SL: VAPNIK - CHERVONENKIS THEORY

We consider the problem of binary classification: predict the unknown label $Y \in \{0,1\}$ of $X \in \mathbb{R}^d$, based on a learning sample $\mathcal{L}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, where each (X_i, Y_i) is iid, with distribution $\mathbb{P}_{X,Y}$.
To do so, we construct from $\mathcal{L}_n$ a function $f_n : \mathbb{R}^d \to \{0,1\}$, which predicts Y using $f_n(X)$. The performance of f_n is evaluated in terms of the conditional expectation

$$R(f_n) := \mathbb{E}\{\ell(Y, f_n(X)) \mid \mathcal{L}_n\},$$ (aka the RISK of f_n)

where $\ell : \{0,1\} \times \{0,1\} \to \mathbb{R}_+$ denotes a loss function, which incurs a cost for mislabelling the variable Y. In the context of binary classification, it is customary to consider the 0-1 loss $\ell_{0\text{-}1}(y, f) = \mathbb{1}(y \neq f)$, which incurs a unit cost per error. The risk of f_n is then the probability of misclassification:

$$R(f_n) = \mathbb{E}\{\mathbb{1}(Y \neq f_n(X)) \mid \mathcal{L}_n\} = \mathbb{P}(Y \neq f_n(X) \mid \mathcal{L}_n).$$

f_n is usually constructed by minimization of the empirical risk $\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell_{0\text{-}1}(Y_i, f(X_i))$ over a class $\mathcal{F}$ of candidate functions

$$\boxed{f_n \in \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}_n(f)}$$

aka the Empirical Risk Minimizer.

- The risk of f_n is usually compared to Bayes Risk $R^* = R(f^*)$, where $f^* \in \operatorname*{argmin}_f R(f) = \mathbb{E}\{\ell_{0\text{-}1}(Y, f(X))\}$ is given by $f^*(x) = +1$ if $\mathbb{P}(Y=1 \mid X=x) \geq 1/2$, and 0 otherwise, leading to the notion of excess risk

$$\mathcal{E}(\hat{f}_n) := R(f_n) - R^*,$$

which can be further decomposed into a sum of two terms:

$$\mathcal{E}(\hat{f}_n) = \underbrace{\left\{R(f_n) - \inf_{f \in \mathcal{F}} R(f)\right\}}_{\text{estimation error}} + \underbrace{\left\{\inf_{f \in \mathcal{F}} R(f) - R^*\right\}}_{\text{approximation error}}$$

- Vapnik-Chervonenkis (VC) theory is dealing with the estimation error: for a given class $\mathcal{F}$ of candidate functions, can we get theoretical guarantees that with high probability, the estimation error remains small. More formally, can we construct a function $n_{\mathcal{F}} : (0,1)^2 \to \mathbb{N}$ such that $\forall (\varepsilon, \delta) \in (0,1)^2$, $\forall n \geq n_{\mathcal{F}}(\varepsilon, \delta)$, $\forall \mathbb{P}_{X,Y}$,

$$R(f_n) - \inf_{f \in \mathcal{F}} R(f) \leq \varepsilon \quad \text{with probability} \geq 1 - \delta.$$

"PAC" learnability
Probably Approximately Correct
$(\geq 1-\delta)$ $(\leq \varepsilon)$

- Remark: definitions in textbooks differ, but the function $n_{\mathcal{F}}(\varepsilon, \delta)$ is usually required to be at most polynomial in $1/\varepsilon$ and $1/\delta$.

- Notation: we write $\bar{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} R(f)$

We have the following decomposition: (3)

$$R(f_n) = R(f_n) + \underbrace{\hat{R}_n(f_n) - \hat{R}_n(f_n)} + R(\bar{f}) - R(\bar{f})$$

$f_n = \text{ERM}$
$\Rightarrow \forall f \in \mathcal{F},\ \hat{R}_n(f_n) \leq \hat{R}_n(f)$.
In particular,
$\hat{R}_n(f_n) \leq \hat{R}_n(\bar{f})$.

$$R(f_n) \leq R(f_n) + \hat{R}_n(\bar{f}) - \hat{R}_n(f_n) + R(\bar{f}) - R(\bar{f}).$$

group terms together

$$= R(\bar{f}) + \{\hat{R}_n(\bar{f}) - R(\bar{f})\} + \{R(f_n) - \hat{R}_n(f_n)\}$$

"sup out" $\bar{f}$ and f_n:

(***)

$$\boxed{R(f_n) \leq R(\bar{f}) + 2 \sup_{f \in \mathcal{F}} |R(f) - \hat{R}_n(f)|}$$

We got rid of the term $\hat{R}_n(f_n)$: difficult to handle, since both f_n and $\hat{R}_n$ depend on the same $\mathcal{L}_n$.

We are loosing a lot of information by taking the supremum over $\mathcal{F}$. However, it turns out to be a surprisingly accurate tool, as we shall see

$\Rightarrow$ The estimation error can be bounded by controlling the size of $\sup_{f \in \mathcal{F}} |R(f) - \hat{R}_n(f)|$. For a fixed $f \in \mathcal{F}$, since $\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell_{0\text{-}1}(Y_i, f(X_i)) \xrightarrow{a.s.} R(f)$, we need to study how fast $\hat{R}_n$ concentrates around its mean.

- Preliminary (not fully satisfactory) answers: (4)

↘ Answer #1: Use Markov / Chebyshev inequality:

$$\mathbb{P}(|X - \mathbb{E}X| \geq \varepsilon) \leq \frac{\operatorname{Var} X}{\varepsilon^2}, \text{ for any RV } X.$$

Taking $X = \hat{R}_n(f)$, we obtain

$$\mathbb{P}(|\hat{R}_n(f) - R(f)| \geq \varepsilon) \leq \frac{\operatorname{Var} \hat{R}_n(f)}{\varepsilon^2} = \frac{\sigma_L^2}{n\varepsilon^2},$$

where $\sigma_L^2 := \operatorname{Var}\{\ell(Y, f(X))\}$.

↳ We obtain a rate of decay of n^{-1}. We can obtain faster rates.

↘ Answer #2: Use the Central Limit Theorem (CLT),

$$\frac{n^{1/2}(\hat{R}_n(f) - R(f))}{\sigma_L} \xrightarrow{d} Z \sim \mathcal{N}(0,1),$$

where for $x > 0$,

$$\mathbb{P}(|Z| > x) = 2\int_x^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

$\frac{u}{x} > 1$

$$\leq 2 \int_x^{+\infty} \frac{u}{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

$$= \sqrt{\frac{2}{\pi}} \frac{1}{x} e^{-\frac{x^2}{2}},$$

since

$$\int_x^{+\infty} u e^{-u^2/2} du = \left[-e^{-u^2/2}\right]_x^{+\infty} = e^{-x^2/2}$$

Thus $\mathbb{P}(|\hat{R}_n(f) - R(f)| \geq \varepsilon)$

$$= \mathbb{P}\left(\frac{n^{1/2}|\hat{R}_n(f) - R(f)|}{\sigma_L} \geq \frac{n^{1/2}\varepsilon}{\sigma_L}\right)$$

$$\simeq \mathbb{P}\left(|Z| \geq \frac{n^{1/2}\varepsilon}{\sigma_L}\right)$$

So that

$$\mathbb{P}(|\hat{R}_n(f) - R(f)| \geq \varepsilon) \lesssim \sqrt{\frac{2}{\pi}}\, \frac{\sigma_L}{n^{1/2}\varepsilon} \exp\left\{-\frac{1}{2}\frac{n\varepsilon^2}{\sigma_L}\right\}.$$

The tail is expected to shrink exponentially fast : e^{-n}. We make this result precise next.

- Summary : To control the estimation error, it is enough to control the quantity $\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)|$, and

for this we need

(i) to know how fast $\hat{R}_n(f)$ concentrates around its mean $R(f)$ (red term)

(ii) how large the class $\mathcal{F}$ is ; in order to take care of the supremum (blue term). If $\mathcal{F}$ contains finitely many elements, this shouldn't be to difficult to handle. However, in most practical cases, the class $\mathcal{F}$ contains uncountably many elements { ex: class of linear functions $x \mapsto \beta_0 + \beta^t x$; $\beta_0 \in \mathbb{R}$, $\beta \in \mathbb{R}^d$ } $\Rightarrow$ we need to introduce a new notion of class complexity. As we will see later, this is captured by the so-called Vapnik Chervonenkis (VC) dimension of $\mathcal{F}$.

→ Section I treats the case of a finite dictionary $\mathcal{F}$.

→ Section II consider the case of uncountably infinite classes of functions.

→ Section III discusses Structural Risk Minimization (SRM)

→ In Section IV, we briefly indicate how these results can be generalized in the context of a general loss.

I. LEARNING WITH A FINITE $\mathcal{F}$

I.1. Hoeffding's inequalities.

Theorem (Hoeffding)

Let $X_1, \dots, X_n$ be independent, bounded RVs, such that $X_i \in [a_i, b_i]$ with probability one.

Put $S_n = \sum_{i=1}^n X_i$. Then, $\forall \varepsilon > 0$,

(i) $\mathbb{P}(S_n - \mathbb{E}S_n \geq \varepsilon) \leq \exp\left\{-\frac{2\varepsilon^2}{\sum(b_i-a_i)^2}\right\}$

(ii) $\mathbb{P}(S_n - \mathbb{E}S_n \leq -\varepsilon) \leq \exp\left\{-\frac{2\varepsilon^2}{\sum(b_i-a_i)^2}\right\}$

(iii) $\mathbb{P}(|S_n - \mathbb{E}S_n| \geq \varepsilon) \leq 2\exp\left\{-\frac{2\varepsilon^2}{\sum(b_i-a_i)^2}\right\}$

proof : Let X be any RV, and $s > 0$. Using Markov inequality,

$$\mathbb{P}(X \geq \varepsilon) = \mathbb{P}(e^{sX} \geq e^{s\varepsilon}) \leq e^{-s\varepsilon}\,\mathbb{E}(e^{sX}).$$

Taking $X = \sum_{i=1}^n (X_i - \mathbb{E}X_i) = S_n - \mathbb{E}S_n$, we have

$$\begin{aligned}
\mathbb{P}(S_n - \mathbb{E}S_n \geq \varepsilon) &\leq e^{-s\varepsilon}\,\mathbb{E}\left\{\exp\left(s\left(\sum[X_i - \mathbb{E}X_i]\right)\right)\right\} \\
&= e^{-s\varepsilon}\,\mathbb{E}\left\{\prod_{i=1}^n \exp(s[X_i - \mathbb{E}X_i])\right\} \\
&\overset{\text{independence}}{=} e^{-s\varepsilon}\prod_{i=1}^n \mathbb{E}\left\{\exp(s[X_i - \mathbb{E}X_i])\right\}.
\end{aligned}$$

We need a bound for this term.

Hoeffding's Lemma

Let Y be a RV such that $\mathbb{E}Y = 0$, and $a \leq Y \leq b$ almost surely. Then

$$\mathbb{E}\{e^{sY}\} \leq \exp\left\{\frac{s^2(b-a)^2}{8}\right\}$$

Making use of Hoeffding's lemma, we obtain

$$\mathbb{P}(S_n - \mathbb{E}S_n \geq \varepsilon) \leq e^{-s\varepsilon} \prod_{i=1}^{n} \exp\left\{ s^2 \frac{(b_i-a_i)^2}{8} \right\}$$

$$= \exp\left\{ -s\varepsilon + \frac{s^2}{8}\sum_{i=1}^{n}(b_i-a_i)^2 \right\}$$

Minimization of $\{\dots\}$ with respect to s yields

$s = 4\varepsilon / \sum(b_i-a_i)^2$.

