VC → says how we can set PAC-learnability

3.3 The VC dimension

- We now learn about the one characteristic of a hypothesis class $\mathcal{H}$ determining its PAC learnability
- This characteristic is its VC dimension which relates to the growth of $\tau_{\mathcal{H}}$
- As preparation for the VC dimension we need:

A hypothesis class H shatters a set of n points if, for every possible labeling (classification) of these points (e.g., 0 or 1 in binary classification), there exists at least one hypothesis in H that correctly classifies all of them.

Definition 3.20: Shattering

Let $M \subseteq \mathcal{X}$ be finite and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, $|\mathcal{Y}| = 2$, be a class of binary hypotheses. Then we say $\mathcal{H}$ shatters the set M if its restriction $\mathcal{H}_M$ to M satisfies

$$\mathcal{H}_M = \mathcal{Y}^{|M|} \qquad \Longleftrightarrow \qquad |\mathcal{H}_M| = 2^{|M|}$$

Can I always separate the points, no matter how they are labeled?" If yes, then that set of points is shattered.

Interpretation: $\mathcal{H}$ shatters a set M if any arbitrary binary labelling of the elements $x \in M$ can be recovered by a hypothesis $h \in \mathcal{H}$.

Example 3.21: Heaviside classifiers

Let $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$ and consider again

$$\mathcal{H} = \{\mathbb{1}_{[a,+\infty)} : a \in \mathbb{R}\}$$

Which sets M does $\mathcal{H}$ shatter?

- For $M = \{x_1\} \subset \mathbb{R}$ we have

  $$\mathcal{H}_M = \{[0], [1]\} = \mathcal{Y}, \qquad |\mathcal{H}_M| = 2.$$

  Hence, any singleton set $M = \{x_1\}$, $x_1 \in \mathcal{X}$, is shattered by $\mathcal{H}$.

- For $M = \{x_1, x_2\} \subset \mathbb{R}$, $x_1 < x_2$, we have

  $$\mathcal{H}_M = \{[0, 0], [0, 1], [1, 1]\} \subset \mathcal{Y}^2, \qquad |\mathcal{H}_M| = 3 < 2^2.$$

  Thus, any set $M = \{x_1, x_2\} \subset \mathbb{R}$, $x_1 \neq x_2$, is not shatted by $\mathcal{H}$.

The meaning of shattering

- If $\mathcal{H}$ shatters M, then $\mathcal{H}_M$ contains all possible $2^{|M|}$ binary bit patterns of $|M|$ bits.
- W.r.t. the approximation error on $M \subseteq \mathcal{X}$ this is a good thing, but for learning this is a disadvantage!
- For instance, if we get $m < |M|$ training data (x_i, y_i) with $x_i \in M$, $i = 1, \ldots, m$, then we have at least

  $$2^{|M|-m}$$

  hypotheses $h \in \mathcal{H}$ which minimize the empirical risk $\mathcal{R}_s$ – and which one of them is the real one?
- Thus, regarding PAC learnability (with arbitrary distribution μ on $\mathcal{D}$), it is rather bad if $\mathcal{H}$ shatters large sets M, since we would then need a lot of training data to learn with high probability a good hypothesis from them.

The VC Dimension

Definition 3.22:

The VC (Vapnik–Chervonenkis) dimension of a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ is

$$\mathrm{VCD}(\mathcal{H}) := \sup\{|M| : \mathcal{H} \text{ shatters } M \subseteq \mathcal{X}\}.$$

To determine the VC dimension of a class $\mathcal{H}$, e.g., $\mathrm{VCD}(\mathcal{H}) = d$, we need to

1. find a set $M \subseteq \mathcal{X}$ with $|M| = d$ which is shattered by $\mathcal{H}$
2. and show that no set $M' \subseteq \mathcal{X}$ with $|M'| = d+1$ is shattered by $\mathcal{H}$.

