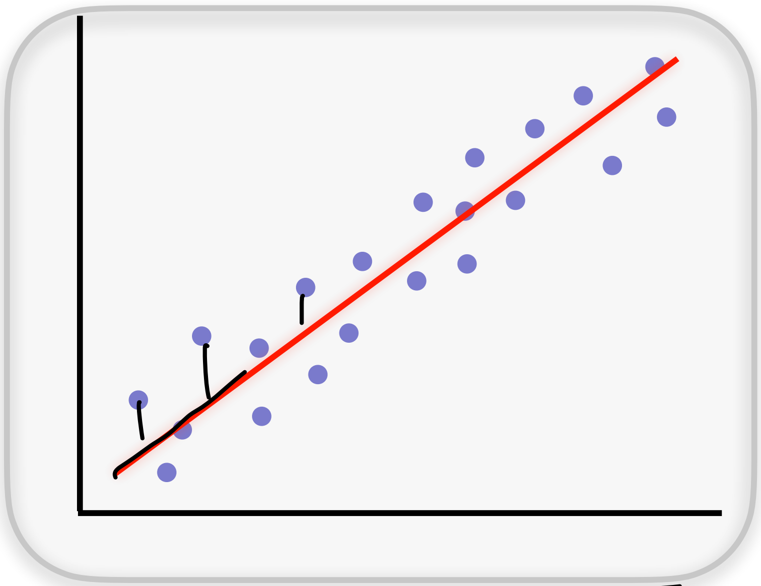
Approximation Theory

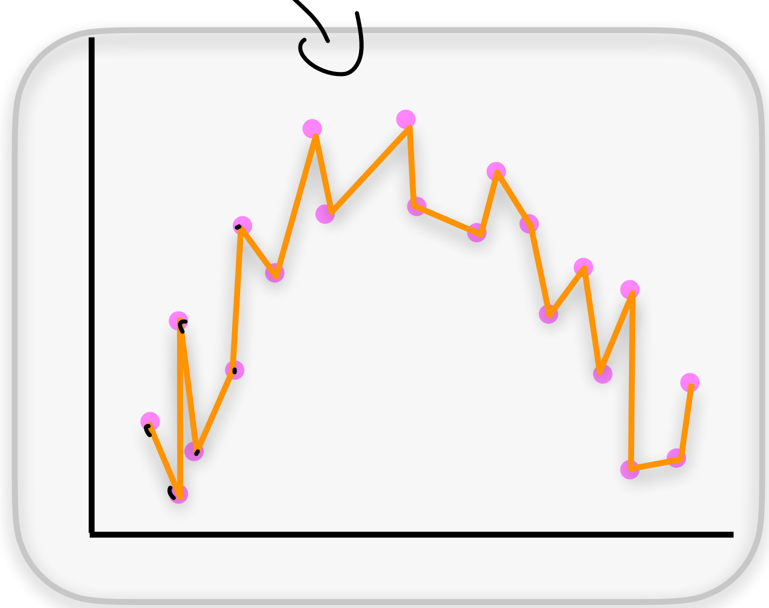
UNIVERSITY *of* WASHINGTON

Expressivity / Representation Power

linear

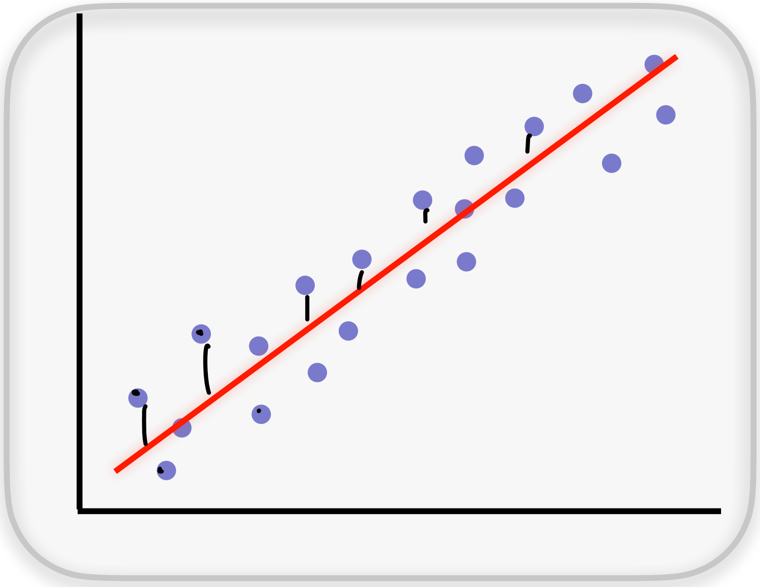
strong

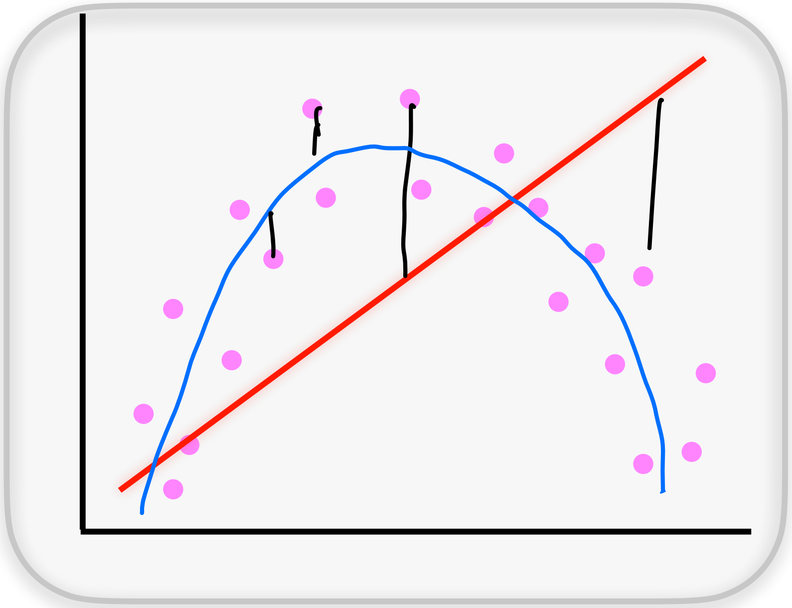




Expressive: Functions in class can represent "complicated" functions.

