Empirical processes, VC dimension

Let $X_1, \dots, X_n$ be iid on $\mathcal{X}$, $\mathcal{F}: \mathcal{X} \to \mathbb{R}$ a function class.

The associated <u>empirical process</u> $\{Z_f\}_{f \in \mathcal{F}}$ is

$$Z_f = \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X_i)$$

Example (Glivenko-Cantelli ULLN):

$$\mathcal{X} = \mathbb{R}, \quad \mathcal{F} = \{ \mathbb{1}_{(-\infty, t]} : t \in \mathbb{R} \}$$

$$Z_f = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \le t\} - \mathbb{P}[X_i \le t]$$

$$= \underbrace{F_n(t)}_{\text{empirical CDF}} - \underbrace{F(t)}_{\text{CDF}} \xrightarrow{P} 0 \quad \text{as } n \to \infty, \text{ by LLN.}$$

Thm (Glivenko-Cantelli): $\sup_{f \in \mathcal{F}} |Z_f| \xrightarrow{P} 0$ as $n \to \infty$.

How to bound expectation / tail of $\sup_{f \in \mathcal{F}} |Z_f|$?

Example (classification risk): $(X_1, Y_1), \dots, (X_n, Y_n)$ iid, where

$X_i \in \mathcal{X}$, $Y_i \in \{0,1\}$. $\mathcal{F}$ = class of functions $\mathcal{X} \to \{0,1\}$.

Empirical risk: $\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{f(X_i) \ne Y_i\}$

True risk: $\mathbb{P}[f(X_i) \ne Y_i]$

How to bound $\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{f(X_i)\neq Y_i\} - \mathbb{P}[f(X_i)\neq Y_i]\right|$?

Note: These function classes are not smooth and do not have finite covers in $\|\cdot\|_\infty$. E.g., for any finite covering net $\mathcal{N}\subseteq\mathcal{F}=\{\mathbb{1}_{(-\infty,t]}: t\in\mathbb{R}\}$, $\exists f\in\mathcal{F}$ s.t. $\min_{g\in\mathcal{N}}\|f-g\|_\infty = 1$

$$\Rightarrow N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) = \infty \text{ for any } \varepsilon < 1.$$

Lemma (Rademacher symmetrization): For any independent r.v.'s $X_1,\ldots,X_n$ and function class $\mathcal{F}$,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X_i) \leq 2\sup_{f\in\mathcal{F}}\sum_{i=1}^n \varepsilon_i f(X_i)$$

where $\varepsilon_1,\ldots,\varepsilon_n\in\{+1,-1\}$ are i.i.d. Rademacher variables.

Proof: Let $X_1',\ldots,X_n'$ be independent copies of $X_1,\ldots,X_n$. Then

$$\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X_i)$$

$$= \mathbb{E}_X\sup_{f\in\mathcal{F}}\mathbb{E}_{X'}\sum_{i=1}^n f(X_i) - f(X_i')$$

$$\leq \mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(X_i) - f(X_i') = \mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n \varepsilon_i\left(f(X_i) - f(X_i')\right),$$

because $\{f(X_i) - f(X_i')\}_{f\in\mathcal{F}} \overset{L}{=} \{\varepsilon_i(f(X_i) - f(X_i'))\}_{f\in\mathcal{F}}$. Now

$$\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^n \varepsilon_i \left(f(X_i) - f(X_i') \right)$$

$$\leq \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^n \varepsilon_i f(X_i) + \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^n (-\varepsilon_i) f(X_i') = 2 \sup_{f \in \mathcal{F}} \sum_{i=1}^n \varepsilon_i f(X_i). \blacksquare$$

Idea: Bound $\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \sum_{i=1}^n \varepsilon_i f(X_i)$ conditional on any fixed $(X_1, \ldots, X_n) \in \mathcal{X}^n$. Note: Conditional on $(X_1, \ldots, X_n)$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i) - \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i g(X_i)$$

is $\|f-g\|^2_{L^2(P_n)} := \frac{1}{n} \sum_{i=1}^n \left(f(X_i) - g(X_i) \right)^2$ - subgaussian. So

$$\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \leq \frac{C}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\mathcal{F}, \|\cdot\|_{L^2(P_n)}, \varepsilon)} \, d\varepsilon$$

Covering $\mathcal{F}$ on only n points is easier than on all of $\mathcal{X}$, e.g. for $\mathcal{F} = \{ \mathbb{1}_{(-\infty, t]} : t \in \mathbb{R} \}$, there exists a net $\mathcal{N}$ of $n+1$ functions s.t. $\|f - \pi(f)\|_{L^2(P_n)} = 0 \quad \forall f \in \mathcal{F}$.

