Training data $(x_i, y_i)$ $i = 1, \dots, n$

algo : $\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) + \Omega(\theta)$ $\quad + \Omega(\theta)$

model class $f_\theta$ $\longrightarrow$ linear models, $f_\theta(x) = \varphi(x)^T \theta$

regularizer $\Omega(\theta)$

explicit: $\ell_2$

implicit: kernel $\ell_2$

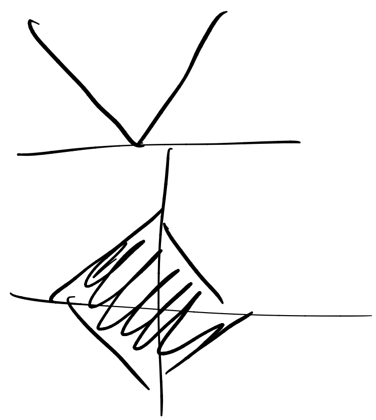
chapter 8

$\Omega(\theta) = \|\theta\|_1 = \sum_{i=1}^{d} |\theta_i|$ $\longleftarrow$ $\ell_1$

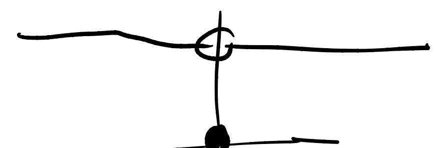
↳ sparsity inducing

1D :

2D



$\Omega(\theta) = \sum 1_{\theta_i \neq 0}$ — $\ell_0$



high dimensional result : $\frac{\sigma^2 d}{n} \longrightarrow \frac{\sigma^2 k \log d}{n}$

# of non zero components of $\theta_*$

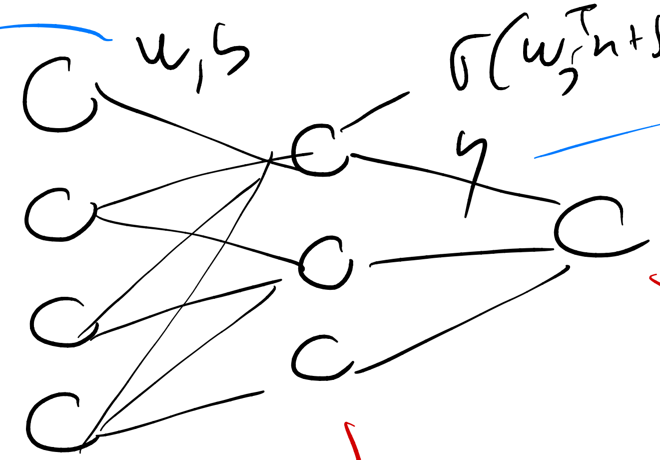
price of adaptivity

**Neural network:** $x \in \mathbb{R}^d$

$$f(x) = \sum_{j=1}^{m} \eta_j \, \sigma(w_j^T x + b_j)$$

"bias" "intercept" "constant term"

