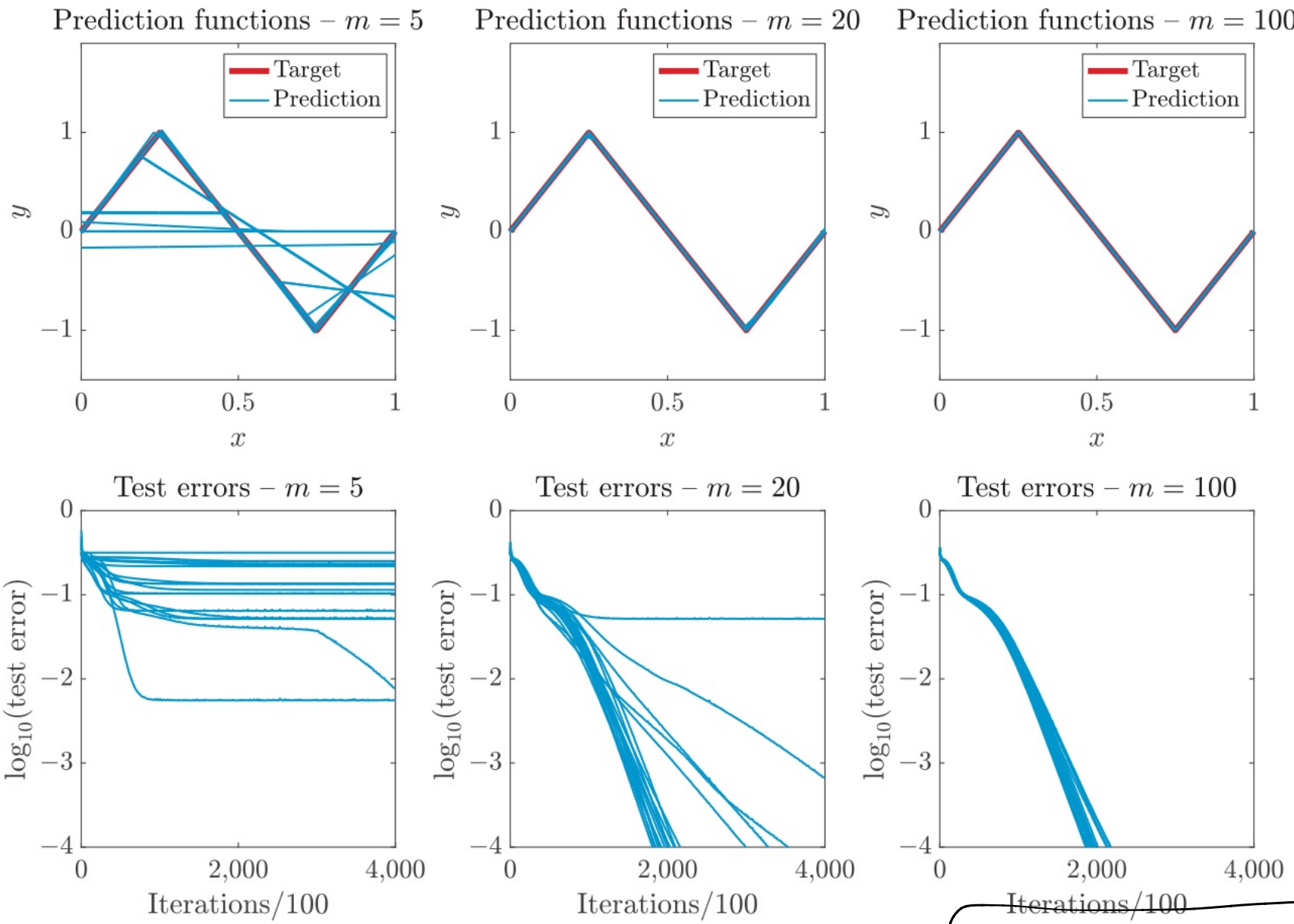
$$f(x) = \sum_{j=1}^{m} \eta_j \sigma(w_j x + b_j) +$$



over parameterized

=> Chapter 12

Exam: December 11 $9^h \rightarrow 11^h 30$

2 sides of A4 page of notes

short answers only.

