6 – Chaining

Recap: – Massart's Finite Lemma (MFL):

$$\sup_{\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(z_i)^2 \le k^2 \Rightarrow \hat{R}(\mathcal{F}) \le k\sqrt{\frac{2\log|\mathcal{F}|}{n}}$$

– Massart's "Infinite" Lemma:

$$\sup_{\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(z_i)^2 \le k^2 \Rightarrow \hat{R}(\mathcal{F}) \le k\sqrt{\frac{2\log s(\mathcal{F},n)}{n}}$$

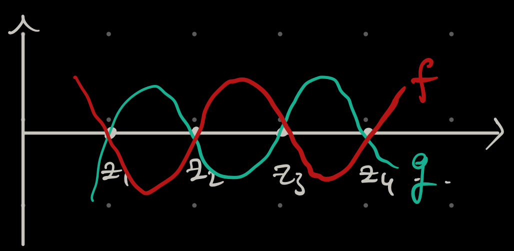
This lecture: 1– Revisit covering (discretization)

2– Want to bound RC instead of excess risk or empirical process.

3– Covering is over $\mathcal{F}$ directly, which requires a distance.

Metric: $f, g \in \mathcal{F}$, $\quad d(f,g) = \left(\frac{1}{n}\sum_{i=1}^{n}(f(z_i)-g(z_i))^2\right)^{1/2}$

where z_i's are data points.



– $d(f,g)=0$

– f and g are treated as the same fnc, because they behave the same over data.

Notation: Define $f = \frac{1}{\sqrt{n}}[f(z_1)\ f(z_2)\ \cdots\ f(z_n)]^T \in \mathbb{R}^n$

$$\Rightarrow \|f\|_2^2 = \frac{1}{n}\sum_{i=1}^{n} f(z_i)^2, \qquad d(f,g) = \|f-g\| \qquad (\|\cdot\|_2 = \|\cdot\|)$$

f is vector definition of the fnc f.

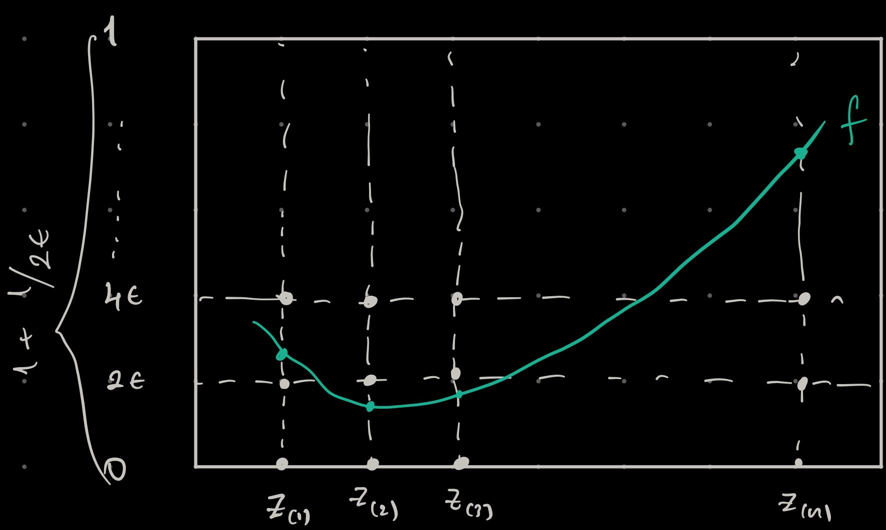
Revisit: Massart's Finite Lemma: $\hat{R}(\mathcal{F}) \le \sup_{\mathcal{F}} \|f\| \sqrt{\frac{2\log|\mathcal{F}|}{n}}$

Recall: An ε-cover of $\mathcal{F}$ wrt a metric d is a set $\mathcal{N}_\epsilon = \{g_1 \dots g_m\}$ s.t. $\forall f \in \mathcal{F}$, $\exists g \in \mathcal{N}_\epsilon$ s.t. $d(f,g) \le \epsilon$.

