Generalization for neural networks

Framework Binary classification

input space $X = \mathbb{R}^n$ output space $Y = \{0, 1\}$

probability distribution P on $Z = X \times Y$

sample $(z_1 = (x_1, y_1), \dots, z_m = (x_m, y_m)) \in Z^m$ sample size m

Hypothesis class H: neural network

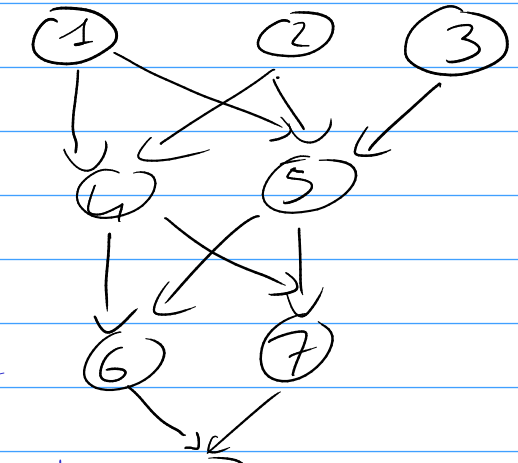
units = neurons

network
- feed-forward

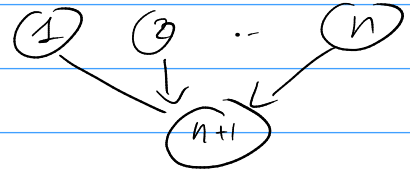
activation function σ_r for unit r

typically: $\sigma_r(u) = \frac{1}{1 + e^{-u}}$

or $\sigma_r(u) = \operatorname{sgn}(u) = \begin{cases} 1 \text{ if } u \geq 0 \\ 0 \text{ otherwise} \end{cases}$



Perceptron: network with 1 layer



Weights: for each unit r of layer $1, 2 \dots$ the outputs of the previous layer are mixed by a linear combination before passing through the activation function σ_r

if unit r is feeded with entries $z_1, \dots, z_d$ from the previous layer, its output is

$$\sigma_r\left(\sum_{j=1}^{d} w_j z_j - \theta\right)$$

where $(w_1, \dots w_d)$ are the corresponding connections weights and θ is the threshold of unit r.

→ for every $x \in \mathbb{R}^n$, the network produces a prediction $\hat{y}(x)$ which is the output of the last unit

The goal is to find the parameters (w_r, θ_r) for all units r of the network such that $\mathbb{P}(\hat{y}(x) = y)$ is as large as possible.

Linear threshold networks = feed-forward networks with activation function $\sigma_r(v) = \operatorname{sgn}(v)$ for every unit r of the network

For every hypothesis $h \in H$, we will study its generalization error

$$\text{er}_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}(h(x) \neq y)$$

we will rely on the training error of h:

$$\text{er}_{S}(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{(h(x_i) \neq y_i)} \qquad S = (z_1, \dots z_m)$$

Best achievable performance: $\text{opt}_{\mathcal{D}}(H) = \inf_{h \in H} \text{er}_{\mathcal{D}}(h)$ = target

Formal definition of learning: Probably Approximately Correct learning

A learning algorithm L for H is a function

$$L : \bigcup_{m=1}^{\infty} Z^m \to H$$

such that for all tolerance $\varepsilon > 0$ and all risk $\delta > 0$, there exists a integer $m_L(\varepsilon, \delta)$ such that if $m \geq m_L(\varepsilon, \delta)$,

$\forall P$, $P^m\left(err_P(L(S)) < opt_P(H) + \varepsilon \right) \geq 1 - \delta$

We say that H is learnable if there exists a learning algorithm for H.

The smallest sample size $m_L(\varepsilon, \delta)$ reachable by an algorithm L is $m_H(\varepsilon, \delta)$ = sample complexity of H.

