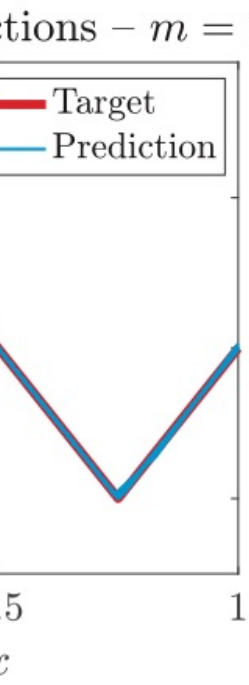
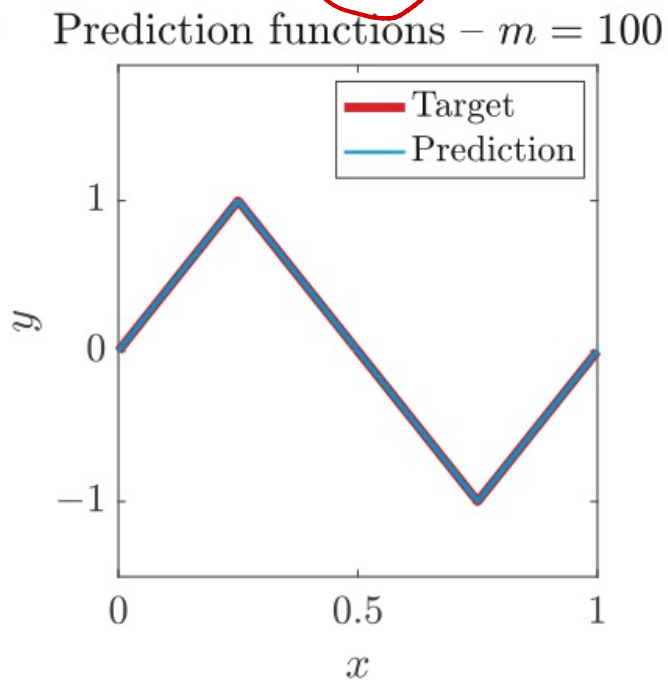
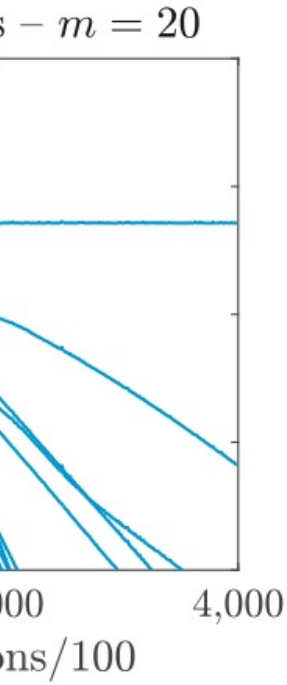
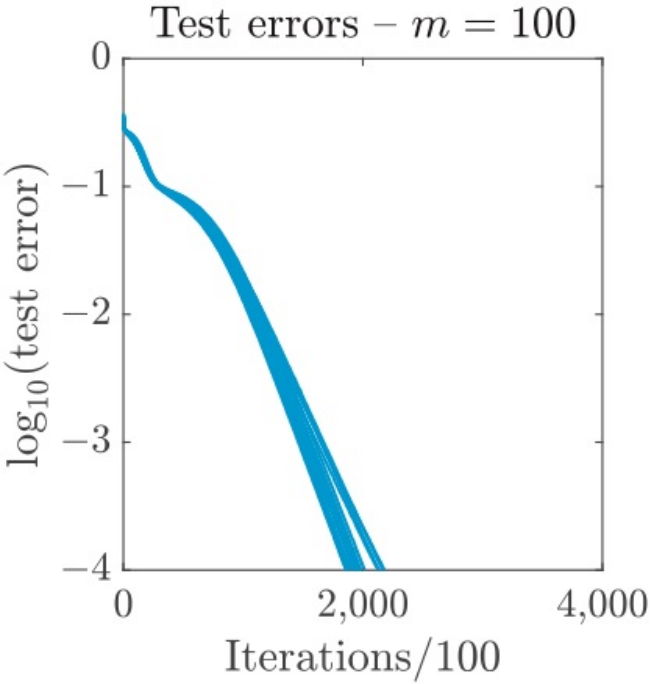

tions – $m = 20$
Target
Prediction
.5
1
x


Prediction functions – $m = 100$
Target
Prediction
1
0
−1
y
0
0.5
1
x


s – $m = 20$
000
4,000
ons/100


Test errors – $m = 100$
0
−1
−2
−3
−4
$\log_{10}$(test error)
0
2,000
4,000
Iterations/100

