STA 3000 Lecture 7

Quantity of interest

$$\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X)\right|.$$

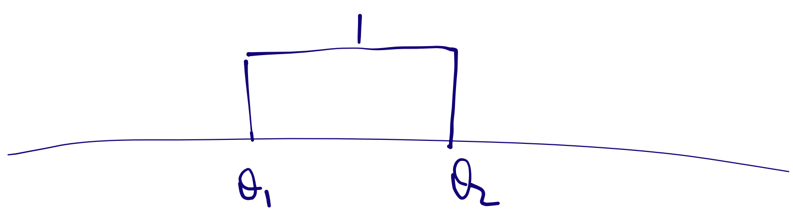
- Symmetrization (Rademacher complexity)
- Dudley chaining $\quad |f(x)| \le F(x) \quad \forall f$

$$\mathbb{E}[\sup \cdots] \le c\sqrt{\frac{\mathbb{E}[F(X)^2]}{n}} \cdot \int_0^1 \sqrt{\log \sup_{Q} N(\varepsilon\|F\|_{L^2(Q)}, \mathcal{F}, L^2(Q))}\, d\varepsilon$$

---

e.g. $\mathcal{F} := \{ 1_{\{x \in [\theta_1, \theta_2]\}} : \theta_1, \theta_2 \in \mathbb{R} \}$

covering in (θ_1, θ_2)



$\Downarrow$

covering in $\mathcal{F}$.

Vapnik – Chervonenkis (Binary functions $\mathcal{X} \to \{0,1\}$)

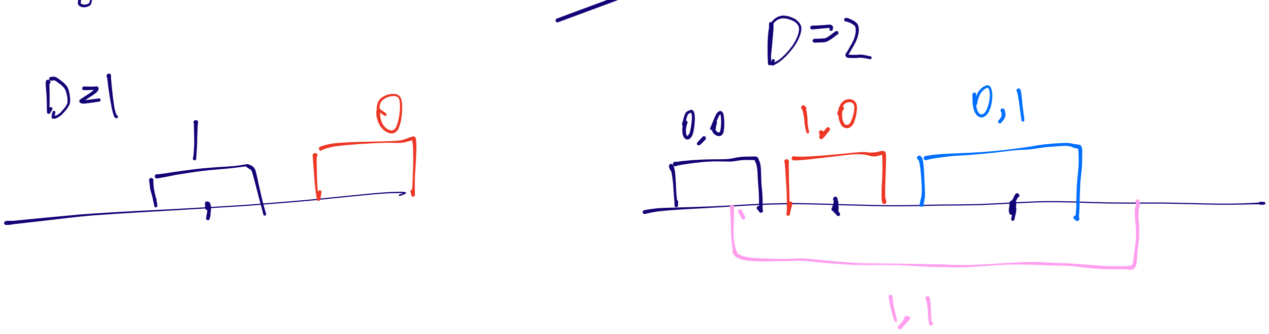
Def We say $(x_1, x_2 \cdots x_n)$ is shattered by $\mathcal{F}$, if

$$\left\{ \left(f(x_1), f(x_2), \dots, f(x_n)\right) \cancel{: f \in \mathcal{F}} \right\} = \{0,1\}^n.$$

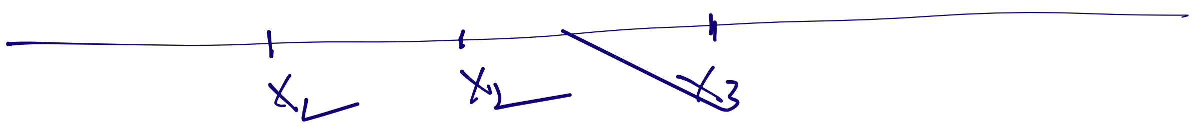
Def. $VC(\mathcal{F}) :=$ largest value of D s.t.

$\exists\, x_1, x_2, \dots, x_D \in \mathcal{X}$ shattered by $\mathcal{F}$.

e.g. indicator of line segment



$(1,0,1)$ is not achievable. $D=3$.