The smallest $m_L(\varepsilon, \delta)$ reachable by algorithm L is called the sample complexity of L.

$$\hookrightarrow \quad err_P(L(S)) = opt_P(H) + \left(err_P(L(S)) - opt_P(H) \right)$$

approximation error

estimation error

Empirical Risk Minimization:

$$L(h) = \underset{h \in H}{\arg\min}\ \hat{er}_z(h)$$

since $\mathbb{E}[\hat{er}_z(h)] = er_p(h)$: for every $i \in \{1, \dots, n\}$,

$$\mathbb{E}[\mathbb{1}\{h(x_i) \neq y_i\}] = P(h(x_i) \neq y_i) = er_p(h)$$

Prop 1. For every $h \in H$, $P^m(\hat{er}_z(h) > er_p(h) + \varepsilon) \leq \exp(-2m\varepsilon^2)$

Proof. $\hat{er}_z(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{h(x_i) \neq y_i\}$

$\sim \frac{1}{m} B(er_p(h))$

Rem: $P(\hat{er}_z(h) < er_p(h) - \varepsilon) \leq \exp(-2m\varepsilon^2)$

Prop 2: if $|H| < \infty$, then H is learnable.

indeed, $P^m(\max_{h \in H} |\hat{er}_z(h) - er_p(h)| \geq \varepsilon) = P^m(\bigcup_{h \in H} |\hat{er}_z(h) - er_p(h)| \geq \varepsilon)$

$$\leq |H| \max_{h \in H} P^m(|\hat{er}_z(h) - er_p(h)| \geq \varepsilon) \leq 2|H| e^{-2m\varepsilon^2}$$

$$\Rightarrow P^m(er_p(L(z)) \geq \hat{er}_z^{h \in H}(L(z)) + \varepsilon) \leq 2|H| e^{-2m\varepsilon^2}$$

→ if $2|\mathcal{H}| \exp(-2m\varepsilon^2) \leq \delta$

that is if $m \geq \sqrt{\frac{\ln 2|\mathcal{H}|/\delta}{2m}}$,

$$err_P(L(S)) \leq \hat{err}_S(L(S)) + \varepsilon$$

$$= \min_{h \in \mathcal{H}} \hat{err}_S(h) + \varepsilon$$

$$\leq \hat{err}_S(h^*) + \varepsilon \quad \text{where } h^* \in \operatorname{argmin}_{h \in \mathcal{H}} err_P(h)$$

$$\leq (err_P(h^*) + \varepsilon) + \varepsilon$$

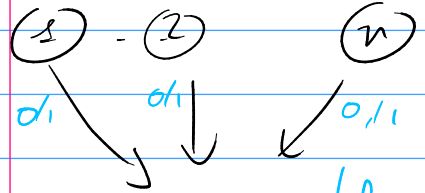
$$= opt_P(\mathcal{H}) + 2\varepsilon$$

→ with proba $\geq 1-\delta$, L returns $L(S)$ s.t.

$$err_P(L(S)) \leq opt_P(\mathcal{H}) + \sqrt{\frac{2}{m} \ln\left(\frac{2|\mathcal{H}|}{\delta}\right)}$$

→ the ERM has sample complexity at most $\frac{2}{\varepsilon^2} \ln\left(\frac{2|\mathcal{H}|}{\delta}\right)$

→ for perceptrons: binary-weight perceptrons: coefficients $\in \{0,1\}$ for the n entry units



→ $|H| = 2^n (n+1)$

$$m_L(\varepsilon, \delta) = \frac{2}{\varepsilon^2}\left(n \ln(2) + \ln(n+1) - \ln\frac{2}{\delta}\right)$$

- k-bit perception: each coefficient and the threshold has k bits: $0.b_1 \dots b_k$ where $b_i \in \{0,1\}$

→ $|H| \leq \left(2^k\right)^{n+1} = 2^{k(n+1)}$

$$\hookrightarrow m_L(\varepsilon, \delta) = \frac{2}{\varepsilon^2}\left(k(n+1)\ln(2) + \ln\frac{2}{\delta}\right)$$

→ How to deal with continuity?

Homework: restricted model

- $y_i = t(x_i)$ for some $t \in H$

  $\Leftrightarrow \operatorname{opt}(H) = 0$ realized by hypothesis $t \in H$

Thm: There is a learning algo L for H in the restricted model with sample complexity $m_L(\varepsilon, \delta) \leq \frac{1}{\varepsilon} \log\left(\frac{|H|}{\delta}\right)$

Generalization for infinite hypotheses classes.

