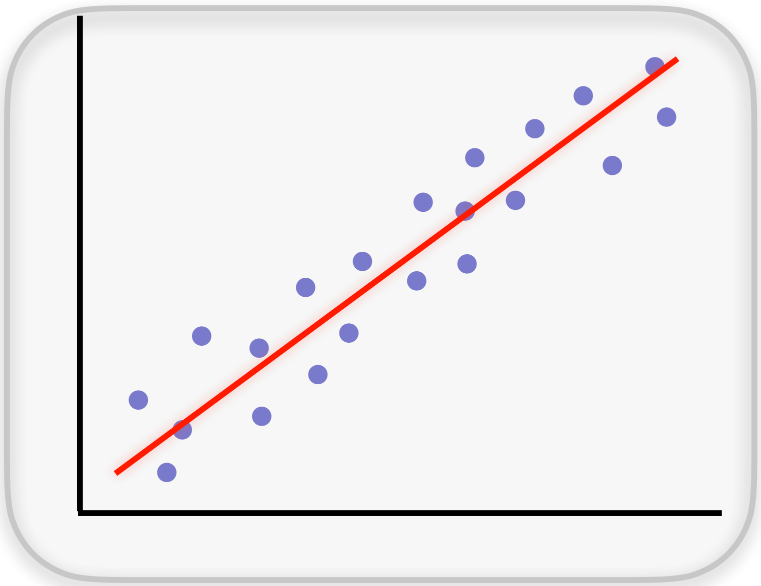
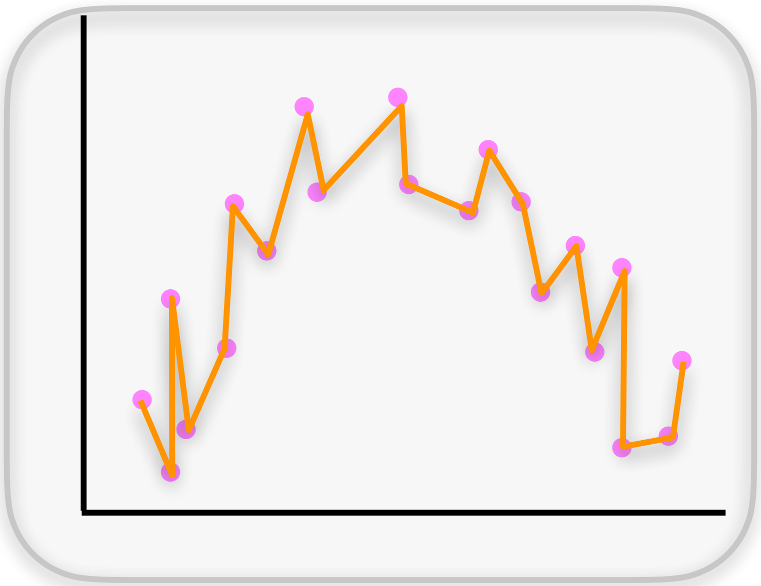
Approximation Theory