<u>Covering number of $\mathcal{F}$</u>: $N(\epsilon, \mathcal{F}, d) = \min\{|\mathcal{N}_\epsilon| : \mathcal{N}_\epsilon \text{ is a } \varepsilon\text{-cover of } \mathcal{F}\}$

<u>Metric entropy of $\mathcal{F}$</u> $\triangleq \log N(\epsilon, \mathcal{F}, d)$

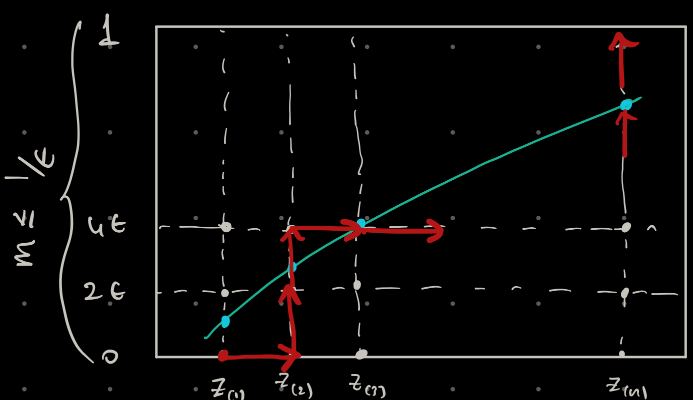
Ex: (all fncs) $\mathcal{F} = \{f : \mathcal{Z} \to [0,1]\}$



- One pnt at every vertical line.
- $|\mathcal{N}_\epsilon| = \left(1 + \frac{1}{2\epsilon}\right)^n$
  $\le \left(\frac{1}{\epsilon}\right)^n$

$\Rightarrow N(\epsilon, \mathcal{F}, d) \le \left(\frac{1}{\epsilon}\right)^n$
(exponential in n).

Ex: (non-decreasing fncs): $\mathcal{F} = \{f : \mathcal{Z} \to [0,1] \text{ non-decreasing}\}$



- Only include $g \in \mathcal{N}_\epsilon$ that are non-decreasing.
- $\forall g \in \mathcal{N}_\epsilon$, we need $n-1$ $\rightarrow$ and m $\uparrow$

- $k_i = \#\uparrow$ at level $z_{(i)}$: $k_1 + k_2 + \dots + k_n = m$.
- $\#$ non-decreasing $g = \left| \left\{ (k_1 \dots k_n) : \sum_{i=1}^{n} k_i = m,\ k_i \in \mathbb{N} \right\} \right|$

$$= \sum_{k_1+k_2 \dots k_n = m} \binom{m}{k_1 k_2 \dots k_n} = n^m \leq n^{1/\epsilon}$$

$$\Rightarrow N(\epsilon, \mathcal{F}, d) \leq n^{1/\epsilon} \quad (\text{poly in } n).$$

Theorem (Discretization): Let $\mathcal{F} = \{f\}$ s.t. $f : \mathcal{Z} \to \mathbb{R}$.
Then for $K = \sup_{\mathcal{F}} \|f\|$, we have

$$\hat{\mathcal{R}}(\mathcal{F}) \leq \underbrace{K\sqrt{\frac{2\log N(\epsilon, \mathcal{F}, d)}{n}}}_{\mathrm{I}} + \underbrace{\epsilon}_{\mathrm{II}}, \quad \forall \epsilon > 0.$$

Remark: I is due to MFL. $\mathrm{I} \uparrow$ w/ $\epsilon \downarrow$.
II is due to discretization $\mathrm{II} \downarrow$ w/ $\epsilon \downarrow$.
$\Rightarrow$ trade-off $\Rightarrow$ we need optimize over ϵ.

proof: Let $f = \frac{1}{\sqrt{n}} [f(z_1) \dots f(z_n)]^T$ and $\sigma = \frac{1}{\sqrt{n}} [\sigma_1 \dots \sigma_n]^T$.
$\|\sigma\| = 1$.