Def: the growth function of H is

$$\Pi_H : \mathbb{N} \to \mathbb{N}$$

$$m \mapsto \max\{|H_{|S}| : S \subset X \text{ and } |S| = m\}$$

clearly $\Pi_H(m) \leq 2^m = |\{0,1\}^m|$

Def a dichotomy (by H) of a finite subset S is one of the functions $f : S \to \{0,1\}$ s.t. $f = h_{|S}$ for some $h \in H$

Def: H shatters S if $\Pi_H(S) = 2^m$

the VC-dimension of H is the largest size of a subset $S \subset X$ shattered by H.

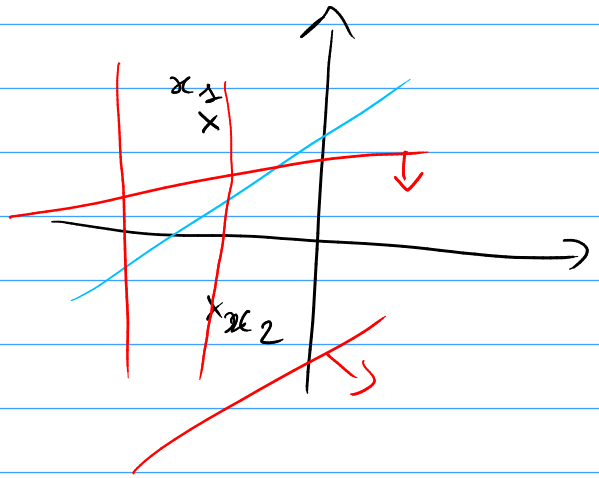
[V.C : Vapnik-Chervonenkis]

Sauer's Lemma:

$$\Pi_H(m) \leq \sum_{i=1}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d$$ polynomial in m

where $\binom{m}{i} = \frac{m(m-1)\dots(m-i+1)}{i!}$ $\quad \binom{m}{0} = 1$ $\quad \binom{m}{i} = 0$ if $i > m$

Ex: $n = 2$ $\quad H$ = perceptrons $= \{x \to \mathrm{sgn}(w_1 x_1 + w_2 x_2 - \theta) : (w_1, w_2, \theta) \in \mathbb{R}^3\}$



$|S| = 1$ $\quad S = \{x_1\}$

$f_1(x_1) = 0$
$f_2(x_1) = 1$
$\Pi_H(1) = 2 = 2^1$

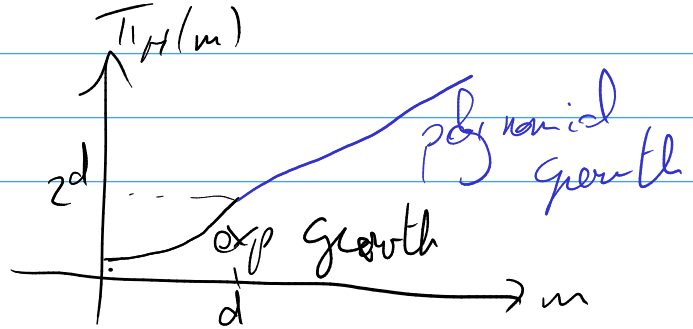
$|S| = 2$ $\quad S = \{x_1, x_2\} \to \{0,1\}$

$x_1 \to$	1	1	0	0
$x_2 \to$	1	0	1	0

$\Pi_H(2) = 4 = 2^2$

$\to$ $VC(H) \geq 2$

Home work: find $VC(H)$.



