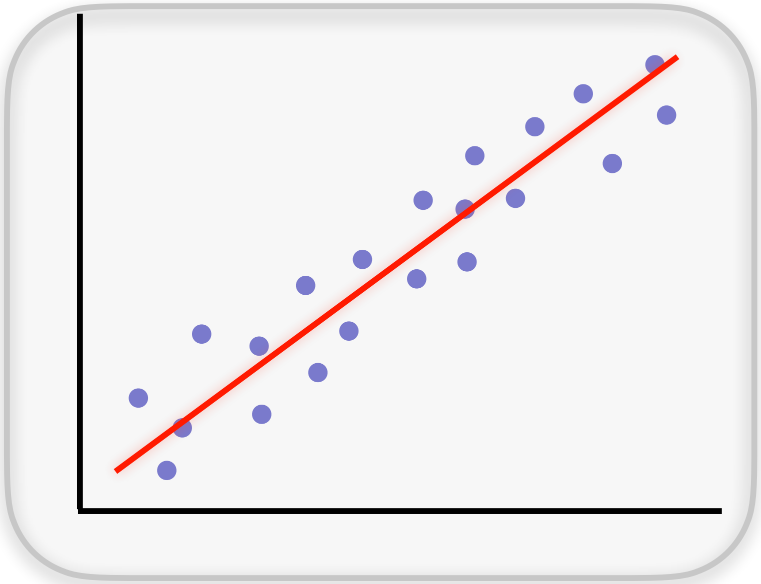
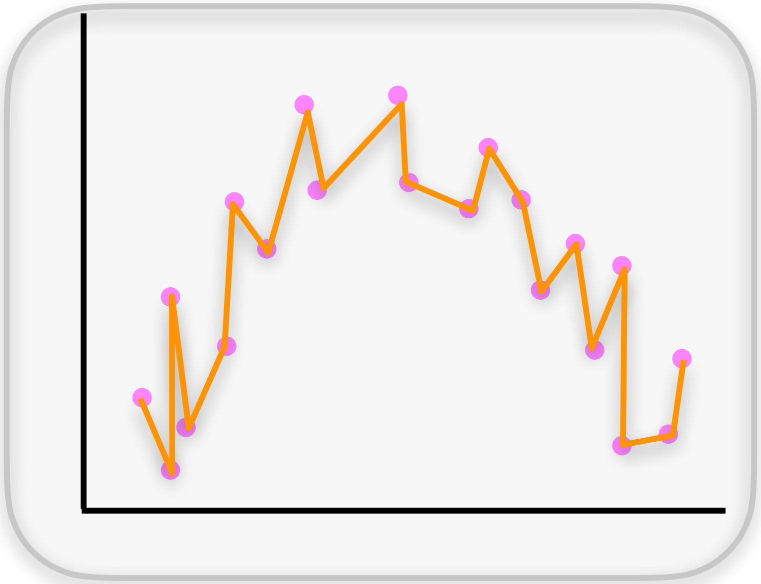
Approximation Theory

Expressivity / Representation Power

$\mathcal{F}$: linear, NN, . . .

