3.13 Covering number (metric entropy)

Covering number : counts the number of ε-balls needed to cover the hypothesis class.

Goal : try to have shattering coefficients for real-valued functions.

[Definition 12] metric space

- A metric space $(\mathcal{X}, \rho)$: set $\mathcal{X} \supseteq \mathcal{F}$ and ρ is a metric.
- $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, non-negative, symmetric, satisfy the triangle inequality and evaluate to 0 iff its arguments are equal.
- If $\rho(f, f') = 0$ is possible, the we say ρ is pseudometric.

[Definition 13] ball

Let $(\mathcal{X}, \rho)$ be a metric space, Define ε-ball as

$$B_\varepsilon(f) := \{ f' \in \mathcal{X} : \rho(f, f') \le \varepsilon \}$$

[Definition 14] covering number

(i) An ε-cover of a set $\mathcal{F} \subseteq \mathcal{X}$ w.r.t. ρ is a finite subset $C = \{f_1, \cdots, f_m\} \subseteq \mathcal{X}$ s.t. $\mathcal{F} \subseteq \bigcup_{j=1}^{m} B_\varepsilon(f_j)$.

(ii) Define the ε-covering number of $\mathcal{F}$ w.r.t. ρ to be :

$$N(\varepsilon, \mathcal{F}, \rho) := \min\{ m : \exists \{f_1, \cdots, f_m\} \subseteq \mathcal{X},\ \mathcal{F} \subseteq \bigcup_{j=1}^{m} B_\varepsilon(f_j) \}$$

(iii) The metric entropy of $\mathcal{F}$ is $\log N(\varepsilon, \mathcal{F}, \rho)$

As $\varepsilon \downarrow$, $N(\varepsilon, \mathcal{F}, \rho) \uparrow$. What is the tradeoff ?

[Example 7] all functions

- Let $\mathcal{F} = \mathcal{X}$ be all functions from $\mathbb{R}$ to $[0,1]$
- [illegible]

- g - $L_2(P_n)$, on $z_1, \dots, z_n$.

- In order to cover F. fix any $f \in F$.

For each z_i, For a segmentation of $[0,1]$: $Y = \{2\varepsilon, 4\varepsilon, \dots, 1\}$.

For $f(z_i) \in [0,1]$, we can pick $g(z_i) \in Y$ s.t. $|f(z_i) - g(z_i)| \le \varepsilon$.

$g(z)$ for $z \neq z_i$ can be chosen arbitrarily. Averaging over all z_i,

we get $\rho(f,g) \le \varepsilon$. We just need to calculate the possible permutation of Y.

Furthermore. $|Y| = \frac{1}{2\varepsilon}$, so

$$N(\varepsilon, F, L_2(P_n)) \le \left(\frac{1}{2\varepsilon}\right)^n$$

the metric entropy is $O(n \log(1/\varepsilon))$, which is too large.

To see this, by Massart's finite lemma, $\hat{R}_n(F) \sim O\left(\sqrt{\frac{n\log(1/\varepsilon)}{n}}\right) = O(1/\varepsilon)$,

not going to zero.

[Example 8] non-decreasing function

- Let $F = \{f : \mathbb{R} \to [0,1], f \text{ is non-decreasing}\}$

- Let $z_1, \dots, z_n$ be n fixed points (in an increasing order)

- $Y = \{\varepsilon, 2\varepsilon, \dots, 1\}$. Fix any function $f \in F$. For each $y \in Y$, consider z_i for which $f(z_i) \in [y-\varepsilon, y]$. Set $g(z_i) = y$ for these points. Note: g is non-decreasing across $z_1, \dots, z_n$ and g satisfies $\rho(f,g) \le \varepsilon$.

- Count the number of possible g. Key observation: each g is non-decreasing, we can associate each level $y \in Y$ with leftmost point z_i for $g(z_i) = y$; the choice of leftmost points for each level unique defines g. Thus:

$$N(\varepsilon, F, L_2(P_n)) = O(n^{1/\varepsilon})$$

and the metric entropy is $O(\frac{1}{\varepsilon}\log n)$, better than example 7.

[Theorem 14] Simple discretization

Let F be a family of functions mapping $Z \to [-1,1]$

$\hat{R}_n(F)$ is bounded by :

$$\hat{R}_n(F) \le \inf_{\varepsilon > 0} \left(\sqrt{\frac{2\log N(\varepsilon, F, L_2(P_n))}{n}} + \varepsilon \right)$$

Preparation :

(i) We will also assume $Z_{1:n}$ are constant. and write $E[A]$ instead of $E[A|Z_{1:n}]$

(ii) To simplify notation, we write $\|f\|$ for $\|f\|_{L_2(P_n)}$ and $\langle f,g\rangle$ for $\langle f, g\rangle_{L_2(P_n)}$

(iii) Overload notation, let $\sigma : Z \to \{-1,+1\}$ be defined as a function : $\sigma_i = \sigma(z_i)$ and we can write

$$\hat{R}_n(F) = E[\sup_{f\in F} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i)]$$
$$= E[\sup_{f\in F} \langle \sigma, f\rangle]$$

(iv) Note that $\|\sigma\| = 1$

(v) think about $f \in F$ as $\frac{1}{\sqrt{n}}[f(z_1), \cdots, f(z_n)]$

σ as $\frac{1}{\sqrt{n}}[\sigma_1, \cdots, \sigma_n]$

Pf of Theorem 14:

Fix $\varepsilon > 0$ and let C be an ε-cover of F, then

$$\hat{R}_n(F) = E[\sup_{f\in F} \langle \sigma, f\rangle]$$
$$= E[\sup_{g\in C} \sup_{f\in F\cap B_\varepsilon(g)} \langle \sigma, g\rangle + \langle \sigma, f-g\rangle]$$
$$\le E[\sup_{g\in C} \frac{1}{n}\langle \sigma, g\rangle + \varepsilon]$$

$\langle \sigma, f-g\rangle \le \|\sigma\|_1 \|f-g\|_2 \le \varepsilon$

$$= \hat{R}_n(C) + \varepsilon$$
$$\le \sqrt{\frac{2\log N(\varepsilon, F, L_2(P_n))}{n}} + \varepsilon$$

Massart's finite lemma □

[Example 9] non-decreasing functions (with simple discretization)

Let F be all non-decreasing functions from $Z = \mathbb{R}$ to $[0,1]$.