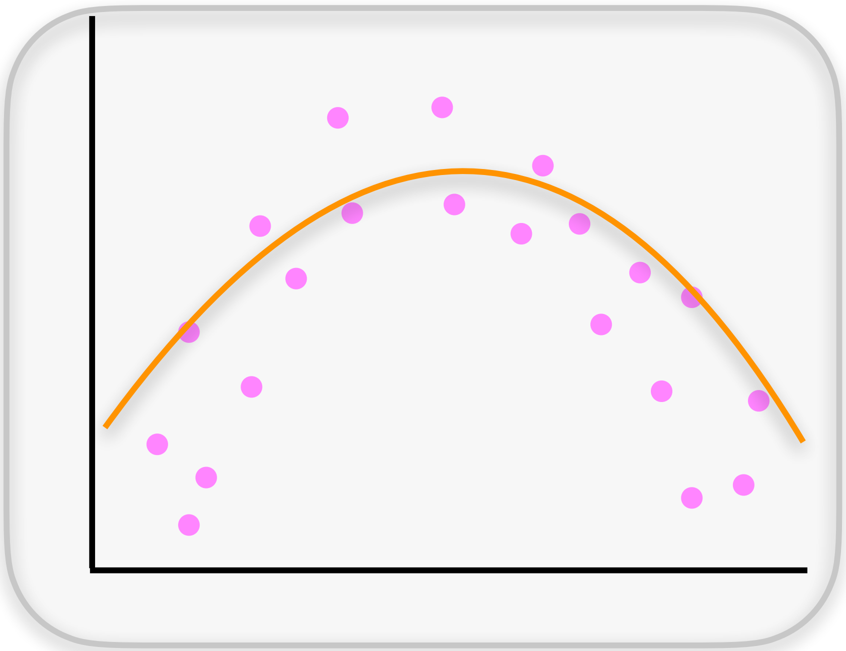
Linear Function





best linear fit

Review: generalized linear regression



Transformed data: $h_0(x) = 1$

$h_1(x) = x$

$h_2(x) = x^2$

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix} \quad h_3(x) = x^3$$

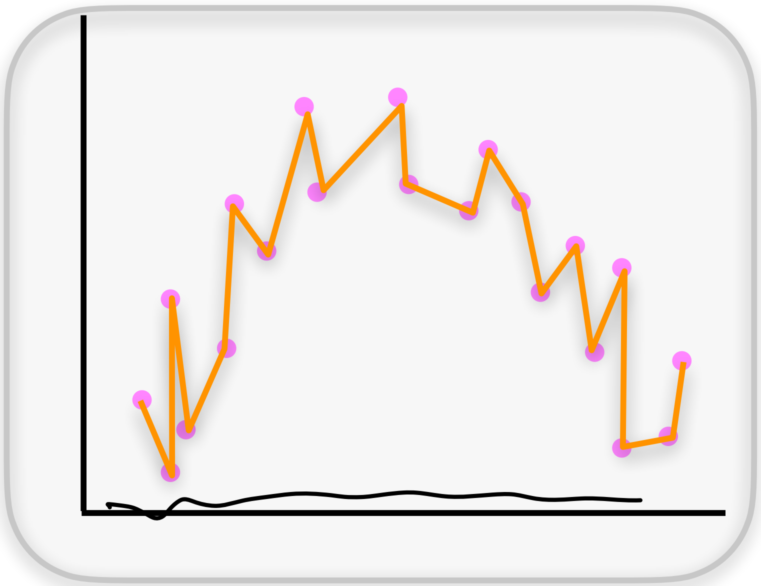
$\vdots$

Hypothesis: linear in h

$$y_i \approx h(x_i)^T w$$

more powerful $X \cdot w$

Review: Polynomial Regression



$$h(x) = \begin{pmatrix} 1 \\ x \\ \vdots \\ x^p \end{pmatrix}$$

$$f(x) = \langle w, h(x) \rangle$$

$$w \in \mathbb{R}^{p+1}$$

Lagrange's Interpolation Theorem

Given a data set $\{(x_i, y_i)\}_{i=1}^{n}$

$\exists$ polynomial P of degree $\underline{n-1}$

s.t. $y_i = P(x_i)$

(condition: no (x_i, y_i), $(x_{i'}, y_{i'})$ s.t.
$x_i = x_{i'}$
$y_i \neq y_{i'}$)

Approximation Theory Setup

No training

- Goal: to show there exists a neural network that has small error on training / test set.

Sanity

- Set up a natural baseline:

$$\inf_{\underbrace{f \in \mathcal{F}}_{NN}} L(f) \text{ v.s. } \inf_{g \in \text{ continuous functions}} L(g)$$

$$\mathcal{F} \subseteq \{\text{continuous functions}\}$$

Example

$\rho : \mathbb{R}_+$

(i) loss $\ell(f(x), y) = \ell(y \cdot f(x))$, ρ-Lipshitz

$$|\ell(z) - \ell(z')| \leq \rho |z - z'|, \ z, z' \in \mathbb{R}$$

e.g. hinge loss