Neural Networks :

Empirical risk minimization :

$$\min_{f \in \mathcal{F}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))}_{\hat{R}(f)} + \Omega(f)$$

losses : ① $\ell(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$

② $\ell(y, \hat{y}) = 1_{y \neq \hat{y}}$ for classif.

convex surrogates : logistic loss

$\ell(y, f(x)) = \log(1 + e^{-y f(x)})$ $y \in \{-1, 1\}$

function spaces : ① linear functions, $f(x) = \theta^\top \varphi(x)$ → feature vector

↳ parameters

ℓ_2 - penalty for dimension independent bounds

ℓ_1 - penalty for model selection

② kernel methods : $f(x) = \langle \theta, \varphi(x) \rangle$ $\varphi(x) \in \mathcal{H}$ Hilbert space

made possible using $\langle \varphi(x), \varphi(x') \rangle = k(x, x')$

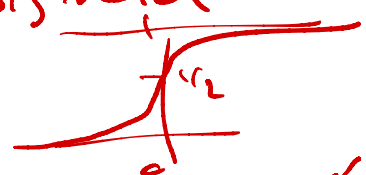
$\varphi(x)$ fixed

Single hidden layer neural network "Shallow"

$x \in \mathbb{R}^d$

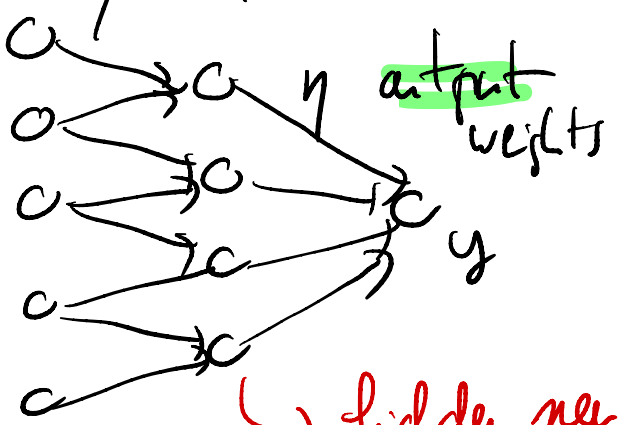
$$f(x) = \sum_{j=1}^{m} \eta_j \, \sigma(w_j^T x + b_j)$$

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



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

w_{ij} input weights



↳ hidden neurons $\sigma(w_j^T x + b_j)$

- No activation function at the output level.

$$f(x) = \eta^T \varphi(x) \quad \text{where } \varphi(x) = \sigma(w_j^T x + b_j)$$

- cross-entropy loss ⇔ maximum likelihood.

$$g(x) = \sigma(f(x)) = \frac{1}{1+e^{-f(x)}} \quad \text{sigmoid} \quad \text{“} \mathbb{P}(y=1|x)$$

$$\Rightarrow \ell(y, g(x)) = -1_{y=1} \log g(x) - 1_{y=-1} \log(1-g(x))$$

$$= -\log\left(\frac{1}{1+e^{-y f(x)}}\right)$$

$$= \log(1+e^{-y f(x)})$$

Lemma $\sigma(-t) = 1 - \sigma(t)$

Optimization: goal: $\min_\theta \mathbb{E}\, \ell(y, f_\theta(x)) = R(f_\theta)$

$\min_\theta \frac{1}{n} \sum \ell(y_i, f_\theta(x_i)) = \hat{R}(f_\theta)$

with $f_\theta(x) = \sum_{j=1}^{m} v_j \sigma(w_j^T x + b_j)$.