Plugging the covering number of F into Theorem 14:

$$\hat{R}_n(F) \le \inf_{\varepsilon>0}\left(\sqrt{\frac{2\cdot O(\frac{\log n}{\varepsilon})}{n}} + \varepsilon\right)$$

Solving for minimal:

$$\hat{R}_n(F) = O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{3}}\right)$$

[Theorem 15] chaining (Dudley's theorem)

Let F be a family of functions mapping Z to $\mathbb{R}$.

The empirical Rademacher complexity can be upper bounded by:

$$\hat{R}_n(F) \le 12\int_0^\infty \sqrt{\frac{\log N(\varepsilon, F, L_2(P_n))}{n}}\, d\varepsilon.$$

Pf: Let $\varepsilon_0 = \sup_{f\in F}\|f\|$ be the maximum norm of a function $f \in F$, which will serve the coarsest resolution.

Let $\varepsilon_j = 2^{-j}\varepsilon_0$ for $j = 1, \cdots, m$ be successively finer resolutions.

For each $j = 0, \cdots, m$, let C_j be an ε_j-cover of F

Fix any $f \in F$

Let $g_j \in C_j$ be s.t. $\|f - g_j\| \le \varepsilon_j$; take $g_0 = 0$.

Let's decompose f :

$$f = f - g_m + \underbrace{g_0}_{=0} + \sum_{j=1}^{m}(g_j - g_{j-1})$$

By Massart's finite lemma:

$$\hat{R}_n(B) \le \left(\sup_{b\in B}\|b\|\right)\sqrt{\frac{2\log|B|}{n}} \qquad (1)$$

Let's bound some norms:

$$\|f - g_m\| \le \varepsilon_m$$

$$\|g_j - g_{j-1}\| \le \|g_j - f\| + \|f - g_{j-1}\| \le \varepsilon_j + \varepsilon_{j-1} = 3\varepsilon_j$$

Now compute $\hat{R}_n(F)$:

Now compute $\hat{R}_n(F)$:

$$\begin{aligned}
\hat{R}_n(F) &= E\Big[\sup_{f\in F}\langle\sigma, f\rangle\Big] \\
&= E\Big[\sup_{f\in F}\langle\sigma, f-g_m\rangle + \sum_{j=1}^{m}\langle\sigma, g_j-g_{j-1}\rangle\Big] \\
&\le \varepsilon_m + E\Big[\sup_{f\in F}\sum_{j=1}^{m}\langle\sigma, g_j-g_{j-1}\rangle\Big] \\
&\le \varepsilon_m + \sum_{j=1}^{m} E\Big[\sup_{f\in F}\langle\sigma, g_j-g_{j-1}\rangle\Big] \\
&\le \varepsilon_m + \sum_{j=1}^{m} E\Big[\sup_{g_j\in C_j,\, g_{j-1}\in C_{j-1}}\langle\sigma, g_j-g_{j-1}\rangle\Big] \\
&\le \varepsilon_m + \sum_{j=1}^{m} 3\varepsilon_j\sqrt{\frac{2\log|C_j||C_{j-1}|}{n}} \qquad \text{(by (1))} \\
&\le \varepsilon_m + \sum_{j=1}^{m} 6\varepsilon_j\sqrt{\frac{\log|C_j|}{n}} \qquad \cdots \text{ since } |C_j|\ge|C_{j-1}| \\
&= \varepsilon_m + \sum_{j=1}^{m} 12(\varepsilon_j-\varepsilon_{j+1})\sqrt{\frac{\log|C_j|}{n}} \qquad \cdots \text{ since } \varepsilon_j = 2(\varepsilon_j-\varepsilon_{j+1}) \\
&\le 12\int_0^{\varepsilon_0}\sqrt{\frac{\log N(\varepsilon, F, L_2(P_n))}{n}}\,d\varepsilon
\end{aligned}$$

$j\uparrow \Rightarrow C_j\uparrow$

$m\to\infty$, $\varepsilon_m\to 0$. □

[Example 10] non-decreasing functions (with chaining)

Let F be all non-decreasing functions from $\mathbb{Z}$ to $[0,1]$.

Note that $\|f\|\le 1$ for all $f\in F$, so the coarest resolution is $\varepsilon_0 = 1$, then by Th. 15:

$$\begin{aligned}
\hat{R}_n(F) &\le 12\int_0^1\Big(\sqrt{\frac{O(\frac{\log n}{\varepsilon})}{n}}\Big)\,d\varepsilon \\
&= O\Big(\sqrt{\frac{\log n}{n}}\Big)\int_0^1\sqrt{\frac{1}{\varepsilon}}\,d\varepsilon \\
&= O\Big(\sqrt{\frac{\log n}{n}}\Big)
\end{aligned}$$

Remark: (i) compare with $O\big((\frac{\log n}{n})^{\frac{1}{3}}\big)$ in example 9, $O\big(\sqrt{\frac{\log n}{n}}\big)$ is better.

(ii) Chaining is better than simple discretization because Chaining use Massart's finite lemma on $C_j - C_{j-1}$, bounded by $3\varepsilon_j$, while SD use MFL on C, bounded by $\sup\|f\|$.