input weights $\mathbb{R}^{(d+1)m}$




output weights $\eta \in \mathbb{R}^m$

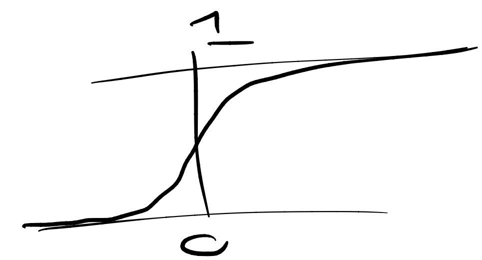
output layer

hidden layer

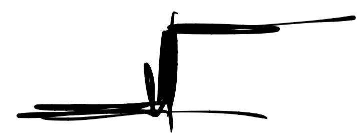
input layer

**activation functions:**

$$\sigma(u) = \frac{1}{1+e^{-u}} \quad \text{sigmoid}$$

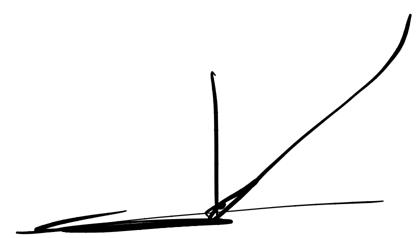



$$\sigma(u) = \mathbb{1}_{u>0}$$



Rectified Linear Unit

$$\sigma(u) = \max\{u, 0\} = u_+$$



**One property:** **not a polynomial**

log loss + $f(x) = \sigma(g(x))$

$\hookrightarrow \sum_j u_j \sigma(w_j^T x + b_j)$

$\Rightarrow$ logistic loss on $g(x)$

optimization error : $\min_{\Theta} \mathbb{E}\left[\ell\left(y_i, \sum_j u_j \sigma(w_j^T x_i + b_j)\right)\right]$

$\Theta = \{(w_j, b_j, u_j)\}$

## Neural tangent kernel (Chizat, B, Cyglan 2019)

Mean field regime $\mathcal{N}(0, \frac{1}{m^2})$

$$f(x) = \sum_{j=1}^{m} u_j (w_j^T x + b_j)_+ = g(x, \Theta) \quad \text{(Bietti B, 622)}$$

$\hookrightarrow \mathcal{N}(0, \frac{1}{m}) \Rightarrow \text{std} \sim \frac{1}{\sqrt{m}}$

$m \to +\infty$, parameters move infinitesimally

$$f(x) = g(x, \Theta_0) + (\Theta - \Theta_0)^T \text{Jacobian}\, g(x, \Theta_0)$$

Lazy regime

Focus on RELU: $\sigma(u) = u_+$

$\sigma(\lambda u) = \lambda \sigma(u)$ if $\lambda > 0$

positive homogeneity

$$f(x) = \sum_{j=1}^{m} a_j (w_j^\top x + b_j)_+ \qquad \begin{pmatrix} w_j \\ b_j \end{pmatrix} \longrightarrow \frac{1}{\alpha_j} \begin{pmatrix} w_j \\ b_j \end{pmatrix}$$

$$a_j \longrightarrow \alpha_j a_j$$

$\alpha_j > 0$

Assumption $\begin{pmatrix} w_j \\ b_j \end{pmatrix}$ are bounded

$$\|w_j\|_2^2 + b_j^2 = 1$$

(if not normalized, take $\alpha_j = \|w_j\|_2^2 + b_j^2$ )

implicit bias $\quad \inf_{\alpha_j} \sum \frac{\|w_j\|_2^2 + b_j^2}{\alpha_j} + a_j^2 \alpha_j$

$$= \sum_j |a_j| \cdot \sqrt{\|w_j\|_2^2 + b_j^2}$$

(Chizat + B, 2020)

$\Rightarrow$ diagonal linear network

$w = u \circ v$

Estimation error: $\mathcal{F} = \{ f_\theta(x) = \sum_{j=1}^{m} \eta_j (w_j^T x + b_j)_+ ,$

$\forall j: \|w_j\|_2^2 + b_j^2 = 1$

$\|\eta\|_1 \leq D \}$

Estimation error $\leq 2 \sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| + \underbrace{\text{optim. error}}_{0}$

$$\mathbb{E} \sup_{f \in \mathcal{F}} R(f) - \hat{R}(f) \leq 2 \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \cdot \ell(y_i, f(x_i))$$

$\hookrightarrow$ rademacher r.v $\in \{-1, 1\}$

$$\leq 2 G \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \cdot f(x_i)$$

Lipschitz constant of the loss

$$\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)=\mathbb{E}\sup_{\substack{a\\ w,b}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\sum_{j=1}^{m}a_j(w_j^Tx_i+b_j)_+$$

