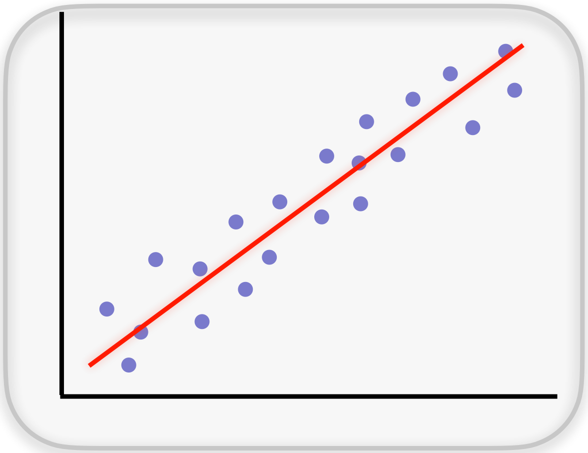
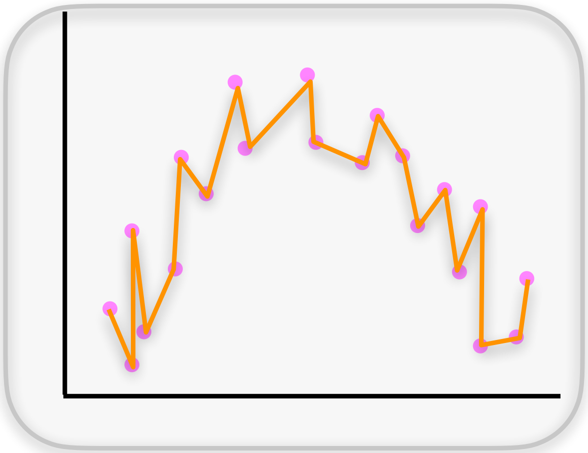
Approximation Theory

UNIVERSITY of WASHINGTON

Expressivity / Representation Power

$f \in \mathcal{F}$

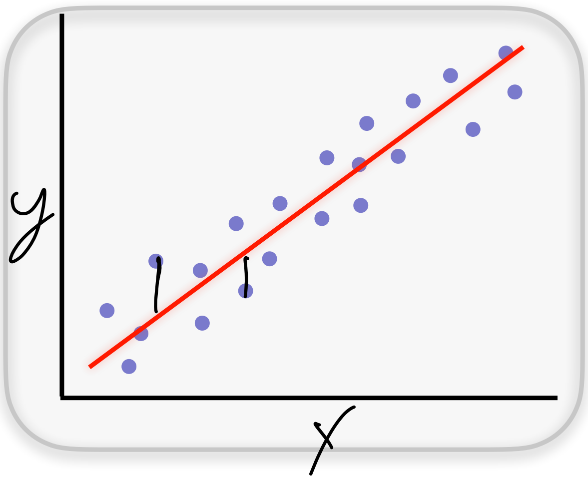


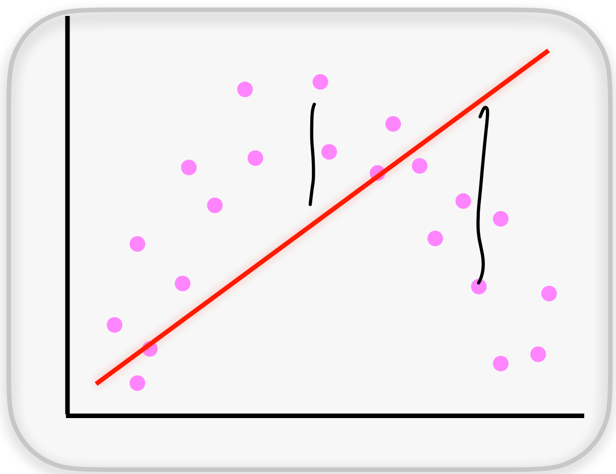


Expressive: Functions in class can represent "complicated" functions.