Example 3.23: Heaviside classifiers

Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{0,1\}$. Then we have for

$$\mathcal{H} = \{\mathbb{1}_{[a,+\infty)} : a \in \mathbb{R}\}, \qquad \mathrm{VCD}(\mathcal{H}) = 1.$$

V. Vapnik (*1936)

A. Chervonenkis (1938–2014)

Example 3.24: Interval hypotheses

Let $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0,1\}$ and $\mathcal{H} = \{\mathbb{1}_{[a,b]} \colon a < b \in \mathcal{X}\}$. Then $\mathrm{VCD}(\mathcal{H}) = 2$, because

1. For $M = \{x_1, x_2\} \subset \mathbb{R}$, $x_1 < x_2$, we have

$$\mathcal{H}_M = \{[0,0],[0,1],[1,0],[1,1]\}\,, \qquad |\mathcal{H}_M| = 4 \qquad 2^m$$

2. but for any $M = \{x_1, x_2, x_3\} \subset \mathbb{R}$, $x_1 < x_2 < x_3$, we have

$$\mathcal{H}_M = \{[0,0,0],[0,0,1],[0,1,0],[1,0,0],[1,1,0],[0,1,1],[1,1,1]\}\,, \qquad |\mathcal{H}_M| = 7.$$

$$\mathcal{H}_M \not\ni [1,0,1]$$

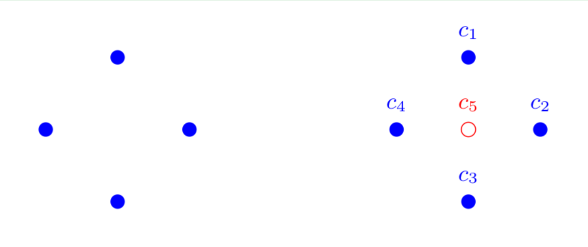
Example 3.25: All hypotheses for finite $\mathcal{X}$

Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ and $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$. Then $\mathrm{VCD}(\mathcal{H}) = n$, because

1. by construction we have $\mathcal{H}_{\mathcal{X}} = \mathcal{Y}^{\mathcal{X}}$;
2. since there is no subset $M \subseteq \mathcal{X}$ of cardinality $n+1$, there holds $\mathrm{VCD}(\mathcal{H}) = n$.

Example 3.26: Rectangle hypotheses

Let $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{H} = \{\mathbb{1}_{[a_1,b_1]\times[a_2,b_2]} \colon (a_1,a_2) < (b_1,b_2) \in \mathcal{X}\}$. Then we have $\mathrm{VCD}(\mathcal{H}) = 4$, because:



Source: S. Shalev–Schwartz, S. Ben–David, "Understanding Machine Learning" (2014)

1. the set $M = \{p_1, p_2, p_3, p_4\} \subset \mathcal{X}$ (left) is shattered by $\mathcal{H}$
2. but for $M = \{c_1, c_2, c_3, c_4, c_5\} \subset \mathcal{X}$ with $c_i = (c_{i,1}, c_{i,2}) \in \mathbb{R}^2$ where

$$c_{1,2} = \max_{i=1,\dots,5} c_{i,2}, \;\; c_{3,2} = \min_{i=1,\dots,5} c_{i,2}, \;\; c_{4,1} = \min_{i=1,\dots,5} c_{i,1}, \;\; c_{2,1} = \max_{i=1,\dots,5} c_{i,1},$$

   as in the righthand side figure we have $[1,1,1,1,0] \notin \mathcal{H}_M$, since no $h \in \mathcal{H}$ exists with $h(c_i) = 1$ for $i = 1,\dots,4$ but $h(c_5) = 0$.