$\theta = (v_j, w_j, b_j)$

$m(d+2)$

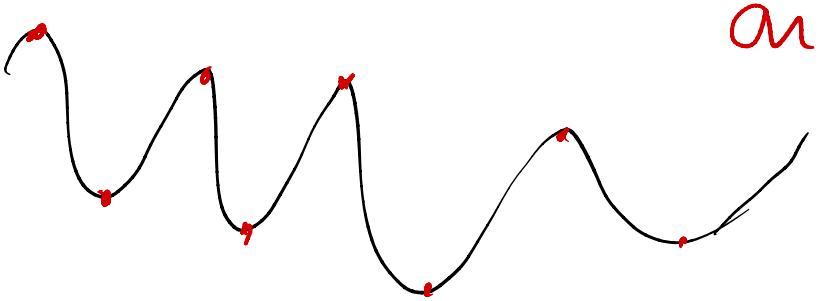
$\theta \in \mathbb{R}$

3 algorithms: GD on $\hat{R}(f_\theta)$ with multiple passes $\frac{1}{\sqrt{t}}$ on train error

SGD on $\hat{R}(f_\theta)$ with multiple passes $\frac{1}{\sqrt{t}}$ on test error

SGD on $R(f_\theta)$ with a single pass $\frac{1}{\sqrt{n}}$ error

Non convexity $\Rightarrow$ only guarantees on gradient norm



Hope: over parameterization m large

Estimator error : ReLU Rectified linear Unit

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

positively-homogeneous : $\sigma(\lambda u) = \lambda \sigma(u)$ if $\lambda \geq 0$

$$\eta_j (w_j^\top x + b_j)_+ = \alpha_j \eta_j \left(\frac{w_j^\top}{\alpha_j} x + \frac{b_j}{\alpha_j} \right)_+ \qquad \forall \alpha_j > 0$$

η_j ↑ $\alpha_j \eta_j$; $\frac{w_j^\top}{\alpha_j} x + \frac{b_j}{\alpha_j}$

consider ℓ_2-penalties a weights → "weight decay" explicit / implicit

$$\sum_{j=1}^m \eta_j^2 + \|w_j\|^2 + \frac{b_j^2}{R^2}$$

Assumption: inputs $\|x\|_2 \leq R$ almost surely

purely after rescaling $= \sum_j \alpha_j^2 \eta_j^2 + \frac{1}{\alpha_j^2} \left(\|w_j\|_2^2 + \frac{b_j^2}{R^2} \right)$ → min over α → $\sum_j |\eta_j| \sqrt{\|w_j\|_2^2 + \frac{b_j^2}{R^2}}$

lemma $\min_{\beta} a\beta + \frac{c}{\beta} = 2\sqrt{ac}$

ADD CONSTRAINT : $\|w_j\|^2 + \frac{b_j^2}{R^2} = 1, \forall j$ ⇒ $\|\eta\|_1$

penalty on weights $\sum_{j=1}^{m} |a_j| \sqrt{\|w_j\|_2^2 + \frac{b_j^2}{B^2}}$

choice of normalization: $\sqrt{\|w_j\|_2^2 + \frac{b_j^2}{B^2}} = 1$ } input weights

$\rightarrow$ penalty $= \sum_{j=1}^{m} |a_j| = \|a\|_1$ } output weights

Goal $= F = \left\{ f(x) = \sum_{j=1}^{m} a_j (w_j^T x + b_j)_+ , \ \|w_j\|_2^2 + \frac{b_j^2}{B^2} = 1 \ \forall j, \ \|a\|_1 \le D \right\}$

Recall: linear models

$F = \{ f(x) = \theta^T \varphi(x), \ \|\theta\|_2 \le D \}$

Tool: Rademacher complexity

$$\mathcal{F} = \left\{ f_\theta(x) = \sum_{j=1}^{m} v_j (w_j^T x + b_j)_+ , \quad \|w_j\|_2^2 + \frac{b_j^2}{B^2} = 1 \; \forall j \right\}$$

$$\|v\|_1 \le D$$

Rademacher average

$$R_n = \mathbb{E}_{data}\left[\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \ell(y_i, f_\theta(x_i)) \right]$$

ε : rademacher r.v $P(\varepsilon = \pm 1) = \frac{1}{2}$, iid

$$\mathbb{E}\left[R(f_{\hat{\theta}}) - \inf_{\theta \in \Theta} R(f_\theta) \right] \le 4 R_n$$

(From Chap. 4)

Estimation error

$f_{\hat{\theta}}$: empirical risk minimizer

$$R_n = \mathbb{E}_{data}\left[\sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \ell(y_i, f_\theta(x_i))\right]$$