Linear Function

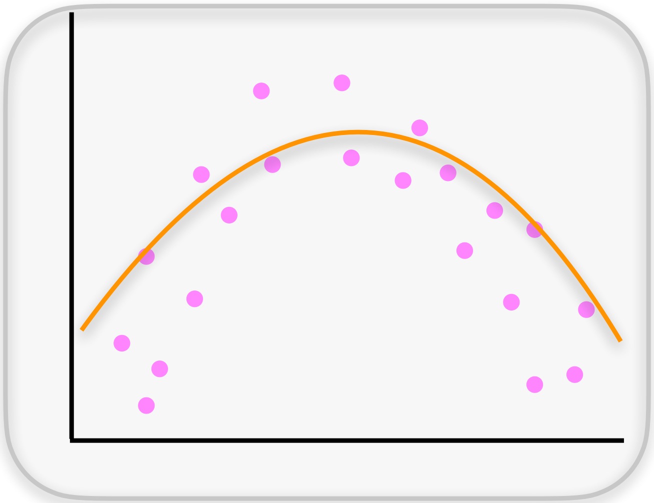
large bias





best linear fit

Review: generalized linear regression



Transformed data:

$h_1(x) = 1$

$h_2(x) = x$

$h_3(x) = x^2$

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

Hypothesis: linear in h

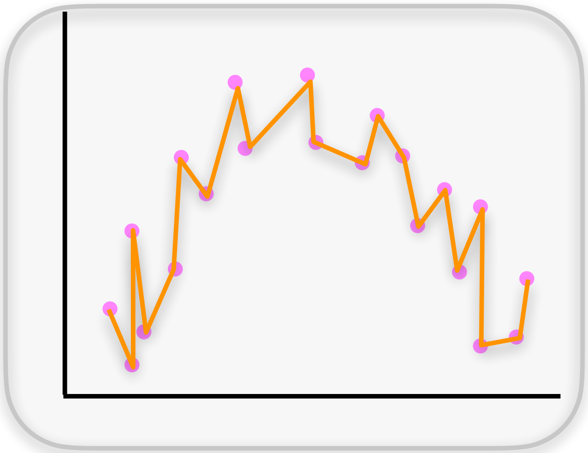
$$y_i \approx h(x_i)^T w$$

Review: Polynomial Regression

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

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

$$w \in \mathbb{R}^n$$



Lagrange's Interpolation Theorem

$\{(x_i, y_i)\}_{i=1}^{n}$, $\exists$ polynomial f of degree $(n-1)$, $\forall i$, $y_i = f(x_i)$

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:

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

Example

① $\ell(f(x), y) = \ell(yf(x))$, ρ-Lipshitz

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

e.g. hinge loss

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

$\rho = 1$

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

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

Decomposition

f, g

$$\int \left(\ell(y f(x)) - \ell(y g(x)) \right) d\mu(x,y)$$

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

$$\leq \int \rho \left| y f(x) - y g(x) \right| d\mu(x,y)$$

$$\leq \rho \int \left| f(x) - g(x) \right| d\mu(x,y)$$

(assume $|y| \leq 1$)

Specific Setups

- "Average" approximation: given a distribution μ

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

- "Everywhere" approximation

$$\|f-g\|_\infty = \sup_{x\in S} |f(x)-g(x)| \geq \|f-g\|_\mu$$

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

$$\leq \int_X \sup_x |f(x)-g(x)|\, d\mu(x) = \|f-g\|_\infty \cdot \int_X d\mu(x)$$

$$= \|f-g\|_\infty$$

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!