In general, for any class $\mathcal{F} : \mathcal{X} \to \{0, 1\}$, there are only 2^n possible values of $(f(X_1), \ldots, f(X_n)) \Rightarrow$ there exists a net $\mathcal{N}$ of 2^n functions s.t. $\|f - \pi(f)\|_{L^2(P_n)} = 0 \quad \forall f \in \mathcal{F}$.

VC Dimension

Def: Let $\mathcal{F}$ be a class of functions $f: \mathcal{X} \to \{0,1\}$.

$\Lambda \subseteq \mathcal{X}$ is shattered by $\mathcal{F}$ if $\{f|_\Lambda : f \in \mathcal{F}\}$ contains all $2^{|\Lambda|}$ possible functions $\Lambda \to \{0,1\}$. The VC dimension is

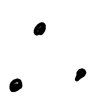
$$vc(\mathcal{F}) = \sup_{\Lambda \subseteq \mathcal{X} \text{ shattered by } \mathcal{F}} |\Lambda|.$$

Equivalently: Let $\mathcal{A}$ be a class of subsets of $\mathcal{X}$. $\Lambda \subseteq \mathcal{X}$ is shattered by $\mathcal{A}$ if $\forall S \subseteq \Lambda$, there exists $A \in \mathcal{A}$ s.t. $\Lambda \cap A = S$.

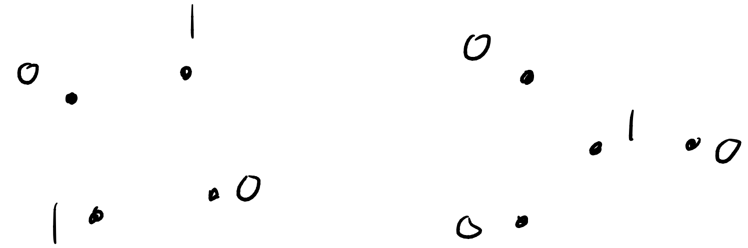
Example: $\mathcal{F} = \{\mathbb{1}_{(-\infty, t]} : t \in \mathbb{R}\}$ on $\mathbb{R}$. Any singleton $\{x\}$ is shattered, any two points $\{x, y\}$ are not shattered. (If $x < y$, cannot have $f(x) = 1$, $f(y) = 0$.) So $vc(\mathcal{F}) = 1$.

Example: $\mathcal{F} = \{\mathbb{1}_{[a,b]} : a, b \in \mathbb{R}\}$ on $\mathbb{R}$. Any two points $\{x, y\}$ are shattered, any 3 points are not shattered. (If $x < y < z$, cannot have $f(x) = 1$, $f(y) = 0$, $f(z) = 1$.) So $vc(\mathcal{F}) = 2$.

Example: $\mathcal{F} = \{$indicators of closed half planes on $\mathbb{R}^2\}$.

3 points  in general position are shattered, any

4 points are not shattered. So $vc(\mathcal{F}) = 3$. (See Homework 9, Problem 1 for more general statement.)



Example: $\mathcal{X} = \{x_1, x_2, x_3\}$. Represent $f : \mathcal{X} \to \{0,1\}$ by binary strings.

$\mathcal{F} = \{001, 010, 100, 111\}$ (number of values 1 is odd).

Here $\{x_1, x_2\}$ is shattered, $\{x_1, x_2, x_3\}$ is not. So $vc(\mathcal{F}) = 2$.

Thm: Let $\mathcal{F} : \mathcal{X} \to \{0,1\}$, let P be any probability measure on $\mathcal{X}$, and define $\|f-g\|^2_{L^2(P)} = \mathbb{E}_P (f(X) - g(X))^2$. Then for a universal constant $C > 0$ and any $\varepsilon \in (0,1)$,

$$N(\mathcal{F}, \|\cdot\|_{L^2(P)}, \varepsilon) \leq \left(\frac{C}{\varepsilon}\right)^{C \cdot vc(\mathcal{F})}.$$

Lemma (Pajor): Suppose $\mathcal{X}$ is a finite set, and $\mathcal{F}$ is any class of functions $f : \mathcal{X} \to \{0,1\}$. Then

$|\mathcal{F}| \leq$ number of subsets of $\mathcal{X}$ shattered by $\mathcal{F}$.

(including empty set).

Proof: Induction on $|X|$.