$$\ell(yf(x)) = \max\{0, 1 - yf(x)\}$$

1 - Lipshitz

- $$L(f) = \int \ell(yf(x)) \, d\mu(x, y)$$

$\mu(x, y)$ distribution on (x, y)

Decomposition

$f \in \mathcal{F}$
$g \in \{\text{continuous functions}\}$

$$L(f) - L(g)$$
$$= \int \left(\ell(yf(x)) - \ell(y\, g(x)) \right) d\mu(x,y)$$
$$\leq \int \left| \ell(yf(x)) - \ell(y\, g(x)) \right| d\mu(x,y)$$
$$\leq \int \rho \left| yf(x) - y\, g(x) \right| d\mu(x,y)$$
$$= \int \rho \cdot |y| \, |f(x) - g(x)| \, d\mu(x,y)$$
$$\leq \rho \int |f(x) - g(x)| \, d\mu(x)$$

Assume $|y| \leq 1$

Specific Setups

$f \in \mathcal{F}$
$g \in$ continuous functions

- "Average" approximation: given a distribution μ

$$\|f-g\|_\mu = \int_x |f(x)-g(x)|\,d\mu(x)$$

- "Everywhere" approximation

$$\|f-g\|_\infty = \sup_x |f(x)-g(x)| \geq \|f-g\|_\mu$$

$$\|f-g\|_\mu = \int_X |f(x)-g(x)|\,d\mu(x)$$
$$\leq \int_X \sup_{\tilde{x}} |f(\tilde{x})-g(\tilde{x})|\,d\mu(x)$$
$$= \|f-g\|_\infty \int_X d\mu(x) = \|f-g\|_\infty$$

Polynomial Approximation

continuous

$\left(\frac{1}{\epsilon}\right)^d$

Theorem (Stone-Weierstrass): for any function f, we can approximate it on any compact set Ω by a sufficiently high degree polynomial: for any $\epsilon > 0$, there exists a polynomial p of sufficient high degree, s.t.,

$$\max_{x \in \Omega} |f(x) - p(x)| \leq \epsilon.$$

Intuition: Taylor expansion!