UNIVERSITY *of* WASHINGTON

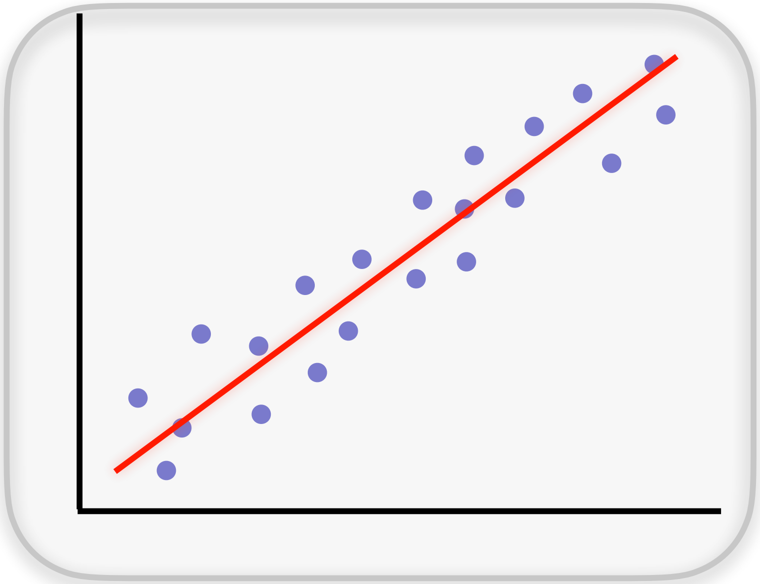
Expressivity / Representation Power

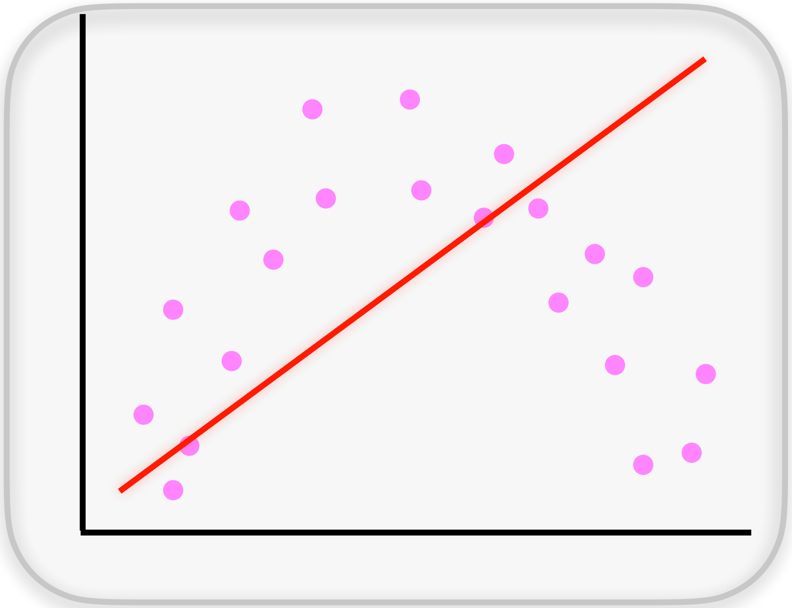




Expressive: Functions in class can represent "complicated" functions.

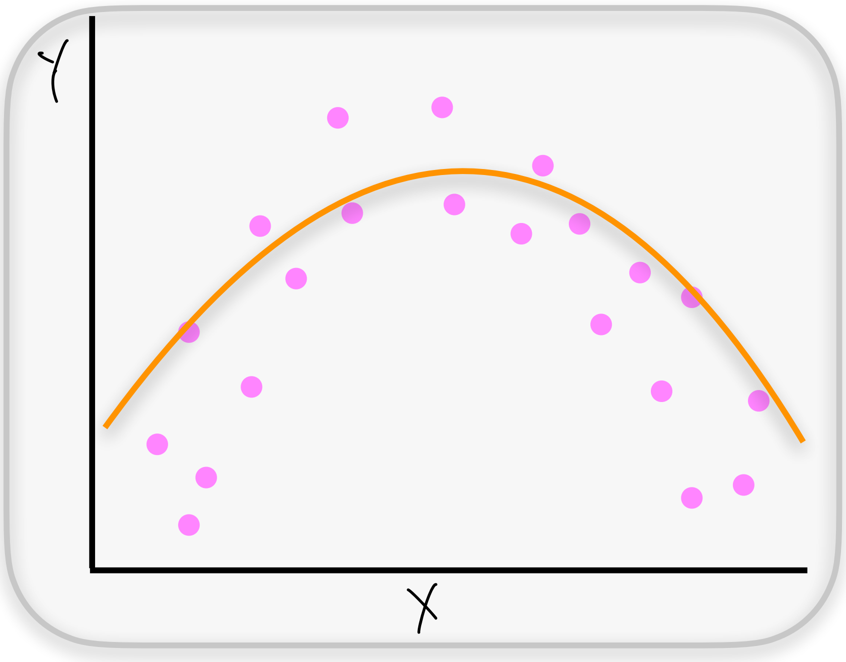
Linear Function





best linear fit

Review: generalized linear regression



Transformed data:

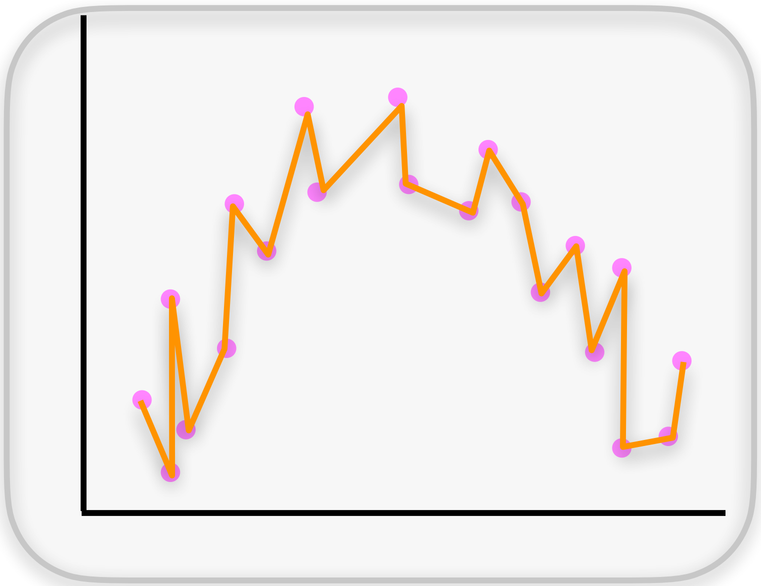
$h_1(x) = 1$
$h_2(x) = x$
$h_3(x) = x^2$

