STA 3000 F Lecture 8

VC-dim is useful for binary functions ($\mathcal{X} \to \{0,1\}$).

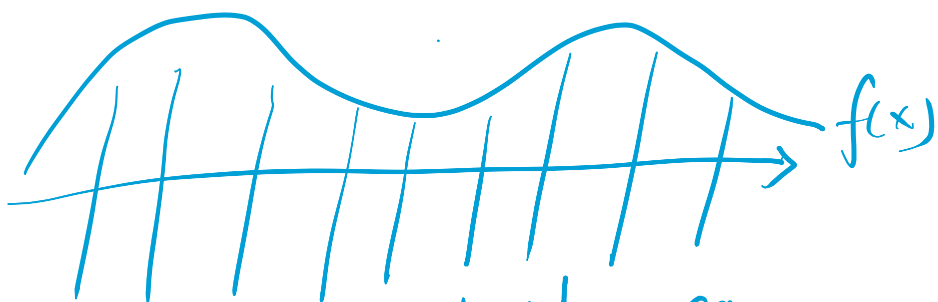
How about real-valued functions.

Def. VC subgraph dimension.

$$\mathcal{F} \subseteq (\mathcal{X} \to \mathbb{R})$$

Convert to binary class

$$\left\{ (t,x) \mapsto \mathbb{1}\{t \le f(x)\} \right\}$$ is a binary class



Indicator of shaded area.

$VC(\mathcal{F}) :=$ VC-dim of the augmented subgraph function class

e.g. $\mathcal{F} := \{x \mapsto \theta^T x : \theta \in \mathbb{R}^d\}$.

$$t \ge f_\theta(x) = \theta^T x \iff \begin{bmatrix} -t \\ x \end{bmatrix} \cdot \begin{bmatrix} -1 \\ \theta \end{bmatrix} \le 0.$$

$\mathrm{VC}(\mathcal{F}) \le d+1.$

e.g. $\mathcal{F} := \{ x \mapsto \varphi(\theta^T x) : \theta \in \mathbb{R}^d \}$

for cts and strictly increasing φ.

$$t \ge f_\theta(x) \iff \begin{bmatrix} \varphi^{-1}(t) \\ x \end{bmatrix} \begin{bmatrix} 1 \\ -\theta \end{bmatrix} \ge 0.$$

$\mathrm{VC}(\mathcal{F}) \le d+1.$

Thm.

$$\sup_Q N\left(\varepsilon \|F\|_{L^2(Q)}; \mathcal{F}, L^2(Q)\right) \le \left(\frac{c_1}{\varepsilon}\right)^{c_2 \cdot \mathrm{VC}(\mathcal{F})}$$

where c_1, c_2 are universal constants.

F is an envelop function of class $\mathcal{F}$.

Remark: this is exactly the covering/packing # that appeared in main empirical process bound.

Recall:

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i}^{n} f(X_i) - \mathbb{E} f(X)\right|\right] \le C \cdot \sqrt{\frac{\mathbb{E}[F(X)^2]}{n}} \int_0^1 \sqrt{\log \sup_Q N(\varepsilon \|F\|_{L^2(Q)}; \mathcal{F}, L^2(Q))}\, d\varepsilon$$

$$\leq C \cdot \sqrt{\frac{\mathbb{E}[F(X)^2]}{n}} \cdot \int_0^1 \sqrt{C_2 VC(\mathcal{F}) \cdot \log\left(\frac{C_1}{\varepsilon}\right)}\, d\varepsilon$$

$$\leq C' \cdot \|F\|_{L^2(\mathbb{P})} \cdot \sqrt{\frac{VC(\mathcal{F})}{n}}.$$

Proof of "VC subgraph $\to$ covering/packing":

Idea: convert covering/packing of subgraph class to that of $\mathcal{F}$ itself.