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2 + \cdots$$

$$f(x) \approx \langle w, \phi(x) \rangle$$

$$\phi(x) = (1, x - x_0, (x - x_0)^2, \ldots)$$

$$w = \left(f(x_0), f'(x_0), \frac{f''(x_0)}{2}, \ldots\right)$$

Kernel Method

fixed (can be infinite dimension)

$x \mapsto \phi(x)$, $f(x) = \langle w, \phi(x) \rangle$ for some w

one only needs to evaluate

$$K(x, x') = \langle \phi(x), \phi(x') \rangle$$

Polynomial kernel

d-dim input

$$\phi(x) = (1, x_1, x_2, \ldots, x_d, x_1 x_2, \ldots, x_d^2, \ldots x_d^p)$$

Gaussian Kernel

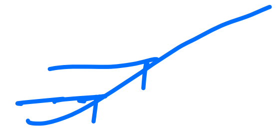
1-dim

$$K(x, x') = \exp\left(-\frac{|x-x'|^2}{2\sigma^2}\right)$$

$$\phi(x) = e^{-\frac{x^2}{2\sigma^2}} \left(1, \sqrt{\frac{1}{1!}} \frac{x}{\sigma}, \sqrt{\frac{1}{2!}} \left(\frac{x}{\sigma}\right)^2, \ldots\right)$$

$\longrightarrow$ Gaussian Kernel has inf strong representation power

1D Approximation

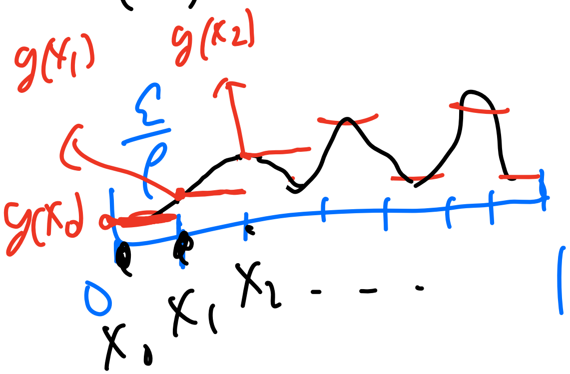


Theorem: Let $g : [0,1] \to R$, and ρ-Lipschitz. For any $\epsilon > 0$, $\exists$ 2-layer neural network f with $\lceil \frac{\rho}{\epsilon} \rceil$ nodes, threshold activation: $\sigma(z) : z \mapsto \mathbf{1}\{z \geq 0\}$ such that

$$\sup_{x \in [0,1]} |f(x) - g(x)| \leq \epsilon.$$

Proof of 1D Approximation

Pf



Let $m \triangleq \lceil \frac{\rho}{\varepsilon} \rceil$, $x_i \triangleq \frac{(i-1)\varepsilon}{\rho}$

$$f(x) = \sum_{i=1}^{m} a_i \cdot \mathbb{1}\{x - x_i \geq 0\}$$

$a_1 = g(x_0)$, $a_i = g(x_i) - g(x_{i-1})$, $i = 1, \ldots$

- if $x < x_1$, $\mathbb{1}\{x - x_i \geq 0\} = 0$, $i = 1, \ldots, m$

$$f(x) = g(x_0)$$

- if $x_1 \leq x < x_2$, $\mathbb{1}\{x - x_i \geq 0\} = 0$, $i = 2, \ldots, m$

$$f(x) = g(x_0) + g(x_1) - g(x_0) = g(x_1)$$

$x_i \leq x \leq x_{i+1}$

$\Rightarrow f(x) = g(x_{i+1})$

$\forall x$,

say $x_i \leq x$, closest

$i = 0, \ldots, m$

$$|g(x) - f(x)| = |g(x) - f(x_i)|$$

$$\leq |g(x) - g(x_i)| + |g(x_i) - f(x_i)|$$

0

$$\leq \rho \cdot |x - x_i|$$

$$\leq \rho \cdot \frac{\varepsilon}{\rho} = \varepsilon$$

□

Multivariate Approximation

$(\frac{\epsilon}{\rho})$-Lip

Theorem: Let g be a continuous function that satisfies $\|x - x'\|_\infty \leq \delta \Rightarrow |g(x) - g(x')| \leq \epsilon$ (Lipschitzness). Then there exists a 3-layer ReLU neural network with $O(\frac{1}{\delta^d})$ nodes that satisfy

