Lecture 8 Uniform Bound

IEMS 402 Statistical Learning

Northwestern

Ref

https://people.eecs.berkeley.edu/~bartlett/courses/281b-sp08/18.pdf
https://people.eecs.berkeley.edu/~bartlett/courses/281b-sp08/19.pdf

Uniform Bound

Recall

$$L(\hat{\theta}) - L(\theta^*) = \underbrace{L(\hat{\theta}) - \hat{L}(\hat{\theta})}_{①} + \underbrace{\hat{L}(\hat{\theta}) - \hat{L}(\theta^*)}_{②} + \underbrace{\hat{L}(\theta^*) - L(\theta^*)}_{③}.$$

$$\leq 2 \sup_{\theta \in \Theta} | L(\theta) - L(\hat{\theta}) |$$

Uniform Bound

Bound $\sup_{\theta \in \Theta} | L(\theta) - L(\hat{\theta}) |$

Why can't we use Chernoff/CLT?

Uniform Bound

Bound $\sup_{\theta \in \Theta} |L(\theta) - L(\hat{\theta})|$

Why can't we use Chernoff/CLT?

Uniform Bound:

$$\Pr\left[\forall \theta \in \Theta, |\hat{L}(\theta) - L(\theta)| \geq \varepsilon'\right] \leq \sum_{\theta \in \Theta} \Pr\left[|\hat{L}(\theta) - L(\theta)| \geq \varepsilon'\right].$$

Finite Hypothesis Class

Theorem 4.1. *Suppose that our hypothesis class $\mathcal{H}$ is finite and that our loss function ℓ is bounded in $[0,1]$, i.e. $0 \leq \ell((x,y),h) \leq 1$. Then $\forall \delta$ s.t. $0 < \delta < \frac{1}{2}$, with probability at least $1-\delta$, we have*

$$|L(h) - \hat{L}(h)| \leq \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2n}} \qquad \forall h \in \mathcal{H}. \tag{4.9}$$

As a corollary, we also have

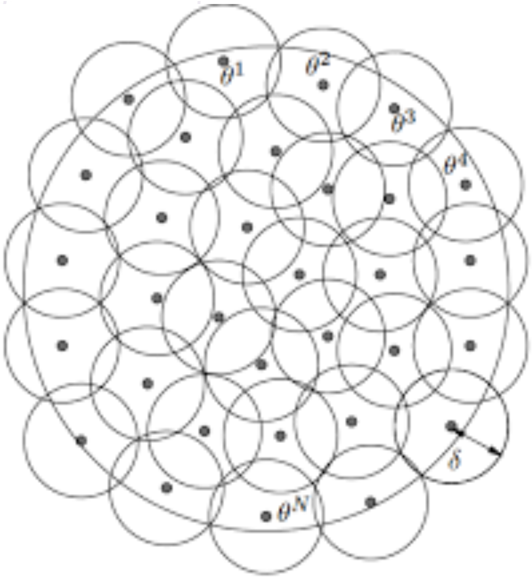
$$L(\hat{h}) - L(h^*) \leq \sqrt{\frac{2(\ln|\mathcal{H}| + \ln(2/\delta))}{n}}. \tag{4.10}$$

Finite Hypothesis Class

Infinite Hypothesis Class

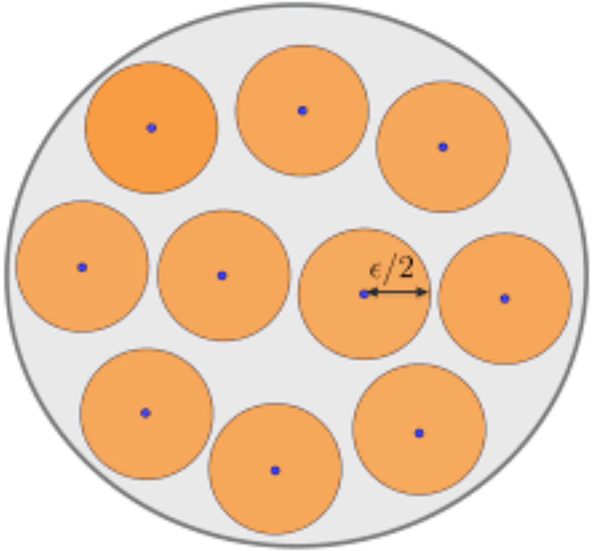
Epsilon Cover

Definition 14.1 (ϵ-covering). Let $(V, \|\cdot\|)$ be a normed space, and $\Theta \subset V$. $\{V_1, ..., V_N\}$ is an ϵ-covering of Θ if $\Theta \subset \cup_{i=1}^{N} B(V_i, \epsilon)$, or equivalently, $\forall \theta \in \Theta$, $\exists i$ such that $\|\theta - V_i\| \leq \epsilon$.



Epsilon Packing

Definition 14.2 (ϵ-packing)**.** Let $(V, \|\cdot\|)$ be a normed space, and $\Theta \subset V$. $\{\theta_1, ..., \theta_M\}$ is an ϵ-packing of Θ if $\min_{i \neq j} \|\theta_i - \theta_j\| > \epsilon$ (notice the inequality is strict), or equivalently $\cap_{i=1}^{M} B(\theta_i, \epsilon/2) = \emptyset$.