$\|a\|_1\leq D$ normalized

$$=D\,\mathbb{E}\sup_{j\in\{1,\dots,m\}}\sup_{\{w_j,b_j\}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(w_j^Tx_i+b_j)_+\right|$$

$$=D\,\mathbb{E}\sup_{\substack{w,b\\ \|w\|_2^2+b^2=1}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(w^Tx_i+b)_+\right|$$

lemma:

$$\sup_{\|a\|_1\leq D}a^Tz=D\cdot\|z\|_\infty$$

$$a^Tz=\sum a_jz_j\leq\sum|a_j|\max_i|z_i|\leq\|a\|_1\|z\|_\infty.$$

$$\leq 2D\,\mathbb{E}\sup_{\|w\|_2^2+b^2=1}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\underbrace{(w^Tx_i+b)}_{\binom{w}{b}^T\binom{x_i}{1}}\right|$$

contraction principle

$$=2D\,\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\binom{x_i}{1}\right\|_2$$ by Cauchy Schwarz

$$\leq 2D\sqrt{\mathbb{E}\left\|\phantom{\frac{1}{n}}\right\|_2^2}=2D\sqrt{\mathbb{E}\frac{1}{n^2}\sum\left\|\binom{x_i}{1}\right\|^2}$$

(assuming $\|x_i\|\leq 1$ a.s)

$$\leq\frac{2D}{\sqrt{n}}\sqrt{2}\sim\frac{D}{\sqrt{n}}$$

if $\forall j \in \{1, \dots, m\}$, $\|w_j\|_2^2 + s_j^2 = 1$

$$\|a\|_1 \leq D$$

Uniform deviations were bounded by $\frac{D}{\sqrt{n}}$

Done in practice by "weight decay"

(penalizing by $\sum_j \|w_j\|_2^2 + a_j^2 + s_j^2$)

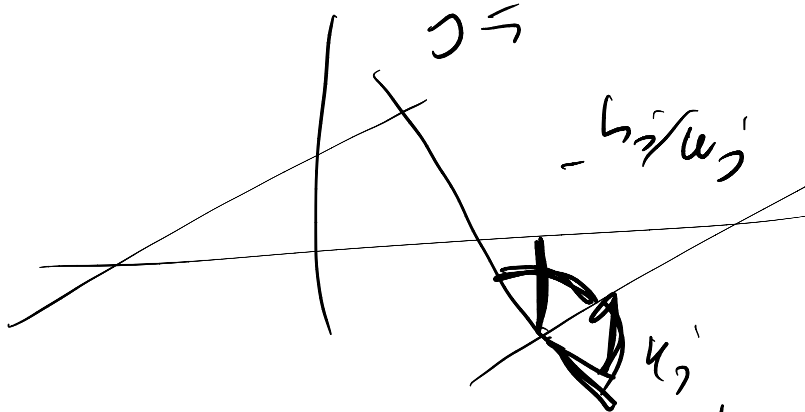
$\Leftrightarrow \|a\|_1$

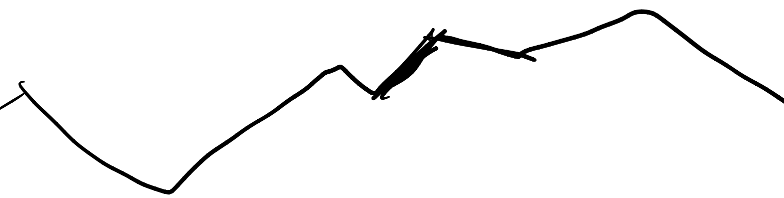
Approximation error

- Universality:
- Precise bound on approximation error ) $m$ infinite
- How big $m$ should be? $\Rightarrow$ link with Frank Wolfe