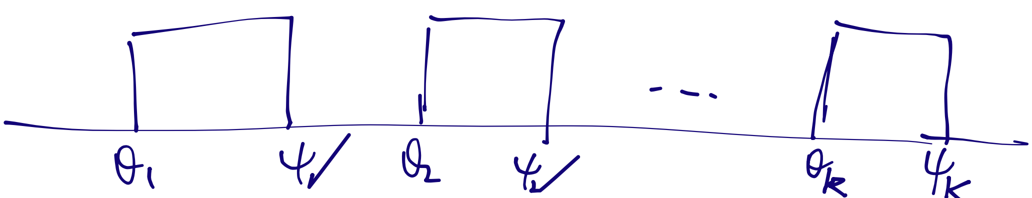


$VC(\mathcal{F}) = 2.$

---

e.g. $\mathcal{F} := \left\{ \mathbb{1}\left\{ x \in \bigcup_{i=1}^{K} [\theta_i, \psi_i] \right\} \right\}.$

2K label changes



2K+2 points

$x_1 \quad x_2, \quad - \quad \text{———} \quad x_{2K+2}.$

$1 \quad 0 \quad 1 \quad 0 \quad - \quad - \quad - \quad 1 \quad 0$

$$VC(\mathcal{F}) < 2K+2.$$

Sauer's lemma. If $\underline{VC(\mathcal{F}) = D}$ (for $n \geq D$).

Then $\left| \left\{ \left(f(x_1), \cdots, f(x_n)\right) : f \in \mathcal{F} \right\} \right| \leq \boxed{\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{D}}$

$\underbrace{\binom{n}{0} + \cdots + \binom{n}{D}}_{\leq \left(\frac{en}{D}\right)^D}$

$\Phi_n(D).$

Proof: Induction $n+D$.

$n=1$ trivial.

Suppose the conclusion holds for $(n-1, D-1)$ and $(n-1, D)$.

$\underbrace{\boxed{x_1, x_2, \cdots, x_{n-1}}}\,\Big|\, x_n$

Consider binary sequence $b_1 b_2 \cdots b_{n-1}$

- $b_1 \text{———} b_{n-1}$'s not achived by $\mathcal{F}(x_1, \cdots, x_{n-1})$. ignore it
- ① $b_1 \cdots b_{n-1} 0,\ b_1 \cdots b_{n-1} 1 \in \mathcal{F}(x_1, \cdots x_n).$
- ② Only one of them

Construct two sub classes $\mathcal{F}_1$, $\mathcal{F}_2$ as follows:

(1) $f_1, f_2 \in \mathcal{F}$ $\quad f_1(x_1 \cdots x_n) = b_1 b_2 \cdots b_{n-1} 0$

$\quad f_2(x_1, \cdots x_n) = b_1 \cdots b_{n-1} 1$

Add f_1 to $\mathcal{F}_1$

Add f_2 to $\mathcal{F}_2$.

(2) $f(x_1, \cdots x_n) = b_1 \cdots b_{n-1} *$

Add f to $\mathcal{F}_1$.

We regard $\mathcal{F}_1$ and $\mathcal{F}_2$ as function classes with domain $(x_1, x_2, \cdots x_n)$.

- $|\mathcal{F}(x_1, \cdots x_n)| = |\mathcal{F}_1| + |\mathcal{F}_2|$.
- $|\mathcal{F}_1| = |\mathcal{F}(x_1 \cdots x_{n-1})| \leq \Phi_{n-1}(D)$.
- $\forall f \in \mathcal{F}_2$, $\exists \tilde{f} \in \mathcal{F}_1$ s.t. f and $\tilde{f}$ disagree only on x_n.

Claim: $VC(\mathcal{F}_2) \leq D-1$.

$$\boxed{\begin{array}{l}\text{Suppose } (x_{k_1}, x_{k_2} \cdots x_{k_D}) \text{ shattered by } \mathcal{F}_2. \\ \underbrace{(x_{k_1}, x_{k_2}, \cdots x_{k_D}, x_n)}_{(D+1)} \text{ shattered by } \mathcal{F}_1 \\ \text{length.}\end{array}}$$

$$|\mathcal{F}_2| \le \Phi_{n-1}(D-1).$$

So $$|\mathcal{F}(x_1, \cdots x_n)| \le \Phi_n(D-1) + \Phi_{n-1}(D-1).$$

---

eg. $VC(\mathcal{F}_1) \le D_1$, $VC(\mathcal{F}_2) \le D_2$.

$$\mathcal{F} = \{\max(f_1, f_2) : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}.$$

Given $(x_1, x_2, \cdots x_n)$.