uniform average approximation

- curse of dimensionality
- tight $\Omega(\frac{1}{\delta^d})$

$$\int_{[0,1]^d} |f(x) - g(x)| dx = \|f - g\|_1 \leq \epsilon$$

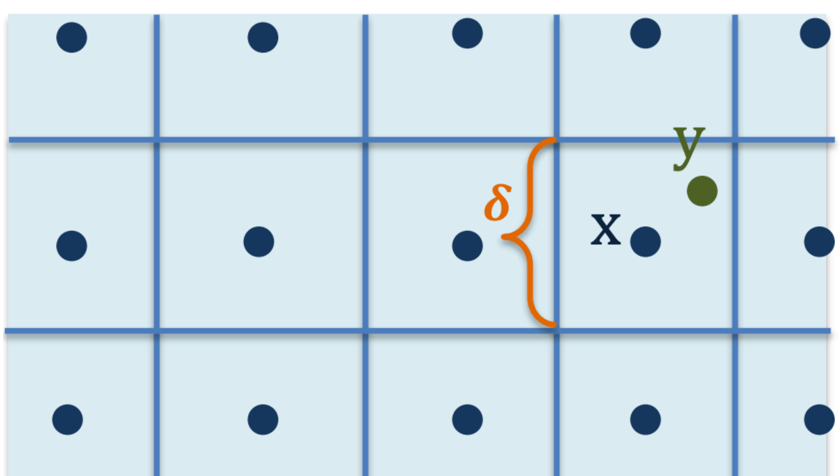
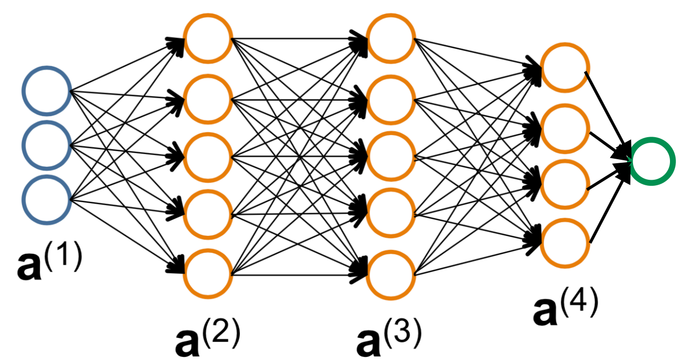


Figure credit to Andrej Risteski



Partition Lemma

Lemma: let g, δ, ϵ be given. For any partition P of $[0,1]^d$, $P = (R_1, \ldots, R_N)$ with all side length smaller than δ, there exists $(\alpha_1, \ldots, \alpha_N) \in \mathbb{R}^N$ such that

everywhere approximating

$$\sup_{x \in [0,1]^d} |g(x) - h(x)| \leq \epsilon \text{ with } h(x) := \sum_{i=1}^{N} \alpha_i \mathbf{1}_{R_i}(x).$$

