$$T_3 = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{\theta_0}(X_i) - E_X[\mathcal{L}_{\theta_0}(X)]$$ concentration of measure inequs.

$$T_1 = E_X[\mathcal{L}_{\hat\theta}(X)] - \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{\hat\theta}(X_i)$$

challenging because $\hat\theta$ is a random quantity

$$\mathcal{F} = \{ \mathcal{L}_\theta(\cdot) \mid \theta \in \Omega_0 \}$$

$$T_1 \le \sup_{\theta \in \Omega_0} \left| \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_\theta(X_i) - E_X[\mathcal{L}_\theta(X)] \right| = \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$$

$$E(\hat\theta, \theta^*) \le 2\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}.$$

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

12.3. Uniform laws via Rademacher complexity

Function class $\mathcal{F}$

$x_1^n = (x_1, \dots, x_n)$

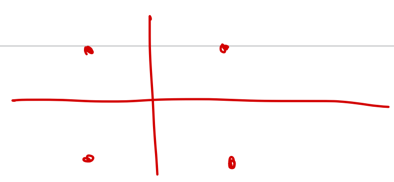
$$\mathcal{F}(x_1^n) = \{ f(x_1), f(x_2), \dots, f(x_n) \mid f \in \mathcal{F} \}$$

empirical Rademacher complexity

$$\mathcal{R}(\mathcal{F}(x_1^n)/n) = E_\varepsilon\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i) \right|\right]$$

$\{\varepsilon_k\}_{k=1}^n$ ... i.i.d. sequence of Rademacher RVs, taking ± 1 with equal prob.

$$\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i) = \frac{1}{n} \underbrace{(\varepsilon_1 \ \ \varepsilon_2 \dots \varepsilon_n)}_{\text{total of } 2^n \text{ realizations}} \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix}$$



$X_1^n \doteq \{X_i\}_{i=1}^n$

$$R_n(\mathcal{F}) = R(\mathcal{F}(X_1^n)/n) = E_X[R(\mathcal{F}(X_1^n)/n)]$$

$$= E_{\varepsilon, X}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right]$$

Theorem 12.10. For any b-uniformly bounded function class, i.e., $\|f\|_\infty \le b$, $\forall f\in\mathcal{F}$, any positive integer $n\ge 1$ and any $\delta \ge 0$, we have

$$\|P_n - P\|_{\mathcal{F}} \le 2R_n(\mathcal{F}) + \delta$$

with prob. $\ge 1 - e^{-\frac{n\delta^2}{2b^2}}$. Hence, as long as $R_n(\mathcal{F}) = o(1)$, we have $\|P_n - P\|_{\mathcal{F}} \xrightarrow{a.s.} 0$.

sufficient condition for $\|P_n - P\|_{\mathcal{F}} \xrightarrow{a.s.} 0$ is : $R_n(\mathcal{F}) \xrightarrow{n\to\infty} 0$

Prop. 12.11. For every b-uniformly bounded function class $\mathcal{F}$, any integer $n\ge 1$ and any $\delta \ge 0$, we have

$$\|P_n - P\|_{\mathcal{F}} \ge \frac{1}{2}R_n(\mathcal{F}) - \frac{\sup_{f\in\mathcal{F}} |E_P[f]|}{2\sqrt{n}} - \delta$$

with prob. $\ge 1 - e^{-\frac{n\delta^2}{2b^2}}$.