Example 3.27: Cuboid and polygon hypotheses

The example of rectangle hypotheses can be generalized to $\mathcal{X} = \mathbb{R}^d$:

$$\mathcal{H} = \{\mathbb{1}_Q : Q = \prod_{j=1}^{d} [a_j, b_j],\ (a_1, \ldots, a_d) < (b_1, \ldots, b_d) \in \mathcal{X}\}.$$

Then we have $\mathrm{VCD}(\mathcal{H}) = 2d$. Moreover for $\mathcal{X} = \mathbb{R}^2$ and

$$\mathcal{H} = \{\mathbb{1}_P : P \text{ is the surface inside a convex polygon with } n \text{ corners}\}$$

we have $\mathrm{VCD}(\mathcal{H}) = 2n + 1$.

$2^{m-1} < n = |\mathcal{H}| < 2^m$

Example 3.28: Finite hypotheses classes

For finite $\mathcal{H}$ we have $\mathrm{VCD}(\mathcal{H}) \leq \lfloor \log_2(|\mathcal{H}|) \rfloor$, because in order to shatter $M \subseteq \mathcal{X}$, $2^{|M|}$ hypotheses are needed.

This upper bound is attained, see example 3.25. However the VC dimension of finite classes can be quite far from it, see example 3.23 with $\mathcal{H} = \{\mathbb{1}_{[a_i,\infty)} : i = 1, \ldots, n\}$.

Infinite VC-Dimension

- Definition 3.22 does not exclude the case $\mathrm{VCD}(\mathcal{H}) = +\infty$. But can this happen?

Proposition 3.29:

Let $\mathcal{X} = \mathbb{R}$ and

$$\mathcal{H} = \{h_\theta(x) = \lceil \sin(\theta\pi x) \rceil : \theta \in \mathbb{R}\},$$

where $\lceil -1 \rceil := 0$. Then we have $\mathrm{VCD}(\mathcal{H}) = +\infty$.

- **Short explanation:** For any number of points $x_1, \ldots, x_d \in \mathbb{R}$ and any binary labelling $y = (y_1, \ldots, y_d) \in \{0,1\}^d$ of these, we find a sufficiently "fast oscillation" $h_\theta(x) = \sin(\theta\pi x)$ with $|\theta| \gg 1$ such that $h_\theta(x_i) > 0$ if $y_i = 1$ and $h_\theta(x_i) \leq 0$ otherwise.

$\Rightarrow$ Even simple hypotheses classes described by only one parameter $\theta \in \mathbb{R}$ can have an infinite VC dimension.

When we do have a soem data set m. we can set the Hypothesis H so it could perfectly classify the data labeling. it is called shattering. now when we could shatter a data set with size M with our H, when we find the largest set that is being shattered by the H, then it is the VCD. when we do have d pair of points according to our condition, we can make labels of them. if the hypothesis can realize all of them then the VCD satisfies and it would be d of the points.

e.g: (0,1) -> we have x1,x2. now if we could satisfy this, 2^m rule(shattering) and find the largest set of shattered Ms, we have VCD: (0,0),(0,1),(1,0),(1,1). we have 2^2 = 4 that is shattered, all of the realizations are holed so the VCD(H) = 2, d = 2 here.

Now lets see if we could see for this condition that x1<x2:
(0,0),(0,1),(1,0),(1,1). It will not hold the condition here and shattering will be failed because when we have this condition that x1<x2, we cannot set 2^2 = 4 cuz (1,0) will not work. Only 3 realizations are being realized and so the VCD(H) := 1 (d=1 will always be shattered. then if the
d+1 not work or shattered under conditions (at least one label is not realizable) it will not work.)
So since the classifier fails at d + 1, then the VCD(H) = d.

The VCD is the largest d such that all 2^d labellings be realizable.

But if the VCD(H) = infinity, e.g the Sin function will lead to infinity all time,because by choosing a large sufficiently Teta in our sine function, we can this function to oscillate quickly and therefore, we can find a Teta that correctly classify all possible labeling for any given points and therefore, the hypothesis class can shatter arbitrarily large sets.
if the VCD(H) leads to infinity, then the number of samples required to generalize well grows arbitrarily large.
In this case, learning is impossible with a finite dataset because the number of hypotheses the model must consider is too large.
The generalization error does not decrease with more training data in a controlled manner and therefore since this exceeds the bound for the.