HW solutions online by Dec.

Data (x_i, y_i) $i \in 1, \dots, n$

$$\hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) + \Omega(\theta)$$

$y \in \mathbb{R}$ square loss $(y_i - f_\theta(x_i))^2$

$y \in \{-1, 1\}$ 0-1 loss $1_{y_i \neq f_\theta(x_i)} \Rightarrow$ logistic loss $\log(1 + e^{-y_i f_\theta(x_i)})$

Empirical minimizers:

Model: $f_\theta : \mathcal{X} \to \mathbb{R}$

- linear models: $f_\theta(x) = \theta^\top \varphi(x)$ $\varphi(x) \in \mathbb{R}^d$
  chapter 3, 4, 5, 8
- kernel methods (positive definite)
  chapter 7
  $f_\theta(x) = \langle \theta, \varphi(x) \rangle$ $\varphi(x) \in \mathcal{H}$
  kernel $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$.
  - linear
  - polynomial
  - gaussian

  | linear in advance

- Today: neural networks

  "feature" learning

Neural networks: single hidden layer fully connected $x \in \mathbb{R}^d$

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

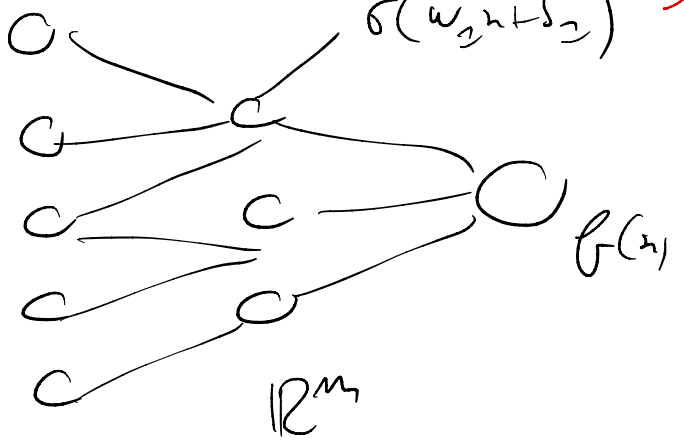
$j = 1, \dots, m$

η_j: output weight $\in \mathbb{R}$

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

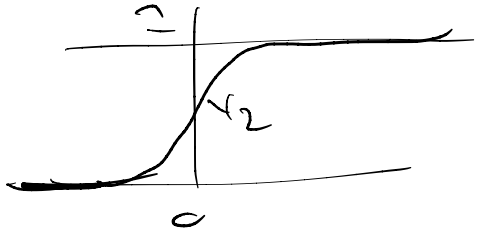
s_j: "biases" = "intercept" = constant term

$\sigma(w_j^T x + s_j)$: hidden neurons



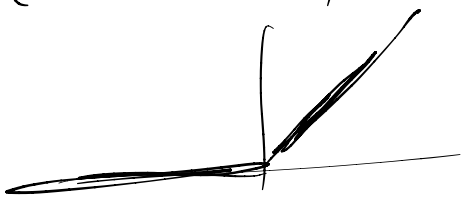
σ: activation function

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



Rectified linear unit: ReLU

$\sigma(u) = \max(0, u)$



No activation at the last layer often for classification: $f(x) = \sum \eta_j \sigma(w_j^T x + s_j)$ $g(x) = \text{sigmoid}(f(x)) \in [0,1]$

logistic loss on $f(x)$ $\in$ { cross-entropy loss $\in$ maximum likelihood $p(y=1) = g(x)$

$-1_{y_i=1} \log g(x_i) - 1_{y_i=-1} \log(1 - g(x_i))$

Algorithms : minimize $\hat{R}(f_\theta)$ with $f_\theta(x) = \sum_{j=1}^{m} u_j \sigma(w_j^T x + b_j)$

$\theta = (w, u, b) \in \mathbb{R}^{m \times (d+2)}$

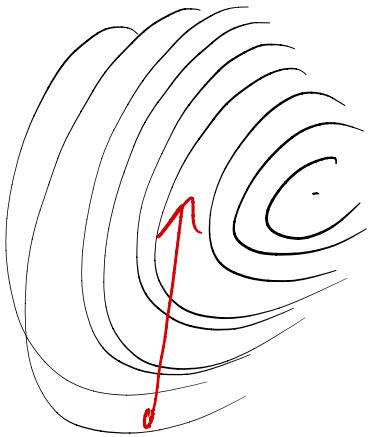
using SGD or ADAM or ...

two types : single pass → directly minimizing the test risk

multiple passes ↳ not enough data

↳ minimize training risk need to control overfitting

Convex



Non Convex

0 gradient ⇒ global optimum.