Levels of escalation in terms of upper-bounding Rademacher complexity

1. $|\mathcal{F}| < \infty$ union bounds won't consider
2. polynomial discrimination function classes ✓
3. VC (Vapnik–Chervonenkis) dimension ✓
4. metric entropy ✓ (+ chaining arguments) ✗

Empirical process theory

12.4. Function classes with polynomial discrimination

recall that $\mathcal{F}(x_1^n) = \{f(x_1), f(x_2), \dots, f(x_n) \mid f\in\mathcal{F}\}$

$$R_n(\mathcal{F}) = E_{\varepsilon,X}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i)\right|\right]$$

$\mathcal{F}$ contains binary-valued functions

$f \in \mathcal{F}$: $f(x) = \pm 1$

$$|\mathcal{F}(x_1^n)| = |\{\underbrace{f(x_1)}_{\pm 1}, \underbrace{f(x_2)}_{\pm 1}, \ldots, \underbrace{f(x_n)}_{\pm 1} \mid f\in\mathcal{F}\}| = 2^n$$

$|\mathcal{F}(x_1^n)|$ is polynomial in n

Def. 12.2 (Polynomial discrimination). A class of functions $\mathcal{F}$ with domain $\mathcal{X}$ has polynomial discrimination of order $\nu \geq 1$, if for each $n \in \mathbb{N}$ and collection $x_1^n = (x_1, \ldots, x_n)$ of n points in $\mathcal{X}$, the set $\mathcal{F}(x_1^n)$ satisfies

$$|\mathcal{F}(x_1^n)| \leq (n+1)^\nu .$$

Lemma 12.13. Suppose that $\mathcal{F}$ has polynomial discrimination of order ν. Then, for all $n\in\mathbb{N}$ and any collection of points $x_1^n = (x_1, \ldots, x_n)$, we have

$$E_\varepsilon\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i)\right|\right] \leq 4 D(x_1^n) \sqrt{\frac{\nu \log(n+1)}{n}} ,$$

where $D(x_1^n) = \sup_{f\in\mathcal{F}} \sqrt{\frac{\sum_{i=1}^n f^2(x_i)}{n}}$ is the ℓ_2-radius of the set $\mathcal{F}(x_1^n)/\sqrt{n}$.

Proof.

$$E_\varepsilon\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i)\right|\right] = \frac{1}{n} E_\varepsilon\left[\max_{\theta\in\mathcal{F}(x_1^n)} |\langle \varepsilon, \theta\rangle|\right]$$

$$\varepsilon = (\varepsilon_1\ \varepsilon_2 \ldots \varepsilon_n)$$

the set $\mathcal{F}(x_1^n)$ has $\leq (n+1)^\nu$

will apply

Lemma 2.4. For $n \geq 2$, let $\{X_i\}_{i=1}^n$ be a set of zero-mean RVs, each sub-Gaussian with parameter σ. Then,

$$\mathbb{E}\left[\max_{i=1,\dots,n} |X_i|\right] \leq 2\sigma\sqrt{\log(n)}.$$

sub-Gaussian $\quad \mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\sigma^2\lambda^2}{2}}, \forall \lambda \in \mathbb{R}$

$\langle \varepsilon, \theta \rangle$ is sub-Gaussian (with parameter $\sigma = \sup_{\theta \in \mathcal{F}(x_1^n)} \sqrt{\sum_{i=1}^n \theta_i^2}$)

(used independence of ε_i)

$$\mathbb{E}_\varepsilon\left[e^{\lambda\langle \varepsilon, \theta\rangle}\right] = \prod_{i=1}^n \mathbb{E}_{\varepsilon_i}\left[e^{\lambda \varepsilon_i \theta_i}\right]$$

(evaluation of expectation)

$$= \prod_{i=1}^n \frac{e^{\lambda\theta_i} + e^{-\lambda\theta_i}}{2}$$

$e^x + e^{-x} \leq 2e^{\frac{x^2}{2}}$

$$\leq e^{\frac{\lambda^2}{2}\sum_{i=1}^n \theta_i^2}$$

$$\leq e^{\frac{\lambda^2}{2}\underbrace{\left\{\sup_{\theta \in \mathcal{F}(x_1^n)} \sum_{i=1}^n \theta_i^2\right\}}_{\sigma^2}}$$

$$R(\mathcal{F}(x_1^n)/n) \leq 2D(x_1^n)\sqrt{\frac{2\log(n+1)}{n}}. \qquad \square$$