Uniform Convergence and Learnability:

Thm: Suppose that $H : X \to \{0,1\}$ has finite VC-dimension $d \geq 1$.
Let L be an ERM learning algorithm.
Then L is a learning algorithm for H with sample complexity

$$m_L(\varepsilon, \delta) \leq \frac{64}{\varepsilon^2}\left(2d \ln\frac{12}{\varepsilon} + \ln 4/\delta\right) = \frac{64}{\varepsilon^2} \ln\left(\left(\frac{12}{\varepsilon}\right)^{2d} \frac{4}{\delta}\right)$$

$$|H| = \left(\frac{12}{\varepsilon}\right)^{2d}$$

(Recall: for finite H:

$$m_L(\varepsilon, \delta) \leq \frac{2}{\varepsilon^2} \ln\left(\frac{2|H|}{\delta}\right)$$

Sauer Lemma: $\Pi_H(m) \leq \left(\frac{em}{d}\right)^d$

Compare with the bounds obtained above for binary-weight and k-bits perceptrons:

Uniform Convergence Theorem.

Thm: $\mathbb{P}^m\left(|er_p(h) - \hat{er}_p(h)| \geq \varepsilon \text{ for some } h \in H\right) \leq 4\, \Pi_H(2m) \exp\left(-\frac{m\varepsilon^2}{8}\right)$

Finite H: $\leq 2|H| \exp(-2m\varepsilon^2)$

From this it is easy to prove the previous theorem as in the finite case:

if $\forall h \in H,\ |er_p(h) - \hat{er}_S(h)| \leq \varepsilon$, then

$$er_p(L(S)) \leq \hat{er}_S(L(S)) + \varepsilon$$

$$= \min_{h \in H} \hat{er}_S(L(S)) + \varepsilon$$

$$\leq \hat{er}_S(h^*) + \varepsilon \quad \text{for any } h^* \text{ s.t. } er_p(h^*) \leq opt_p(H) + \alpha \quad (\alpha > 0)$$

$$\leq er_p(h^*) + \varepsilon + \varepsilon \leq opt_p(H) + 2\varepsilon + \alpha$$

true for any $\alpha > 0 \rightarrow er_p(L(S)) \leq opt_p(H) + 2\varepsilon$

if $\varepsilon^2 \geq \frac{8}{m} \ln\left(\frac{\Pi_H(2m)}{\delta}\right)$, then $4\Pi_H(2m)\exp\left(-\frac{m\varepsilon^2}{8}\right) \leq \delta$

and with prob. $\geq 1-\delta$ $er_p(L(S)) \leq opt_p(H) + \frac{\sqrt{32\left(d \ln(2em/d) + \ln(4/\delta)\right)}}{m}$

it suffices that $m \geq \frac{32}{\varepsilon^2}\left(d \ln m + d \ln \frac{2e}{d} + \ln \frac{4}{\delta}\right)$

→ Sauer's Lemma

since $\ln x \le \alpha x - \ln \alpha - 1$ for all $\alpha, x > 0$,

$$\frac{32}{\varepsilon^2} \ln m \le \frac{32d}{\varepsilon^2}\left(\frac{\varepsilon^2}{64d} m + \ln \frac{64d}{\varepsilon^2} - 1\right)$$

$$= \frac{m}{2} + \frac{32d}{\varepsilon^2} \ln \frac{64d}{e\varepsilon^2}$$

$\rightarrow$ it suffices that $m \ge \frac{m}{2} + \frac{32}{\varepsilon^2}\left(d \ln \frac{128}{\varepsilon^2} + \ln \frac{4}{\delta}\right)$

so $m \ge \frac{64}{\varepsilon^2}\left(2d \ln \frac{12}{\varepsilon} + \ln \frac{4}{\delta}\right)$ suffices.