$\mathbb{1}_{R_i}(x) = \mathbb{1}\{x \in R_i\}$

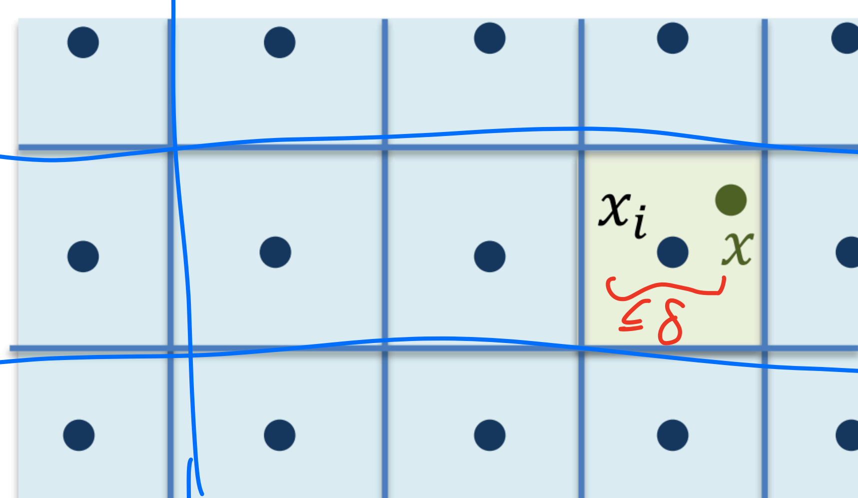
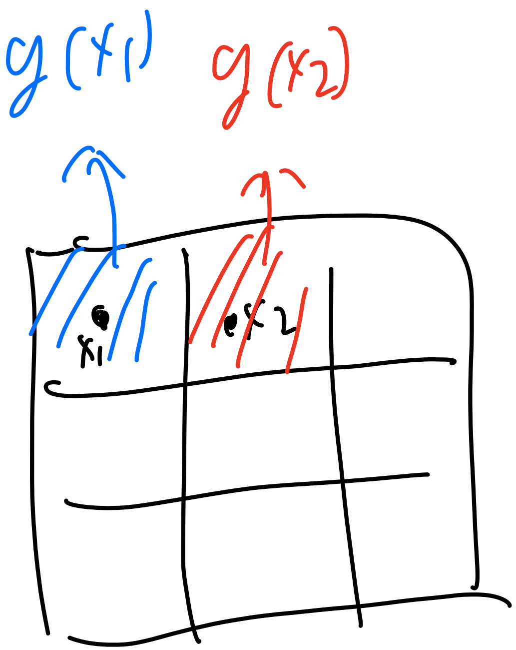


Figure credit to Andrej Risteski

Proof of Partition Lemma

Pf: For each Ω_i, pick $\forall x_i \in \Omega_i$, set $\alpha_i \doteq g(x_i)$



$$\sup_{x\in[0,1]^d} |g(x)-h(x)|$$

$$= \sup_{i\in\{1,\dots,N\}} \sup_{x\in\Omega_i} |g(x)-h(x)|$$

$$\le \sup_{i\in\{1,\dots,N\}} \sup_{x\in\Omega_i} \left(|g(x)-g(x_i)| + |g(x_i)-h(x_i)| \right)$$

0

$$\le \sup_{i\in\{1,\dots,N\}} \sup_{x\in\Omega_i} (\varepsilon + 0)$$

$$= \varepsilon$$

$\square$

Proof of Multivariate Approximation Theorem

Idea: $h(x) = \sum_i \alpha_i \mathbb{1}_{R_i}(x)$ in the lemma

1) use a 2-layer NN to approximate

$$x \mapsto \mathbb{1}_{R_i}(x)$$

2) find a linear combination to represent h

$$\Rightarrow \quad \|f - g\|_1 \le \|f - h\|_1 + \underbrace{\|h - g\|_1}_{\le \varepsilon}$$

Let $f(x) = \sum_{i=1}^{N} \alpha_i f_i(x)$ $\quad \alpha_i = g(x_i)$, f_i to approximate $\mathbb{1}_{R_i}(x)$

$$\|f - h\|_1 = \|\sum_i \alpha_i (\mathbb{1}_{R_i} - f_i)\|_1$$

$$\le \sum_{i=1}^{N} |\alpha_i| \, \|\mathbb{1}_{R_i} - f_i\|_1$$

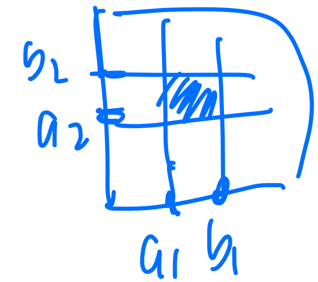
want to show $\quad \|\mathbb{1}_{R_i} - f_i\|_1 \le \dfrac{\varepsilon}{\sum_{i=1}^{N} |\alpha_i|}$

what if $\sum_{i=1}^{N} |\alpha_i| = 0 \to g(x_i) = 0$, $|g(x)| \le \varepsilon$, use 0-function