$$D(x_1^n) = \sup_{f \in \mathcal{F}} \sqrt{\frac{\sum_{i=1}^n f^2(x_i)}{n}}$$

$$R_n(\mathcal{F}) = \mathbb{E}_X\left[R(\mathcal{F}(X_1^n)/n)\right] \leq 4\,\mathbb{E}_{X_1^n}\left[D(X_1^n)\right]\sqrt{\frac{2\log(n+1)}{n}}.$$

b- uniformly bounded function classes

$$R_n(\mathcal{F}) \leq 4b\sqrt{\frac{2\log(n+1)}{n}}, \text{ for all } n \in \mathbb{N}$$

Corollary 12.4. (Classical Glivenko–Cantelli). Let $F(t) = P[X \leq t]$ be the CDF of a random variable $X \sim P$, and let $\hat{F}_n$ be the empirical CDF based on n i.i.d. samples $X_i \sim P$. Then,

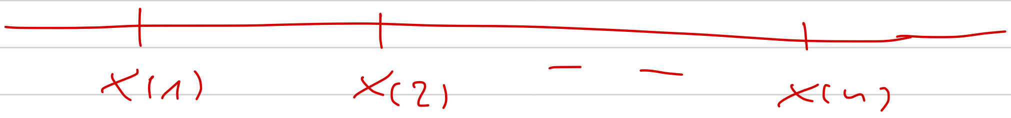
$$P\left[\|\hat{F}_n - F\|_\infty \geq 8\sqrt{\frac{\log(n+1)}{n}} + \delta\right] \leq e^{-n\delta^2/2}, \quad \delta \geq 0$$

and hence $\|\hat{F}_n - F\|_\infty \xrightarrow{a.s.} 0$.

Proof $\mathcal{F}(x_1^n) = \{f(x_1), f(x_2), \ldots, f(x_n) \mid f \in \mathcal{F}\}$

$x_1^n = (x_1, \ldots, x_n) \in \mathbb{R}^n$, consider the set $\mathcal{F}(x_1^n)$ of $\{0,1\}$-valued indicator functions of the half-intervals $(-\infty, t]$, $t \in \mathbb{R}$.

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$



divides $\mathbb{R}$ up into $n+1$ intervals

for a given $t \in \mathbb{R}$, $\mathbb{1}_{(-\infty, t]}(x) = \begin{cases} 1, & x \leq t \\ 0, & \text{else} \end{cases}$

$t \leq x_{(1)}$: $(f(x_1), f(x_2), \ldots, f(x_n)) = (0, 0, \ldots, 0)$

$x_{(1)} \leq t \leq x_{(2)}$: — " — $= (1, 0, \ldots, 0)$

$x_{(2)} \leq t \leq x_{(3)}$: — " — $= (1, 1, 0, \ldots, 0)$

$\vdots$

$(1, 1, \ldots, 1)$

$$|\mathcal{F}(x_1^n)| \leq (n+1) = (n+1)^\nu, \quad \nu = 1$$

with $\nu = 1$, Lemma 12.13.

$$E_\varepsilon\left[\sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right|\right] \leq 4\sqrt{\frac{\log(n+1)}{n}}$$

E w.r.t. X $\Rightarrow$ $R_n(\mathcal{F}) \leq 4\sqrt{\frac{\log(n+1)}{n}}$

apply Thm. 12.10. $\square$

12.5 Vapnik-Chervonenkis dimension

$$\mathcal{F}(x_1^n) = \{f(x_1), \ldots, f(x_n) \mid f \in \mathcal{F}\}$$

$$|\mathcal{F}(x_1^n)| \leq 2^n$$

just had an example with $|\mathcal{F}(x_1^n)| \le (n+1)$

Def 12.15. (Shattering and VC dimension). Given a class $\mathcal{F}$ of binary-valued functions, we say that the set $x_1^n = (x_1, \dots, x_n)$ is shattered by $\mathcal{F}$ if $|\mathcal{F}(x_1^n)| = 2^n$.