ε : rademacher r.v $\quad \mathbb{P}(\varepsilon = \pm 1) = \frac{1}{2}$, iid

Extra-assumption : loss is G-Lipschitz cont. wrt $|\cdot|$: $|\ell(y,u) - \ell(y,v)| \le G|u-v|$

$$R_n \le G\, \mathbb{E}_{\substack{data\\ \varepsilon}}\left[\sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f_\theta(x_i)\right]$$

$f_\theta(x_i) = \sum_{j=1}^{m} \eta_j (w_j^T x_i + b_j)_+$

$\|\eta\|_1 \le D$

$\|w_j\|_2^2 + b_j^2 = 1$, $w_j \in \mathbb{R}^d$

↓ contraction principle

Lemme : $\sup_{\|\eta\|_1 \le D} \eta^T z = D\,\|z\|_\infty$

$$= GD\,\mathbb{E}\left[\sup_{\|w_j\|_2^2 + \frac{b_j^2}{R^2} = 1} \sup_{j\in[1,m]} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i (w_j^T x_i + b_j)_+\right|\right]$$

$\forall j$

uniformly in $j \in [1,m]$

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

contraction principle

$$\le 2GD\,\mathbb{E}_{\varepsilon}\left[\sup \left|\frac{1}{n}\sum_{i=1} \varepsilon_i (w^T x_i + b)\right|\right]$$

$$R_n \le 2G\mathcal{D}\, \mathbb{E}_{\varepsilon} \mathbb{E}_{data} \left[\sup_{\|w\|_2^2 + \frac{b^2}{R^2} = 1} \frac{1}{n} \sum_i \varepsilon_i (w^T x_i + b) \right]$$

$$= \frac{1}{n} \sum_i \varepsilon_i \begin{pmatrix} x_i \\ R \end{pmatrix}^T \begin{pmatrix} w \\ b/R \end{pmatrix}$$

Lemma: $\sup_{\|v\|_2 = 1} v^T z = \|z\|_2$

$$\left\| \begin{pmatrix} w \\ b/R \end{pmatrix} \right\|_2 = 1$$

$$= 2G\mathcal{D}\, \mathbb{E}_{data} \mathbb{E}_{\varepsilon} \left\| \frac{1}{n} \sum_i \varepsilon_i \begin{pmatrix} x_i \\ R \end{pmatrix} \right\|_2$$

$$\le 2G\mathcal{D}\, \mathbb{E} \left[\left\| \frac{1}{n} \sum_i \varepsilon_i x_i \right\|_2 + \left| \frac{1}{n} \sum \varepsilon_i R \right| \right]$$

Jensen's

$$\le 2G\mathcal{D} \left(\sqrt{\mathbb{E} \left\| \frac{1}{n} \sum_i \varepsilon_i x_i \right\|_2^2} + R \sqrt{\mathbb{E} \left| \frac{1}{n} \sum_i \varepsilon_i \right|^2} \right)$$

$$\mathbb{E} \left\| \frac{1}{n} \sum_i \varepsilon_i x_i \right\|_2^2 = \frac{1}{n^2} \sum_i \mathbb{E} \|x_i\|_2^2 \le \frac{R^2}{n}$$

$$\mathbb{E} \left| \frac{1}{n} \sum_i \varepsilon_i \right|^2 = \frac{1}{n}$$

$$\le 2G\mathcal{D} \left[\frac{R}{\sqrt{n}} + \frac{R}{\sqrt{n}} \right) = \frac{4G\mathcal{D}R}{\sqrt{n}}$$

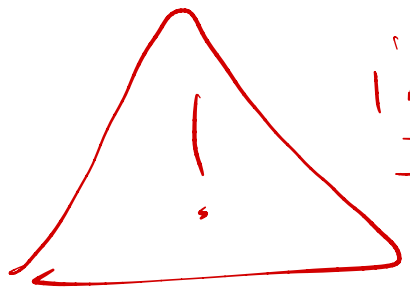
Summary $\mathcal{F} = \left\{ f(x) = \sum_{j=1}^{m} v_j (w_j^T x + b_j)_+ , \ \|w_j\|_2^2 + \frac{b_j^2}{R^2} \leq 1 \ \forall j \right\}$

$\|v\|_1 \leq B$

Estimation error for the ERM on $\mathcal{F}$

$\leq 4 \cdot \frac{4 GRB}{\sqrt{n}} = 16 \frac{GRB}{\sqrt{n}}$

comes from link between estimation error & Rademacher



indep of m

Approximation error :