$\mathbb{E}\hat{R}(\theta_t) - \hat{R}(\theta_s) \leq \frac{C}{\sqrt{t}}$

$\min_{t \in (0,T)} \mathbb{E}\|\nabla \hat{R}(\theta_t)\|^2 \leq \frac{1}{\sqrt{T}}$

Error decomposition: $R(\hat{f}) - R_* = \underbrace{R(\hat{f}) - \inf_{g \in \mathcal{F}} R(g)}_{\text{estimation error}} + \underbrace{\inf_{g \in \mathcal{F}} R(g) - R^*}_{\text{approximation error}}$

$\hat{f} \in \mathcal{F}$

for $\hat{f}$ an approximate ERM

$\leq$ 2 · uniform deviations + optimization error $) = 2 \sup_{f \in \mathcal{F}} |\hat{R}(f) - R(f)|$

$\hat{R}(\hat{f}) - \inf_{f \in \mathcal{F}} \hat{R}(f) \leq \delta$

empirical risk — expected risk

---

consider $\sigma(u) = \text{ReLU}(u) = (u)_+ = \max\{0, u\}$.

homogeneity $\sigma(\lambda u) = \lambda \sigma(u)$ for $\lambda > 0$

$$f(x) = \sum_{j=1}^{m} \alpha_j a_j (w_j^T x + b_j)_+ \quad \text{if } \alpha_j > 0$$

α_j

$$\mathcal{F} = \left\{ \sum_{j=1}^{m} a_j (w_j^T x + b_j)_+ \text{ s.t. } \|w_j\|_2^2 + b_j^2 = 1 \text{ and } \|a\|_2 \leq D \right\}$$

penalty on weights: ℓ_2-norm^2: $\frac{1}{2} \sum_{j=1}^{m} \frac{a_j^2}{\alpha_j^2} + \frac{\|w_j\|_2^2 + b_j^2}{\alpha_j^2}$

(= weight decay)

(= "implicit bias")

of SGD

see Ch-12

optimum: $|a_j| = \sqrt{\|w_j\|_2^2 + b_j^2}$

$\inf_{\alpha_j} 2 \sqrt{a_j^2} \sqrt{\|w_j\|_2^2 + b_j^2}$

$\Rightarrow \boxed{\sum_{j=1}^{m} |a_j| \sqrt{\|w_j\|_2^2 + b_j^2}}$

$\mathcal{F} = \left\{ \sum_{j=1}^{m} a_j (w_j^T x + b_j)_+ \text{ s.t. } \|w_j\|_2^2 + |b_j|^2 = 1 \text{ and } \|a\|_2 \le D^{R^2} \right\}$

assume $\|x_i\|_2 \le R$

If $\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{R}(f)$

$R(\hat{f}) - \min_{g \in \mathcal{F}} R(g) \le \sup_{f \in \mathcal{F}} \hat{R}(f) - R(f) + \sup_{f \in \mathcal{F}} R(f) - \hat{R}(f)$

estimation error

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

data ε

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

Rademacher complexity

average of loss ℓ G-Lipschitz

$\le 2G\, \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i \varepsilon_i f(x_i)$

(typically $\frac{D}{\sqrt{n}}$)

