Idea. $\mathbb{P}(E_1 \cup E_2) \le \mathbb{P}(E_1) + \mathbb{P}(E_2)$

Lecture 8 Uniform Bound

IEMS 402 Statistical Learning

Northwestern

Ref

https://raw.githubusercontent.com/tengyuma/cs229m_notes/main/master.pdf section 4.1-4.3

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

http://www.stat.yale.edu/~yw562/teaching/598/lec14.pdf

Uniform Bound

Recall

ERM

$\theta^* = \arg\min_\theta L(\theta)$

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

$\mathbb{E}_{\mathbb{P}} - \mathbb{E}_{\hat{\mathbb{P}}}$

$\hat{\theta}$ and $\hat{\mathbb{P}}$ are correlated !!

Optimisation

$\mathbb{E}_{\mathbb{P}}$ $\mathbb{E}_{\hat{\mathbb{P}}}$

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

generalization

$\mathbb{E}_{\mathbb{P}} - \mathbb{E}_{\hat{\mathbb{P}}}$ This can be bounded by concentration

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:

use Chernoff bound for each hypothesis

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

sum it up together
↳ upper bound of $\sup_\theta |L(\theta) - \hat{L}(\theta)|$

Finite Hypothesis Class

Chernoff

$|\Theta| < \infty$

then it's subgaussian

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.* $0 \le \ell((x,y),h) \le 1$. *Then* $\forall \delta$ *s.t.* $0 < \delta < \frac{1}{2}$ *, with probability at least* $1-\delta$, *we have*

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

$$\sqrt{\frac{\ln|\#\text{number of hyp}|}{n}}$$

As a corollary, we also have

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

Finite Hypothesis Class

$\mathbb{P}(E_1 \cup E_2 \cup \cdots \cup E_k) \leq \sum_{i=1}^{k} \mathbb{P}(E_i)$

E_i in our proof is

uniform bound

$\sum_{\theta \in \Theta}$

$\mathbb{P}\left(|\hat{L}(h) - L(h)| > \varepsilon \right)$

using chernoff bound $\leq 2\exp(-2n\varepsilon^2)$

$\leq |\Theta| \exp(-n\varepsilon^2)$

$\Downarrow$

$\varepsilon = \sqrt{\frac{\ln|\Theta|}{n}}$

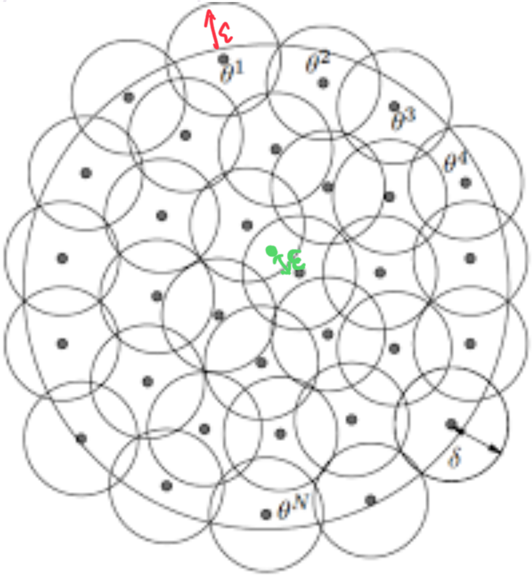
"rate distortion function"

Infinite Hypothesis Class

Epsilon Cover

#(covering)

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 $\|\theta - V_i\| \leq \epsilon$.



use finite hypothesis class to approximate infinite hypothesis class!

black dot $\xleftrightarrow{\varepsilon\text{- distance.}}$ all the hypothesis

bias/approximation

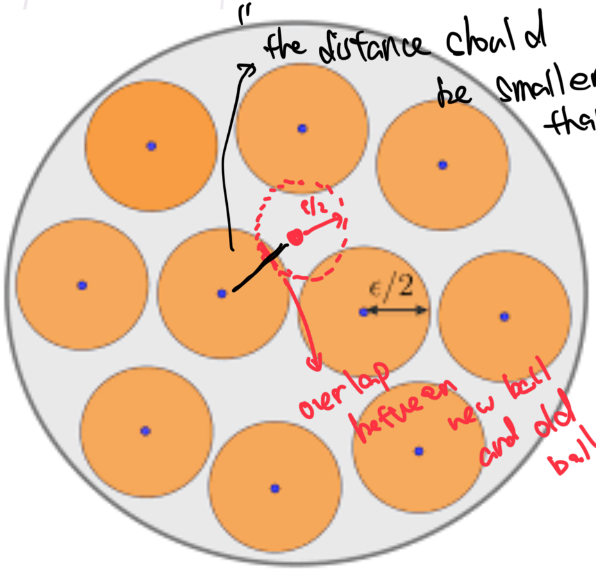
$$\varepsilon + \sqrt{\frac{\ln N}{n}}$$

generalization / variance.