- $\hat{\mathcal{R}}(\mathcal{F}) = \mathbb{E}\left[\sup_{\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \,\middle|\, Z_{1:n} \right] = \mathbb{E}\left[\sup_{\mathcal{F}} \langle \sigma, f \rangle \,\middle|\, Z_{1:n} \right]$

- Let $\mathcal{N}_\epsilon$ be an ϵ-cover of $\mathcal{F}$

$\Rightarrow \forall f \in \mathcal{F},\ \exists g \in \mathcal{N}_\epsilon$ s.t. $\|f - g\| \leq \epsilon$.

$$- \sup_{f\in\mathcal{F}} \langle \sigma, f\rangle = \langle \sigma, g\rangle + \langle \sigma, f-g\rangle$$

$$\le \langle \sigma, g\rangle + \underbrace{\|\sigma\|}_{=1}\cdot\underbrace{\|f-g\|}_{\lesssim \epsilon} \quad \text{(by CS)}$$

$$\le \max_{g\in\mathcal{N}_\epsilon} \langle \sigma, g\rangle + \epsilon$$

$$\Rightarrow \hat{R}(\mathcal{F}) \le \mathbb{E}\Big[\max_{g\in\mathcal{N}_\epsilon} \langle \sigma, g\rangle \,\Big|\, z_{1:n}\Big] + \epsilon = \hat{R}(\mathcal{N}_\epsilon) + \epsilon$$

$$\text{(by MFL)} \quad \le \sup_{\mathcal{N}_\epsilon} \|g\| \sqrt{\frac{2\log|\mathcal{N}_\epsilon|}{n}} + \epsilon \qquad \forall\, \mathcal{N}_\epsilon$$

$$\text{(minimize over } \mathcal{N}_\epsilon) \quad = \sup_{\mathcal{F}} \|f\| \sqrt{\frac{2\log N(\epsilon,\mathcal{F},d)}{n}} + \epsilon . \qquad \boxtimes$$

Ex (all fncs): $N(\epsilon,\mathcal{F},d) \le (1/\epsilon)^n$

$\Rightarrow$ by the previous thm $\hat{R}(\mathcal{F}) \lesssim \sqrt{\frac{\cancel{n}\cdot \log 1/\epsilon}{\cancel{n}}} + \epsilon$

(no generalization!)

Ex (non-decreasing fncs): $N(\epsilon,\mathcal{F},d) \le n^{1/\epsilon}$

$\Rightarrow$ by the previous thm $\hat{R}(\mathcal{F}) \lesssim \sqrt{\frac{\log n}{\epsilon n}} + \epsilon$

$\Rightarrow$ Optimize over ϵ: $\epsilon = \sqrt{\frac{\log n}{\epsilon n}} \Rightarrow \epsilon = \left(\frac{\log n}{n}\right)^{\frac{1}{3}}$

$\Rightarrow$ slow rate!

$\Rightarrow$ generalization

Theorem (Dudley's thm): $\mathcal{F}$ is a family of fncs $f: \mathcal{Z} \to \mathbb{R}$

$$\hat{R}(\mathcal{F}) \le 12 \int_0^\infty \sqrt{\frac{\log N(\epsilon, \mathcal{F}, d)}{n}}\, d\epsilon .$$

Remarks: 1- When ϵ is large, $N(\epsilon, \mathcal{F}, d) = 1$. Not really integrating to ∞ for a compact $\mathcal{F}$.

2- Discretization error ϵ is gone!