Base case $X = \{x\}$. ϕ always shattered. If $\mathcal{F}$ contains both $f(x)=0$ and $f(x)=1$, then $\{x\}$ also shattered, so lemma holds.

Inductive step: Let $X = X_0 \cup \{x\}$ and suppose lemma holds for X_0. Let $S(\mathcal{F})$ = number of sets shattered by $\mathcal{F}$. Write

$$\mathcal{F} = \underbrace{\{f \in \mathcal{F} : f(x) = 0\}}_{=\mathcal{F}_0} \cup \underbrace{\{f \in \mathcal{F} : f(x) = 1\}}_{=\mathcal{F}_1}$$

so $|\mathcal{F}| = |\mathcal{F}_0| + |\mathcal{F}_1| \leq S(\mathcal{F}_0) + S(\mathcal{F}_1)$ by induction hypothesis applied to X_0.

- Suppose $\Lambda \subseteq X_0$ shattered by exactly one of $\mathcal{F}_0, \mathcal{F}_1$. Then, trivially, Λ is also shattered by $\mathcal{F}$.
- Suppose $\Lambda \subseteq X_0$ shattered by <u>both</u> $\mathcal{F}_0, \mathcal{F}_1$. Then Λ and $\Lambda \cup \{x\}$ are both shattered by $\mathcal{F}$.

These sets Λ, $\Lambda \cup \{x\}$ are distinct, so $S(\mathcal{F}) \geq S(\mathcal{F}_0) + S(\mathcal{F}_1)$, hence $|\mathcal{F}| \leq S(\mathcal{F})$. ∎

Cor (Sauer-Shelah): If $|X| = m$ and $vc(\mathcal{F}) = d$, then

$$|\mathcal{F}| \le \sum_{k=0}^{d} \binom{m}{k} \le \left(\frac{em}{d}\right)^d.$$

Proof: Number of sets shattered by $\mathcal{F}$ is at most $\sum_{k=0}^{d} \binom{m}{k}$, since largest such set has size d. $\sum_{k=0}^{d} \binom{m}{k} \le \left(\frac{em}{d}\right)^d$ is elementary. ∎

Proof of Thm: By packing number upper bound,

$$N(\mathcal{F}, \|\cdot\|_{L^2(P)}, \varepsilon) \le D(\mathcal{F}, \|\cdot\|_{L^2(P)}, \varepsilon).$$

Let $\mathcal{D}$ be an ε-packing of largest size, i.e.

$$\mathbb{E}_P (f(X) - g(X))^2 > \varepsilon^2 \quad \forall f \ne g \in \mathcal{D}.$$

Idea: Apply probabilistic method — sample m points $X_1, \dots, X_m \overset{iid}{\sim} P$. By Hoeffding's ineq., for any $f, g \in \mathcal{F}$,

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} (f(X_i) - g(X_i))^2 - \mathbb{E}_P (f(X) - g(X))^2\right| \ge \frac{\varepsilon^2}{2}\right]$$
$$\le 2e^{-cm\varepsilon^4}, \quad \text{because } |f(X) - g(X)| \le 1.$$

Take union bound over all pairs $f \ne g \in \mathcal{D}$:

On an event of probability $\geq 1 - |\mathcal{D}|^2 e^{-cm\varepsilon^4}$,

$$\frac{1}{m}\sum_{i=1}^{m}(f(X_i) - g(X_i))^2 \geq \frac{\varepsilon^2}{2} \text{ for every } f \neq g \in \mathcal{D}.$$

For $m = C\varepsilon^{-4}\log|\mathcal{D}|$, this probability is positive,

This shows in particular: there exist $X_1,, X_m \in \mathcal{X}$ s.t. $f|_{\{X_1,, X_m\}}$ are distinct for all $f \in \mathcal{D}$.

Let d_m be VC-dimension of $\mathcal{D}$ on $\{X_1,, X_m\}$.

$$\Rightarrow |\mathcal{D}| \leq \left(\frac{em}{d_m}\right)^{d_m} = \left(\frac{C\varepsilon^{-4}\log|\mathcal{D}|}{d_m}\right)^{d_m}$$

Apply $\frac{\log|\mathcal{D}|}{2d_m} = \log|\mathcal{D}|^{\frac{1}{2d_m}} \leq |\mathcal{D}|^{\frac{1}{2d_m}}$

$$\Rightarrow |\mathcal{D}| \leq (C\varepsilon^{-4})^{2d_m} \leq \left(\frac{C'}{\varepsilon}\right)^{C'd_m} \quad \forall \varepsilon \in (0,1).$$

Finally note that $d_m \leq$ VC-dim of $\mathcal{F}$ on $\mathcal{X}$. ∎

Thm: Let $X_1,\dots,X_n \in \mathcal{X}$ be iid, $\mathcal{F}$ a class of functions $f: \mathcal{X} \to \{0,1\}$. Then for a universal constant $C \geq 0$

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

Proof: Let $\bar{\mathcal{F}} = \{1-f : f \in \mathcal{F}\}$.