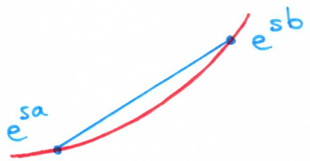
For this choice of s, we obtain,

$$\mathbb{P}(S_n - \mathbb{E}S_n \geq \varepsilon) \leq \exp\left\{ -\frac{2\varepsilon^2}{\sum(b_i-a_i)^2} \right\},$$

as required. We prove (ii) and (iii) in a similar way. ■

- Proof of Hoeffding's lemma.

By convexity of the exponential,
$\forall y \in [a,b]$,



$$e^{sy} \leq \lambda e^{sa} + (1-\lambda) e^{sb}, \quad 0 \leq \lambda \leq 1.$$

Taking $\lambda = \frac{b-y}{b-a}$; $1-\lambda = \frac{y-a}{b-a}$,

$$e^{sy} \leq \left(\frac{b-y}{b-a}\right) e^{sa} + \left(\frac{y-a}{b-a}\right) e^{sb}$$

Taking $\mathbb{E}\{\dots\}$

$$\mathbb{E}\{e^{sY}\} \leq \mathbb{E}\left(\frac{b-Y}{b-a}\right) e^{sa} + \mathbb{E}\left(\frac{Y-a}{b-a}\right) e^{sb}$$

$$= \frac{b}{b-a} e^{sa} - \frac{a}{b-a} e^{sb}$$

Put $p := -\frac{a}{b-a}$

$1-p = \frac{b}{b-a}$

$$= (1-p) e^{sa} + p e^{sb}$$

$$= [1 - p + p e^{s(b-a)}] e^{sa}$$

$$\Rightarrow \mathbb{E}\{e^{sY}\} \leq [1-p+pe^{s(b-a)}] e^{-ps(b-a)}$$

Put $u := s(b-a)$, and define

$\varphi(u) := -pu + \log(1-p+pe^u)$

Then $\mathbb{E}\{e^{sY}\} \leq e^{\varphi(u)}$.

We now optimize the upper bound.

Consider a Taylor expansion of φ;

$$\varphi(u) = \underset{0}{\underset{\parallel}{\varphi(0)}} + u\varphi'(0) + \frac{1}{2}u^2\varphi''(v), \text{ for some } v \in [0,u]$$

with

$$\varphi'(u) = -p + \frac{pe^u}{1-p+pe^u} \Rightarrow \varphi'(0) = 0$$

$$\varphi''(u) = \frac{pe^u(1-p+pe^u) - p^2e^{2u}}{(1-p+pe^u)^2}$$

$$= \frac{pe^u}{1-p+pe^u}\left(1 - \frac{pe^u}{1-p+pe^u}\right)$$

Put $\rho := \frac{pe^u}{1-p+pe^u}$

$$= \rho(1-\rho) \leq 1/4$$

We conclude that $\varphi(u) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}$,

and $\mathbb{E}\{e^{sY}\} \leq \exp\left\{\frac{s^2(b-a)^2}{8}\right\}$ follows. ■

I.2. Oracle Inequalities for finite classes $\mathcal{F}$

A direct application of Hoeffding's inequalities yield:

$$\mathbb{P}(|\hat{R}_n(f) - R(f)| \geq \varepsilon)$$

Fixed function in $\mathcal{F}$

$$= \mathbb{P}\left(\left| \frac{1}{n}\sum_{i=1}^{n} \ell_{01}(Y_i, f(X_i)) - \underset{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{\ell_{01}(Y_i, f(X_i))\}}{\underset{\parallel}{\mathbb{E}\,\ell_{01}(Y, f(X))}} \right| \geq \varepsilon \right)$$

$$\Rightarrow \mathbb{P}\left(|\hat{R}_n(f) - R(f)| \geq \varepsilon\right)$$
$$= \mathbb{P}\left(\left|\sum_{i=1}^{n} \ell_{0\text{-}1}(Y_i, f(X_i)) - \mathbb{E}\,\ell_{0\text{-}1}(Y_i, f(X_i))\right| \geq n\varepsilon\right)$$
$$\leq 2\exp\left\{-\frac{2(n\varepsilon)^2}{n}\right\}$$
$$= 2\exp\{-2n\varepsilon^2\}$$

For a finite class of functions $\mathcal{F}$ with $|\mathcal{F}|$ elements, we get

$$\mathbb{P}\left(\max_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)| \geq \varepsilon\right)$$
$$= \mathbb{P}\left(\bigcup_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)| \geq \varepsilon\right)$$

"Union Bound" (sub-additivity)

$$\leq \sum_{f\in\mathcal{F}} \mathbb{P}\left(|\hat{R}_n(f) - R(f)| \geq \varepsilon\right)$$
$$\leq 2|\mathcal{F}|\exp\{-2n\varepsilon^2\}.$$

Put $\delta := 2|\mathcal{F}|\exp\{-2n\varepsilon^2\}$.

Then $\log\delta = \log(2|\mathcal{F}|) - 2n\varepsilon^2$

$$\varepsilon = \left(\log\left\{\frac{2|\mathcal{F}|}{\delta}\right\} / 2n\right)^{1/2}, \text{ and}$$

(1)
$$\boxed{\forall f\in\mathcal{F},\quad |\hat{R}_n(f) - R(f)| \leq \sqrt{\frac{\log\left(\frac{2|\mathcal{F}|}{\delta}\right)}{2n}} \quad \text{with probability} \geq 1-\delta.}$$

Moreover, $R(f_n) \leq R(\bar{f}) + 2\max_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)|$

(2)
$$\boxed{R(f_n) \leq R(\bar{f}) + \sqrt{\frac{2}{n}\log\left(\frac{2|\mathcal{F}|}{\delta}\right)} \quad \text{w.p.} \geq 1-\delta}$$

Equation (2) is known as an oracle inequality. The logarithmic dependence on $|\mathcal{F}|$ implies that we can increase the size of $|\mathcal{F}|$ exponentially fast with n, and maintain the same accuracy.

The bound (2) does not depend on the underlying distribution $\mathbb{P}_{X,Y}$: it is a distribution free bound $\rightarrow$ AGNOSTIC learning.

↳ It may be used to answer questions such as: "How many observations do we need in order to achieve a certain level of accuracy".

A consequence of (2) is that finite classes of functions are PAC learnable.

We may consider a variant of (1) and use a one-sided inequality; using the 2nd Hoeffding's inequality:

$$\mathbb{P}\left(\max_{f\in\mathcal{F}} \{R(f) - \hat{R}_n(f)\} \geq \varepsilon\right) \leq |\mathcal{F}|\exp\{-2n\varepsilon^2\}.$$

$\Leftrightarrow$

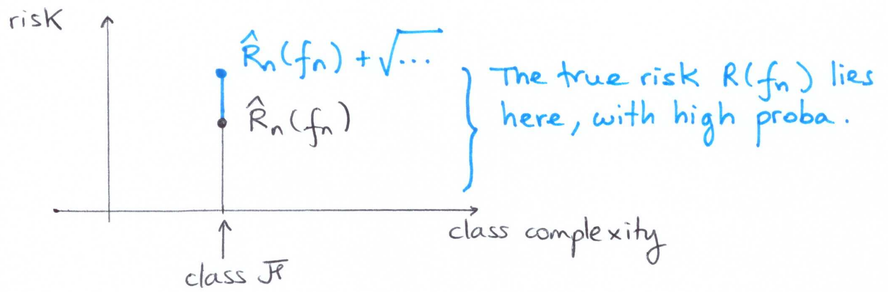
$$\max_{f\in\mathcal{F}} \{R(f) - \hat{R}_n(f)\} \leq \varepsilon \quad \text{w.p.} \geq 1 - |\mathcal{F}|\exp\{-2n\varepsilon^2\}.$$

Put $\delta = |\mathcal{F}|e^{-2n\varepsilon^2}$; $\varepsilon = \sqrt{\frac{\log\frac{|\mathcal{F}|}{\delta}}{2n}}$,

(3)
$$\boxed{\forall f\in\mathcal{F},\ \forall\delta>0,\quad R(f) \leq \hat{R}_n(f) + \sqrt{\frac{\log(|\mathcal{F}|/\delta)}{2n}} \quad \text{with probability} \geq 1-\delta}$$

In particular, true for $f = f_n$; the ERM.

Relation (3) may be used to correct the training error $\hat{R}_n(f_n)$ by an amount equal to $\left(\log\left(\frac{|\mathcal{F}|}{\delta}\right)/2n\right)^{1/2}$ to get a more reliable estimate of the test error ($\equiv$ the true risk $R(f_n)$).



↳ We discuss this further in Section III, when introducing Structural Risk Minimization (SRM).

- **Remark**: Inequality for $\mathbb{E}\{R(f_n)\} - R(\bar{f})$.

Recall from page 3 that $R(f_n) - R(\bar{f}) \leq 2\sup|\hat{R}_n(f) - R(f)|$

$$\Rightarrow \mathbb{E}\{R(f_n)\} - R(\bar{f}) \leq 2\,\mathbb{E}\{\underbrace{\sup|\hat{R}_n(f) - R(f)|}_{\text{We bound this term}}\}$$

Since $\mathcal{F}$ is finite, we can enumerate all its elements: $\mathcal{F} = \{f_1, \dots, f_{|\mathcal{F}|}\}$. Put $\begin{cases} Z_j = R(f_j) - \hat{R}_n(f_j) \\ Z_{|\mathcal{F}|+j} = -Z_j \end{cases}$

Then

$$\mathbb{E}\left\{\max_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)|\right\} = \mathbb{E}\left\{\max_{1\leq j\leq 2|\mathcal{F}|} Z_j\right\}$$

$$= \frac{1}{s}\log\exp\left\{s\,\mathbb{E}\max_j Z_j\right\}$$

$$\mathbb{E}\left\{\max_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)|\right\} \leq \frac{1}{s}\log\mathbb{E}\, e^{s\max Z_j}$$

$$\leq \frac{1}{s}\log\mathbb{E}\sum_{j=1}^{2|\mathcal{F}|} e^{sZ_j}$$

$$= \frac{1}{s}\log\sum_{j=1}^{2|\mathcal{F}|}\mathbb{E}\, e^{sZ_j}$$

We bound each term $\mathbb{E}\, e^{sZ_j}$ using Hoeffding's lemma.