$\le 2G \times \frac{4DR}{\sqrt{n}}$

$$\mathcal{F} = \left\{ \sum_{j=1}^{m} \eta_j (w_j^T x + b_j)_+ \text{ s.t. } \|w_j\|_2^2 + b_j^2 \le 1 \text{ and } \|\eta\|_2 \le D \sqrt{R^2} \right\}$$

assume $\|x\|_2 \le R$

$$\|\eta\|_1 = \sum_{j=1}^{m} |\eta_j|$$

$$\mathbb{E}_{\substack{\varepsilon \\ \text{Data}}} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i \varepsilon_i f(x_i)$$

$$= \mathbb{E} \sup_{\eta, w, b} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \sum_{j=1}^{m} \eta_j (w_j^T x_i + b_j)_+$$

$$= \mathbb{E} \sup_{w, b} \sup_{\|\eta\|_2 \le D} \sum_{j=1}^{m} \eta_j \left[\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (w_j^T x_i + b_j)_+ \right]$$

$\sup_{\|\eta\|_2 \le 1} \eta^T \beta = \|\beta\|_2 \le 1$

proof: $\eta^T \beta = \sum_j \eta_j \beta_j \le \sum_j |\eta_j| \cdot |\beta_j| \le \|\beta\|_2 \sum_j |\eta_j| \le \|\beta\|_2 \cdot 1$

$$= D \, \mathbb{E} \sup_{\substack{\|w\|^2 + b^2 = 1 \\ R^2}} \sup_{j \in \{1, \dots, m\}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (w_j^T x_i + b_j)_+ \right| \quad \text{using lemma}$$

$$\{1, \dots, m\}$$

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

$$\le 2D \, \mathbb{E} \sup_{\|w\|_2^2 + \frac{b^2}{R^2} = 1} \left| \frac{1}{n} \sum_i \varepsilon_i (w^T x_i + b) \right| \quad \text{using the contraction principle with 1-Lipschitz for}$$

$$\left\| \begin{pmatrix} w \\ b/R \end{pmatrix} \right\|_2 = 1 \qquad \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \begin{pmatrix} w \\ b/R \end{pmatrix}^T \begin{pmatrix} x_i \\ R \end{pmatrix} \right|$$

$$\in \mathbb{R}^{d+1}$$

$$= 2D \, \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \begin{pmatrix} x_i \\ R \end{pmatrix} \right\|_2 \le 4 \frac{DR}{\sqrt{n}}$$