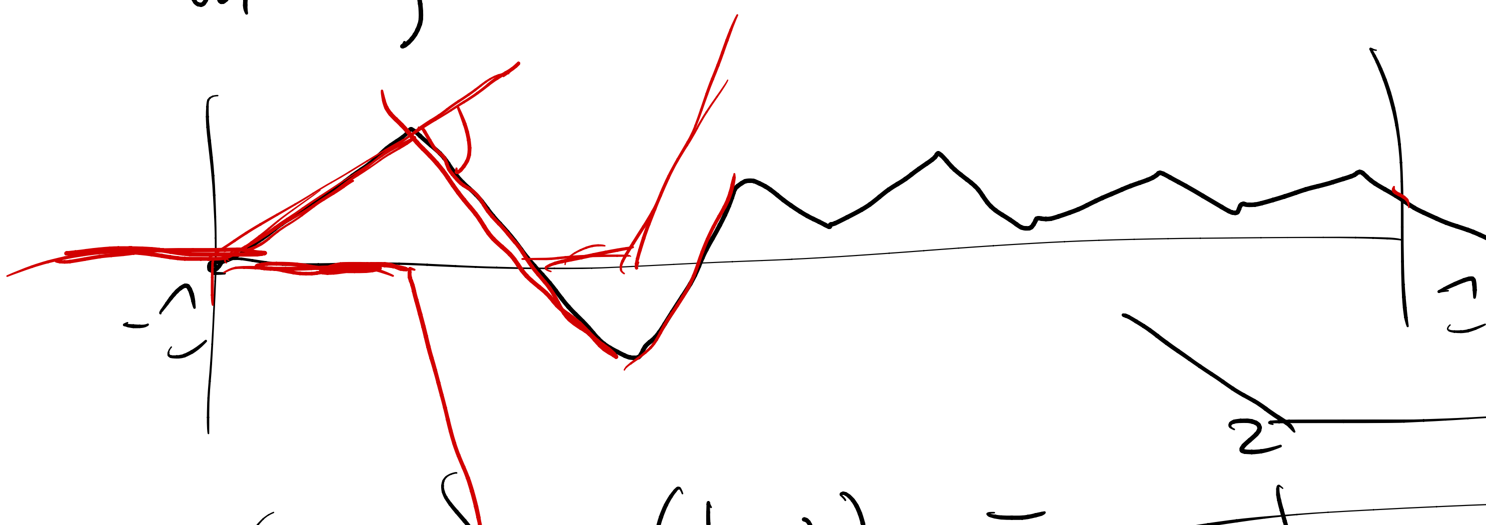
$$\boxed{m \simeq M}$$

Universality in 1D : $\sum_{j=1}^{m} a_j (w_j x + b_j)_+$

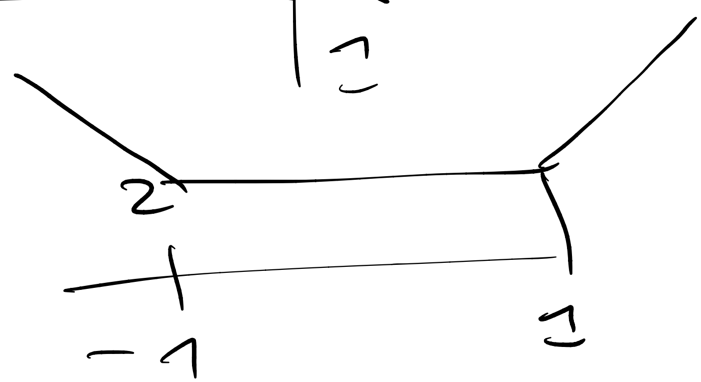


Result 1 : a piecewise affine continuous fct with m kinks can be exactly recovered by m+2 relus on any interval [-1, 1]




$(x+1)_+ + (1-x)_+ =$




Result 2

<u>Result</u> (Leshno, 1993)

NN with one hidden layer are universal iff $\sigma$ is not a polynomial.