$$|\mathcal{F}(x_1, x_2, \cdots x_n)| \le |\mathcal{F}_1(x_1, \cdots x_n)| \cdot |\mathcal{F}_2(x_1, \cdots, x_n)|.$$

$$\le \left(\frac{en}{D_1}\right)^{D_1} \cdot \left(\frac{en}{D_2}\right)^{D_2}$$

$$\le (e\, n)^{D_1 + D_2}$$

$D = \underline{VC(\mathcal{F})}$ $$2^D \le (e \cdot D)^{D_1 + D_2}$$

$$D \le C \cdot (D_1 + D_2).$$

---

eg. linear threshold function.

$$\mathcal{F} = \{\mathbb{1}_{\{\theta^T x \ge 0\}} : \theta \in \mathbb{R}^d\}$$

$(x_1, x_2, \cdots, x_{d+1})$ $\forall$ binary sequence $b_1, b_2, \cdots, b_{d+1}$

$\exists \theta$, s.t. $b_i = \mathbb{1}\{\theta^T x_i \geq 0\}$ $(\forall i)$.

On the other hand $\exists \lambda_1, \lambda_2, \dots \lambda_{d+1}$ not all 0.

$$\sum_{i=1}^{d+1} \lambda_i x_i = 0.$$

$$b_i = \begin{cases} 1 & \lambda_i > 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$\sum_{i=1}^{D+1} \lambda_i \theta^T x_i = 0$$

$\underline{\lambda_i \theta^T x_i} \geq 0.$ $\forall i$

So $\lambda_i \theta^T x_i = 0$ $(\forall i)$.

Naive application of VC dim

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i)\right] \leq C \sqrt{\frac{\log |\mathcal{F}(x_1, \dots, x_n)|}{n}}$$

$$\leq C \sqrt{\frac{D \log(n/D)}{n}}.$$ suboptimal

Thm (VC dim bound for covering/packing). ($\mathcal{F}$ binary)

$$\forall Q, \quad \log N(\varepsilon; \mathcal{F}, \|\cdot\|_{L^2(Q)}) \leq c \cdot VC(\mathcal{F}) \cdot \log(1/\varepsilon).$$

Proof — $f_1, f_2, \dots, f_N$ max ε-packing of $\mathcal{F}$ under $\mathbb{L}^2(\mathbb{Q})$.

$$\varepsilon^2 \le \|f_i - f_j\|_{L^2(Q)}^2 = \int_{\mathcal{X}} (f_i(x) - f_j(x))^2 \, d\mathbb{Q}(x)$$

$$= \mathbb{Q}(f_i(X) \neq f_j(X)). \quad (\forall i,j \in [N]).$$

Draw iid samples $x_1, x_2, \dots, x_n \overset{iid}{\sim} \mathbb{Q}$.

$$A = \{(f_i(x_1), f_i(x_2), \dots f_i(x_n)) : i \in [N]\}$$

- $\forall i,j \in [N]$

$$\mathbb{P}((f_i(x_1), f_i(x_2), \dots, f_i(x_n)) = (f_j(x_1), f_j(x_2) \dots, f_j(x_n))).$$

$$\le (1-\varepsilon^2)^n \le \exp(-c\varepsilon^2 n).$$

- $\mathbb{P}(\exists i,j \in [N] : f_i, f_j$ not distinguishable under $(x_1, \dots x_n))$.

$$\le \binom{N}{2} \cdot \exp(-c\varepsilon^2 n) \le N^2 \exp(-c\varepsilon^2 n).$$

$$n = \frac{4 \log N}{\varepsilon^2}. \quad \mathbb{P}(\dots) \le \frac{1}{2}.$$

$\exists x_1, x_2 \dots x_n$ s.t. $(f_i : i \in [N])$ are pairwise distinguishable on $(x_1, \dots x_n)$.

$$|A| = N \le \left(\frac{en}{D}\right)^D. \qquad (D = VC(\mathcal{F})).$$

$$N \le \left(\frac{4e \log N}{D\, \varepsilon^2}\right)^D.$$

$$\log N \le D \cdot \left\{ c + \log\log N + \log\left(\tfrac{1}{\varepsilon}\right) \right\}$$

Solving for N,

$$\log N \le c \cdot D \log\left(\tfrac{1}{\varepsilon}\right).$$

$\mathcal{F}$ = set of linear classifiers in $\mathbb{R}^d$.