For $f, g \in \mathcal{F}$

$$\|f-g\|_{L^2(Q)}^2 = \int_{\mathcal{X}} |f(x) - g(x)|^2 \, dQ(x)$$

$$\leq 2 \int_{\mathcal{X}} F(x) \cdot |f(x) - g(x)| \, dQ(x).$$

$$|f(x) - g(x)| = \int_{-F(x)}^{F(x)} \left| \mathbb{1}\{t \leq f(x)\} - \mathbb{1}\{t \leq g(x)\} \right| dt$$

So we have

$$\|f-g\|_{L^2(Q)}^2 \leq 2 \iint_{(x,t):|t| \leq F(x)} F(x) \cdot \left| \mathbb{1}\{t \leq f(x)\} - \mathbb{1}\{t \leq g(x)\} \right| dt \, dQ(x)$$

$$\leq 2 \cdot \sqrt{\iint_{|t| \leq F(x)} \left(\sqrt{F(x)}\right)^2 dt\, dQ(x)}$$

$$\cdot \sqrt{\iint_{|t| \leq F(x)} \left(\sqrt{F(x)}\right)^2 \left(\mathbb{1}_{\{t \leq f(x)\}} - \mathbb{1}_{\{t \leq g(x)\}}\right)^2 dt\, dQ(x)}$$

$$= 2\sqrt{2} \cdot \|F\|_{L^2(Q)} \cdot \sqrt{\iint_{|t| \leq F(x)} \left(\mathbb{1}_{\{t \leq f(x)\}} - \mathbb{1}_{\{t \leq g(x)\}}\right)^2 \cdot F(x)\, dt\, dQ(x)}$$

Define $d\tilde{Q}(x,t) \propto F(x)\, dt\, dQ(x)$.

on the domain $\{(t,x) : |t| \leq F(x)\}$.

Note that

$$\iint_{|t| \leq F(x)} F(x)\, dt\, dQ(x) = 2 \cdot \|F\|_{L^2(Q)}^2$$

So $$d\tilde{Q}(x,t) = \frac{F(x)\, dt\, dQ(x)}{2\|F\|_{L^2(Q)}^2}.$$

Putting them together,

$$\|f - g\|_{L^2(Q)}^2 \leq 4 \cdot \|F\|_{L^2(Q)}^2 \sqrt{\iint \left(\mathbb{1}_{\{t \leq f(x)\}} - \mathbb{1}_{\{t \leq g(x)\}}\right)^2 d\tilde{Q}}$$

So we have

$$N\left(2\|F\|_{L^2(Q)}\sqrt{\varepsilon};\ \mathcal{F}, L^2(Q)\right)$$

$$\leq N\left(\varepsilon;\ \{\mathbb{1}_{\{t\leq f(x)\}} : f\in\mathcal{F}\}, L^2(\tilde{Q})\right).$$

$$\leq \left(\frac{c}{\varepsilon}\right)^{c' VC(\mathcal{F})}$$

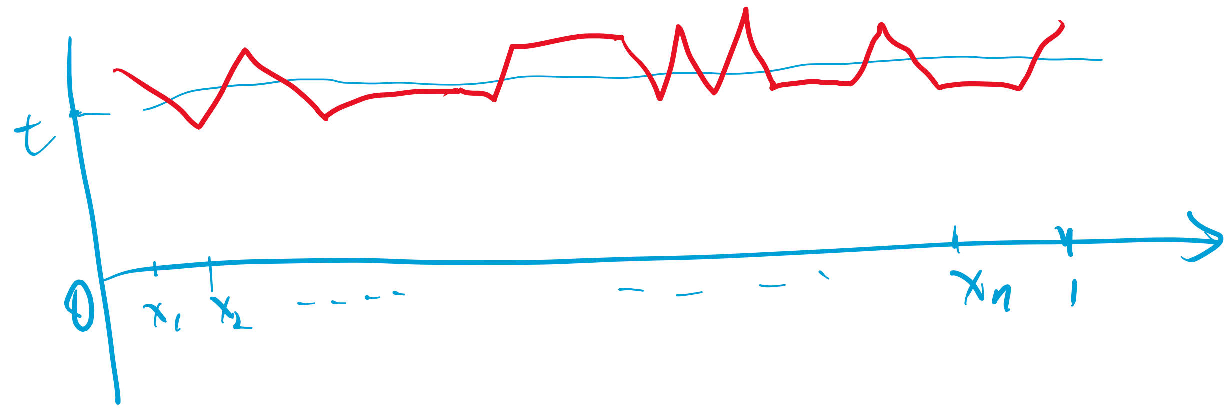
You can take an $L^2(\tilde{Q})$ covering of $\{\mathbb{1}_{\{t\leq f(x)\}} : f\in\mathcal{F}\}$, and map it to $L^2(Q)$ covering of $\mathcal{F}$.