## Approximation properties with infinitely many neurons

$$f(x) = \sum_{j=1}^{m} \eta_j (w_j^T x + b_j)_+ \qquad (w_j, b_j) \in K$$

$\|\eta\|_1$

→ corresponds to $\nu = \sum_{j=1}^{m} \eta_j \delta_{(w_j, b_j)}$

$\int |d\nu| = \|\eta\|_1$

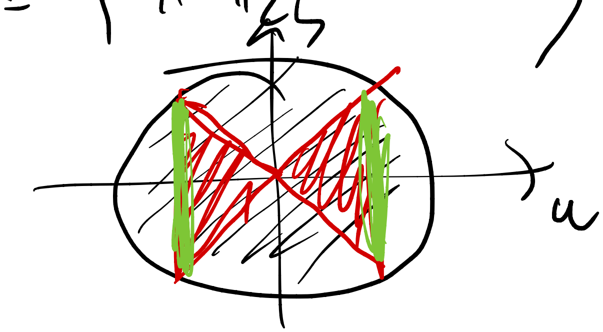
$$f(x) = \int_K (w^T x + b)_+ \, d\nu(w, b) \qquad p(w,b)\, dw\, db$$

constraint on total variation $\int |d\nu(w,b)| = \int |p(w,b)| \, dw\, db$

<u>NB</u>: if $\nu$ has a density $p$ wrt Lebesgue

$K = \{ \|w\|_2^2 + b^2 = 1 \}$ or $\{ \|w\|_2^2 + b^2 \le 1 \}$

$K = \{ \|w\|_2 = \frac{1}{\sqrt{2}}, \ |b| \le \frac{1}{\sqrt{2}} \}$



Definition: $\varphi$ : $B$ ball of radius $1 \longrightarrow \mathbb{R}$

$$\gamma(\varphi) = \inf \int_K |dv(u,s)| \quad \text{s.t.} \quad \forall x \in B,\ \varphi(x) = \int (u^T x + s)_+ \, dv(u,s)$$

Prop: $\gamma$ is a norm on $\mathcal{F} = \{ \varphi,\ \gamma(\varphi) < +\infty \}$

"variation norm"

$\|u\|_2 \leq D \Rightarrow$ estimation error $\leq \dfrac{D}{\sqrt{n}}$

$\gamma(\varphi) \leq D$

1D: $\varphi : [-1,1] \to \mathbb{R}$

$$\int (x+s)_+ \, dv_+(s) + \int (-x+s)_+ \, dv_-(s)$$

$K = \{ \|u\|_2 = 1,\ |s| \leq 1 \}$

$= \{-1\} \times [-1,1] \cup \{+1\} \times [-1,1]$

1D: Taylor formula with integral remainder

f is $C^2$

$$f(x) = f(-1) + f'(-1)(x+1) + \int_{-1}^{x} f''(s)(x-1)\,ds$$

$$\int_{-1}^{1} f''(s)\,(x-s)_+\,ds$$

$$f(x) = f(-1) + \int_{-1}^{x} f'(s)\,ds$$

$$1 = \frac{(x+1)_+ + (1-x)_+}{2} \text{ on } (-1,1)$$

$$x = \frac{(x+1)_+ - (1-x)_+}{2} \text{ on } (-1,1)$$

$$\gamma(f) \leq |f(-1)| + |f'(-1)| + \int_{-1}^{1} |f''(s)|\,ds$$

going multiple dimensions

$$f(x) = \frac{1}{(2\pi)^d} \int \hat{f}(\omega)\, e^{i\omega^T x}\, d\omega$$

Fourier inversion formula

$$\gamma(f) \leq \frac{1}{(2\pi)^d} \int |\hat{f}(\omega)|\, \underbrace{\gamma\left(x \mapsto e^{i\omega^T x}\right)}_{\left(1 + \|\omega\|_2^2\right)} d\omega$$

kernel norm : $\int \frac{|\hat{f}(\omega)|^2}{\hat{q}(\omega)}$

Summary = if $f^*(x) = g^*(w^T x)$

adaptivity