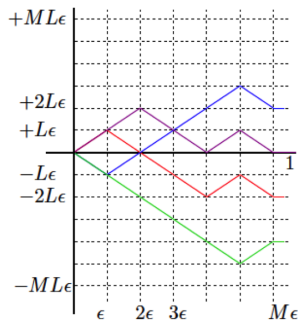


Figure 5-2. The function class $\{f_\beta, \beta \in \{-1, +1\}^M\}$ used to construct a packing of the Lipschitz class \mathcal{F}_L . Each function is piecewise linear over the intervals $[0, \epsilon]$, $[\epsilon, 2\epsilon]$, \dots , $[(M-1)\epsilon, M\epsilon]$ with slope either $+L$ or $-L$. There are 2^M functions in total, where $M = \lceil 1/\epsilon \rceil$.

- For any pair $\beta \neq \beta' \in \{-1, 1\}^M$ there is at least one interval where they have the same starting point.
- So $\|f(\beta) - f(\beta')\|_\infty \geq 2L\epsilon$
- $f_\beta \in \mathcal{F}_L$ for all $\beta \in \{-1, 1\}^M$
- So f_β forms a $2L\lfloor 1/\epsilon \rfloor$ packing.
- Making $\epsilon L = \delta$ we see

$$N(\delta; \mathcal{F}_L, \|\cdot\|_\infty) \geq M(2L\epsilon; \mathcal{F}_L, \|\cdot\|_\infty) = 2^{\lfloor \frac{1}{\epsilon} \rfloor} = 2^{\lfloor \frac{L}{\delta} \rfloor}$$

- Also the set f_β also form a suitable covering of the original functions, and this gives the upper bound.

Acknowledgment

This lecture was very much based on Martin Wainwright's unpublished book and Peter Bartlett's notes.