$$\text{gen err} \le \mathbb{E}\left[\sup_{\cancel{f \in \mathcal{F}}} \left| \frac{1}{n} \sum_{1}^{n} f(X_i) - \cancel{\mathbb{E} f(X)} \right| \right]$$

$$\le \frac{C}{\sqrt{n}} \cdot \int_0^1 \sqrt{\log \sup_Q N(\varepsilon; \mathcal{F}, \|\cdot\|_{L^2(Q)})}\, d\varepsilon.$$

$$\le \frac{C}{\sqrt{n}} \int_0^1 \sqrt{c' \cdot d \log\left(\tfrac{1}{\varepsilon}\right)}\, d\varepsilon$$

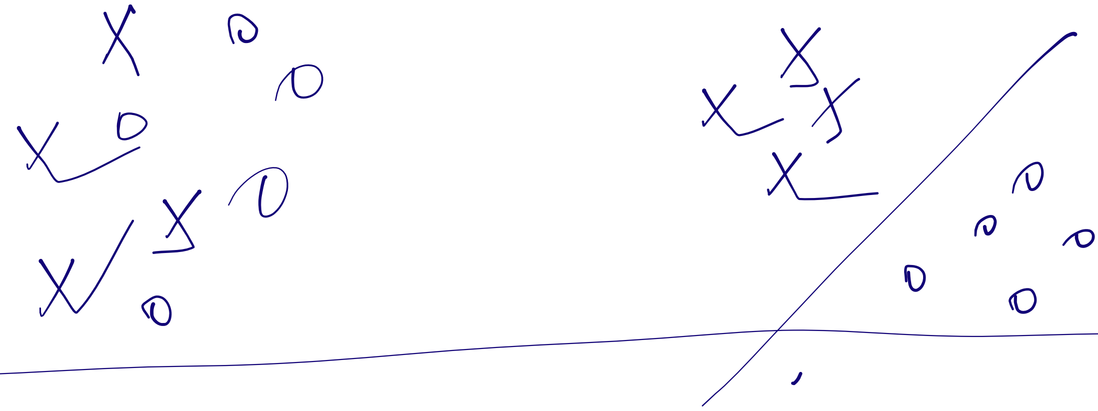
$$= c'' \sqrt{\frac{d}{n}}.$$

P. Bartlett

S. Mendelson '05

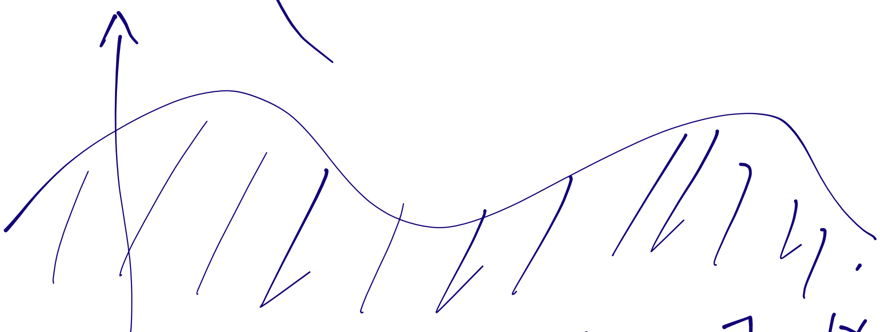
"Local Rademacher Complexities"

(Tight when data are not separable).



Def: VC subgraph dimension $\mathcal{F} : \mathcal{X} \to \mathbb{R}$.

$$VC(\mathcal{F}) := VC\left(\{(t,x) \mapsto \mathbb{1}\{t \le f(x)\} : f \in \mathcal{F}\}\right).$$



equivalently. largest D st. $\exists \, (x_i, t_i)_{i=1}^{D}$

$$\left\{ \left(\mathbb{1}_{t_i \ge f(x_i)}\right)_{i=1}^{D} \cancel{: f \in \mathcal{F}} \right\}$$

achieves all the binary configurations.

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

$$t_i \ge f_\theta(x_i) \iff \begin{bmatrix} -t_i \\ x_i \end{bmatrix}^T \begin{bmatrix} 1 \\ +\theta \end{bmatrix} \ge 0.$$

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

eg φ is strictly <u>increasing</u>

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

$$t_i \geq f_\theta(x_i) \iff \begin{bmatrix} \varphi^{-1}(t_i) \\ x_i \end{bmatrix} \begin{bmatrix} 1 \\ -\theta \end{bmatrix} \geq 0.$$

Thm. $\sup_Q N\left(\varepsilon \|F\|_{L^2(Q)}; \mathcal{F}, L^2(Q)\right) \leq \left(\frac{C_1}{\varepsilon}\right)^{C_2 VC(\mathcal{F})}$ $(\forall \varepsilon > 0)$.

where F is envelope for $\mathcal{F}$.