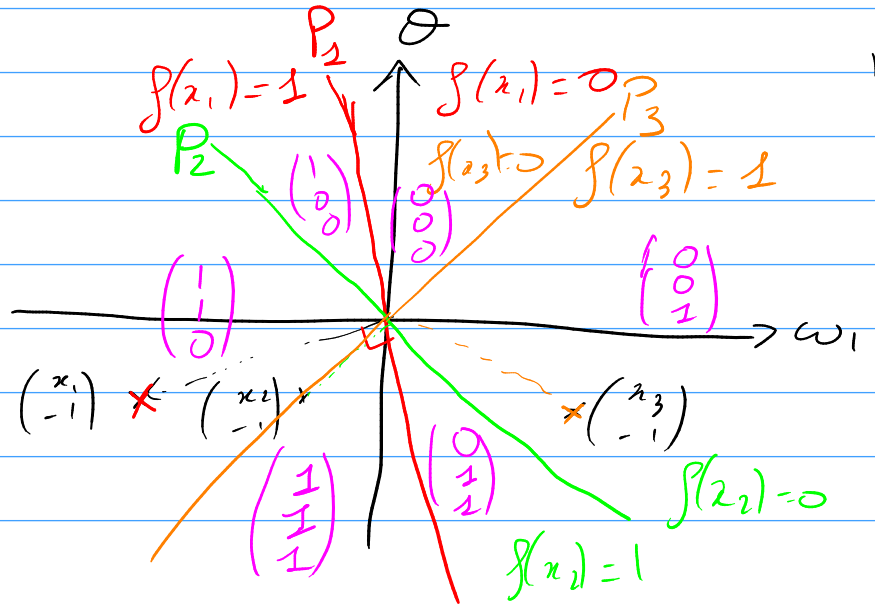
Growth Function and VC-dimension for neural networks

I - Perceptrons:

Th: Let N be the real-weight perceptron with $n \in \mathbb{N}$ (real) inputs and H be the set of functions it computes:

$$H = \{ x \in \mathbb{R}^n \to \operatorname{sgn}(w_1 x_1 + \dots + w_n x_n - \theta) : (w_1, \dots, w_n, \theta) \in \mathbb{R}^{n+1} \}$$

Then $\Pi_H(m) = 2 \sum_{k=0}^{n} \binom{m-1}{k}$



$n = 1$ $\quad P_j = \{ \omega_1 x_j - \theta = 0 \}$

$\operatorname{sgn}(\omega_1 x_1 - \theta)$

$n = 1$

$\Pi_H(3) = 6 = 2 \left(\sum_{k=0}^{1} \binom{2}{k} \right)$

$= 2(1 + 2)$ ✓

$\{x_1, x_2, x_3\}$ is shattered by H.

Lemma: For a set $S = \{a_1, \dots a_m\} \subseteq \mathbb{R}^n$, let $P_1, \dots P_m$ be the hyperplanes $P_i = \{(\omega, 0) \in \mathbb{R}^{n+1} : \omega^T x_i - 0 = 0\}$

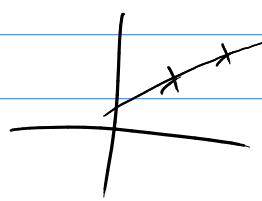
Then $|H_{1S}| = CC\left(\mathbb{R}^{n+1} \setminus \bigcup_{i=1}^{m} P_i\right)$

↑ nbr of connected components -

Proof: Homework 3 -

Remark: $x_i = \begin{pmatrix} x_i \\ -1 \end{pmatrix} \rightarrow P_i = \{v \in \mathbb{R}^{n+1} : a^T z_i = 0\}$

Def: a set of points in $\mathbb{R}^n$ is in general position if no subset of $k+1$ points lies on a $(k-1)$-plane for all $k = 1, \dots n$ -

ex:  - $\mathbb{R}^n$ - not in general position

Rep: The $(a_i)_{1 \le i \le m}$ are in general position iff every subset of up to $n+1$ points in $\{r_1, \dots r_m\}$ is linearly independant.

Lemma: for $m, d \in \mathbb{N}$, suppose $T = \{r_1, \dots r_m\} \subset \mathbb{R}^d$ has every subset of no more than d points linearly independant.

Let $P_i = \{v \in \mathbb{R}^d : v^T a_i = 0\}$ for $i = 0, \dots m$ and define $C(T) = CC\left(\mathbb{R}^d \setminus \bigcup_{i=1}^m P_i\right)$

Then $C(T)$ depends only on m and d; $C(T) = C(m,d)$ and for all $m, d \ge 1$: $C(m,d) = 2 \sum_{k=0}^{d-1} \binom{m-1}{k}$

[then we apply the Lemma with $d = n+1$].