f_{target}

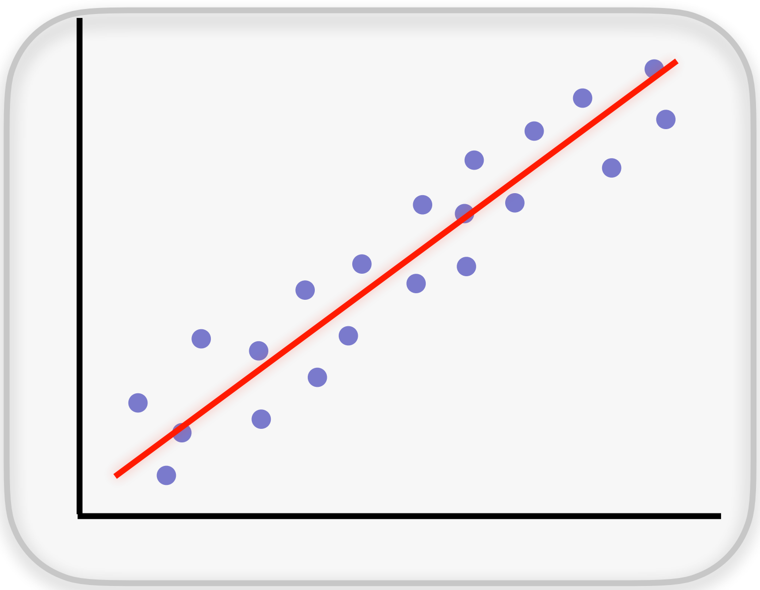


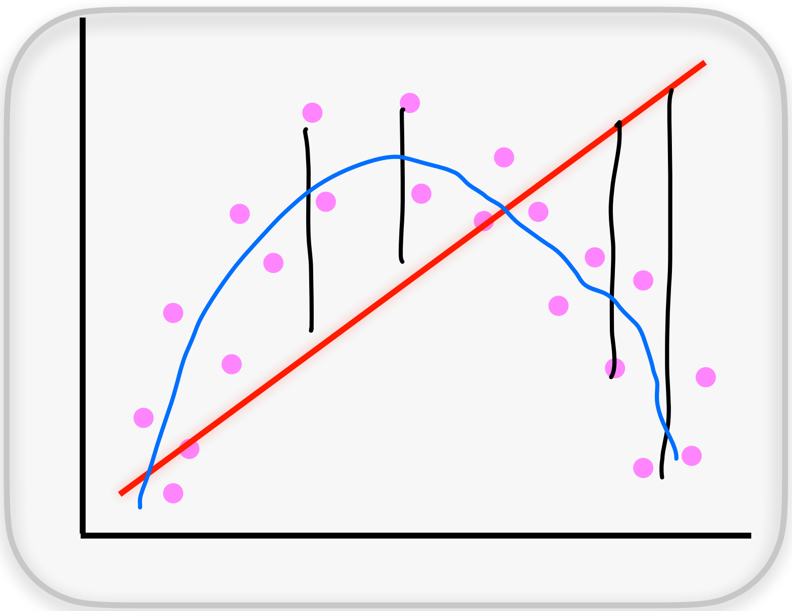


Expressive: Functions in class can represent "complicated" functions.

Linear Function

large bias

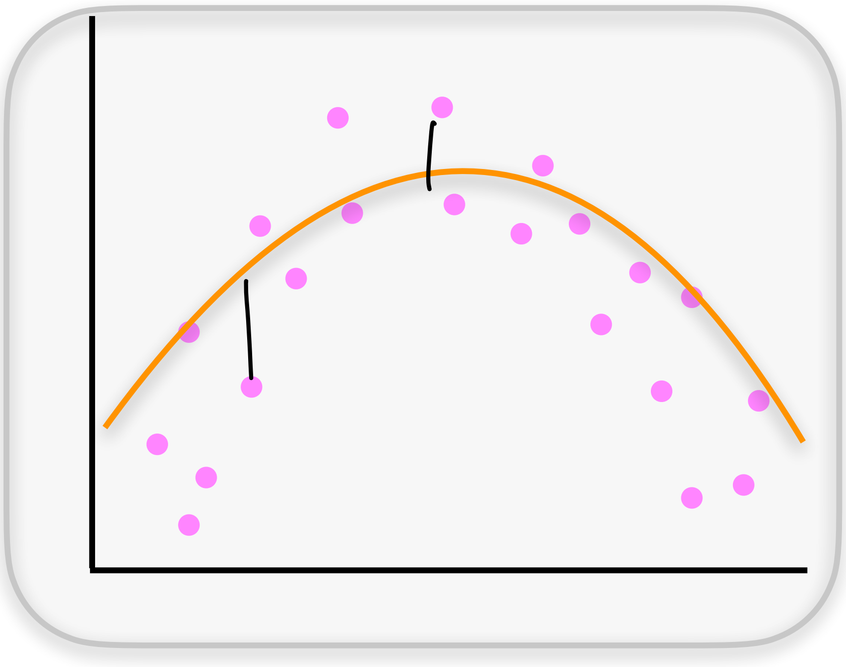




best linear fit

Review: generalized linear regression

smaller bias



Transformed data:

$h_1(x) = 1$

$h_2(x) = x$

$h_3(x) = x^2$

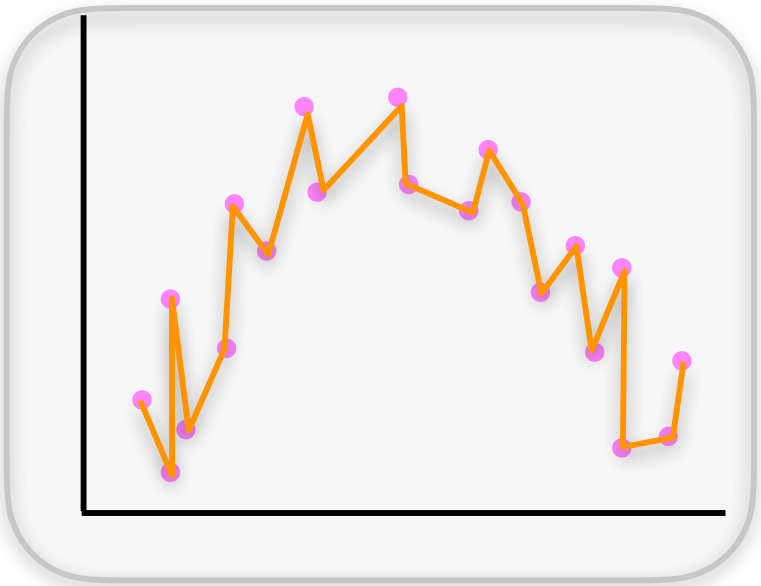
$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$

Hypothesis: linear in h

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

Review: Polynomial Regression

0 bias



$$h(x) = \begin{pmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^p \end{pmatrix}$$

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

n data points

$p \geq n-1$

Lagrange's Interpolation Theorem

Given a data $\{(x_i, y_i)\}_{i=1}^{n}$, $\exists$ polynomial P of degree $n-1$ s.t. $y_i = f(x_i)\ \forall i$

Approximation Theory Setup

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

Sanity check

- Set up a natural baseline:

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

L : loss

$\mathcal{F}$: NN function class

Example

(1) loss $\quad \ell(f(x), y) = \ell(y \cdot f(x))$, ρ-Lipschitz

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

e.g. hinge loss

$$\ell(y f(x)) = \max\{0, 1 - y \cdot f(x)\}$$

1 — Lip

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

$\mu(x, y)$: distribution over (x, y)

Decomposition

$$L(f) - L(g)$$

$$= \int \left(\ell(yf(x)) - \ell(yg(x)) \right) d\mu(x,y)$$

$$\leq \int \left| \ell(yf(x)) - \ell(yg(x)) \right| d\mu(x,y)$$

$$\leq \int \rho \cdot \left| yf(x) - yg(x) \right| d\mu(x,y)$$

(assume $|y| \leq 1$)

$$\leq \rho \cdot \underline{\int |f(x) - g(x)| \, d\mu(x)}$$

Specific Setups

- "Average" approximation: given a distribution μ

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

simple

- "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

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!

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

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

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

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

Kernel Method

can be inf-dim, fixed

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

only need evaluate

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

Polynomial kernel

d-dim

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

if p is large

$\to$ strong approx power

Gaussian Kernel

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

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

Kernels have strong approximation power

1D Approximation

$g \in G$: ρ – Lipschitz

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

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

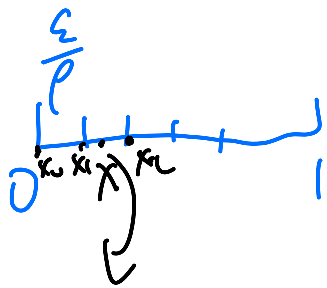
Proof of 1D Approximation

Pf :

Let $m \triangleq \left\lceil \frac{\rho}{\epsilon} \right\rceil$, $x_j \triangleq \frac{(j-1)\,\epsilon}{\rho}$