Since $Z_j = \frac{1}{n}\sum_{i=1}^{n}\{\mathbb{E}[\ell(Y_i, f_j(X_i))] - \ell(Y_i, f_j(X_i))\}$, $1\leq j\leq|\mathcal{F}|$

we get

$$\mathbb{E}\{e^{sZ_j}\} = \mathbb{E}\left\{\exp\frac{s}{n}\sum_{i=1}^{n}[\mathbb{E}\,\ell(Y_i, f_j(X_i)) - \ell(Y_i, f_j(X_i))]\right\}$$

$$= \mathbb{E}\left\{\prod_{i=1}^{n}\exp\left(\frac{s}{n}[\mathbb{E}\,\ell(Y_i, f_j(X_i)) - \ell(Y_i, f_j(X_i))]\right)\right\}$$

independence

$$= \prod_{i=1}^{n}\mathbb{E}\left\{\frac{s}{n}[\mathbb{E}\,\ell(Y_i, f_j(X_i)) - \ell(Y_i, f_j(X_i))]\right\}$$

The 0-1 loss takes values in $\{0,1\}$ $\Rightarrow$ the term $\frac{1}{n}[\dots] \in [-\frac{1}{n}, \frac{1}{n}]$. In addition, since both Z_j and $-Z_j$ appear when we sum all the terms from 1 to $2|\mathcal{F}|$, necessary one $\frac{1}{n}[\dots]$ lies in $[0, \frac{1}{n}]$, while the other lies in $[-\frac{1}{n}, 0]$. Applying Hoeffding's lemma yields

$$\mathbb{E}\left\{\exp\frac{s}{n}[\dots]\right\} \leq \exp\left\{\frac{s^2}{8}\left(\frac{1}{n}\right)^2\right\} = \exp\left\{\frac{s^2}{8n^2}\right\}$$

$$\Rightarrow \mathbb{E}\left\{\max|\hat{R}_n(f) - R(f)|\right\} \leq \frac{1}{s}\log\left\{2|\mathcal{F}|\exp\left(\frac{s^2}{8n}\right)\right\}$$

$$= \frac{\log(2|\mathcal{F}|)}{s} + \frac{s}{8n}$$

Optimize with respect to s: $s = 2\sqrt{2n\log(2|\mathcal{F}|)}$

(4) $$\boxed{\mathbb{E}\{R(f_n)\} \leq R(\bar{f}) + \sqrt{\frac{2\log(2|\mathcal{F}|)}{n}}}$$

- Remark: Alternatively, to get a bound on the expected estimation error, we can use relation (3) page 10,

$$\forall f \in \mathcal{F} \quad \forall \delta > 0 \quad R(f) \leq \hat{R}_n(f) + \underbrace{\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{2n}}}_{=:\, \mathcal{E}(\mathcal{F},n,\delta)} \quad \text{w.p.} \geq 1-\delta$$

Inequality holds true in particular for the empirical risk minimizer $f = f_n$:

$$R(f_n) \leq \hat{R}_n(f_n) + \mathcal{E}(\mathcal{F},n,\delta) \quad \text{w.p.} \geq 1-\delta$$

$$\leq \hat{R}_n(\bar{f}) + \mathcal{E}(\mathcal{F},n,\delta) \quad \text{by definition of } f_n$$

Let A = event on which this equality holds: $\mathbb{P}(A) \geq 1-\delta$.

$$\mathbb{E}\, R(f_n) - R(\bar{f}) = \mathbb{E}\{ R(f_n) - \hat{R}_n(\bar{f}) \}$$

$$= \mathbb{E}\{ R(f_n) - \hat{R}_n(\bar{f}) \mid A \} \overbrace{\mathbb{P}(A)}^{\leq 1}$$

$$+ \underbrace{\mathbb{E}\{ R(f_n) - \hat{R}_n(\bar{f}) \mid \bar{A} \}}_{\leq 1} \underbrace{\mathbb{P}(\bar{A})}_{\leq \delta}$$

$$\leq \mathbb{E}\{ R(f_n) - \hat{R}_n(\bar{f}) \mid A \} + \delta$$

$$\leq \mathcal{E}(\mathcal{F},n,\delta) + \delta$$

The choice of δ is arbitrary. Since $\mathcal{E}(\mathcal{F},n,\delta)$ is of order $n^{-1/2}$, take $\delta = n^{-1/2}$.

$$\Rightarrow \mathbb{E}\{R(f_n)\} - R(\bar{f}) \leq \sqrt{\frac{\log|\mathcal{F}| + \frac{1}{2}\log n}{2n}} + \sqrt{\frac{1}{n}}$$

$\sqrt{x} + \sqrt{y} \leq \sqrt{2}\sqrt{x+y} \quad \forall x,y > 0$

$$\leq \sqrt{\frac{\log|\mathcal{F}| + \frac{1}{2}\log n + 2}{n}}$$

$$\boxed{\mathbb{E}\{R(f_n)\} - R(\bar{f}) = O\left(\sqrt{\frac{\log(n|\mathcal{F}|)}{n}}\right)}$$

We are "loosing" a $\log n$ factor!

II. VAPNIK-CHERVONENKIS THEORY

II.1. Step I: Concentration Inequalities.

In this section, we prove the bounded difference inequality (McDiarmid); a concentration inequality that generalizes Hoeffding's inequality.

Theorem (McDiarmid, 1989)

Let $g: \mathcal{X}^n \to \mathbb{R}$, and constants $c_1, \dots, c_n \geq 0$ such that

$$\sup_{x_1,\dots,x_n,x_i'} | g(x_1,\dots,x_i,\dots,x_n) - g(x_1,\dots,x_i',\dots,x_n) | \leq c_i$$

$\forall i \in \{1,\dots,n\}$ — such a g is said to satisfy the bounded difference assumption.

Then for any independent RVs $X_1,\dots,X_n$, $\forall \varepsilon > 0$,

$$\mathbb{P}\left(| g(X_1,\dots,X_n) - \mathbb{E}\{g(X_1,\dots,X_n)\} | > \varepsilon \right) \leq 2\exp\left\{ -\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2} \right\}$$

proof: Put $V_i := \mathbb{E}\{ g(X_1,\dots,X_n) \mid X_1,\dots,X_i \} - \mathbb{E}\{ g(X_1,\dots,X_n) \mid X_1,\dots,X_{i-1} \}$

$\{V_i\}_{i=1,\dots,n}$ is called a Martingale Difference Sequence, since $\mathbb{E}\{ V_i \mid X_1,\dots,X_{i-1} \} = 0$.

Note that

$$g(X_1,\dots,X_n) - \mathbb{E}\{g(X_1,\dots,X_n)\} = \sum_{i=1}^n V_i$$

Telescoping sum.

Fix $\varepsilon > 0$, and $s > 0$.

$$\begin{aligned}\mathbb{P}\big(g(X_1,..,X_n) - \mathbb{E}\,g(X_1,..,X_n) > \varepsilon\big) &= \mathbb{P}\Big(\sum_{i=1}^n V_i > \varepsilon\Big)\\ &\leqslant e^{-s\varepsilon}\,\mathbb{E}\Big\{\exp\Big(s\sum_{i=1}^n V_i\Big)\Big\}\\ &= e^{-s\varepsilon}\,\mathbb{E}\Big\{\prod_{i=1}^n \exp(sV_i)\Big\} \qquad (*)\end{aligned}$$

sequence of dependent RVs.

To get a bound on $\mathbb{E}\{\prod_{i=1}^n \exp(sV_i)\}$, we introduce

$$L_i := \inf_x \Big(\mathbb{E}\{g(X_1,..,X_n) \mid X_1,..,X_{i-1}, X_i = x\}\Big) - \mathbb{E}\{g(X_1,..,X_n) \mid X_1,..,X_{i-1}\}$$

$$U_i := \sup_{x'} \Big(\mathbb{E}\{g(X_1,..,X_n) \mid X_1,..,X_{i-1}, X_i = x'\}\Big) - \mathbb{E}\{g(X_1,..,X_n) \mid X_1,..,X_{i-1}\}.$$

We see that $L_i \leqslant V_i \leqslant U_i$ a.s.. Moreover,

$$\begin{aligned}U_i - L_i &= \sup_x(\dots) - \inf_{x'}(\dots)\\ &= \sup_{x,x'} \int \big\{g(X_1,..,X_{i-1}, x, x_{i+1},..,x_n) - g(X_1,..,X_{i-1}, x', x_{i+1},..,x_n)\big\}\, d\mathbb{P}(x_{i+1},..,x_n)\\ &\leqslant c_i. \quad \text{(by assumption)}\end{aligned}$$

Thus, $\mathbb{E}\{\prod_{i=1}^n \exp(sV_i)\}$

$$\begin{aligned}&= \mathbb{E}\Big\{\mathbb{E}\Big(\prod_{i=1}^n e^{sV_i} \mid X_1,..,X_{n-1}\Big)\Big\}\\ &= \mathbb{E}\Big\{\prod_{i=1}^{n-1} e^{sV_i}\,\mathbb{E}\big(e^{sV_n} \mid X_1,..,X_{n-1}\big)\Big\}\end{aligned}$$

$V_i =$ function of $X_1,..,X_i$

We can use Hoeffding's lemma to bound the term $\mathbb{E}(e^{sV_n} \mid X_1,..,X_{n-1})$ since

- $\mathbb{E}(V_n \mid X_1,..,X_{n-1}) = 0$
- $V_n \in [L_n, U_n]$ a.s.; an interval of length bounded by c_n.

$$\Rightarrow \mathbb{E}\big(e^{sV_n} \mid X_1,..,X_{n-1}\big) \leqslant \mathbb{E}\Big(\exp\frac{s^2c_n^2}{8}\Big).$$

Thus,

$$\begin{aligned}\mathbb{E}\Big\{\prod_{i=1}^n e^{sV_i}\Big\} &\leqslant e^{\frac{s^2c_n^2}{8}}\,\mathbb{E}\Big\{\prod_{i=1}^{n-1} e^{sV_i} \mid X_1,..,X_{n-1}\Big\}\\ &= e^{\frac{s^2c_n^2}{8}}\,\mathbb{E}\Big\{\prod_{i=1}^{n-2} e^{sV_i}\,\mathbb{E}\big(e^{sV_{n-1}} \mid X_1,..,X_{n-2}\big)\Big\}\\ &\leqslant \exp\Big\{\frac{s^2}{8}(c_{n-1}^2 + c_n^2)\Big\}\,\mathbb{E}\Big\{\prod_{i=1}^{n-2} e^{sV_i}\Big\}\\ &\;\;\vdots\\ &\leqslant \exp\Big\{\frac{s^2}{8}\sum_{i=1}^n c_i^2\Big\}.\end{aligned}$$

$$\Rightarrow \mathbb{P}\big(g(X_1,..,X_n) - \mathbb{E}\,g(X_1,..,X_n) > \varepsilon\big) \leqslant \exp\Big\{-s\varepsilon + \frac{s^2}{8}\sum_{i=1}^n c_i^2\Big\}$$

$(*)$ page 15.

Optimise the upper bound with respect to s gives $s = 4\varepsilon / \sum_{i=1}^n c_i^2$, from which we obtain

$$\mathbb{P}\big(g(X_1,..,X_n) - \mathbb{E}\,g(X_1,..,X_n) > \varepsilon\big) \leqslant \exp\Big(\frac{-2\varepsilon^2}{\sum c_i^2}\Big).$$

A similar bound can be obtained for $\mathbb{P}(\dots < -\varepsilon)$; which gives the desired double sided inequality of the theorem. ∎

- Back to our learning problem, recall that the estimation error is bounded by $2 \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|$.
- Consider the function

$$(x_1,y_1),\dots,(x_n,y_n) \mapsto \sup_{f \in \mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}(y_i \neq f(x_i)) - R(f) \right|$$

It satisfies the bounded difference assumption, since changing one of the (x_i, y_i) changes the function only by $1/n \Rightarrow c_i = 1/n \ \forall i$.

$$\Rightarrow \mathbb{P}\left(\left| \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| - \mathbb{E}\left(\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \right) \right| > \varepsilon \right) \leq \underbrace{2\exp(-2n\varepsilon^2)}_{=:\delta}$$

$$\Leftrightarrow \varepsilon = \sqrt{\frac{\log(2/\delta)}{2n}}.$$

$$\left| \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| - \mathbb{E}\left\{ \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \right\} \right| \leq \sqrt{\frac{\log(2/\delta)}{2n}} \quad \text{w.p.} \geq 1-\delta.$$

(•)

$$\boxed{\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \leq \mathbb{E}\left\{ \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \right\} + \sqrt{\frac{\log(2/\delta)}{2n}}}$$

with probability larger than $1-\delta$.

We only need to focus on the expectation, to get a bound in probability. Also, using this approach, we can only hope for a bound in $O(n^{-1/2})$.

II.2. Step II: Symmetrization & Rademacher complexity.

To bound $\mathbb{E}\{\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|\}$, we use a general technique known as symmetrization. In addition to the learning sample $\mathcal{L}_n = \{(X_1,Y_1),\dots,(X_n,Y_n)\}$, we consider another independent copy $\mathcal{L}'_n = \{(X'_1,Y'_1),\dots,(X'_n,Y'_n)\}$, where $(X'_i, Y'_i) \sim \mathbb{P}_{X,Y}$, iid.