How about infinite VC dim?

Motivating e.g.

$$\mathcal{F} = \left\{ f: [0,1]\to[0,1] \,\middle|\, |f(x)-f(y)| \leq |x-y|,\ \forall x,y \right\}$$

Any possible label sequence can be shattered.



$t_1 = t_2 \cdots = t_n = t.$

Any binary seq can be realized

$$VC(\mathcal{F}) = +\infty.$$

Solution: fat-shattering dim.

Def (ε-fat-shattering dim)

$fat_\varepsilon(\mathcal{F})$ is largest D s.t.

$\exists\ (x_i, t_i)_{i=1}^{D}$ that is ε-shattered by $\mathcal{F}$.

"ε-shattered": $\forall$ binary seq $b_1, b_2, \ldots, b_D$

$$\exists f \in \mathcal{F}.\ \text{s.t.}\quad f(x_i) \begin{cases} < t_i, & b_i = 0 \\ \geq t_i + \varepsilon, & b_i = 1 \end{cases}$$

"Scale-sensitive combinatorial dim".

Why useful:

Thm (Mendelson-Vershynin)

If $\mathcal{F}$ is unif bdd by 1,

$$\sup_{Q} M(\varepsilon; \mathcal{F}, L^2(Q)) \le \left(\frac{1}{\varepsilon}\right)^{c_1 \mathrm{fat}_{c_2 \varepsilon}(\mathcal{F})}.$$

So that by Dudley chaining integral, we get

$$\sup_{f \in \mathcal{F}} |(P_n - P) f| \le \frac{C}{\sqrt{n}} \int_0^1 \sqrt{\mathrm{fat}_{c_2 \varepsilon}(\mathcal{F}) \cdot \log\left(\frac{1}{\varepsilon}\right)}\, d\varepsilon.$$

(Improved version by Rudelson-Vershynin)

under some mild assumptions

$$\sup_{Q} M(\varepsilon; \mathcal{F}, L^2(Q)) \le \exp\left(c_1 \mathrm{fat}_{c_2 \varepsilon}(\mathcal{F}) \right)$$

so that we have

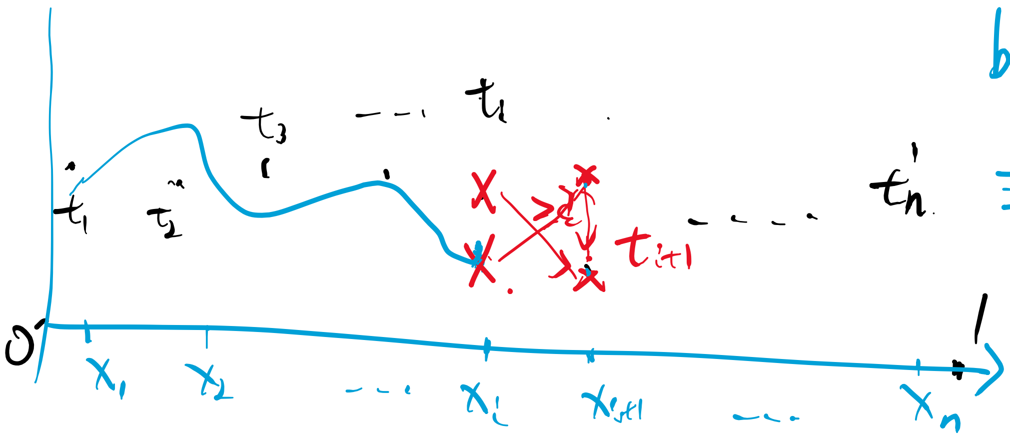
$$\sup_{f \in \mathcal{F}} |(P_n - P) f| \le \frac{C}{\sqrt{n}} \int_0^1 \sqrt{\mathrm{fat}_{c_2 \varepsilon}(\mathcal{F})}\, d\varepsilon.$$

Fat-shattering dim of Lip funcs (in 1D).