Proof of Proposition 3.29 for the interested:

- Let $x \in (0,1)$ have the binary representation $0.x_1x_2x_3\dots$ meaning $x = \sum_{i=1}^{\infty} x_i \cdot 2^{-i}$ with $x_i \in \{0,1\}$.
- Then

$$\sin(2^m \pi x) = \sin\left(2\pi\left(\sum_{i=1}^{m-1} x_i \cdot 2^{m-1-i} + \sum_{i=m}^{\infty} x_i \cdot 2^{m-1-i}\right)\right)$$
$$= \sin\left(2\pi \sum_{i=m}^{\infty} x_i \cdot 2^{m-1-i}\right).$$

- If $x_m = 0$ but $x_{m+i} \neq 0$ for an $i \in \mathbb{N}$, then $h_{2^m}(x) = 1$ because

$$2\pi(0.5x_m + 0.25x_{m+1} + 0.125x_{m+2}\dots) \in (0,\pi).$$

- If $x_m = 1$, then $h_{2^m}(x) = 0$, because

$$2\pi(0.5x_m + 0.25x_{m+1} + 0.125x_{m+2}\dots) \in [\pi, 2\pi).$$

- In summmary, we have $h_{2^m}(x) = 1 - x_m$ for $x \in (0,1)$ with $x_{m+i} \neq 0$ for an $i \in \mathbb{N}_0$.

- We choose now $n \in \mathbb{N}$ points $x^{(j)} \in (0,1)$ with binary representations

$$
\begin{aligned}
x^{(1)} &= 0.\ 0\ 1\ 0\ 0\ \ldots\ 1\ 1,\\
x^{(2)} &= 0.\ 0\ 0\ 1\ 0\ \ldots\ 1\ 1,\\
x^{(3)} &= 0.\ 0\ 0\ 0\ 1\ \ldots\ 1\ 1,\\
&\ldots\\
x^{(n-1)} &= 0.\ 0\ 0\ 0\ 0\ \ldots\ 1\ 1,\\
x^{(n)} &= 0.\ 0\ 0\ 0\ 0\ \ldots\ 0\ 1.
\end{aligned}
$$

$\Rightarrow$ Each of the 2^n possible binary labellings of $x^{(1)}, \ldots, x^{(n)}$ is represented by one of their $m = 1, \ldots, 2^n$ binary coefficients (i.e., the columns above).

- For instance, the labelling $y^{(1)} = \ldots = y^{(n)} = 1$ is reproduced by $h_{2^1}(x^{(j)})$.
- The labelling of $x^{(2)}, \ldots, x^{(n)}$ by 1 and $x^{(1)}$ by 0 corresponds to $h_{2^2}(x^{(j)})$...
- Each binary labelling of $x^{(1)}, \ldots, x^{(n)}$ corresponds to a hypothesis $h_{2^k}(x)$, $k \in \{1, \ldots, 2^n\}$, and hence, $M = \{x^{(1)}, \ldots, x^{(n)}\}$ is shattered by $\mathcal{H}$.
- Since n was arbitrary, we obtain $\mathrm{VCD}(\mathcal{H}) = +\infty$. $\square$

VC dimension and PAC learnability

- How does an infinite VC dimension relate to PAC learnability? By the "No-Free-Lunch" theorem we conclude:

Corollary 3.30:

A class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with $\mathrm{VCD}(\mathcal{H}) = +\infty$ is not PAC-learnable with respect to the 0-1 loss (under the realizability assumption).