3- For the above non-decreasing fncs

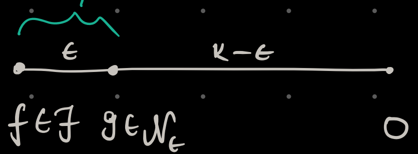
$$\hat{R}(\mathcal{F}) \le 12 \int_0^1 \sqrt{\frac{\log n}{\epsilon n}}\, d\epsilon = 12 \sqrt{\frac{\log n}{n}} \int_0^1 \frac{1}{\sqrt{\epsilon}}\, d\epsilon$$

$$= 24 \sqrt{\frac{\log n}{n}} \quad \Rightarrow \text{ faster rate!}$$

proof (by chaining):

discretization error

Previously,

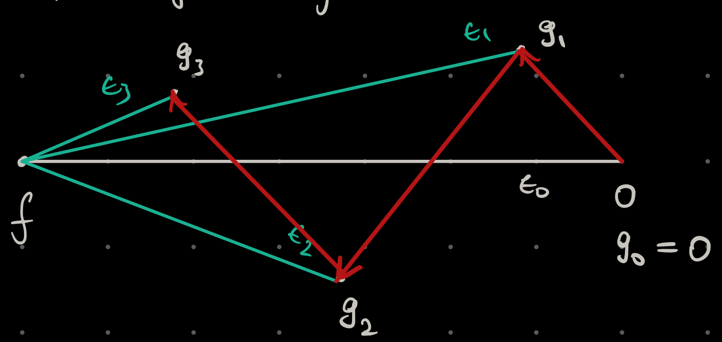


$$\Rightarrow K \sqrt{\frac{2 \log |\mathcal{N}_\epsilon|}{n}} + \epsilon$$

MFL

Now, choose a chain of ϵ-Nets: $\epsilon_j = 2^{-j} \epsilon_0 \qquad \epsilon_0 = \sup_{\mathcal{F}} \|f\|$

f, $\exists g_j \in \mathcal{N}_{\epsilon_j}$



- MFL is applied to $g_{j+1} - g_j$
- $\epsilon_j \downarrow 0$, so does $\|g_{j+1} - g_j\|$

- Let $\epsilon_0 = \sup_{\mathcal{F}} \|f\|$ and $\epsilon_j = 2^{-j}\epsilon_0$, $\mathcal{N}_{\epsilon_j}$ are ϵ_j-covers.

- $\forall f \in \mathcal{F}$, find $g_j \in \mathcal{N}_{\epsilon_j}$ for $j = 1 \dots$ s.t. $\|f - g_j\| \le \epsilon_j$.
  and $g_0 = 0$.

- $f = \underbrace{f - g_m}_{\text{vanishing as } m \to \infty} + \underbrace{\sum_{j=1}^{m} g_j - g_{j-1}}_{\text{forms a chain}}$ (by triangle inequality $\|g_j - g_{j-1}\| \le \epsilon_j + \epsilon_{j-1} = 3\epsilon_j$)

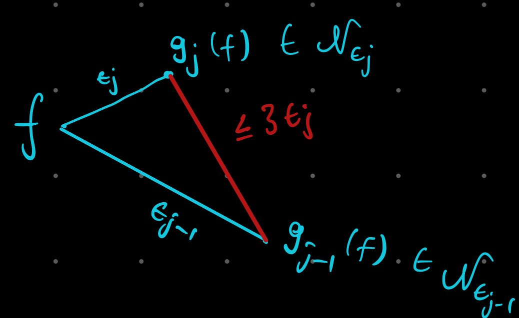
- Recall MFL: $\hat{\mathcal{R}}(\mathcal{G}) \le \sup_{\mathcal{G}} \|g\| \cdot \sqrt{\frac{2\log|\mathcal{G}|}{n}}$

- $\hat{\mathcal{R}}(\mathcal{F}) = \mathbb{E}\left[\sup_{\mathcal{F}} \langle \sigma, f\rangle \mid z_{1:n}\right]$

$$= \mathbb{E}\left[\sup_{\mathcal{F}} \left\{ \underbrace{\langle \sigma, f - g_m\rangle}_{\text{by CS } \le \underbrace{\|\sigma\|}_{=1} \underbrace{\|f - g_m\|}_{\le \epsilon_m}} + \left\langle \sigma, \sum_{j=1}^{m} g_j - g_{j-1} \right\rangle \right\} \Bigm| z_{1:n}\right]$$

$$\le \epsilon_m + \mathbb{E}\left[\sup_{\mathcal{F}} \left\langle \sigma, \sum_{j=1}^{m} g_j - g_{j-1}\right\rangle \Bigm| z_{1:n}\right]$$