Suppose $(x_i, t_i)_{i=1}^{n}$ being ε - shattered.

(assume $x_1 < x_2 < \cdots < x_n$ w.l.o.g.).

$\forall$ binary seq $b_1, b_2, \ldots, b_n$ $\exists f_b \in \mathcal{F}$. s.t.



$$f_b(x_i) \begin{cases} < t_i & (b_i = 0) \\ \geq t_i + \varepsilon & (b_i = 1). \end{cases}$$

Fix. $b_1, b_2, \ldots, b_{i-1}$

Let f_+ be the function obtained by $b_i = 0$, $b_{i+1} = 1$

f_- — — — — $b_i = 1$, $b_{i+1} = 0$.

$$f_+(x_i) \geq t_i + \varepsilon, \qquad f_+(x_{i+1}) \leq t_{i+1}$$

$$f_-(x_i) \leq t_i, \qquad f_-(x_{i+1}) \geq t_{i+1} + \varepsilon$$

By Lip condition, $|f_+(x_i) - f_+(x_{i+1})| \leq |x_{i+1} - x_i|$

$$|f_-(x_i) - f_-(x_{i+1})| \leq |x_{i+1} - x_i|.$$

So we have

$$|x_{i+1} - x_i| \geq \max\{t_i + \varepsilon - t_{i+1},\ t_{i+1} + \varepsilon - t_i\}$$
$$= \varepsilon + |t_i - t_{i+1}| \geq \varepsilon.$$

So we have $n \leq \frac{1}{\varepsilon}$.

and therefore $\mathrm{fat}_\varepsilon(\mathcal{F}) \leq \frac{1}{\varepsilon}$.

Original problem: $L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l_\theta(X_i)$.

$$\hat{\theta}_n := \arg\min\{L_n(\theta)\} \qquad \theta^* = \arg\min\{L(\theta)\}.$$

$$L(\hat{\theta}_n) - L(\theta^*) \leq 2\sup_{\theta\in\Theta} |L_n(\theta) - L(\theta)|.$$

(e.g. $\leq C\sqrt{\frac{VC(\mathcal{F})}{n}}$).

sup over Θ can be too conservative.

$\hat{\theta}_n$ may be close to θ^*

If we know this a priori,

then we can probably take

sup over $\Theta \cap \{$neighbors of $\theta^*\}$.

Chicken-egg problem?

Thm (Localization).

Assume $L(\theta) - L(\theta^*) \geq \|\theta - \theta^*\|^2$.

Suppose that we can show

$$\mathbb{E}\left[\sup_{\substack{\theta \in \Theta \\ \|\theta - \theta^*\| \leq u}} \left|(P_n - P)\cdot(l_\theta - l_{\theta^*})\right|\right] \leq \phi_n(u).$$

satisfying $\phi_n(cx) \leq c^\alpha \phi_n(x)$

$(\forall\ c > 1, x > 0)$ for some $\alpha < 2$.

Then for smallest δ_n satisfying $\phi_n(\delta_n) \leq \delta_n^2$

$\forall \varepsilon > 0$, $\exists$ constant $C_\varepsilon > 0$ (depending only on ε)

s.t. $\|\hat{\theta}_n - \theta^*\| \leq C_\varepsilon \cdot \delta_n$ w.p. $1 - \varepsilon$.

δ_n : "critical radius".

Proof: $L(\hat{\theta}_n) - L(\theta^*) \geq \|\hat{\theta}_n - \theta^*\|^2.$

$$= L(\hat{\theta}_n) - L_n(\hat{\theta}_n) + \underbrace{L_n(\hat{\theta}_n) - L_n(\theta^*)}_{\leq 0} + L_n(\theta^*) - L(\theta^*)$$

$$\leq \left(L_n(\theta^*) - L_n(\hat{\theta}_n)\right) - \left(L(\theta^*) - L(\hat{\theta}_n)\right)$$

$$= (P_n - P)\left(\ell_{\theta^*} - \ell_{\hat{\theta}_n}\right).$$

Consider the tail prob

$$\mathbb{P}\left(\|\hat{\theta}_n - \theta^*\| \geq 2^M \delta_n\right)$$

$$= \sum_{j > M} \mathbb{P}\left(2^{j-1}\delta_n \leq \|\hat{\theta}_n - \theta^*\| \leq 2^j \delta_n\right).$$