Proof of Multivariate Approximation Theorem

(2) construct f_i

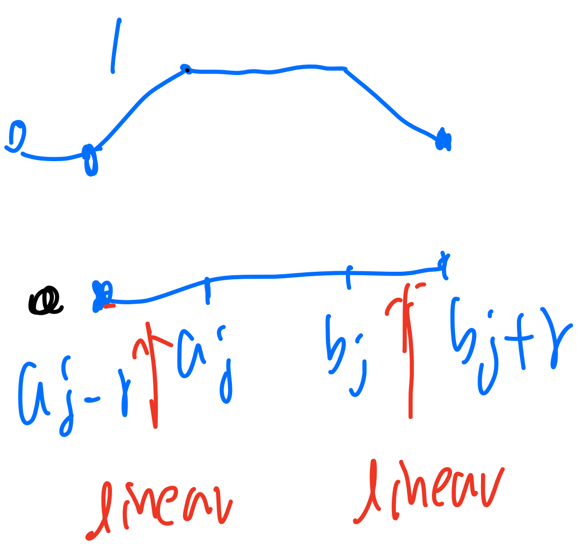
Recall: $\Omega_i \triangleq [a_1,b_1]\times[a_2,b_2]\cdots,[a_d,b_d]$



☆ bump function

given $r>0$, define (σ: ReLU)

$$g_{r,j}(z)=\sigma\left(\frac{z-(a_j-r)}{r}\right)-\sigma\left(\frac{z-a_j}{r}\right)-\sigma\left(\frac{z-b_j}{r}\right)+\sigma\left(\frac{z-(b_j+r)}{r}\right)$$



$z\in[a_j,b_j]$, $g_{r,j}(z)=1$

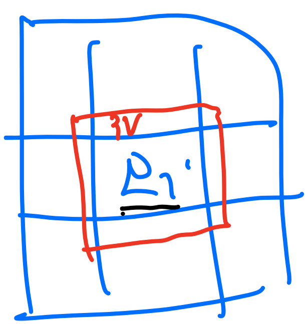
$z\notin[a_j-r,b_j+r]$, $g_{r,j}(z)=0$

☆ $r\to 0$, $g_{r,j}\to \mathbb{1}_{[a_j,b_j]}$

Proof of Multivariate Approximation Theorem

Define $g_r(x) = 6\left(\sum_{j=1}^{d} g_{r,j}(x^j) - (d-1)\right)$

$x = \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^d \end{pmatrix}$ $\quad g_r(x) = \begin{cases} 1 & \text{if } x \in \Omega_i \\ 0 & \text{if } x \notin [a_1 - r, b_1 + r] \times \cdots \times [a_d - r, a_d + r] \end{cases}$



Since $r \to 0$, $g_{r,j} \to \mathbb{1}_{[a_j, b_j]}$

$\to g_r \to \mathbb{1}_{\Omega_i}$

$\exists\, r$ with $\| g_r - \mathbb{1}_{\Omega_i} \|_1 \le \frac{\varepsilon}{\sum_{i=1}^{N} |\alpha_i|}$

Let $f_i = g_r$

$f = \sum_{i=1}^{N} \alpha_i f_i$

$\square$

Universal Approximation

Definition: A class of functions $\mathscr{F}$ is universal approximator over a compact set S (e.g., $[0,1]^d$), if for every continuous function g and a target accuracy $\epsilon > 0$, there exists $f \in \mathscr{F}$ such that

$$\sup_{x \in S} |f(x) - g(x)| \leq \epsilon$$

Stone-Weierstrass Theorem

Theorem: If $\mathscr{F}$ satisfies

1. Each $f \in \mathscr{F}$ is continuous.

2. $\forall x, \exists f \in \mathscr{F}, f(x) \neq 0$

3. $\forall x \neq x', \exists f \in \mathscr{F}, f(x) \neq f(x')$

4. $\mathscr{F}$ is closed under multiplication and vector space operations,

Then $\mathscr{F}$ is a universal approximator:

$$\forall g : S \to R, \epsilon > 0, \exists f \in \mathscr{F}, \|f - g\|_\infty \leq \epsilon.$$

Example: cos activation

Example: cos activation

Other Examples

Exponential activation

ReLU activation

Curse of Dimensionality

- Unavoidable in the worse case

- Barron's theory

Recent Advances in Representation Power

- Depth separation
- Analyses of different architectures
  - Graph neural network
  - Attention-based neural network
- Finite data approximation
- …