- To start, we express the term $R(f)$ as a conditional expectation:

$$\begin{aligned} R(f) &= \mathbb{E}\, \mathbb{1}(Y \neq f(X)) \\ &= \mathbb{E}\left\{ \frac{1}{n}\sum_{i=1}^n \mathbb{1}(Y'_i \neq f(X'_i)) \right\} \\ &= \mathbb{E}\left\{ \frac{1}{n}\sum_{i=1}^n \mathbb{1}(Y'_i \neq f(X'_i)) \mid \mathcal{L}_n \right\} \\ &= \mathbb{E}\left\{ \hat{R}'_n(f) \mid \mathcal{L}_n \right\}, \end{aligned}$$

where $\hat{R}'_n(f)$ denotes the empirical risk of f based on $\mathcal{L}'_n$: $\hat{R}'_n(f) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(Y'_i \neq f(X'_i))$.

- $\mathbb{E}\left\{ \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \right\}$

$$= \mathbb{E}\left\{ \sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - \mathbb{E}(\hat{R}'_n(f) \mid \mathcal{L}_n) \right| \right\}$$

$\mathcal{L}_n$- measurable

$$= \mathbb{E}\left\{ \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left(\hat{R}_n(f) - \hat{R}'_n(f) \mid \mathcal{L}_n \right) \right| \right\}$$

Jensen

$$\leq \mathbb{E}\left\{ \sup_{f \in \mathcal{F}} \mathbb{E}\left(|\hat{R}_n(f) - \hat{R}'_n(f)| \mid \mathcal{L}_n \right) \right\}$$

Next, note that

$$\forall f \in \mathcal{F}, \quad \mathbb{E}\left(|\hat{R}_n(f) - \hat{R}'_n(f)| \,\middle|\, \mathcal{L}_n\right) \leq \underbrace{\mathbb{E}\left(\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - \hat{R}'_n(f)| \,\middle|\, \mathcal{L}_n\right)}_{\text{the right-hand side is independent of } f}$$

$$\Rightarrow \sup_{f\in\mathcal{F}} \mathbb{E}\left(|\hat{R}_n(f) - \hat{R}'_n(f)| \,\middle|\, \mathcal{L}_n\right) \leq \mathbb{E}\left(\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - \hat{R}'_n(f)| \,\middle|\, \mathcal{L}_n\right),$$

and we get

$$\mathbb{E}\left\{\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)|\right\} \leq \mathbb{E}\left\{\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - \hat{R}'_n(f)|\right\}$$

$$= \mathbb{E}\left\{\sup_{f\in\mathcal{F}} \frac{1}{n}\left| \sum_{i=1}^{n} \underbrace{\left(\mathbb{1}(Y_i \neq f(X_i)) - \mathbb{1}(Y'_i \neq f(X'_i)\right)}_{\text{A symmetric random variable}}\right|\right\}$$

$\Rightarrow$ it has the same distribution as $\sigma_i\,(\mathbb{1}(\dots) - \mathbb{1}(\dots))$, where $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2$.

σ_i is known as a Rademacher RV in the litterature

$$\leq \mathbb{E}\left\{\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i\left(\mathbb{1}(Y_i \neq f(X_i)) - \mathbb{1}(Y'_i \neq f(X'_i))\right)\right|\right\}$$

$$\leq \frac{2}{n}\,\mathbb{E}\left\{\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^{n} \sigma_i\, \mathbb{1}(Y_i \neq f(X_i))\right|\right\}.$$

↳ The sample $\mathcal{L}'_n$ has disappeared from the final expression of the upper bound. For this reason, it is referred to as a <u>ghost sample</u> in the litterature.

Note that the random variable $\sigma_i \mathbb{1}(Y_i \neq f(X_i))$ has zero mean $\rightarrow$ expect the upper bound to vanish as n tends to $+\infty$. Without the introduction of Rademacher RVs, this would not be the case.

To get rid of the dependence of the learning sample $\mathcal{L}_n$ on the upper bound, we "sup-out" the variables $(x_1, y_1), \dots, (x_n, y_n)$, and consider the worst case scenario:

$$\boxed{\mathbb{E}\left\{\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)|\right\} \leq 2 \sup_{(x_1,y_1),\dots,(x_n,y_n)} \underbrace{\mathbb{E}\left\{\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbb{1}(y_i \neq f(x_i))\right|\right\}}}$$

(••)

Alternatively, instead of "supping out" (x_i, y_i), we may consider the conditional expectation with respect to $(X_i, Y_i) = (x_i, y_i)$.

The upper bound (without the factor 2) is known as **RADEMACHER COMPLEXITY** of the class $\mathcal{F}$. We denote it $\mathcal{R}_S(\mathcal{F})$, for $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$.

↓

It captures the ability of a class of functions $\mathcal{F}$ to reproduce $y_i = f(x_i)$, for arbitrary choices of (x_i, y_i).

aka how much "representation power" $\mathcal{F}$ has; aka richness of $\mathcal{F}$.

Loosely speaking, $R_s(\mathcal{F})$ is large whenever we misclassify an observation (x_i, y_i) associated with $\sigma_i = +1$ (note that we have no control over the variables σ_i; and that $\mathbb{E}(\dots)$ is taken with respect to the distribution of $\sigma_1, \dots, \sigma_n$ only), and correctly classify observations (x_i, y_i) associated with $\sigma_i = -1$. Since these values are arbitrary, $R_s(\mathcal{F})$ is large if, given $x_1, \dots, x_n$, we can find an $f \in \mathcal{F}$ that will either correctly classify or misclassify $y_1, \dots, y_n \Rightarrow$

"large $R_s(\mathcal{F})$" $\Leftrightarrow$ "rich class $\mathcal{F}$"
$\Leftrightarrow$ "large upper bound on the estimation erro".

This observation motivates the notion of shattering coefficients.

II.3. Step III: Shattering & VC dimension.

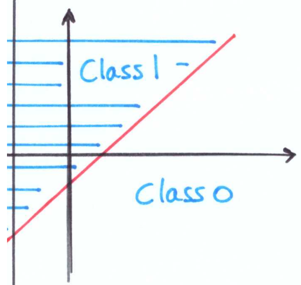
A class $\mathcal{F}$ SHATTERS (pulvérise, in French) points $x_1, \dots, x_n$ if and only if for any $y_1, \dots, y_n \in \{0,1\}^n$, $\exists f \in \mathcal{F}$ which achieves zero training error on $(x_1, y_1), \dots, (x_n, y_n)$; that is $y_i = f(x_i)$, $\forall i = 1, \dots, n$.

Let $S(\mathcal{F}, x_1, \dots, x_n) :=$ number of labelling sequences the class $\mathcal{F}$ induces over $x_1, \dots, x_n$. Since there are exactly 2^n sequences $y_1, \dots, y_n$, we see that necessarily $S(\mathcal{F}, x_1, \dots, x_n) \leq 2^n$.

Put $S(\mathcal{F}, n) = \max\limits_{x_1, \dots, x_n} S(\mathcal{F}, x_1, \dots, x_n) \leq 2^n$

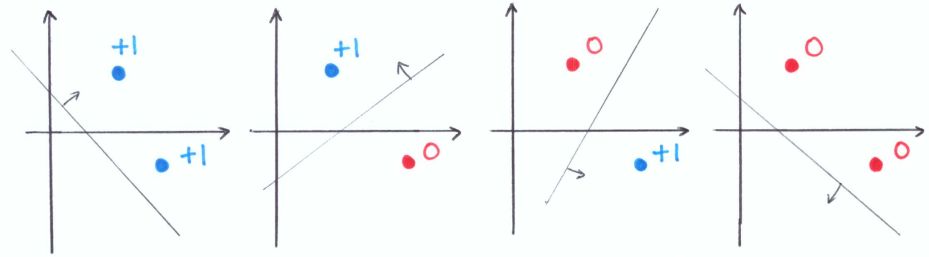
= configuration of n points $x_1, \dots, x_n$ that induces the maximum number of labelling sequences.

× Examples = (i) $\mathcal{F} = \{ x \mapsto \text{sign}(\beta_0 + \beta^t x), \ \beta, x \in \mathbb{R}^2, \beta_0 \in \mathbb{R} \}$

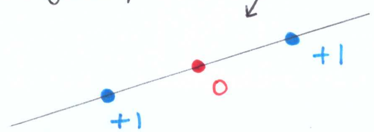


Consider two points $x_1, x_2 \in \mathbb{R}^2$. There are 4 possible labels for x_1 and x_2: $\{(x_1, 0), (x_2, 0)\}$
$\{(x_1, 1), (x_2, 1)\}$
$\{(x_1, 0), (x_2, 1)\}$
$\{(x_1, 1), (x_2, 0)\}$

For each of these 4 cases, it is easy to see that we can find a hyperplane that correctly predicts x_1 and x_2:

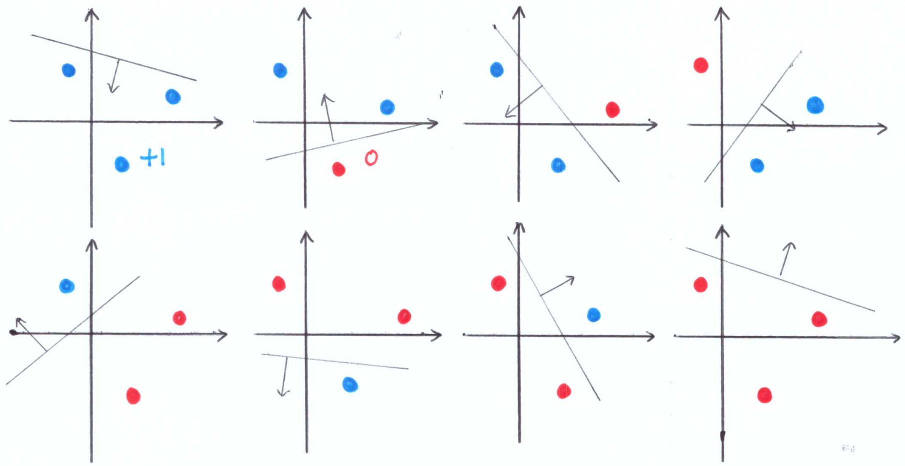


Can $\mathcal{F}$ shatter 3 points?
↳ $\mathcal{F}$ cannot shatter 3 aligned points:

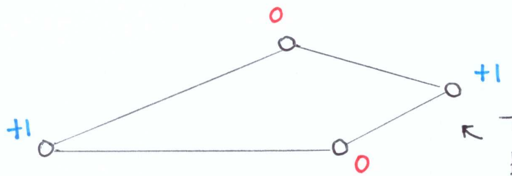


However, $\mathcal{F}$ can shatter any other configuration of 3 points.

Consider 3 non-aligned points.



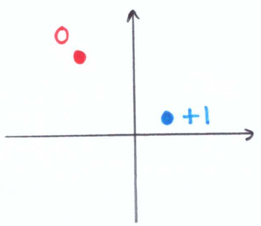
However, $\mathcal{F}$ cannot shatter 4 points:



This configuration cannot be shattered

Observe that the convex hulls of the points labelled +1, and of the points labelled 0, intersect.

(ii) $\mathcal{F} = \{x \mapsto \text{sign}(x^t x - \beta_0),\ x \in \mathbb{R}^d\}$



$\mathcal{F}$ can shatter one point only. With two points, there is always one that is closer to the origin. The configuration on the left cannot be shattered.