$$f(x) = \sum_{j=1}^{m} a_j \cdot \mathbb{1}\{x - x_j \geq 0\}$$

$a_1 = g(0)$, $a_j = g(x_j) - g(x_{j-1})$, $j = 2, \dots, m$



$g(x)$

$f(x) = g(x_1)$

- if $x < x_1$, $\mathbb{1}\{x - x_j \geq 0\} = 0$, $j = 1, \dots, m$

  $f(x) = g(0)$

- if $x_1 \leq x < x_2$, $\mathbb{1}\{x - x_j \geq 0\} = 0$, $j = 2, \dots, m$

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

$|g(x) - f(x)| = |g(x) - f(x_j)|$, $x_j \leq x$ closest

$\leq |g(x) - g(x_j)| + \underbrace{|g(x_j) - f(x_j)|}_{0}$

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

$= \rho \cdot \frac{\epsilon}{\rho} = \epsilon$ $\square$

Multivariate Approximation

$x \in \mathbb{R}^d$, $g(x) \in \mathbb{R}$

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 distribution

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

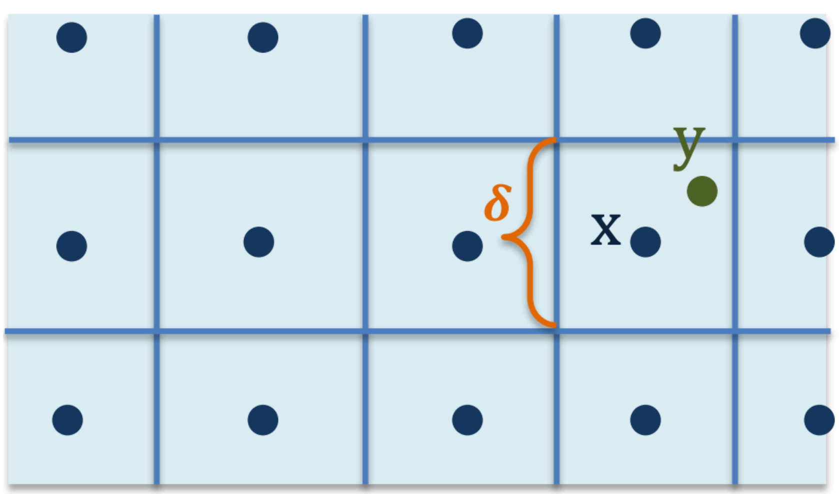
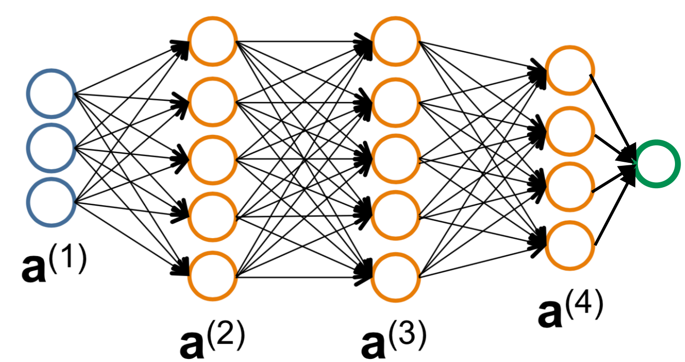


Figure credit to Andrej Risteski

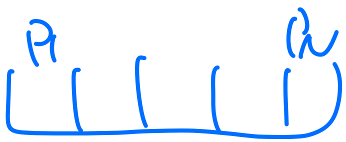


Partition Lemma

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

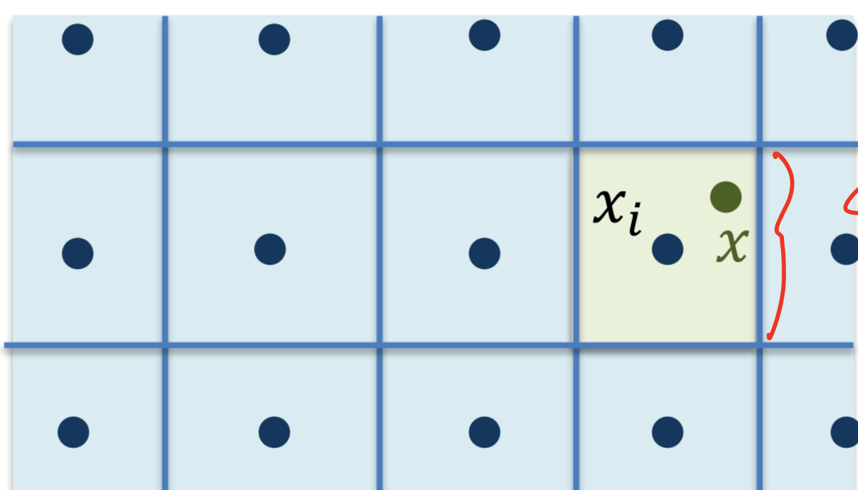
$$\sup_{x \in [0,1]^d} |g(x) - h(x)| \leq \epsilon \text{ with } h(x) := \sum_{i=1}^{N} \alpha_i \mathbf{1}_{R_i}(x).$$