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

$$= \mathbb{E} \sup_{f \in \mathcal{F} \cup \bar{\mathcal{F}}} \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E} f(X_i)$$

$$\leq \mathbb{E}_{\varepsilon, X} \sup_{f \in \mathcal{F} \cup \bar{\mathcal{F}}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \quad \text{(Rademacher sym.)}$$

$$\leq \frac{C}{\sqrt{n}} \mathbb{E}_X \int_0^1 \sqrt{\log N(\mathcal{F} \cup \bar{\mathcal{F}}, \|\cdot\|_{L^2(P_n)}, \varepsilon)}\, d\varepsilon \quad \text{(Dudley's ineq.)}$$

$$\leq \frac{C'}{\sqrt{n}} \mathbb{E}_X \int_0^1 \sqrt{\log 2 \underbrace{N(\mathcal{F}, \|\cdot\|_{L^2(P_n)}, \varepsilon)}_{\leq \left(\frac{C}{\varepsilon}\right)^{C \cdot vc(\mathcal{F})}}}\, d\varepsilon \leq C'' \sqrt{\frac{vc(\mathcal{F})}{n}}.$$

$\blacksquare$

Example: Recall $\mathcal{F} = \{\mathbb{1}_{(-\infty, t]} : t \in \mathbb{R}\}$ has VC-dim $= 1$. Then

$$\mathbb{E} \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i \leq t\} - \mathbb{P}[X_i \leq t] \right| \leq \frac{C}{\sqrt{n}}.$$

Example (excess risk in classification): $(X_i, Y_i)_{i=1}^n$ iid, $Y_i \in \{0,1\}$.

Optimal estimator in $\mathcal{F}$: $f_* = \operatorname{argmin}_{f \in \mathcal{F}} \underbrace{\mathbb{P}[f(X_i) \neq Y_i]}_{\equiv R(f)}$.

Empirical risk minimizer: $\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \underbrace{\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{f(X_i) \neq Y_i\}}_{\equiv R_n(f)}$.

How large is the excess risk $R(\hat{f}) - R(f_*)$?

$$R(\hat{f}) \le R_n(\hat{f}) + \sup_{f \in \mathcal{F}} |R_n(f) - R(f)|$$
$$\le R_n(f_*) + \sup_{f \in \mathcal{F}} |R_n(f) - R(f)|$$
$$\le R(f_*) + 2 \sup_{f \in \mathcal{F}} |R_n(f) - R(f)|$$

Let $\mathcal{L} = \{ \ell(x,y) = \mathbb{1}\{f(x) \neq y\} : f \in \mathcal{F} \}$. Then

$$\mathbb{E} \sup_{f \in \mathcal{F}} |R_n(f) - R(f)| = \mathbb{E} \sup_{\ell \in \mathcal{L}} \left| \frac{1}{n} \sum_{i=1}^n \ell(X_i, Y_i) - \mathbb{E}\,\ell(X_i, Y_i) \right|$$
$$\le \mathbb{E} \frac{C}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\mathcal{L}, \|\cdot\|_{L^2(P_n)}, \varepsilon)}\, d\varepsilon.$$

If $\mathcal{N}$ is ε-net of $\mathcal{F}$, i.e.

$$\frac{1}{n} \sum_{i=1}^n \left(f(X_i) - \pi(f)(X_i) \right)^2 = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{ f(X_i) \neq \pi(f)(X_i) \} \le \varepsilon^2,$$

then $\{ \mathbb{1}\{f(x) \neq y\} : f \in \mathcal{N} \}$ is also ε-net of $\mathcal{L}$.

$$\Rightarrow \mathbb{E} \sup_{f \in \mathcal{F}} |R_n(f) - R(f)| \le C \sqrt{\frac{vc(\mathcal{F})}{n}} \quad \text{(instead of } vc(\mathcal{L})\text{)}.$$

So $\mathbb{E} R(\hat{f}) - R(f_*) \le C \sqrt{\frac{vc(\mathcal{F})}{n}}$.