Sketch of proof: $\mathrm{VCD}(\mathcal{H}) = +\infty$ means there exist sets M of arbitrary finite cardinality which are shattered by $\mathcal{H}$. Thus, $\mathcal{H}$ or $\mathcal{H}_M$ corresponds to all $h\colon M \to \{0,1\}$. However, from the NFL theorem we conclude for $\mathcal{X} = M$

Sample complexity of restriction on M

$$m_{\mathcal{H}_M}(1/8, 1/7) \geq |M|/2 \xrightarrow{|M|\to\infty} \infty.$$

Cardinality of set M

Since $m_{\mathcal{H}}(\epsilon,\delta) \geq m_{\mathcal{H}_M}(\epsilon,\delta)$, $\mathcal{H}$ cannot be PAC-learnable. $\square$

- **Question:** Does the converse also apply?

VC dimension and uniform convergence

- We answer the question by studying the relation between $\mathrm{VCD}(\mathcal{H})$ and the growth function $\tau_{\mathcal{H}}$
- This will relate a finite VC dimension $\mathrm{VCD}(\mathcal{H}) < +\infty$ to uniform convergence

Theorem 3.31:

For a binary hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, $|\mathcal{Y}| = 2$, we have

- If $\mathrm{VCD}(\mathcal{H}) = \infty$, then

$$\tau_{\mathcal{H}}(m) = 2^m$$

- If $\mathrm{VCD}(\mathcal{H}) = d < \infty$, then

$$\tau_{\mathcal{H}}(m) \in \mathcal{O}(m^d) \rightsquigarrow \tau_{\mathcal{H}}(m) \leq \left(\frac{e}{d}\right)^d m^d \in \mathcal{O}(m^d)$$

- If $\tau_{\mathcal{H}}(m) \leq (em)^p$ for a $p > 0$ and all $m > p + 1 > 2$, then

$$\mathrm{VCD}(\mathcal{H}) \in \mathcal{O}\left(p \ln(p)\right)$$

$$\mathrm{VCD}(\mathcal{H}) \leq 16 p \ln(p)$$

Proof of 3,31:

$\Rightarrow$ let $VCD(\mathcal{H}) = +\infty$ then we know for each $m \in \mathbb{N}$ there exists a set $M \subseteq X$ with cardinality $m = |M|$ which is shattered by $\mathcal{H}$. Thus for this m we have cardinality of m points $(\mathcal{H}_M) = 2^m$, which yields the statement: $\overline{\tau}(\mathcal{H}) = \sup\limits_{M \subseteq X, |M| = m} |\mathcal{H}_M| = 2^m \ \forall m \in \mathbb{N}$

$\Rightarrow$ let $VCD(H) =: d < +\infty$, then the Shale-Sane lemma:

$$\tau_H(m) = \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{l}{d}\right)^d m^d$$

where the last inequality is due to technical result (ch 6).

$3 \to$ if $\tau_{(H)} \leq (em)^p = e^p m^p$ & $\forall\ m > p+1 > 2$

we know that VCD(H) must be finte (by contradic
using first statment, it grows exponentially (the
growth function if the VCD $= \infty$). let
$d = VCD(H) < \infty$, then choosing $m = d$ we have
if $d > p+1$: $\tau_H = \underline{\underline{2^d}} \leq e^p d^p$

$$\to d \leq p \cdot \frac{1 + \ln(d)}{\ln(2)}$$

a technical result, yields that:

$$d \leq \frac{4p}{\ln(2)} \ln\left(\frac{2p}{\ln(2)}\right) + \frac{2p}{\ln(2)}$$

which yields to: $d \leq 16\, p \ln(p)$

Corollary 3.32:

A class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, $|\mathcal{Y}| = 2$, with finite VC dimension $\mathrm{VCD}(\mathcal{H}) = d < \infty$ satisfies the uniform convergence condition w.r.t. the 0-1 loss.