(1) universally

(2) bounds on approximation error

$m = +\infty$

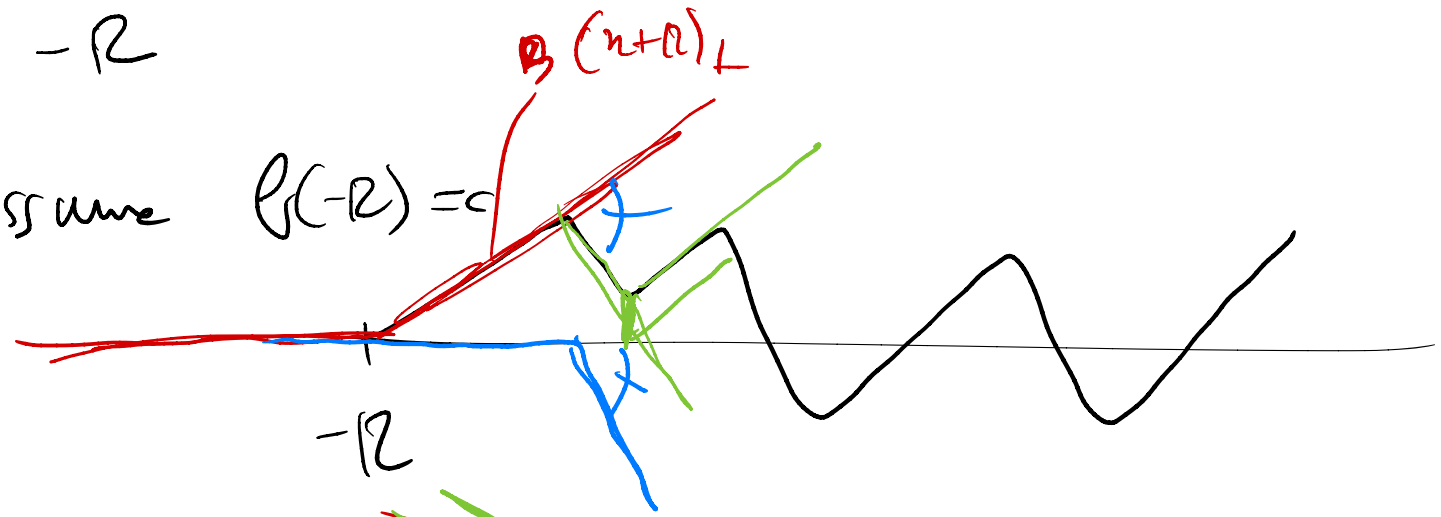
(3) what can be achieved with m finite ?

Universality in $\mathbb{R}$ | Given $f_\theta : \mathbb{R}^d \to \mathbb{R}$ with many layers and many neurons