$(\sup \sum \le \sum \sup)$ $$\le \epsilon_m + \sum_{j=1}^{m} \mathbb{E}\left[\sup_{\mathcal{F}} \langle \sigma, g_j - g_{j-1}\rangle \mid z_{1:n}\right]$$



$$\leq \epsilon_m + \sum_{j=1}^{m} \mathbb{E}\left[\sup \langle \sigma, g_j - g_{j-1} \rangle \mid z_{1:n} \right]$$

$$\mathcal{H}_j \triangleq \left\{ \begin{array}{l} g_j \in \mathcal{N}_{\epsilon_j},\ g_{j-1} \in \mathcal{N}_{\epsilon_{j-1}} \\ \|g_j - g_{j-1}\| \leq 3\epsilon_j \end{array} \right\}$$

$$\boxed{\mathcal{H}_j = \left\{ h = g_j - g_{j-1} : g_j \in \mathcal{N}_{\epsilon_j},\ g_{j-1} \in \mathcal{N}_{\epsilon_{j-1}},\ \|g_j - g_{j-1}\| \leq 3\epsilon_j \right\}}$$

$$\leq \epsilon_m + \sum_{i=1}^{m} \underbrace{\mathbb{E}\left[\sup_{\mathcal{H}_j} \langle \sigma, h \rangle \mid z_{1:n} \right]}$$

$$= \hat{\mathcal{R}}(\mathcal{H}_j) \leq \sup_{\mathcal{H}_j} \|h\| \cdot \sqrt{\frac{2 \log |\mathcal{H}_j|}{n}}$$

$$|\mathcal{H}_j| \leq |\mathcal{N}_{\epsilon_j}| \cdot |\mathcal{N}_{\epsilon_{j-1}}|$$

$$\leq |\mathcal{N}_{\epsilon_j}|^2$$

$$(\text{by MFL}) \leq \epsilon_m + \sum_{j=1}^{m} \underbrace{\left(\sup_{\mathcal{H}_j} \|h\| \right)}_{\leq 3\epsilon_j = 6(\epsilon_j - \epsilon_{j+1})} \cdot \sqrt{\frac{2 \log |\mathcal{N}_{\epsilon_j}|^2}{n}}$$

$$\leq \epsilon_m + 12 \sum_{j=1}^{m} (\epsilon_j - \epsilon_{j+1}) \cdot \sqrt{\frac{\log |\mathcal{N}_{\epsilon_j}|}{n}}$$

$$= \epsilon_m + 12 \sum_{j=1}^{m} \int_{\epsilon_{j+1}}^{\epsilon_j} \sqrt{\frac{\log |\mathcal{N}_{\epsilon_j}|}{n}} \cdot d\epsilon$$

$$\leq \epsilon_m + 12 \sum_{j=1}^{m} \int_{\epsilon_{j+1}}^{\epsilon_j} \sqrt{\frac{\log |\mathcal{N}_{\epsilon}|}{n}} \cdot d\epsilon$$

$$\leq \epsilon_m + 12 \int_{\epsilon_{m+1}}^{\epsilon_0} \sqrt{\frac{\log |\mathcal{N}_{\epsilon}|}{n}} \, d\epsilon \qquad \forall \mathcal{N}_\epsilon$$

$$\left(\begin{array}{c}\text{optimize} \\ \text{over} \\ \mathcal{N}_\epsilon\end{array}\right) = \epsilon_m + 12 \int_{\epsilon_{m+1}}^{\infty} \sqrt{\frac{\log N(\epsilon, \mathcal{F}, d)}{n}} \cdot d\epsilon \qquad \forall m$$

let $m \to \infty$. ∎