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

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

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

$$w = (f(x_0), f'(x_0), \ldots)$$

Kernel Method

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

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

Polynomial kernel

d-dimension x

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

Gaussian Kernel

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

$x, x' \in \mathbb{R}$

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

$\rightarrow$ kernel has inf representation power

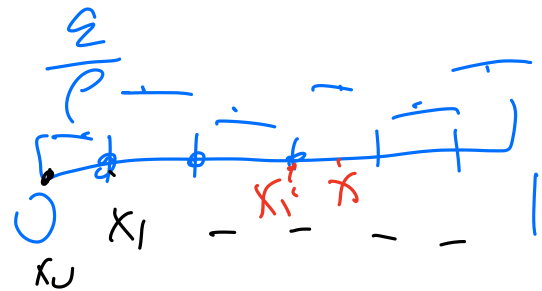
1D Approximation

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

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

g: target function

Proof of 1D Approximation

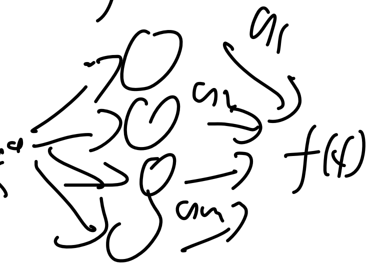


Pf: $m \triangleq \left\lceil \frac{\rho}{\varepsilon} \right\rceil$

$x_i \triangleq \frac{(i-1)\,\varepsilon}{\rho}$

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

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



— if $x < x_1$, $f(x) = a_0 = g(x_0)$

— $x_1 \leq x < x_2$ $\quad f(x) = g(x_1)$

$|g(x) - f(x)| = |g(x) - f(x_i)|$, $x_i \leq x$, closet

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

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

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

Multivariate Approximation

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

ϵ

δ

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

unif diff

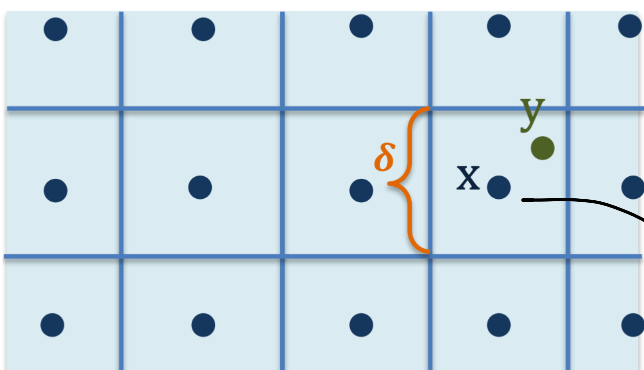
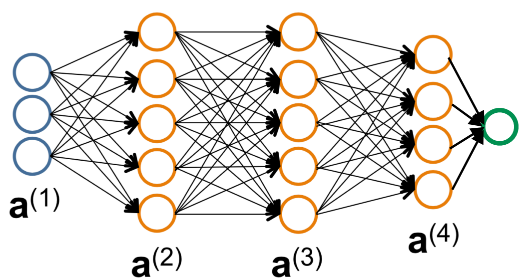


Figure credit to Andrej Risteski



$|g(x) - g(y)| \le \epsilon$

Partition Lemma

generalization of 1d approx

Lemma: let g, δ, ϵ be given. For any partition P of $[0,1]^d$, $d \geq 2$
$P = (R_1, \ldots, R_N)$ with all side length smaller than δ,
there exists $(\alpha_1, \ldots, \alpha_N) \in \mathbb{R}^N$ such that

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

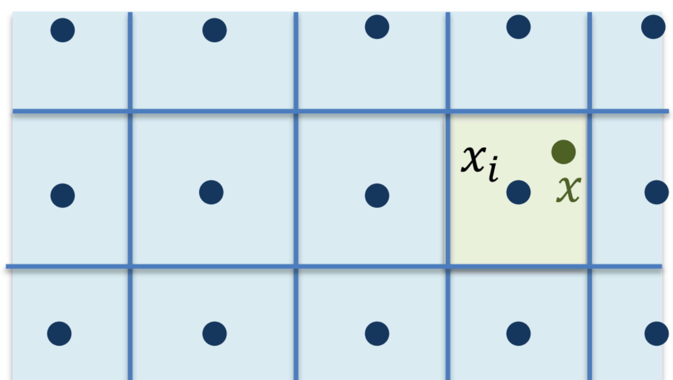
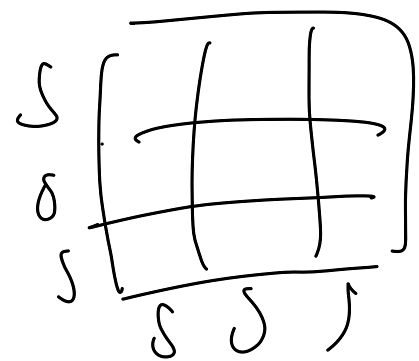


Figure credit to Andrej Risteski



Proof of Partition Lemma

Pf: For each R_i, pick $x_i \in R_i$, s.t.

$$\alpha_i \triangleq g(x_i) \qquad h(x) = \sum \alpha_i \mathbb{1}\{x \in R_i\}$$