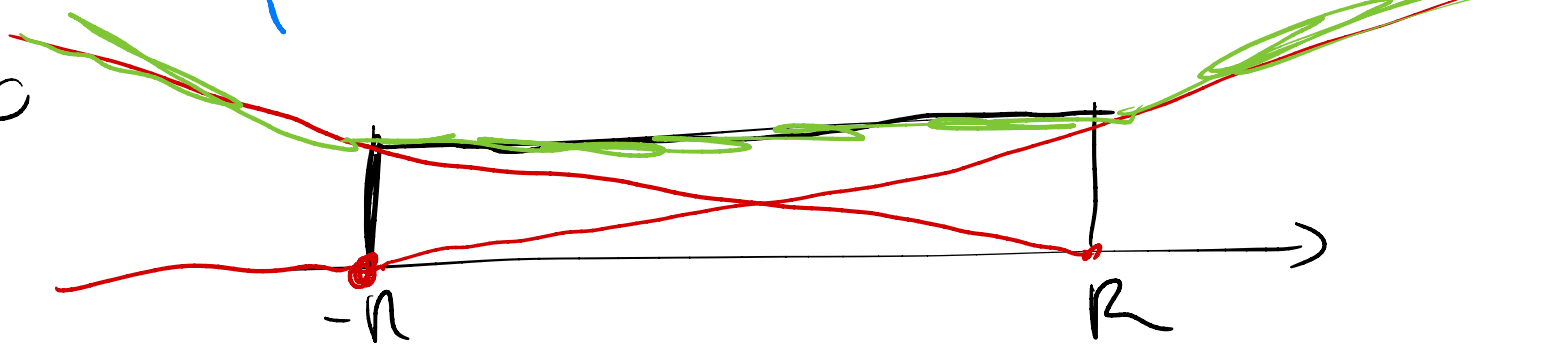
Can we approach f_θ

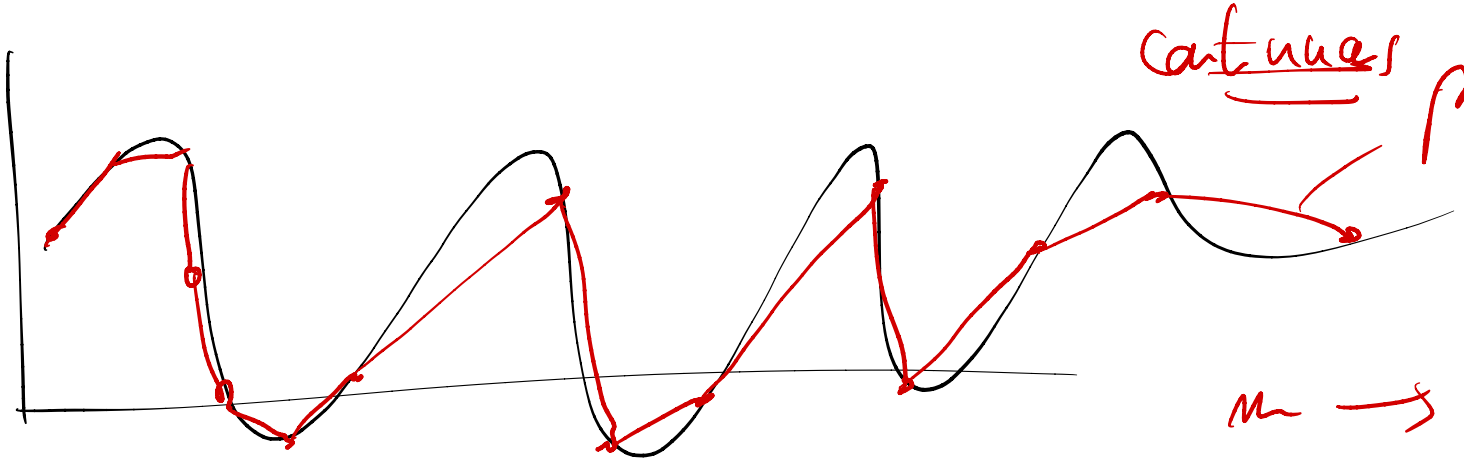
① True for piecewise affine functions

② assume $f(-R) = 0$



③ If $f(-R) \neq 0$

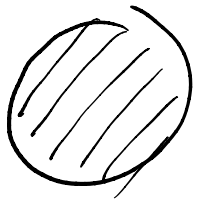




Definition of a function space

$$f(x) = \sum_{j=1}^{m} v_j (w_j^T x + b_j)_+$$

$(w_j, b_j) \in K$ compact set $\subset \mathbb{R}^{d+1}$



$$= \int_K (w^T x + b)_+ \, dv(w,b)$$

$$v(w,b) = \sum_{j=1}^{m} v_j \, \delta_{(w_j, b_j)}$$

↳ Dirac measure at (w_j, b_j)

$$\|v\|_2 = \sum_{j=1}^{m} |v_j| = \text{total variation of } v$$

Consider $f(x) = \int_K (w^T x + b)_+ \, d\nu(w,b)$

penalty $\int |d\nu(w,b)|$ total variation

ex: if ν has density $\mu(w,b)$ with respect to Lebesgue

$\int_K |\mu(w,b)| \, dw \, db$

$\gamma_1(f) = \inf \int |d\nu(w,b)|$ s.t. $f(x) = \int_K (w^T x + b)_+ d\nu(w,b)$