The Vapnik Chervonenkis (VC) dimension of $\mathcal{F}$, denoted $VC(\mathcal{F})$, is defined as the maximum number of points that $\mathcal{F}$ can shatter:

$$\exists (x_1, \dots, x_n)\ \forall (y_1, \dots, y_n) \in \{0,1\}^n \quad y_i = f(x_i).$$

Examples: (i) $\mathcal{F} = \{\text{sign}(\beta_0 + \beta^t x),\ \beta, x \in \mathbb{R}^2,\ \beta_0 \in \mathbb{R}\}$

$$\mathcal{S}(\mathcal{F}, 1) = 2^1$$
$$\mathcal{S}(\mathcal{F}, 2) = 2^2$$
$$\mathcal{S}(\mathcal{F}, 3) = 2^3$$
$$\mathcal{S}(\mathcal{F}, 4) < 2^4 \Rightarrow VC(\mathcal{F}) = 3.$$

Note that equivalently, $VC(\mathcal{F})$ is defined as the largest integer k, such that $\mathcal{S}(\mathcal{F}, k) = 2^k$

$\Rightarrow$ To establish that $VC(\mathcal{F}) = d$, one must

(a) Find a configuration $x_1, \dots, x_d$ of d points that $\mathcal{F}$ can shatter $\rightarrow$ usually not too hard

(b) Show that no $(d+1)$ points $x_1, \dots, x_{d+1}$ can be shattered $\rightarrow$ usually harder

(ii) $\mathcal{F} = \{x \mapsto \text{sign}(x^t x - \beta_0),\ x \in \mathbb{R}^2\} \rightarrow VC(\mathcal{F}) = 1$ since $\mathcal{S}(\mathcal{F}, 2) = 3 < 2^2$.

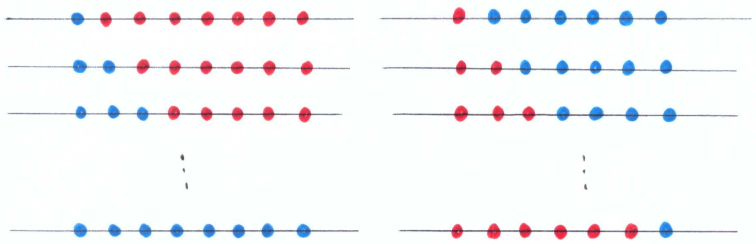
(iii) $\mathcal{F} = \{x \mapsto \mathbb{1}(x > u)$ or $x \mapsto \mathbb{1}(x < u)\}$ $x \in \mathbb{R}$.

It is obvious that $\mathcal{F}$ can shatter up to 2 points. 3 points cannot be shattered:

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

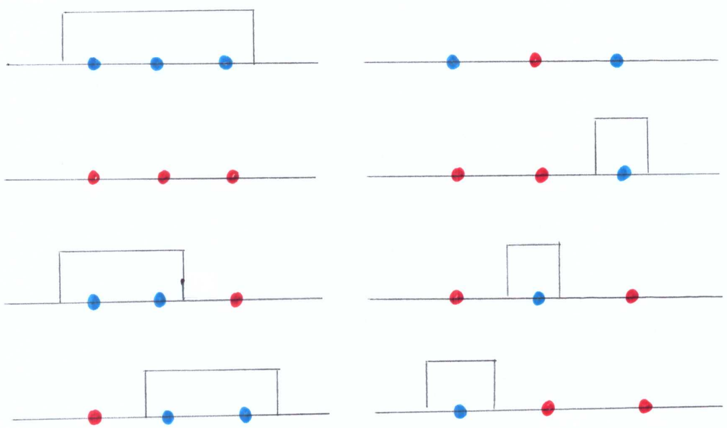
Note that $\forall n$, $\mathcal{S}(\mathcal{F}, n) = 2n \sim$ polynomial
$(= 2^n$ for $n = 1, 2)$.

Indeed,



(iv) $\mathcal{F}$ = class of intervals
$= \{ x \mapsto \mathbb{1}(x \in [a, b]) \,,\ a < b \}$.

Up to 2 points can be shattered



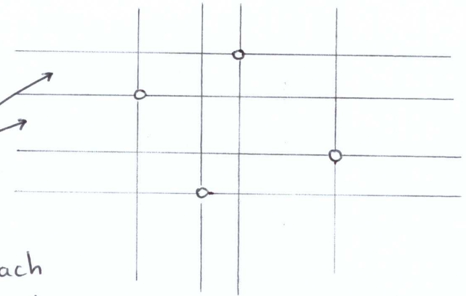
$\mathcal{S}(\mathcal{F}, 3) = 7 < 2^8 \Rightarrow \quad VC(\mathcal{F}) = 2$

For n points, $\mathcal{S}(\mathcal{F}, n) = 1 + \frac{1}{2} n(n+1) \sim$ polynomial
$(= 2^n$ for $n = 1, 2)$

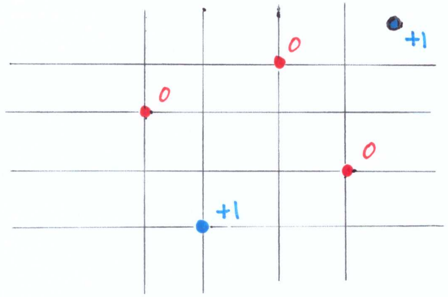
n consecutive ones $\oplus$
n-1 two consecutive +1 $\oplus$ ... / ...

(v) $\mathcal{F}$ = class of rectangles
$= \{ x \mapsto \mathbb{1}(x \in A) \,;\ A = [a,b] \times [c,d],\ a < b,\ c < d,\ x \in \mathbb{R}^2 \}$

Convince yourself that you can shatter up to 4 points



To place the 5th point, there are 25 possible regions, created by 4 non-aligned points. In each case, there is always at least one labelling of the points that rectangles cannot shatter.

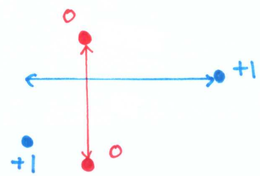


This configuration cannot be shattered: there are two red observations "on the way".

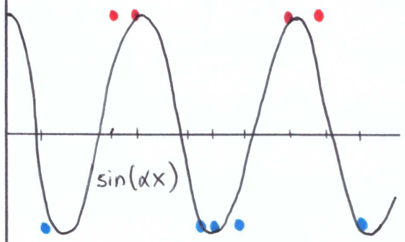
$\Rightarrow VC(\mathcal{F}) = 4$

(vi) $\mathcal{F}$ = class of squares
$= \{ x \mapsto \mathbb{1}(x \in A) \,;\ A = [a, a+b] \times [c, c+b],\ b > 0,\ x \in \mathbb{R}^2 \}$

3 points can be shattered, but not 4. Labelling two opposite points located the furthest away +1 yields a configuration that cannot be shattered.



(vii) $\mathcal{H} : \{ x \mapsto \text{sign} \sin(\alpha x), \quad \alpha > 0, \ x \in \mathbb{R} \}$



Consider labels $y_i \in \{-1, 1\}$.
We show that $\forall n$, there exists $x_1, \dots, x_n$, and $\alpha > 0$ (depending on n), such that $\sin(\alpha x_i) > 0$ if and only if $y_i = +1$.
In other words, we show that $VC(\mathcal{H}) = +\infty$.

"Infinite" representation power; while $\mathcal{H}$ contains only one parameter

Put $z_i := \frac{1 - y_i}{2} \in \{0, 1\}$ $\qquad y_i = +1 \leftrightarrow z_i = 0$, $\quad y_i = -1 \leftrightarrow z_i = +1$

Take $x_1, \dots, x_n$; $x_i = 2^{-i}$

$$\alpha = \pi \left(1 + \sum_{i=1}^{n} 2^i z_i \right)$$

We show that such an α correctly classifies $x_1, \dots, x_n$ irrespectively of their label.

$\forall i = 1, \dots, n$,

$$\alpha x_i = \alpha 2^{-i}$$
$$= \pi \left(2^{-i} + \sum_{j=1}^{n} 2^{j-i} z_j \right)$$
$$= \pi \left(2^{-i} + \sum_{j=1}^{i-1} 2^{j-i} z_j + z_i + \sum_{j=i+1}^{n} 2^{j-i} z_j \right)$$

integer, ≥ 1

a multiple of 2π
$\Rightarrow$ has no effect on $\sin(\alpha x_i)$
$\Rightarrow$ can be dropped

Consider the term

$$\pi \left(2^{-i} + \sum_{j=1}^{i-1} 2^{j-i} z_j + z_i \right) = \pi \left(2^{-i} + z_i + \sum_{k=1}^{i-1} 2^{-k} z_{i-k} \right)$$

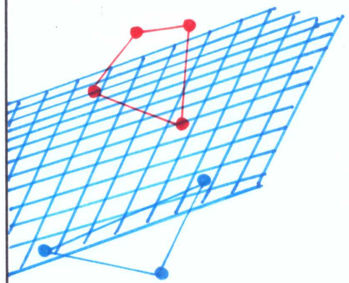
$k = i - j$

$$\Rightarrow \pi z_i < \pi \left(2^{-i} + \sum_{j=1}^{i-1} 2^{j-i} z_j + z_i \right)$$
$$\leq \pi \left(2^{-i} + z_i + \sum_{k=1}^{i-1} 2^{-k} \right)$$
$$= \pi \left(z_i + \underbrace{\sum_{k=1}^{i} 2^{-k}}_{< 1} \right) < \pi (1 + z_i)$$

- If $y_i = +1$, $z_i = 0$, and $0 < \alpha x_i < 1 \pmod{2\pi}$
  $\sin(\alpha x_i) > 0$
  $\text{sign} \sin(\alpha x_i) = +1$
  $\hookrightarrow$ correct prediction.

- If $y_i = -1$, $z_i = +1$, and $\pi < \alpha x_i < 2\pi \pmod{2\pi}$
  $\sin(\alpha x_i) < 0$
  $\text{sign} \sin(\alpha x_i) = -1$
  $\hookrightarrow$ correct prediction.

(viii) $\mathcal{H}$ = class of hyperplanes in $\mathbb{R}^d$
$= \{ x \mapsto \text{sign}(\beta_0 + \beta^t x) ; \ \beta, x \in \mathbb{R}^d, \ \beta_0 \in \mathbb{R} \}$



We proved that the class of hyperplanes in $\mathbb{R}^2$ have VC dimension 3. We now show more generally that in $\mathbb{R}^d$, $VC(\mathcal{H}) = (d+1)$.

$\hookrightarrow$ Consider $(d+1)$ points $x_0, x_1, \dots, x_d$,
where $x_0 := 0$
$x_i := (0, \dots, 0, 1, 0, \dots, 0)^t$
$\uparrow$ i-th coordinate

Take $\beta_0 := \frac{y_0}{2}$ & $\beta = (y_1, \dots, y_d)^t \in \mathbb{R}^d$,
where $y_0, \dots, y_d \in \{-1, 1\}$.
Then $\beta_0 + \beta^t x_i = \frac{y_0}{2} + y_i$, whose sign is equal to the sign of y_i $\Rightarrow$ this choice of (β_0, β) yields a hyperplane

that shatters points $x_0, x_1, \dots, x_d$.

↳ Next, we need to show that no configuration of $(d+2)$ points can be shattered. We need the following lemma.

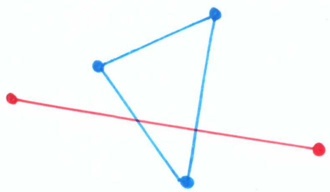
× Radon's lemma.

Any set of $(d+2)$ points in $\mathbb{R}^d$ can be partitioned into two subsets X_1 and X_2 such that the convex hulls of X_1 and X_2 intersect.

proof = Consider

$X = \{x_1, \dots, x_{d+2}\} \subset \mathbb{R}^d$

+ the system of $(d+1)$ linear equations, with $(d+2)$ unknowns:



$$\underbrace{\begin{pmatrix} | & | & & | \\ x_1 & x_2 & \dots & x_{d+2} \\ | & | & & | \\ 1 & 1 & \dots & 1 \end{pmatrix}}_{(d+1)\times(d+2)} \underbrace{\begin{pmatrix} \alpha_1 \\ | \\ \alpha_{d+2} \end{pmatrix}}_{(d+2)\times 1} = 0$$

↳ More unknowns than equations, so there exists a non-zero solution $(\alpha_1^*, -, \alpha_{d+2}^*) = \alpha^*$

Since $\sum_{i=1}^{d+2} \alpha_i^* = 0$, and $\alpha^* \neq 0$, the sets

$I_+ := \{1 \leq i \leq d+2 \mid \alpha_i^* > 0\}$

$I_- := \{1 \leq i \leq d+2 \mid \alpha_i^* \leq 0\}$ are non empty.