$\varepsilon \downarrow$ $\ln N \uparrow$

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 $\cap_{i=1}^{M} B(\theta_i, \epsilon/2) = \emptyset$.



"the distance should be smaller than ε"

biggest packing means → add one more ball the new ball will overlap with one of the old ball

claim: the biggest ε-packing is also a ε-cover !!!

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*

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

Dimension Depedency

Method 1
Volumn Argument

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

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

no matter what is the norm here

this is always tight
but tight only when ε is small enough.

Method 2
Convex hull argument

Midterm

$\sum_{i=1}^{d} |x_i| \leq 1$

l_1 ball

Convex hull argument
"optimal?"

Method 2 is good when ε is large $\rightarrow$ Constant $\times$ data "is small" $\rightarrow$ poly(d)
poly $\rightarrow$ is suboptimal

Method 1 is optimal when ε is small $\rightarrow$ Constant $\times$ data "large" $\rightarrow$ exp(d)
poly $\rightarrow$ is optimal

Volumn Argument

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, $\forall s \geqslant 0$,

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

log covering number

chernoff bound

do uniform bound.

$$\|A\|_{op} = \max \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|=1} \|Ax\|$$

$x \in \mathbb{R}^d$.

$\sqrt{\|A\|_{op}}$ = largest eigenvalue.

Talagrand

Chaining

Dudley's Theorem

Theorem 3.1. Dudley:

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

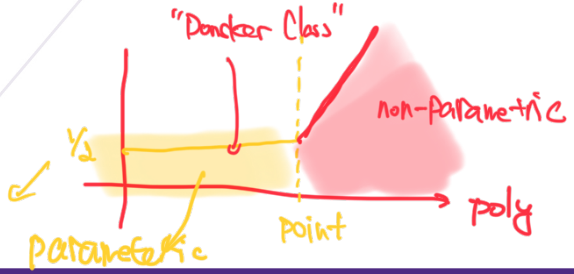
- Before Chaining. Error $\leq \sqrt{\frac{\log \text{Covernumber}(\varepsilon)}{n}} + \varepsilon$

- After chaining: Error $\leq \varepsilon + \int_{\varepsilon}^{1} \sqrt{\frac{\log \text{Covernumber}(t)}{n}} dt$

Try: If log Covernumber$(\varepsilon) \propto \varepsilon^{-\text{poly}}$

"Gil Kur Ph.d. Thesis"

$R_n(f) \propto \sqrt{\frac{1}{n}}$

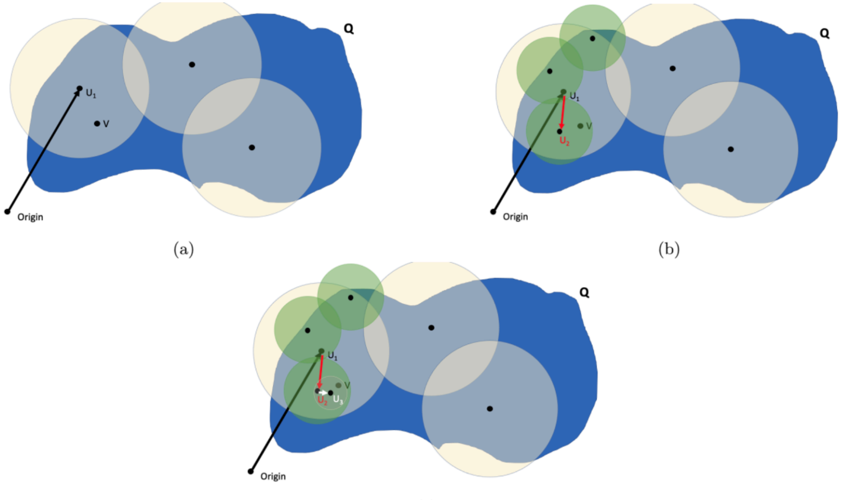


Chaining

"Multiscale"

The Chaining idea is to rewrite f as follows:

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



(c)

$$f = f - \underbrace{f_{fine}}_{} + f_{fine} \qquad \varepsilon \qquad \sqrt{\frac{\log \text{covering num}}{n}}$$

$$= f - f_{fine} + f_{fine} - f_{coarse} + f_{coarse}. \qquad \varepsilon \qquad \varepsilon\sqrt{\frac{\log \text{covering}}{n}}$$

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