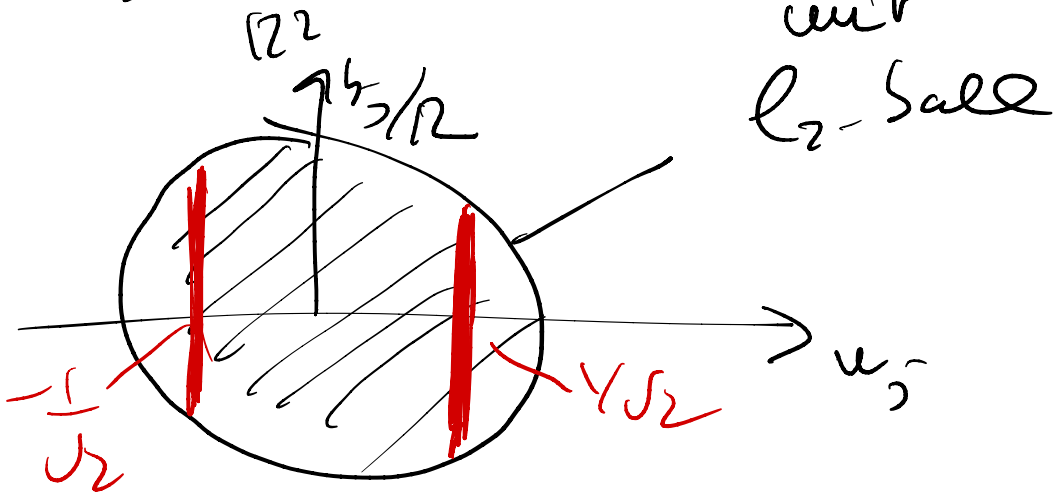
ex: if $f(x) = (w_0^T x + b_0)_+$ with $(w_0, b_0) \in K$

$\gamma_1(f) \leq 1$

$$K = \left\{ (w,s) \in \mathbb{R}^{d+1},\ \|w\|_2 = \frac{1}{\sqrt{2}},\ |s| \le \frac{R}{\sqrt{2}} \right\}$$

Earlier : $\overline{K} = \left\{ \|w\|_2^2 + \frac{s^2}{R^2} \le 1 \right.$

$K \subset \overline{K}$



$$\left\{ \|w\|_2 = \frac{1}{\sqrt{2}},\ \frac{|s|}{R} \le \frac{1}{\sqrt{2}} \right.$$

$$K \subset \frac{K}{\sqrt{2}}$$

$$K = \left\{ (w,s) \in \mathbb{R}^{d+1},\ \|w\|_2 = 1,\ |s| \le R \right\}$$

How to write $f(x) = \int_K (w^T x + s)_+ \, d\nu(w,s)$?

$\int |d\nu(w,s)|$

1D | Section 4.3.3

Taylor's formula with integral remainder

order 1 —

$$f(x) = f(y) + \int_y^x f'(t)\,dt$$

$$f(x) = f(y) + (x-y) f'(y) + \frac{1}{2}(x-y)^2 f''(c)$$

$c \in [x, y]$ — classical remainder

$$+ \int_y^x f''(t)(x-t)\,dt$$

$$+ \int_y^{\infty} f''(t)(x-t)_+\,dt$$

$(x > y)$

$$+ \int_y^{R} f''(t)(x-t)_+\,dt$$

$y = -R$: $f(x) = f(-R) + (x+R) f'(-R) + \int_{-R}^{R} f''(t)(x-t)_+\,dt$

$$\gamma_2(\mathcal{G}) \leq \int_{-R}^{R} |f''(x)| + \begin{pmatrix} f(-R), f(R) \\ f'(-R), f'(R) \end{pmatrix}$$

↑ Section 9.3?

Beyond 1D: $f(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{f}(\omega) e^{i\omega^T x} d\omega$ (inverse Fourier transform)

If $\|x\|_2 \leq R$

e^{iu} function $[-R, R] \to \mathbb{C}$.

$$e^{iu} = \int_{-R}^{R} \mathbb{1}_{(t)}(u-t)_+ dt \Rightarrow f(x) = \frac{1}{(2\pi)^d} \int \hat{f}(\omega) \mathbb{1}_{(t)}(\omega^T x - t)_+ d\omega$$

ADAPTIVITY TO LINEAR LATENT VAR

if $f^*(x) = g^*(w_*^T x)$, then $\gamma_n(f^*)$ "small".

| lipschitz-continuous $\Rightarrow$ CV rate $\frac{1}{n^{1/3}}$

lipschitz-continuous $\Longrightarrow$ CV rate $\frac{1}{n^{1/d}}$

Example: $f^*(x) = s^*(W^T x)$

| $W \in \mathbb{R}^{d \times h}$

h - parameters in W

$$f(x) = \sum_{j=1}^{m} v_j (w_j^T x + b_j)_+ = v^T \varphi(x)$$