Consider

$X_+ := \{x_i \in X \mid i \in I_+\}$

$X_- := \{x_i \in X \mid i \in I_-\}$.

$$\sum_{i=1}^{d+2} \alpha_i^* x_i = 0 \Rightarrow \sum_{i \in I_+} \alpha_i^* x_i = -\sum_{i \in I_-} \alpha_i^* x_i$$

Since $\sum_{j \in I_+} \alpha_j^* > 0$, we have

$$a := \sum_{i \in I_+} \underbrace{\left(\frac{\alpha_i^*}{\sum_{j \in I_+} \alpha_j^*} \right)}_{} x_i = \sum_{i \in I_-} \underbrace{\left(\frac{-\alpha_i^*}{\sum_{j \in I_+} \alpha_j^*} \right)}_{} x_i$$

positive coefficients, that sum to one:

$$\sum_{i \in I_+} \left(\frac{\alpha_i^*}{\sum_{j \in I_+} \alpha_j^*} \right) = \sum_{i \in I_-} \left(\frac{-\alpha_i^*}{\sum_{j \in I_+} \alpha_j^*} \right) = 1$$

Point a lies in the convex hull of X_1 and X_2

$\Rightarrow$ Convex hulls of X_1 and X_2 intersect. ∎

↳ A direct consequence of Radon's lemma is that no configuration of $(d+2)$ points can be shattered by hyperplanes in $\mathbb{R}^d$ (since if two sets of points are separated by a hyperplane, then their convex hulls are also separated by this hyperplane. ∎

Back to our learning problem of page 20: we evaluate Rademacher complexity:

$$\sup_{\substack{(x_1,y_1) \\ (x_n,y_n)}} \mathbb{E} \left\{ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \mathbb{1}(y_i \neq f(x_i)) \right| \right\}$$

For each $(x_1, \dots, x_n)$, denote $\mathcal{F}(x_1, \dots, x_n)$ = smallest subset of $\mathcal{F}$ which gives rise to all possible labellings of $x_1, \dots, x_n$.

(For each value of $(y_1, \dots, y_n)$, take an arbitrary representative in $\mathcal{F}$).

Then $|\mathcal{F}(x_1, \dots, x_n)| \leq \mathcal{S}(\mathcal{F}, n) \leq 2^n$.

A finite class of functions.

Put $Z_f := \sum_{i=1}^{n} \sigma_i \mathbb{1}(y_i \neq f(x_i))$. Then

$$\mathbb{E}\{\exp(s Z_f)\} = \prod_{i=1}^{n} \mathbb{E}\Big\{\exp\big(s \underbrace{\sigma_i \mathbb{1}(y_i \neq f(x_i))}_{\in[-1,1],\ \text{zero mean}}\big)\Big\}$$

Hoeffding's lemma

$$\leq \left[\exp\left(\frac{s^2}{8} 2^2\right)\right]^n = \exp\left(\frac{s^2 n}{2}\right)$$

$$\Rightarrow \mathbb{E}\Big\{\sup_{f\in\mathcal{F}(x_1,\dots,x_n)} |Z_f|\Big\} = \mathbb{E}\Big\{\max_{f\in\mathcal{F}(x_1,\dots,x_n)} |Z_f|\Big\}$$

$$= \frac{1}{s}\log\exp\left(\mathbb{E}\max|Z_f|\right)$$

$$\leq \frac{1}{s}\log \mathbb{E}\{\exp\max|Z_f|\}$$

$$\leq \frac{1}{s}\log \sum_{f\in\bar{\mathcal{F}}} \mathbb{E}\{e^{sZ_f}\}$$

$$\leq \frac{1}{s}\log\left\{|\bar{\mathcal{F}}|\exp\left(\frac{s^2 n}{2}\right)\right\}$$

$$= \frac{\log|\bar{\mathcal{F}}|}{s} + \frac{sn}{2}$$

Put $\bar{\mathcal{F}} := \mathcal{F}(x_1,\dots,x_n) \cup (-\mathcal{F}(x_1,\dots,x_n))$

$|\bar{\mathcal{F}}| \leq 2\,\mathcal{S}(\mathcal{F},n)$

Optimal choice for s is $\sqrt{2n^{-1}\log|\bar{\mathcal{F}}|}$; which gives the bound $\sqrt{2n\log|\bar{\mathcal{F}}|}$. We finally obtain

$$\boxed{\sup_{\substack{(x_1,y_1)\\ \vdots \\ (x_n,y_n)}} \mathbb{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i \mathbb{1}(y_i\neq f(x_i))\right|\right\} \leq \sqrt{\frac{2\log(2\,\mathcal{S}(\mathcal{F},n))}{n}} \quad (\bullet\bullet\bullet)}$$

Remark: At this stage, it is unclear if the bound $(\bullet\bullet\bullet)$ is meaningful: if $\mathcal{S}(\mathcal{F},n) \sim 2^n$ $\forall n$, then the right-hand

side does not vanish. So far, we know that for classes of functions with finite VC dimension, $\mathcal{S}(\mathcal{F},n) < 2^n$ for $n > VC(\mathcal{F})$. However, it may be the case that for some classes $\mathcal{F}$, $\mathcal{S}(\mathcal{F},n) = 2^n - 1$, $\forall n > VC(\mathcal{F})$. The lemma below indicates that fortunately, this never happens.

II.4. Step IV: Sauer lemma.

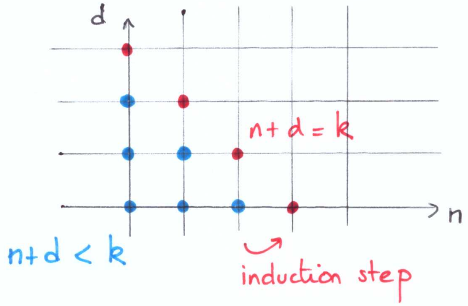
× Sauer lemma. If $d := VC(\mathcal{F}) < +\infty$, then
$$\forall n \geq 1, \quad \mathcal{S}(\mathcal{F},n) \leq \sum_{i=0}^{d}\binom{n}{i}$$

proof: We proceed by induction on $n+d$.

↪ If $(n=0,\ d=0)$

↪ If $(n=1,\ d=0)$ or $(n=0,\ d=1)$, the inequality holds.

Suppose that the inequality holds for $n+d<k$. We want to show that it holds for $n+d=k$. In particular, as we shall see, assuming that the inequality holds for $(n-1,d-1)$ & $(n-1,d)$ is enough to show that it holds for (n,d) [note that the inequality holds also for $d=0$ & any n, and $n=0$ and any d].



Let $\mathcal{F}$ = set of functions $\mathcal{X} \to \{0,1\}$

$S := \{x_1,\dots,x_n\} \subset \mathcal{X}$

$\mathcal{F}_S := \{(f(x_1),\dots,f(x_n)) \mid f\in\mathcal{F}\} \subset \{0,1\}^n$

Put $\mathcal{F}_{1,S} := \{ (y_1, \dots, y_{n-1}) \mid (y_1, \dots, y_{n-1}, 0) \in \mathcal{F}_S$ or $(y_1, \dots, y_{n-1}, 1) \in \mathcal{F}_S \}$

$\mathcal{F}_{2,S} := \{ (y_1, \dots, y_{n-1}) \mid (y_1, \dots, y_{n-1}, 0) \in \mathcal{F}_S$ and $(y_1, \dots, y_{n-1}, 1) \in \mathcal{F}_S \}$.

Ex: $n = 5$.

	$\mathcal{F}_S$					$\mathcal{F}_{1,S}$				$\mathcal{F}_{2,S}$			
	x_1	x_2	x_3	x_4	x_5	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
f_1	0	1	1	0	1	0	1	1	0				
f_2	0	1	1	0	0					0	1	1	0
f_3	0	1	0	1	0	0	1	0	1				
f_4	1	0	1	1	1	1	0	1	1				
f_5	1	0	1	1	0					1	0	1	1
f_6	1	0	1	0	0	1	0	1	0				

labels induced by $f_1, \dots, f_6 \in \mathcal{F}$

↳ Through this example, we see that $|\mathcal{F}_S| = |\mathcal{F}_{1,S}| + |\mathcal{F}_{2,S}|$.

- Note that $VC(\mathcal{F}_{1,S}) \leq VC(\mathcal{F}_S) \leq \underset{VC(\mathcal{F}) = d}{d}$, so that

$$|\mathcal{F}_{1,S}| \leq \mathcal{S}(\mathcal{F}_{1,S}, n-1) \underset{\text{induction hypothesis}}{\leq} \sum_{i=0}^{d} \binom{n-1}{i}$$

- $VC(\mathcal{F}_{2,S}) + 1 \leq VC(\mathcal{F}_S) \leq d \Rightarrow VC(\mathcal{F}_{2,S}) \leq d-1$.

Since if any subset of $n-1$ can be shattered by $\mathcal{F}_{2,S}$, we can add x_n so that $\mathcal{F}_S$ can shatter a strictly larger set of points.

Thus $|\mathcal{F}_{2,S}| \leq \mathcal{S}(\mathcal{F}_{2,S}, n-1) \underset{\text{induction hypothesis}}{\leq} \sum_{i=0}^{d-1} \binom{n-1}{i}$.

$$\Rightarrow |\mathcal{F}_S| = |\mathcal{F}_{1,S}| + |\mathcal{F}_{2,S}| \leq \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i} = \sum_{i=0}^{d} \left\{ \binom{n-1}{i} + \binom{n-1}{i-1} \right\} = \sum_{i=0}^{d} \binom{n}{i},$$

which concludes the proof, since S is arbitrary.

x Consequences: Let $n \geq d$. Then

$$\mathcal{S}(\mathcal{F}, n) \leq \sum_{i=0}^{d} \binom{n}{i} = \sum_{i=0}^{d} \binom{n}{i} \left(\frac{n}{d}\right)^i \left(\frac{d}{n}\right)^i$$
$$\leq \left(\frac{n}{d}\right)^d \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^i$$
$$\leq \left(\frac{n}{d}\right)^d \sum_{i=0}^{n} \binom{n}{i} \left(\frac{d}{n}\right)^i 1^{n-i}$$
$$= \left(\frac{n}{d}\right)^d \left(1 + \frac{d}{n}\right)^n$$
$$\leq \left(\frac{ne}{d}\right)^d$$

since $(1+x) \leq e^x$

For $n \leq d$, $\mathcal{S}(\mathcal{F}, n) = 2^n$ → exp growth
For $n > d$, $\mathcal{S}(\mathcal{F}, n) = O(n^d)$ → polynomial growth

$\mathcal{S}(\mathcal{F}, n)$ can exhibit only two kinds of behaviour. In particular, for classes $\mathcal{F}$ with finite VC dimension, $\mathcal{S}(\mathcal{F}, n)$ eventually grows as a polynomial function of n.

II.5. Step V : VC inequality.

For $n \geqslant d$, we proved that $S(\mathcal{F}, n) \leqslant \left(\frac{ne}{d}\right)^d$, so that $2\, S(\mathcal{F}, n) \leqslant \left(\frac{2ne}{d}\right)^d$, $d \geqslant 1$.

Combining bounds / inequalities (.) page 17
(..) page 20
(...) page 31, (***) p.3

we finally get :

Theorem (VC inequality). (::)

Let $\mathcal{F}$ = family of binary classifiers with finite VC dimension $d \geqslant 1$. Then, $\forall n \geqslant d$,

$$R(\hat{f}_n) - R(\bar{f}) \leqslant 4\sqrt{\frac{2d \log(2en/d)}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}$$

with probability $\geqslant 1-\delta$

VC INEQUALITY

The estimation error is of order $O\left(\sqrt{\frac{\log(n/d)}{n/d}}\right)$ which emphasizes the importance of the ratio n/d for performance.