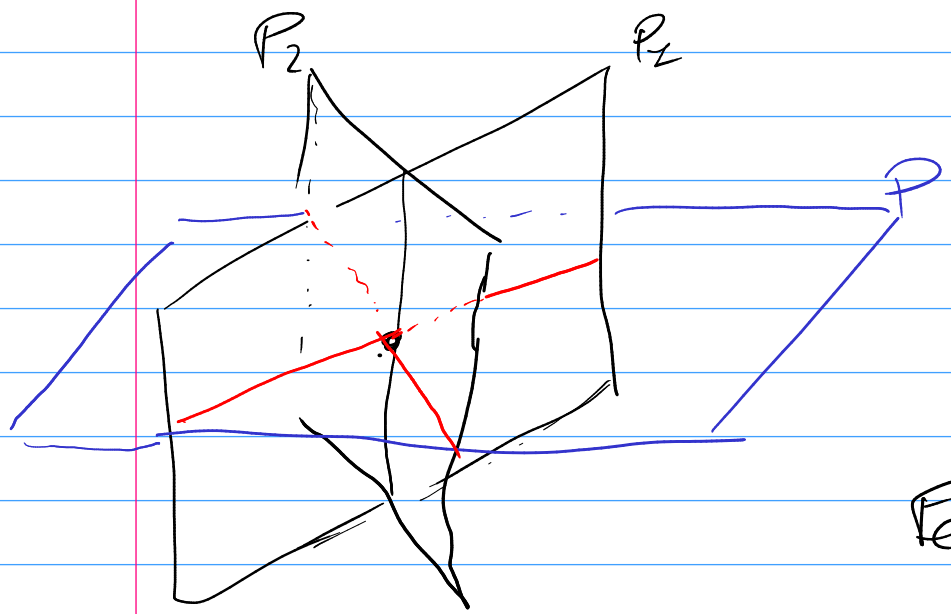
Prof: $c(m,1) = 2$ for $m \geq 1$

$c(1,d) = 2$ for $d \geq 1$

Prof by induction: assume claim is true for all $T \subset \mathbb{R}^j$ $|T| \leq m$, $j \leq d$

Suppose $P_1, \dots P_m$ satisfy the independence condition

and let P be another hyperplane also satisfying the independence conditions.



each intersection of a P_i with P is a $(d-2)$ plane in the $(d-1)$ hyperplane P and all these planes satisfy the independence condition

Each of them cuts a cc of $\mathbb{R}^d \setminus \bigcup_{i=1}^{m} P_i$ in two.

Hence, $C(m+1, d) = C(m, d) + C(m, d-1)$

$$= 2\left(\sum_{k=0}^{d-1} \binom{m-1}{k} + \sum_{k=0}^{d-2} \binom{m-1}{k} \right)$$

$$= 2\left(\binom{m-1}{0} + \sum_{k=1}^{d-1} \left(\binom{m-1}{k} + \binom{m-1}{k-1} \right) \right)$$

$$= 2\left(\binom{m-1}{0} + \sum_{k=1}^{d-1} \binom{m}{k} \right) = 2 \sum_{k=0}^{d-1} \binom{m}{k}$$

→ if $S = \{a_1 .. a_n\}$ is in general position, then

$$|H_{1S}| = C(m, n+1) = 2 \sum_{k=0}^{n} \binom{m-1}{k}$$

if S is not in general position, it can be seen that

$$|H_{1S}| \leq C(m, n+1)$$

→ $H_H(m) = 2 \sum_{k=0}^{n} \binom{m-1}{k} = \begin{cases} 2^m & \text{if } n \geq m-1 \\ 2^m - 2 \sum_{k=n+1}^{m-1} \binom{m-1}{k} < 2^m & \text{otherwise} \end{cases}$

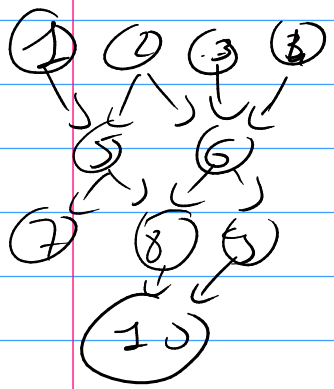
→ $\boxed{VCDim(N) = n+1}$

Nr of parameters $(\omega_1, \omega_2 .. \omega_n, \theta)$

Growth function of linear threshold networks

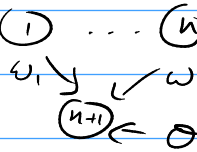
activation function $\varphi(u) = \operatorname{sgn}(u)$

k units $1, 2, \dots, k$ such that the output of i is an input of j only if $i < j$