by Cauchy-Schwarz

same as chap. 4

$3^4 36$

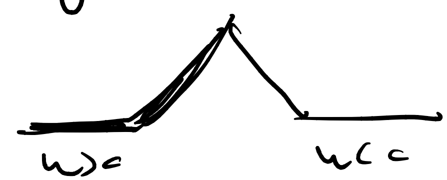
$3^4 56$

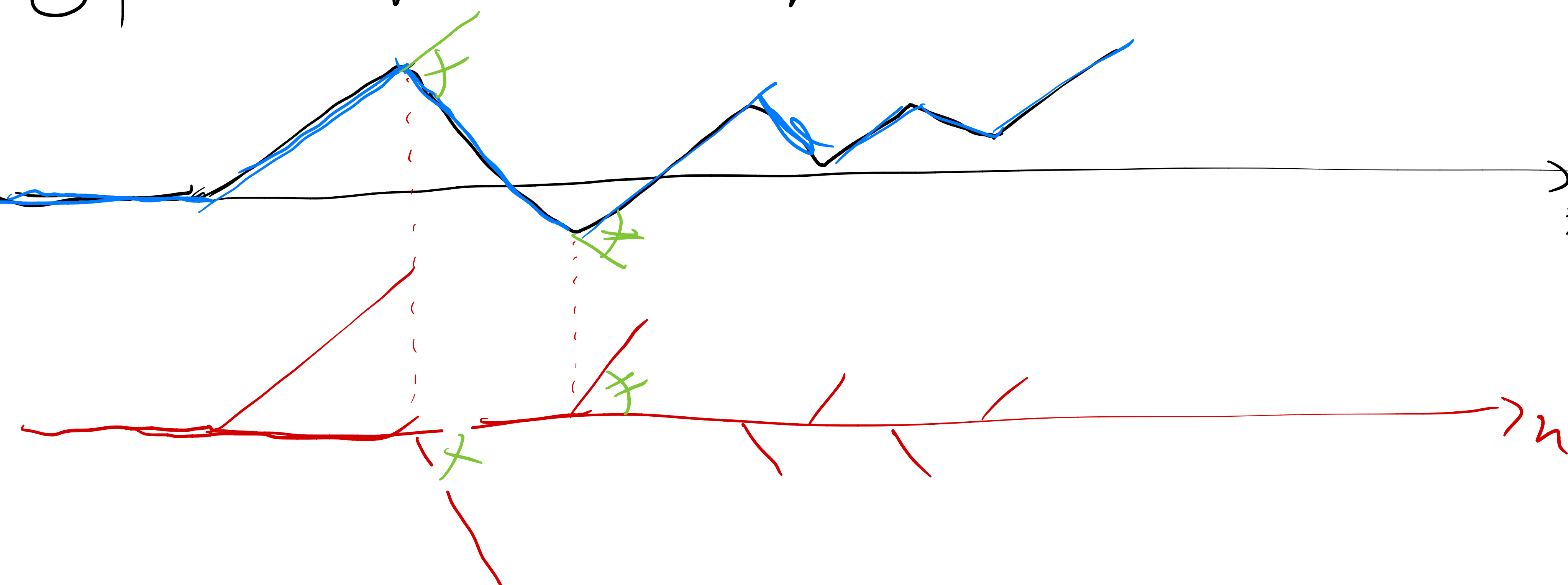
Approximation properties

① universality:
② if $\|a\|_2 \le D$ which functions can I approximate] $m =$ infinity
③ How big m should be? | $m = n$ is enough |

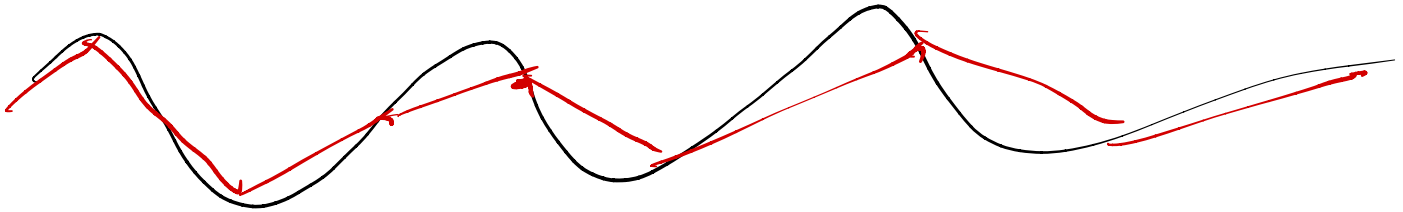
universality: if $f : \mathbb{R}^d \to \mathbb{R}$, can I find $\Theta = (a, w, b)$ s.t. $f_\Theta \approx f$

① in 1D proof by drawing
② piecewise affine continuous function $(wx+b)_+$





(S) for any "reasonable" function, there is a piecewise affine function which is a "close" approximation



$\Rightarrow$ all "reasonable" fcts can be approximated by a neural network

Formulation through measures

$(w_j, b_j) \in K$ compact set

$$f(x) = \sum_{j=1}^{m} \eta_j (w_j^T x + b_j)_+ = \int_K (w^T x + b)_+ \, d\eta(w,b)$$