VC inequality implies that a class of binary classifiers with finite VC dimension is PAC learnable. It turns out that the converse holds as well : if $\mathcal{F}$ is PAC learnable, then it has finite VC dimension (in the context of binary classification). This result is known as the fundamental theorem of Statistical Learning.

Theorem (Fundamental Theorem of Statistical Learning)

Let $\mathcal{F}$ = class of functions $\mathcal{X} \to \{0,1\}$

ℓ = 0-1 loss functions.

Then

$\mathcal{F}$ is PAC learnable $\Leftrightarrow$ $VC(\mathcal{F}) < +\infty$.

proof : $\boxed{\Leftarrow}$ Follows from the VC inequality.

$\boxed{\Rightarrow}$ We show equivalently that $VC(\mathcal{F}) = +\infty \Rightarrow \mathcal{F}$ is not PAC learnable. Suppose by contradiction that $\mathcal{F}$ is PAC learnable. Then there exists an algorithm $\mathcal{A}$, and a function $n_{\mathcal{F}} : (0,1)^2 \to \mathbb{N}$ s.t. $\forall \varepsilon, \delta > 0$, $\forall n \geqslant n_{\mathcal{F}}(\delta, \varepsilon)$, $\forall$ distribution $\mathbb{P}_{X,Y}$,

$$R(\mathcal{A}(\mathcal{L}_n)) \leqslant R(\bar{f}) + \varepsilon, \text{ w.p. } \geqslant 1-\delta. \quad (*)$$

Fix $\varepsilon < 1/8$, $\delta < 6/7$, and take $n \geqslant n_{\mathcal{F}}(\delta, \varepsilon)$.

From the no free lunch theorem p.17 in SL-FOUNDATIONS, there exists $\mathbb{P}_{X,Y}$ and a function $f : \mathcal{X} \to \{0,1\}$ with $R(f) = 0$, such that $R(\mathcal{A}(\mathcal{L}_n)) > 1/8$ with probability $> 1/7$. In the proof of the theorem, we see that this distribution $\mathbb{P}_{X,Y}$ can be supported on a discrete set of size $2n$. Since $VC(\mathcal{F}) = +\infty$, this discrete set can be chosen such that it is shattered by $\mathcal{F}$, so that $R(\bar{f}) = 0$. This contradicts $(*)$, since we showed that $\forall n \geqslant n_{\mathcal{F}}(\delta, \varepsilon)$, there exists a distribution $\mathbb{P}_{X,Y}$ for which

$$R(\mathcal{A}(\mathcal{L}_n)) > \underbrace{R(\bar{f})}_{=0} + 1/8, \text{ w.p. } \geqslant 1/7. \quad \blacksquare$$

III. STRUCTURAL RISK MINIMIZATION (SRM)

Structural Risk Minimization (SRM) is a general technique for model selection. Consider a sequence of classes $\mathcal{F}_1, \mathcal{F}_2, \dots$ of increasing complexity ($\equiv$ increasing VC dimension).
Select $f_n \in \bigcup_k \mathcal{F}_k$ such that

$$\boxed{f_n \in \underset{f \in \bigcup_k \mathcal{F}_k}{\operatorname{argmin}} \left\{ \hat{R}_n(f) + \mathcal{C}(\mathcal{F}_k) \right\}}$$

Penalty term depending on $\mathcal{F}_k$: increases as the representation power of $\mathcal{F}_k$ increases.

A possible way to choose $\mathcal{C}(\mathcal{F}_k)$ is to consider the inequality:

$$\sup_{f \in \mathcal{F}_k} |R(f) - \hat{R}_n(f)| \leq \underbrace{2\sqrt{\frac{2\, d_k \log(2en/d_k)}{n}} + \sqrt{\frac{\log(2/\delta_k)}{2n}}}_{= \mathcal{C}(\mathcal{F}_k, n, \delta_k)}$$

w.p. $\geq 1-\delta_k$

A consequence of relations (.) p. 17, (..) p. 20, (...) p. 31

$$\forall f \in \mathcal{F}_k, \quad R(f) \leq \underbrace{\hat{R}_n(f) + \mathcal{C}(\mathcal{F}_k, n, \delta_k)}_{\text{Pick the upper bound for model selection.}} \quad \text{w.p.} \geq 1-\delta_k.$$

For a collection of candidate models, pick the one that returns the smallest upper bound.

↳ As the number of classes increases, one of the $R(f)$ may lie above the upper band just by chance → we modify the expression of the bound so that it holds simultaneously for all classes.

Put $\mathcal{F} := \bigcup_k \mathcal{F}_k$, $\delta_k = \delta\, 2^{-k}$

& $k(f)$ = smallest integer such that $f \in \mathcal{F}_k$.

We have:

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left\{ |R(f) - \hat{R}_n(f)| - \mathcal{C}(\mathcal{F}_{k(f)}, n, \delta_{k(f)}) \right\} > 0 \right)$$

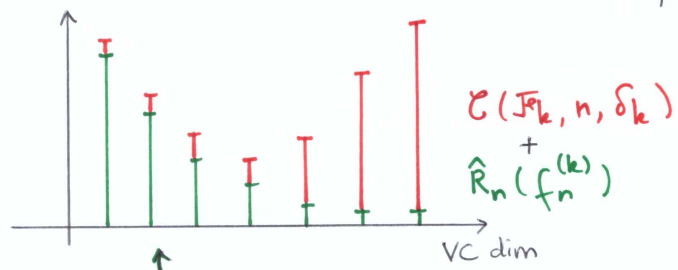
$$= \mathbb{P}\left(\sup_k \sup_{f \in \mathcal{F}_k} \{ \text{——} '' \text{——} \} > 0 \right)$$

$$\leq \sum_k \mathbb{P}\left(\sup_{f \in \mathcal{F}_k} \{ \text{——} '' \text{——} \} > 0 \right)$$

$$\leq \sum_k \delta_k = \delta \sum_k 2^{-k} < \delta.$$

$$\Rightarrow \forall f \in \left(\bigcup_k \mathcal{F}_k \right), \quad |R(f) - \hat{R}_n(f)| \leq \mathcal{C}(\mathcal{F}_{k(f)}, n, \delta_{k(f)})$$

w.p. $\geq 1-\delta$.



In green = training error of the empirical risk minimizer in class $\mathcal{F}_k$:

$$f_n^{(k)} \in \underset{f \in \mathcal{F}_k}{\operatorname{argmin}}\, \hat{R}_n(f)$$

In red = penalty term $\mathcal{C}(\mathcal{F}_k, n, \delta_k)$

Put $\hat{k} = \underset{k \geq 1}{\operatorname{argmin}}\, \hat{R}_n(f_n^{(k)}) + \mathcal{C}(\mathcal{F}_k, n, \delta_k)$, and select

$$f_n = f_n^{(\hat{k})}.$$

IV - LEARNING WITH A GENERAL LOSS

The VC dimension of a class $\mathcal{F}$ requires the binary nature of $f \in \mathcal{F} = \{ f : \mathcal{X} \to \{0,1\} \}$. The concept of VC dimension does not extend naturally to classes of functions taking continuous values, and is thus not suitable for regression problems $\to$ we need other measures of class complexity.

In this section, we assume that the response variable Y is continuous and bounded. Without loss of generality, assume that $\mathbb{P}(Y \in [-1, 1]) = 1$, and consider a family of functions $\mathcal{F} \subset \{ f : \mathcal{X} \to [-1, 1] \}$. Under theses assumptions, the square loss $\ell(y, f(x)) = (y - f(x))^2$ and the absolute loss $\ell(y, f(x)) = |y - f(x)|$ (and many other) are bounded: $0 \leq \ell(y, f(x)) \leq 1 \quad \forall x, y$, $\forall f \in \mathcal{F}$.

$\hookrightarrow$ This assumption allows us to use the bounded difference inequality, so that w.p. $\geq 1-\delta$,

$$\sup_{f \in \mathcal{F}} |R(f) - \hat{R}_n(f)| \leq \mathbb{E}\Big\{ \sup_{f \in \mathcal{F}} |R(f) - \hat{R}_n(f)| \Big\} + \sqrt{\frac{\log(2/\delta)}{2n}}$$

As before, this term is used to bound the estimation error of the empirical risk minimizer f_n:

$$R(f_n) \leq R(\bar{f}) + 2 \sup_{f \in \mathcal{F}} |R(f) - \hat{R}_n(f)|$$

$\Rightarrow$ We only need to bound the expected value, which can be done using symmetrization (see section II.2 p. 18) & introducing Rademacher random variables σ_i:

$$\boxed{\mathbb{E}\Big\{ \sup_{f \in \mathcal{F}} |R(f) - \hat{R}_n(f)| \Big\} \leq 2 \sup_{\substack{(x_1, y_1) \\ \vdots \\ (x_n, y_n)}} \mathbb{E}\Big\{ \sup_{f \in \mathcal{F}} \Big| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(y_i, f(x_i)) \Big| \Big\}}$$

$= 2 \times$ Rademacher complexity $R_S(\ell \circ \mathcal{F})$, $S = \{(x_i, y_i)\}_{i=1,\dots,n}$

If $\mathcal{F}$ contains finitely many elements, proceeding as before, we immediately get

$$R_S(\ell \circ \mathcal{F}) \leq \sqrt{\frac{2 \log(2|\mathcal{F}|)}{n}}$$

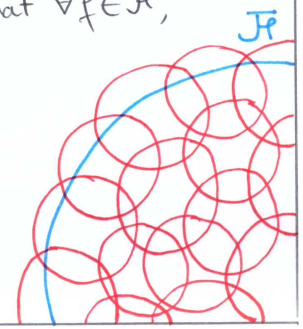
When $\mathcal{F}$ is uncountably infinite, we need somehow to reduce the problem to finite classes of functions, just as we did in binary classification with the concept of shattering ('sup' becomes 'max' + union bound). To do so, we introduce next a new measure of complexity of function classes, known as COVERING NUMBERS.

IV.1. Covering Numbers.

A covering number is an object $\mathcal{N}(\mathcal{F}, d, \varepsilon)$, where

- $\mathcal{F}$ = family of functions
- d = a metric on $\mathcal{F}$
- ε = resolution of the covering of $\mathcal{F}$ under the metric d.

An ε-NET of $(\mathcal{F}, d)$ is a set V such that $\forall f \in \mathcal{F}$, $\exists g \in V$ s.t. $d(f, g) \leq \varepsilon$



The covering number $\mathcal{N}(\mathcal{F}, d, \varepsilon)$ of $(\mathcal{F}, d)$ is

$$\mathcal{N}(\mathcal{F}, d, \varepsilon) = \inf \{ |V| : V \text{ is an } \varepsilon\text{-net} \}$$

We introduce

$$\mathbb{R}_S(\mathcal{F}) := \mathbb{E}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i) \right| \right\},$$

for $S = \{x_1, \dots, x_n\}$, and the empirical ℓ_1 distance

$$d_1^S(f,g) := \frac{1}{n}\sum_{i=1}^n |f(x_i) - g(x_i)|.$$

Theorem: $\forall \mathcal{F} \subset \{ f \mid X \to [-1,1] \}$, $\forall S = \{x_1, \dots, x_n\}$,

$$\mathbb{R}_S(\mathcal{F}) \leq \inf_{\varepsilon>0} \left\{ \varepsilon + \sqrt{\frac{2\log(2\mathcal{N}(\mathcal{F}, d_1^S, \varepsilon))}{n}} \right\}$$

proof: Fix $S = \{x_1, \dots, x_n\}$ and $\varepsilon > 0$.
Let V = minimal ε-net of $(\mathcal{F}, d_1^S)$; $|V| = \mathcal{N}(\mathcal{F}, d_1^S, \varepsilon)$.
$\forall f\in\mathcal{F}$, define $f^\circ \in V$ such that $d_1^S(f, f_0) \leq \varepsilon$.
" a representative of $f\in\mathcal{F}$.

$$\mathbb{R}_S(\mathcal{F}) = \mathbb{E}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i) \right| \right\}$$

$$\leq \mathbb{E}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i (f(x_i) - f_0(x_i)) \right| \right\} + \mathbb{E}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i f_0(x_i) \right| \right\}$$

$$\leq \varepsilon + \mathbb{E}\left\{ \max_{f\in V} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i) \right| \right\}$$

contains finitely many elements: $|V| = \mathcal{N}(\mathcal{F}, d_1^S, \varepsilon)$.
$\Rightarrow$ Proceed as before to get

$$\leq \varepsilon + \sqrt{\frac{2\log(2\mathcal{N}(\mathcal{F}, d_1^S, \varepsilon))}{n}} \quad \forall \varepsilon. \qquad \blacksquare$$

↳ We can obtain a better bound using a technique called chaining.