I denote by W the total number of weights and thresholds

(for a perceptron $k = n+1 = W$ $\quad VC\text{-}dim = k$)



H = the set of functions that an LTN with k units and W parameters

Thm: For $m \geq W$ the growth function of H is

$$\Pi_H(m) \leq \left(\frac{e\,m\,k}{W}\right)^W$$

Hence $VC\,dim(H) \leq 2W \log\left(\frac{2k}{\ln(2)}\right)$

Proof:

- first consider unit (1) (below the 1st layer):
  it is a perceptron, hence the number of dichotomies it computes is at most

$$D_1(S) = \left(\frac{em}{d_1}\right)^{d_1}$$

  where d_1 is the number of parameters attached to unit 1.

- consider now unit ℓ $\quad 2 \le \ell \le h$.
  its output depends on the output of unit $j < \ell$ and on its d_ℓ parameters.

  ⇒ for each of the $D_{\ell-1}(S)$ states of the network on units $1, \dots \ell-1$, it can reach at most

$$D_\ell(S) \le D_{\ell-1}(S)\left(\frac{em}{d_\ell}\right)^{d_\ell}$$

different states

so by induction $D_h(S) \le \prod_{\ell=1}^{h} \left(\frac{em}{d_\ell}\right)^{d_\ell}$

and $\ln D_h(S) \le \sum_{\ell=1}^{h} d_\ell \ln\left(\frac{em}{d_\ell}\right)$

→ $T_H(m) := \max_Q D_Q(S)$ over all possible S:

$$\ln T_H(m) \le \sum_{e=1}^{k} d_e \ln \frac{em}{d_e}$$

$$\frac{1}{W} \ln T_H(m) + \ln\left(\frac{W}{em}\right) \le \sum_{e=1}^{k} \frac{d_e}{W} \ln \frac{em}{d_e} + \ln \frac{W}{em}$$

$$= \sum_{e=1}^{k} \frac{d_e}{W} \ln \frac{em}{d_e} + \sum_{e=1}^{k} \frac{d_e}{W} \ln \frac{W}{em}$$

$$= \sum_{e=1}^{q} \frac{d_e}{W} \ln \frac{W}{d_e} = \sum_{e=1}^{q} p_e \ln 1/p_e = H(p)$$

where $p_e = \frac{d_e}{W}$ $\quad \sum_{e=1}^{n} p_e = \frac{\sum d_e}{W} = 1$

$$\text{but} \quad H(p) = \sum_{e=1}^{n} p_e \ln 1/p_e \le \ln k \quad \left(\text{reached for } \frac{d_e}{W} = 1/k\right)$$

$$\rightarrow T_H(m) \le \left(\frac{emk}{W}\right)^W \quad \square$$

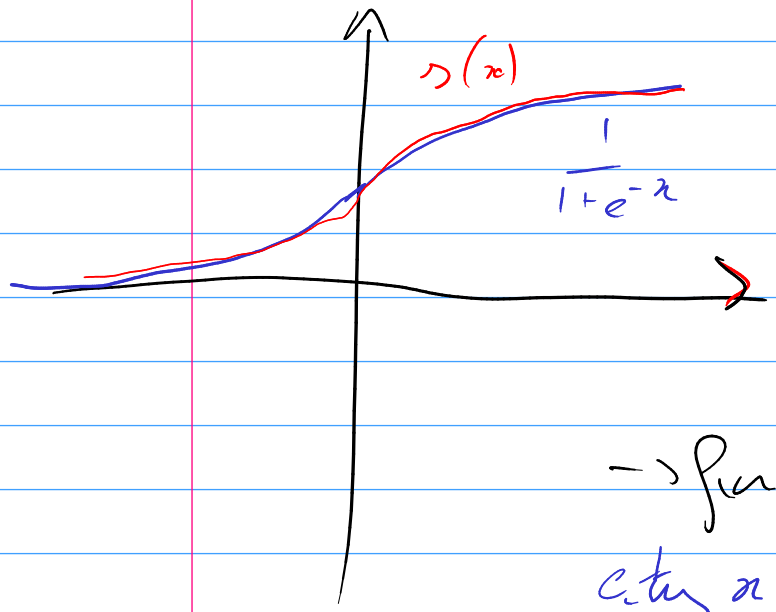
NB: one can prove that there exist Neural architectures N such that $VC\dim(H) \ge 3W/5$