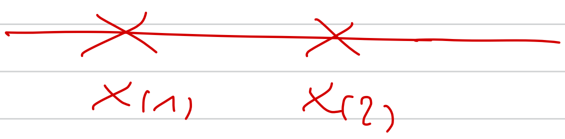
The VC dimension $\nu(\mathcal{F})$ is the largest integer n for which there is a collection $x_1^n = (x_1, \dots, x_n)$ of n points that is shattered by $\mathcal{F}$.

When $\nu(\mathcal{F})$ is finite, $\mathcal{F}$ is said to be a VC class.

Example 12.16. $S_{left} = \{ (-\infty, a] \mid a \in \mathbb{R} \}$, $\mathcal{F}$ = indicator functions on S_{left}

$n=1$: $|\mathcal{F}| = 2 = 2^1$, with $\mathcal{F} = \{0,1\}$

$n=2$:



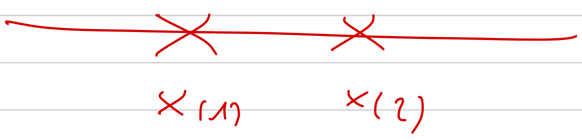
$$\mathcal{F}(x_1^2) = \{ f(x_{(1)}), f(x_{(2)}) \mid f \in \mathcal{F} \}$$

$$\mathcal{F}(x_1^2) = \{ (0,0), (1,0), (1,1) \}$$

$\Rightarrow$ 2-point sets are not being shattered

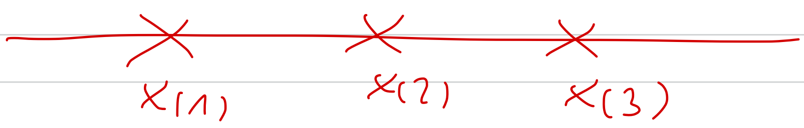
$\Rightarrow \nu(\mathcal{F}) = 1$.

$S_{two} = \{ (a,b] \mid a,b \in \mathbb{R} \text{ s.t. } a<b \}$



$$\mathcal{F} = \{ (0,0), (1,0), (0,1), (1,1) \}$$

$$|\mathcal{F}| = 4 = 2^2$$



$$\mathcal{F} = \{ (1,0,0), (0,1,0), (0,0,1), (1,1,0), (0,1,1), (1,1,1), (0,0,0), \cancel{(1,0,1)} \}$$

$\Rightarrow |\mathcal{F}| = 7 < 2^3$

$\Rightarrow \nu(\mathcal{F}) = 2$.

$$|\mathcal{F}| \le (n+1)^2 \Rightarrow v(\mathcal{F}) = 2.$$

all examples we considered are of finite VC dim. and also have polynomial discrimination.

Q: Does every finite VC class have polynomial discrimination?

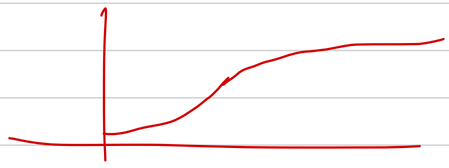
$n=1$ 2, $|\mathcal{F}| = 2^1$

$n=2$ $2^2 = 4$, $|\mathcal{F}| = 2^2$ $|\mathcal{F}| \le (n+1)^2$

$n=3$ $2^3 = 8$, $|\mathcal{F}| = 2^3$

$n=4$ breaks < 16

$\vdots$



Theorem 12.17. (Vapnik–Chervonenkis, Sauer–Shelah). Consider a set class S with $v(S) < \infty$. Then, for any collection of points $P = (x_1, \dots, x_n)$ with $n \ge v(S)$, we have

$$|S(P)| \le \sum_{i=0}^{v(S)} \binom{n}{i} \le (n+1)^{v(S)}.$$