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

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

$$\|x - x_i\|_\infty \le \delta$$

$$\rightarrow |g(x) - g(x_i)| \le \varepsilon \qquad \le \qquad \varepsilon$$

$\square$

Proof of Multivariate Approximation Theorem

Idea: $h(x) = \sum_{i=1}^{N} \underbrace{\alpha_i}_{} \mathbb{1}_{R_i}(x)$ in the lemma

1) use 2-layer NN to approximate

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

2) then find a linear combination to represent h

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

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

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

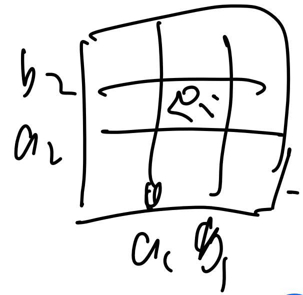
$$\le \sum_{i=1}^{N} |\alpha_i| \cdot \|\mathbb{1}_{R_i} - f_i\|_1$$

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

Proof of Multivariate Approximation Theorem

② construct f_i

$\Omega_i \doteq [a_1,b_1]\times[a_2,b_2]\cdots\times[a_d,b_d]$



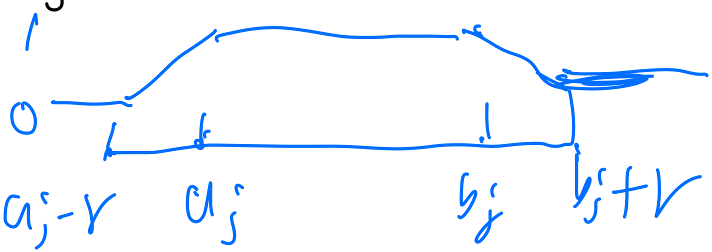
☆ bump function

σ: ReLU, $\sigma(z)=z\cdot\mathbb{1}_{\{z\ge 0\}}$

Given $r>0$,

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

$j=1,\dots,d$ $\quad \mathbb{R}\mapsto\mathbb{R}$



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

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

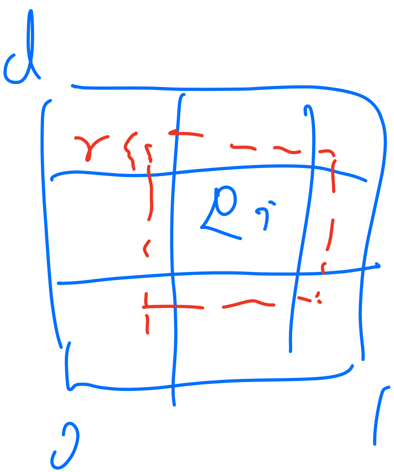
$r\to 0$, $g_{r,j}(z)=\mathbb{1}_{[a_j,b_j]}$

Proof of Multivariate Approximation Theorem

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

$$x = \begin{pmatrix} x^1 \\ \vdots \\ x^d \end{pmatrix}$$

$$g_r(x) = \begin{cases} 1 & \text{if } x \in R_i \\ 0 & \text{if } x \notin [a_1 - r, b_1 + r] \times \cdots \times [a_d - r, b_d + r] \\ [0,1] & \text{o.w.} \end{cases}$$



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

$\exists\, r$ s.t. $\|g_r - 1_{R_i}\|_1 \le \dfrac{\varepsilon}{\sum_i |\alpha_i|}$

Define $f_i = g_r$

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

$\square$

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

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

$h(x) = \sum \alpha_i \mathbb{1}_{R_i}$

Stone-Weierstrass Theorem

Theorem: If $\mathscr{F}$ satisfies

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

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

$g(x) \neq g(x')$

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,

$f_1, f_2 \in \mathcal{F}$, $f_1 \cdot f_2 \in \mathcal{F}$

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

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

Example: cos activation

Example: cos activation

Other Examples

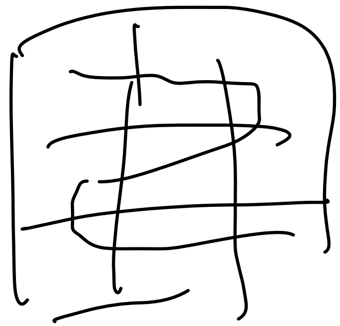
Exponential activation

ReLU activation

Curse of Dimensionality

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

- Unavoidable in the worse case



- Barron's theory

Fourier

Recent Advances in Representation Power

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