What if other activation functions? → conditions are needed!

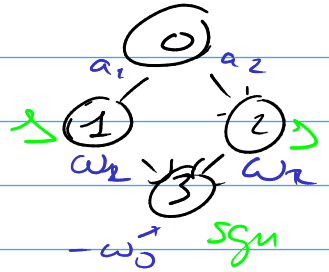
Thm: define $s(x) = \frac{1}{1+e^{-x}} + c x^3 e^{-x^2} \sin(x)$ for $c > 0$

One can show that if c is small enough, $s(x) \xrightarrow[x \to \infty]{} 1$

$s''(x) = \begin{cases} < 0 & \text{if } x > 0 \\ > 0 & \text{if } x < 0 \end{cases}$ $\quad s(x) \xrightarrow[x \to -\infty]{} 0$



We consider the neural net N
one hidden layer
two first-layer computation units
one output unit



→ functions of H are of the form: $x \to \mathrm{Sgn}(\omega_0 + \omega_1 s(a_1 x) + \omega_2 s(a_2 x))$

entry x and parameters $a_1, a_2, \omega_1, \omega_2, \omega_0 \in \mathbb{R}$

then $VC\text{-dim}(H) = +\infty$

Lemma: the class $\mathcal{F} = \{ x \mapsto \mathrm{sgn}(\sin(ax)), a \in \mathbb{R} \}$ of functions defined on $\mathbb{R}$ has $VC\text{-}dim(\mathcal{F}) = \infty$.

This proves the previous theorem since:

$$h_a(x) = s(ax) + s(-ax) - 1$$

($\omega_1 = 1$, $a_1 = a$, $\omega_2 = 1$, $a_2 = -a$, $\omega_0 = -1$)

$$= \underbrace{\frac{1}{1+e^{-ax}} + \frac{1}{1+e^{+ax}}}_{1} - 1$$

$$+ 2c\,(ax)^3 e^{-a^2x^2} \sin(ax)$$

$$= 2c\,(ax)^3 \sin(ax)\, e^{-a^2x^2}$$

for $a > 0$ and $x > 0$, has the same sign as $\sin(ax)$. □

Proof of the Lemma:

For $d \in \mathbb{N}$, choose $a_i = 2^{i-1}$ for $i = 1 \dots d$.

We show that $\{a_1, \dots a_d\}$ is shattered by F.

Since discreteness, this shows that $VCd_0(F) = \infty$

For $(b_1, \dots b_d) \in \{0,1\}^d$ let $c = \sum_{j=1}^{d} 2^{-j} b_j + 2^{-(d+1)} = 0.b_1 b_2 \dots b_d 1$

where $a = 2\pi c$

For all i

$$\operatorname{sgn}(\sin(a a_i)) = \operatorname{sgn}\left(\sin\left(2\pi\left(\sum_{j=1}^{d} 2^{-j} b_j + 2^{-(d+1)}\right) 2^{i-1}\right)\right)$$

$$= \operatorname{sgn}\left(\sin\left(\sum_{j=1}^{i-1} 2^{i-j}\pi\, b_j + \pi b_i + \sum_{j=i+1}^{d} 2^{i-j}\pi\, b_j + 2^{i-d-1}\pi\right)\right)$$

$$= \operatorname{sgn}\left(\sin\left(\pi\left(b_i + \sum_{j=1}^{d-i} 2^{-j} b_{i+j} + 2^{-(d-i+1)}\right)\right)\right)$$

$\in]0,1[$

$$= 1 - b_i$$

□