Covering and Packing Number

Definition 14.3 (Covering number). $N(\Theta, \|\cdot\|, \epsilon) := \min\{n : \exists\epsilon\text{-covering over } \Theta \text{ of size } n\}$.

Definition 14.4 (Packing number). $M(\Theta, \|\cdot\|, \epsilon) := \max\{m : \exists\epsilon\text{-packing of } \Theta \text{ of size } m\}$.

Fact

Theorem 14.1. *Let $(V, \|\cdot\|)$ be a normed space, and $\Theta \subset V$. Then*

$$M(\Theta, \|\cdot\|, 2\epsilon) \overset{(a)}{\leq} N(\Theta, \|\cdot\|, \epsilon) \overset{(b)}{\leq} M(\Theta, \|\cdot\|, \epsilon).$$

Dimension Depedency

Intuition: A d-dimensional set has metric dimension d. ($N(\epsilon) = \Theta(1/\epsilon^d)$.)

Example: $([0,1]^d, l_\infty)$ has $N(\epsilon) = \Theta(1/\epsilon^d)$.

Discretization Theorem

Theorem 1.1. Discretization Theorem:

$$\hat{R}(f) \leq \inf_{\alpha} \left(\alpha + \sqrt{\frac{2 \log N(\alpha, F, L_2(P_n))}{n}} \right)$$

Application

Theorem 3.3 (Subgaussian covariance concentration). *Suppose $A \in \mathbb{R}^{d \times n}$ is a random matrix with columns $a_i \in \mathbb{R}^d$ that are independent, zero-mean, and 1-subgaussian. Further, assume that $\mathbb{E}\left[\frac{1}{n}AA^\top\right] = I_d$.*

Then, $\exists$ universal constant $C > 0$ such that, $\forall s \geqslant 0$,

$$\mathbf{Pr}\left[\left\|\frac{1}{n}AA^\top - I_d\right\|_{op} > \max(\delta, \delta^2)\right] \leqslant 2\exp(-s^2), \text{ for } \delta = C\left(\sqrt{\frac{d}{n}} + \frac{s}{\sqrt{n}}\right).$$

Chaining

Dudley's Theorem

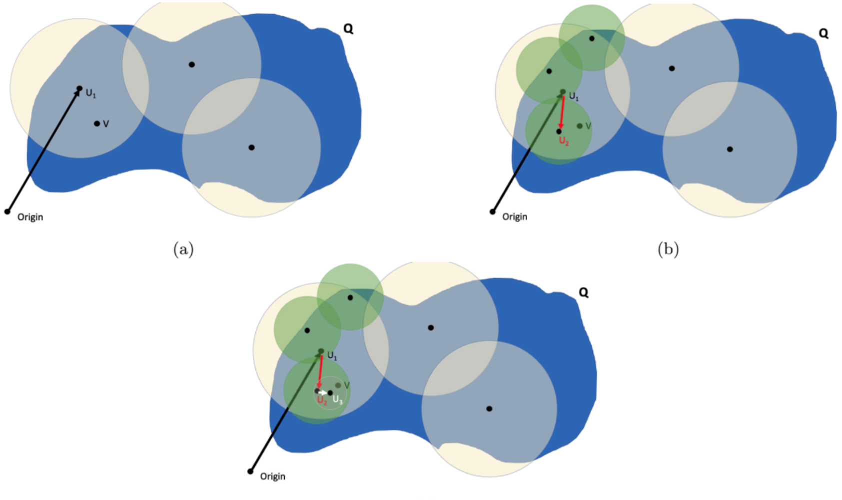
Theorem 3.1. Dudley:

$$\hat{R}(F) \leq 12 \int_0^\infty \frac{logN(\epsilon, F, L_2(P_n))}{n} d\epsilon$$

Chaining

The Chaining idea is to rewrite f as follows:

$$f = f + \sum_{j=1}^{N} (\hat{f}_j - \hat{f}_{j-1}) + \cancel{\hat{f}_0} - \hat{f}_N.$$



(a) (b) (c)

Example

Example. F = the non-decreasing function from $\mathbb{R}$ to $[0,1]$.

We can actually cover such a function uniformly. We only need to approximate it at n points, marked in the figure. If it is within α at each of these points then the L_2 distance will be no more than α. From the approximating points one can produce a non-decreasing function: for each of the α-levels (of which there will be $1/\alpha$), just specify one of the n points at which it increases above that level. From this we can (loosely, but to the right order of magnitude) upper bound the size of the class of estimate functions: $|\hat{F}| \leq n^{1/\alpha}$.

We see that we can cover F in L_2:

$$N(\alpha, F, L_2(P_n)) \leq Cn^{1/\alpha}.$$

1. The Discretization Theorem gives

$$\hat{R}_n(F) \leq c\left(\frac{\log n}{n}\right)^{1/3}$$

2. The Chaining Theorem gives

$$\hat{R}_n(F) \leq 12\int_0^1 \sqrt{\frac{\log n}{\alpha n}}d\alpha = 12\sqrt{\frac{\log n}{n}}\int_0^1 \sqrt{\frac{1}{\alpha}}d\alpha = 24\sqrt{\frac{\log n}{n}}$$

Chaining