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

$$\leq 2 \int F(x) |f(x) - g(x)| \, dQ(x)$$

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

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

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

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

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

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

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

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

So we have

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

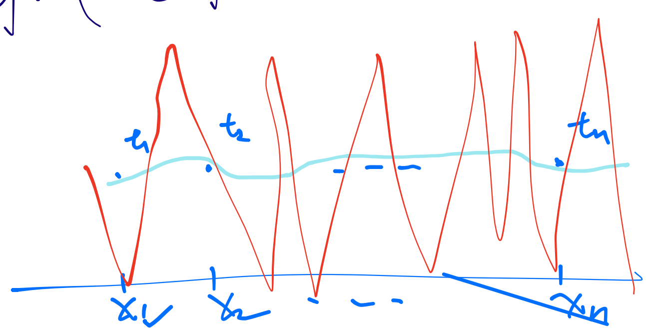
ε - covering of $\{(t,x) \to 1_{t\le f(x)}\ \cancel{f\in\mathcal{F}}\}$ under $\tilde{Q}$

$$\Downarrow$$

$2\cdot\|F\|_{L^2(Q)}\cdot\sqrt{\varepsilon}$ - covering of $\mathcal{F}$ under Q.

Nonparametric classes usually have infinite VC (subgraph) dimension.

Def. (ε-fat-shattering dimension).



$\text{fat}_\varepsilon(\mathcal{F})$ is largest D s.t.
$\exists (x_i, t_i)_{i=1}^D$ ε- shattered by $\mathcal{F}$.

i.e.
$\forall$ binary seq
$b_1, b_2 \cdots, b_D$
$\exists f \in \mathcal{F}$

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

examples :- future lectures on nonpara

Thm (Mendelson - Vershynin). $\mathcal{F}$ bdd by 1

$$\sup_Q M(\varepsilon; \mathcal{F}, L^2(Q)) \leq \left(\frac{1}{\varepsilon}\right)^{c_1 \text{fat}_{c_2\varepsilon}(\mathcal{F})}$$

(where c_1 c_2 are universal consts).

Thm (Rudelson - Vershynin) (Ann. Math.).
under some mild assumptions

$$\sup_Q M(\varepsilon; \mathcal{F}, L^2(Q)) \leq \exp\left(c_1 \text{fat}_{c_2\varepsilon}(\mathcal{F})\right).$$

Original problem

$$\hat{\theta}_n := \arg\min \frac{1}{n}\sum_1^n \ell(\theta; X_i) =: \hat{L}_n(\theta)$$

$$\theta^* = \arg\min \mathbb{E}[\ell(\theta; X)] =: L(\theta).$$

$$L(\hat{\theta}_n) - L(\theta^*) \leq 2 \cdot \sup_{\theta \in \Theta} \left| \hat{L}_n(\theta) - L(\theta) \right| . \lesssim \sqrt{\frac{VC}{n}}.$$

Too conservative.

$$L(\theta) - L(\theta^*) \geq \|\theta - \theta^*\|^2 .$$ (at least locally around θ^*).

$$\|\hat{\theta}_n - \theta^*\| \lesssim \left(\frac{VC}{n}\right)^{1/4} .$$

Thm (main convergence rate thm).

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

Suppose that

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

satisfying $\phi_n(cx) \leq c^\alpha \phi_n(x)$ (for any $c > 1, x > 0$) for some $\alpha < 2$.

Then. for any δ_n that satisfies $\phi_n(\delta_n) \leq \delta_n^2$

$\forall \varepsilon > 0, \exists C_\varepsilon > 0$, s.t. $\|\hat{\theta}_n - \theta^*\| \leq C_\varepsilon \cdot \delta_n$ w.p. $1-\varepsilon$.

Proof. $\|\hat{\theta}_n - \theta^*\|^2 \leq L(\hat{\theta}_n) - L(\theta^*) \leq (P_n - P)(l_{\theta^*} - l_{\hat{\theta}_n})$.

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

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

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

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

$$\leq \frac{1}{2^{2j-2}\delta_n^2}\, \mathbb{E}\left[\sup_{\cdots} \cdots\cdots\right].$$

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

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

Summing up.

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

$$\leq \underbrace{\frac{4\phi_n(\delta_n)}{\delta_n^2}}_{\leq 4} \cdot \frac{2^{-(\alpha-2)M}}{1 - 2^{\alpha-2}} \to 0.$$