Each term $\leq \mathbb{P}\left(\|\hat{\theta}_n - \theta^*\| \leq 2^j \delta_n, \left|(P_n - P)(\ell_{\theta^*} - \ell_{\hat{\theta}_n})\right| \geq 2^{2j-2}\delta_n^2\right)$

$$\leq \mathbb{P}\left(\sup_{\substack{\theta \in \Theta \\ \|\theta - \theta^*\| \leq 2^j \delta_n}} \left|(P_n - P)(\ell_{\theta^*} - \ell_\theta)\right| \geq 2^{2j-2}\delta_n^2\right)$$

$$\leq \frac{1}{2^{2j-2}\delta_n^2} \phi_n(2^j \delta_n).$$

$$\leq \frac{4\phi_n(\delta_n)}{\delta_n^2} \cdot 2^{(\alpha-2)j}.$$

Sum them up, we have

$$\mathbb{P}\left(\|\hat{\theta}_n - \theta^*\| \geq 2^M \delta_n\right)$$

$$\leq \underbrace{\frac{4\,\phi_n(\delta_n)}{\delta_n^2}}_{\leq 4} \cdot \frac{2^{-(\alpha-2)M}}{1-2^{\alpha-2}}$$

choose M large enough, make it $\leq \frac{\varepsilon}{4}$.

Remark: bad dependence on tail prob (ε)

usually sharp in terms of n

and complexity of $\mathcal{F}$.

e.g. "Regular": M-estimators $\Theta = \mathbb{R}^d$.

$$|l_{\theta_1}(x) - l_{\theta_2}(x)| \leq M(x) \cdot \|\theta_1 - \theta_2\|_2$$

Assume that $L(\theta) - L(\theta^*) \geq \|\theta - \theta^*\|_2^2$

then we can compute

$$\mathbb{E}\Big[\sup_{\substack{\theta \in \mathbb{R}^d \\ \|\theta-\theta^*\|_2 \leq u}} \left|(P_n - P)(l_\theta - l_{\theta^*})\right|\Big].$$

$\Big(\forall \theta$ s.t. $\|\theta - \theta^*\|_2 \leq u$, $|l_\theta(x) - l_{\theta^*}(x)| \leq u \cdot M(x)\Big)$.

$$\leq c \cdot \frac{u}{\sqrt{n}} \|M\|_{L^2(P)} \cdot$$

$$\cdot \int_0^1 \sqrt{\log \sup_Q N\left(u \cdot \delta \cdot \|M\|_{L^2(Q)} ; \mathcal{F}_u, \|\cdot\|_{L^2(Q)}\right)} \, d\delta .$$

$$\left(\mathcal{F}_u := \{ l_\theta - l_{\theta^*} : \|\theta - \theta^*\|_2 \leq u \} \right).$$

$$N\left(u \cdot \delta \cdot \|M\|_{L^2(Q)} ; \mathcal{F}_u, \|\cdot\|_{L^2(Q)}\right)$$

$$\left(\|\theta_1 - \theta_2\|_2 \leq \varepsilon, \Longrightarrow \|l_{\theta_1} - l_{\theta_2}\|_{L^2(Q)} \leq \varepsilon \cdot \|M\|_{L^2(Q)} \right).$$

$$\leq N\left(u\delta ; \{\theta : \|\theta - \theta^*\|_2 \leq u\}, \|\cdot\|_2\right)$$

$$\leq \left(\frac{c}{\delta}\right)^d .$$

and we can conclude

$$\int_0^1 \sqrt{\log \sup \cdots\cdots} \, d\delta \leq c' \sqrt{d}.$$

$$\phi_n(u) = c \sqrt{\frac{d}{n}} \cdot \|M\|_{L^2(P)} \cdot u.$$

Solving fixed pt $\quad \delta_n = c \sqrt{\frac{d}{n}} \cdot \|M\|_{L^2(P)}$.