Generalization of 1d theorem



non-parametric regression

Hölder space

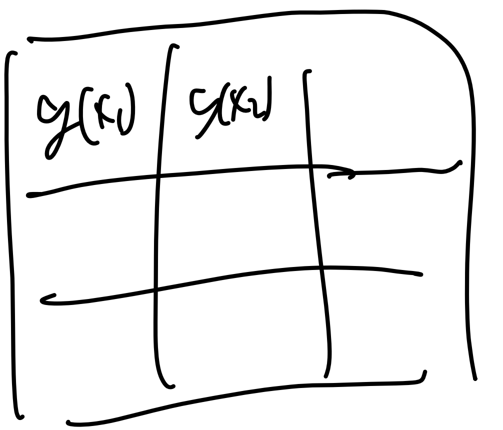


$< \delta$

Figure credit to Andrej Risteski

Proof of Partition Lemma

Pf: For each R_i, pick $x_i \in R_i$, s.t. $\alpha_i \triangleq g(x_i)$



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

$$\le \sup_{i \in \{1,\dots,N\}} \sup_{x \in R_i} \Big(\underbrace{|g(x) - g(x_i)|}_{\varepsilon} + \underbrace{|g(x_i) - h(x)|}_{0} \Big)$$

$$\le \varepsilon + 0$$

$$= \varepsilon$$

$\square$

Proof of Multivariate Approximation Theorem

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

1) use 2-layer NN to approximate $x \mapsto \mathbb{1}_{R_i}(x)$

2) find a linear combination to represent h

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

① Let $f = \sum_{i=1}^{N} \alpha_i f_i$, $\alpha_i = g(x_i)$

goal: f_i to approximate $\mathbb{1}_{R_i}(x)$

$$\|f-h\|_1 = \Big\| \sum_{i=1}^{N} \alpha_i \left(\mathbb{1}_{R_i} - f_i\right) \Big\|_1$$

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

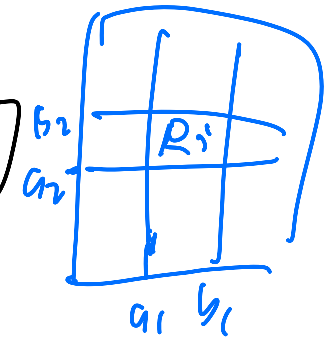
want to show $\|\mathbb{1}_{R_i} - f_i\|_1 \le \dfrac{\varepsilon}{\sum_{i=1}^{N} |\alpha_i|}$

what if $\sum_{i=1}^{N} |\alpha_i| = 0$? $g(x_i) \equiv 0$, $|g(x)| \le \varepsilon$

use 0-function

Proof of Multivariate Approximation Theorem

★ bump function goal: construct f_j

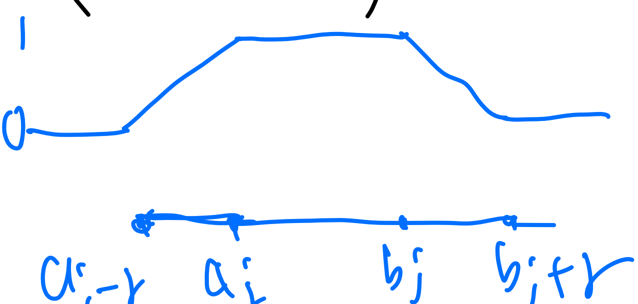


Recall $R_j \triangleq [a_1,b_1]\times[a_2,b_2]\times\cdots\times[a_d,b_d]$

Given $\gamma>0$, define (σ: ReLU)

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

$(j=1,\dots,d)$



smooth approximation to indicator

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

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

$\gamma\to 0$, $\underbrace{g_{\gamma,j}\longrightarrow \mathbb{1}_{[a_j,b_j]}}_{\text{in } L_1}$

Proof of Multivariate Approximation Theorem

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

$x = \begin{pmatrix} \vdots \\ \vdots \\ x^d \end{pmatrix}$ $\quad g_r(x) = \begin{cases} 1 & \text{if } x \in R_i \\ 0 & \text{if } x \notin [a_1-r, b_1+r) \times [a_2-r, b_2+r) \cdots [a_d-r, b_d+r) \\ [0,1] & \text{otherwise} \end{cases}$

Since $r \to 0$, $g_r \to \mathbb{1}_{R_i}$

$\exists\, r$ with $\|g_r - \mathbb{1}_{R_i}\|_1 \le \dfrac{\varepsilon}{\sum_i |\alpha_i|}$

Let $f_i = g_r$

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

$\square$

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
- …