Dirac

with $d\eta(w,b) = \sum_{j=1}^{m} \eta_j \delta_{w_j, b_j}$

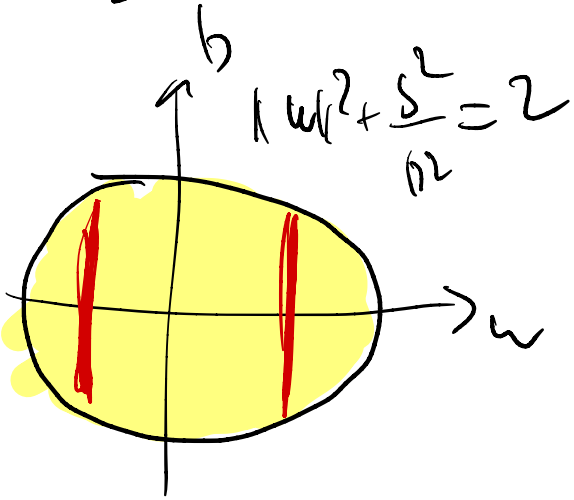
penalty is $\|\eta\| = \sum_j |\eta_j|$

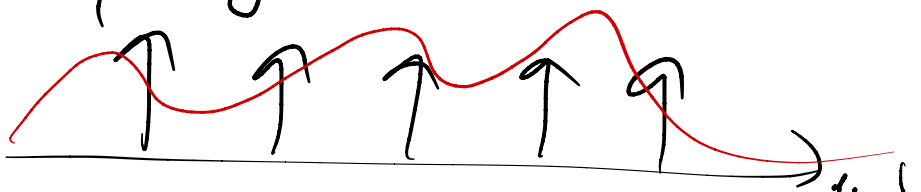
$$= \int |d\eta(w,b)| = \text{total variation}$$

limit when $m \to +\infty$

$$f(x) = \int (w^T x + b)_+ \, d\eta(w,b) \quad \text{with penalty} \int |d\eta(w,b)|$$

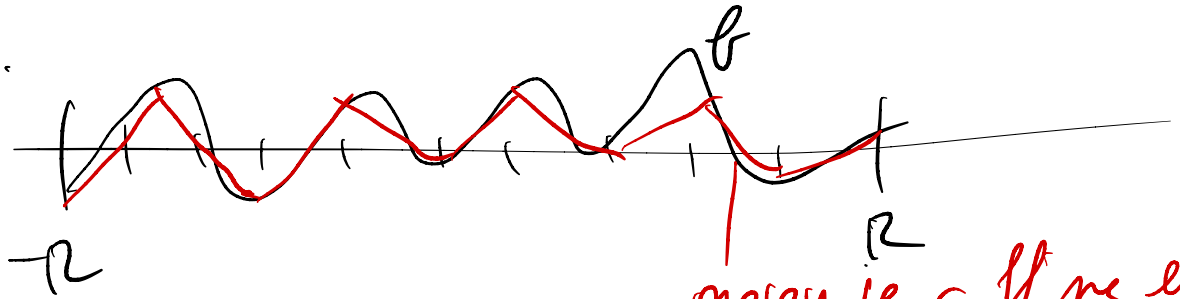
for η any finite measure





$$K = \{\|w\|_2 = 1;\ b \le R\}$$

in 1D : optimal



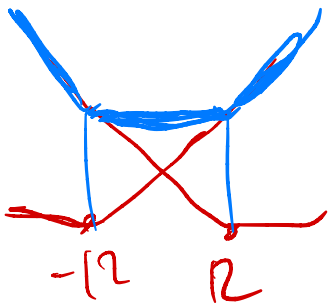
piecewise affine interpolant

optimal : Taylor's formula with integral remainder

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

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

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



$$\frac{f(-R)}{2R}\left[(x+R)_+ + (R-x)_+\right]$$

$$= \int_{[-1,1]\times[-R,R]} (ux+s)_+\,d\gamma(u,s)$$

with $d\gamma(1,s) = \left(\frac{f(-R)}{2R} + f'(R)\right)\delta_{-R} + \frac{f''(s)\,ds}{\text{see beach}}$

$d\gamma(-1,s) = \frac{f(-R)}{2R}\delta_{-R}$

$\int |d\gamma(u,s)| = 1 + \int |f''(s)|\,ds$

From (1) to any dimension

$f : \mathbb{R}^d \to \mathbb{R}$

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

$$e^{iu} = \int (\alpha u + \beta)_+ \, d\eta(\alpha, \beta)$$

$$= \frac{1}{(2\pi)^d} \iint \hat{f}(\omega) (\alpha \omega^T x + \beta)_+ \, d\omega \, d\eta(\alpha, \beta) \Rightarrow \text{universality}$$

with an explicit norm on f

the norm is "better" than for kernels.

why? if $f(x) = h(\omega^T x)$ single ridge model

ex: $h(x_j)$ [See chap 9.3]

the norm will be controlled $\begin{cases} \text{derivatives} \\ \text{dependance on unknown projection} \end{cases}$