IV. 2. Chaining.

Theorem. (DUDLEY INTEGRAL)

$\forall \mathcal{F} \subset \{ f \mid X \to [-1,1] \}$, $\forall S = \{x_1, \dots, x_n\}$,

$$\mathbb{R}_S(\mathcal{F}) \leq \inf_{\varepsilon>0} \left\{ 4\varepsilon + \frac{12}{\sqrt{n}} \int_\varepsilon^{1/2} \sqrt{\log \mathcal{N}(\mathcal{F}, d_2^S, u)}\, du \right\}$$

where

$$d_2^S(f,g) := \left(\frac{1}{n}\sum_{i=1}^n (f(x_i) - g(x_i))^2 \right)^{1/2}$$

proof Consider $S = \{x_1, \dots, x_n\}$, and V_j a minimal $\varepsilon = 2^{-j}$-net of $\mathcal{F}$ under d_2^S, for $j = 1, \dots, N$, where N is an integer to be determined later.
Put $F := \{ (f(x_1), \dots, f(x_n))^t \mid f\in\mathcal{F} \}$.
In this notation,

$$\mathbb{R}_S(\mathcal{F}) = \frac{1}{n}\mathbb{E}\left\{ \sup_{f\in F} |\langle \sigma, f\rangle| \right\}, \text{ where } \sigma = (\sigma_1, \dots, \sigma_n)^t.$$

Write

$$\langle\sigma, f\rangle = \langle\sigma, f - \underbrace{f_N^\circ}_{\in V_N}\rangle + \langle\sigma, f_N^\circ - \underbrace{f_{N-1}^\circ}_{\in V_{N-1}}\rangle + \dots + \langle\sigma, f_1^\circ - \underbrace{f_0^\circ}_{=0}\rangle$$

" chain "

$f_j^\circ \equiv$ a representative of $f\in\mathcal{F}$ in V_j : $d_2^S(f, f_j^\circ) \leq 2^{-j}$.
Thus

$$\mathbb{R}_S(\mathcal{F}) \leq \underbrace{\frac{1}{n}\mathbb{E}\left\{ \sup_{f\in F} |\langle\sigma, f - f_N^\circ\rangle| \right\}}_{\text{I}} + \underbrace{\frac{1}{n}\sum_{j=1}^N \mathbb{E}\left\{ \sup_{f\in F} |\langle\sigma, f_j^\circ - f_{j-1}^\circ\rangle| \right\}}_{\text{II}}.$$

To bound I, note that $|\langle\sigma, f - f_N^\circ\rangle| = \left| \sum_{i=1}^n \sigma_i (f(x_i) - f_N^\circ(x_i)) \right|$

$$\overset{CS}{\Rightarrow} |\langle \sigma, f - f_N^\circ \rangle| \leqslant \left(\sum_{i=1}^n \sigma_i^2 \sum_{i=1}^n \left(f(x_i) - f_N^\circ(x_i)\right)^2 \right)^{1/2}$$
$$= n^{1/2}\left(n^{1/2} d_2^S(f, f_N^\circ) \right),$$

so that $\frac{1}{n}|\langle \sigma, f - f_N^\circ \rangle| \leqslant d_2^S(f, f_N^\circ) \leqslant 2^{-N}$.

- We turn our attention to the second term (II).

$f_j^\circ \in V_j$, where V_j contains $|V_j|$ elements.

$f_{j-1}^\circ \in V_{j-1}$, where V_{j-1} contains $|V_{j-1}| \leqslant \frac{|V_j|}{2}$ elements.

$\Rightarrow$ There are at most $|V_j||V_{j-1}| \leqslant \frac{1}{2}|V_j|^2$ possible differences $f_j^\circ - f_{j-1}^\circ$, and

$$\frac{1}{n}\mathbb{E}\left\{ \sup_{f \in F} |\langle \sigma, f_j^\circ - f_{j-1}^\circ \rangle| \right\} \leqslant \max_{b \in B} \|b\|_2 \frac{\sqrt{2\log(2|B|)}}{n}$$

Where $B := \{ f_j^\circ - f_{j-1}^\circ, f \in \mathcal{F} \}$

$|B| \leqslant \frac{1}{2}|V_j|^2$

(adapt the proof on page 31)

Note that

$$\|f_j^\circ - f_{j-1}^\circ\|_2 = n^{1/2} d_2^S(f_j^\circ, f_{j-1}^\circ)$$
$$\leqslant n^{1/2}\left(d_2^S(f_j^\circ, f) + d_2^S(f, f_{j-1}^\circ) \right)$$
$$\leqslant 3 \cdot 2^{-j} \cdot n^{1/2}.$$

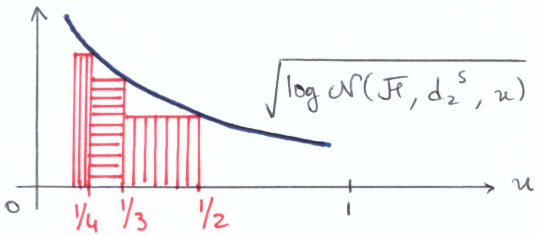
$$\Rightarrow \frac{1}{n}\mathbb{E}\left\{ \sup_{f \in F} |\langle \sigma, f_j^\circ - f_{j-1}^\circ \rangle| \right\} \leqslant 6 \cdot 2^{-j} \sqrt{\frac{\log |V_j|}{n}}$$
$$= 6 \cdot 2^{-j} \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, d_2^S, 2^{-j})}{n}}$$

To bound (II), it remains to sum all these terms $\uparrow$.

With $2^{-j} = 2(2^{-j} - 2^{-j-1})$, we have

$$\text{(II)} \leqslant 6\, n^{-1/2} \sum_{j=1}^N 2^{-j} \sqrt{\log \mathcal{N}(\mathcal{F}, d_2^S, 2^{-j})}$$
$$= 12\, n^{-1/2} \sum_{j=1}^N (2^{-j} - 2^{-j-1}) \sqrt{\log \mathcal{N}(\mathcal{F}, d_2^S, 2^{-j})}$$
$$\leqslant 12\, n^{-1/2} \int_{2^{-(N+1)}}^{1/2} \sqrt{\log \mathcal{N}(\mathcal{F}, d_2^S, u)}\, du$$

$\uparrow$ since



Combining (I) and (II) yields

$$\mathcal{R}_S(\mathcal{F}) \leqslant 2^{-N} + 12\, n^{-1/2} \int_{2^{-(N+1)}}^{1/2} \sqrt{\log \mathcal{N}(\mathcal{F}, d_2^S, u)}\, du$$

Choose $2^{-(N+2)} \leqslant \varepsilon \leqslant 2^{-(N+1)}$

$$\Rightarrow \mathcal{R}_S(\mathcal{F}) \leqslant \varepsilon + \frac{12}{\sqrt{n}} \int_{\varepsilon}^{1/2} \sqrt{\log \mathcal{N}(\mathcal{F}, d_2^S, u)}\, du \quad \blacksquare$$

IV. 3. Back to learning.

We want to bound $R_S(\ell \circ \mathcal{F}) = \sup_{(x_i, y_i)} \mathbb{E}\left\{ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \ell(y_i, f(x_i)) \right| \right\}$

(see page 40)

Assume that the loss function ℓ is Lipschitz in its second argument: $\forall\, y, z \in [-1, 1]$, $|\ell(\cdot, y) - \ell(\cdot, z)| \leqslant L|y - z|$, then it is possible to show that $R_S(\ell \circ \mathcal{F}) \leqslant 2L\, R_S(\mathcal{F})$.

(Talagrand's lemma; see lemma 4.2 in Mohri et al (2012)).

Putting things together, we finally get

$$\boxed{\sup_{f\in\mathcal{F}} |R(f) - \hat{R}_n(f)| \leqslant 4L \sup_{S} \inf_{\varepsilon>0} \left\{ 4\varepsilon + 12 n^{-1/2} \int_{\varepsilon}^{1/2} \sqrt{\log \mathcal{N}(\mathcal{F}, d_2^S, u)}\, du \right. + \sqrt{\frac{\log(2/\delta)}{n}} \quad \text{w.p.} \geqslant 1-\delta}$$

We can obtain more informative bounds for specific choices of $\mathcal{F}$.

Examples: (i) $\mathcal{F} := \{ f : x \mapsto \langle a, x \rangle, \ a \in B_p^d$
$x \in B_q^d$
$p^{-1} + q^{-1} = 1 \}$,

Class of linear functions $\subset \mathbb{R}^d \to \mathbb{R}$ indexed by a finite-dimensional parameter

where B_p^d = unit ball under the ℓ_p norm
$= \{ x \in \mathbb{R}^d \mid \|x\|_p \leqslant 1 \}$.

Hölder $\Rightarrow |f(x)| \leqslant \|a\|_p \|x\|_q \leqslant 1$

One can show that for $1 \leqslant p \leqslant q$ (possibly $= \infty$),
$\mathcal{N}(\mathcal{F}, d_p^S, \varepsilon) \leqslant \mathcal{N}(\mathcal{F}, d_q^S, \varepsilon)$

&

$\mathcal{N}(\mathcal{F}, d_\infty^S, \varepsilon) \leqslant \left(\frac{2}{\varepsilon}\right)^d$ ← bound independent of S

$\left[d_p^S(f,g) := \left(\frac{1}{n} \sum_{i=1}^n |f(x_i) - g(x_i)|^p \right)^{1/p}, \right.$
for $\left. S = \{x_1, \dots, x_n\} \right]$

Thus, Dudley integral can be bounded independently of S by

$$\int_0^{1/2} \sqrt{\log(2/u)^d}\, du = \sqrt{d} \underbrace{\int_0^{1/2} \sqrt{\log(2/u)}\, du}_{<\infty} \leqslant C\sqrt{d},$$

for some $C>0$.

It follows that w.p $\geqslant 1-\delta$, $\exists C>0$ s.t.

$$\sup_{f\in\mathcal{F}} |R(f) - \hat{R}_n(f)| \leqslant C L \sqrt{\frac{d}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

↓ Empirical Risk Minimizer $\hat{f}_n \in \operatorname{argmin}_{f\in\mathcal{F}} \hat{R}_n(f)$ satisfies

$$\boxed{R(\hat{f}_n) - R(\bar{f}) \leqslant C' L \sqrt{\frac{d}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}} \quad \text{w.p} \geqslant 1-\delta}$$

(C' = constant indpt of L, d, n)

The ratio d/n plays an important role in this bound.

(ii) $\mathcal{F} = \{ f : \mathbb{R}^d \to \{-1, 1\} \}$ with finite VC dimension $VC(\mathcal{F})$.

A result by Haussler show that $\mathcal{N}(\mathcal{F}, d_2^S, \varepsilon) \leqslant \left(\frac{C}{\varepsilon}\right)^{VC(\mathcal{F})}$, and it is possible to obtain a risk bound of order $\sqrt{\frac{VC(\mathcal{F})}{n}}$.

Compare with the bound on page 35: we removed the log factor $\log(n / VC(\mathcal{F}))$.