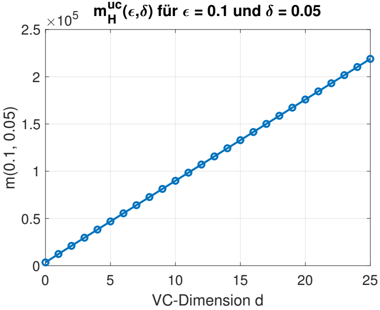
Moreover, an upperbound for $m_{\mathcal{H}}^{\mathsf{uc}}(\epsilon, \delta)$ for $\epsilon, \delta \in (0,1)$ is given by

$$m_{\mathcal{H}}^{\mathsf{uc}}(\epsilon, \delta) \leq \min \left\{ m \in \mathbb{N} \colon m \geq \frac{2\ln(4)}{\epsilon^2} \quad \text{and} \quad 4\left(\frac{2\mathrm{e}m}{d}\right)^d \exp\left(-\frac{\epsilon^2 m}{8}\right) \leq \delta \right\}$$

- We compute numerically for $d = 0, \ldots, 25$ and $\epsilon = 0.1$, $\delta = 0.05$ the integers $m \in \mathbb{N}$ such that

$$4\left(\frac{2\mathrm{e}m}{d}\right)^d \exp\left(-\frac{\epsilon^2 m}{8}\right) \leq \delta$$

- The larger $\mathrm{VCD}(\mathcal{H})$ the more data $m_{\mathcal{H}}^{\mathsf{uc}}(\epsilon, \delta)$ is required!
- In fact: $\quad m_{\mathcal{H}}^{\mathsf{uc}}(\epsilon, \delta) \;\approx\; a\,\mathrm{VCD}(\mathcal{H}) + b$



$\Rightarrow$ VC: $P\left(\sup_{h\in H} \left(R_S(h) - R_\mu(h)\right) > \epsilon\right)$

$$\le 4 \cdot \underbrace{\left(\frac{e}{d}\right)^d (2m)^d}_{} \cdot e^{-\frac{\epsilon^2}{8} m} \le \delta \quad \forall m \ge \frac{2\log}{\epsilon^2}$$

if $VCD(H) = d$ (finite) $\rightarrow$ $\underbrace{\tau_H(m) \le \left(\frac{e}{d}\right)^d m^d}_{}$

Theorem 3.33: **Fundamental theorem of learning** (will come in exam)

For a class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, $|\mathcal{Y}| = 2$, the following statements are equivalent given the 0-1-loss:

1. $\mathcal{H}$ satisfies uniform convergence (UC).
2. $\mathcal{H}$ is (agnostic) PAC-learnable by $A = \mathrm{ERM}_{\mathcal{H}}$.
3. $\mathcal{H}$ is (agnostic) PAC-learnable. (specific learning algorithm is required)
4. $\mathcal{H}$ has finite VC dimension.

we can use ERM as well when we have finite VCD to be sure it's PAC-learnable

Moreover, if $\mathrm{VCD}(\mathcal{H}) < \infty$ then for $A = \mathrm{ERM}_{\mathcal{H}}$ we have for the sample complexity $m_{\mathcal{H}}$

intercept

necessary number of training data

$$c\frac{1}{\epsilon^2}\left(\mathrm{VCD}(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right)\right) \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C\frac{1}{\epsilon^2}\left(\mathrm{VCD}(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right)\right), \qquad (\star)$$

where $c, C < \infty$ are universal constants which are independent of $\mathcal{H}$.

To derive the bounds $(\star)$ refined techniques (Rademacher complexity) are required, see Chapter 28 in "Understanding Machine Learning" (2014) for details.

Conse

if $\forall$ 2 classes, then all data required for it must be suitable (so it could generate more training data)

there exists 2 classes, (hypothesis class)

if

Proof:

we know

1. $\xrightarrow{\text{corollary } 3,8}$ 2.

2. $\xRightarrow{\text{definition } 3,2}$ 3

3. $\xRightarrow{\text{corollary } 3,30}$ 4. (by contradiction)

4. $\xRightarrow{\text{corollary } 3,32}$ 1.