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

Hypothesis: linear in h

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

Review: Polynomial Regression



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

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

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

Lagrange's Interpolation Theorem:

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

$\exists$ polynomial f of degree $(n-1)$ s.t.

$\forall i, \; y_i = f(x_i)$

Assumption: $\forall (x_i, y_i), (x_j, y_j)$

if $x_i = x_j$, $y_i = y_j$

Approximation Theory Setup

Sanity check

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

- Set up a natural baseline: $\geq$

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

$\mathcal{F} \subset$ continuous function

Example

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

$$|\ell(z) - \ell(z')| \le \ell |z - z'|$$

e.g. hinge loss

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

1- Lipshitz

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

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

Decomposition

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

$$= \int \big(\ell(y f(x)) - \ell(y g(x)) \big) \, d\mu(x,y)$$

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

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

(assume $|y| \leq 1$)

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

Specific Setups

- "Average" approximation: given a distribution μ

$$\|f - g\|_{\mu} = \int_{x} |f(x) - g(x)|\, d\mu(x)$$

- "Everywhere" approximation

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

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

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

$$= \|f-g\|_{\infty} \cdot \underbrace{\int_X d\mu(x)}_{=1}$$

Polynomial Approximation

continuous

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

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

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

$f(x_{test}) = y^T K(X,X) K(X, x_{test})$

Polynomial kernel

$\phi(x) = (1, x, x^2, \ldots, x^{10})$

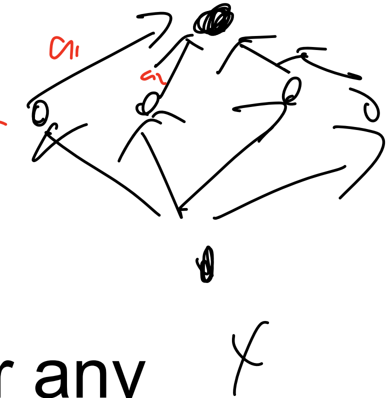
or $\phi(x) = (1, x - x_0, (x - x_0)^2, \ldots)$

Gaussian Kernel

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

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

1D Approximation

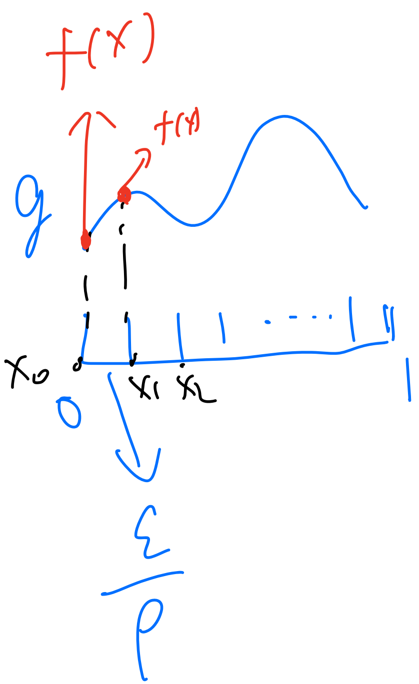


Theorem: Let $g : [0,1] \to R$, and ρ-Lipschitz. For any $\epsilon > 0, \exists$ 2-layer neural network f with $\lceil \frac{\rho}{\epsilon} \rceil$ nodes, threshold activation: $\sigma(z) : z \mapsto \mathbf{1}\{z \geq 0\}$ such that

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

Proof of 1D Approximation

Pf:



$$m \doteq \left\lceil \frac{\rho}{\varepsilon} \right\rceil, \quad x_i = \frac{(i-1)\varepsilon}{\rho}$$

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

$$a_0 = g(0), \quad a_i = g(x_i) - g(x_{i-1})$$

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

$$f(x) = g(0)$$

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

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

$$|g(x) - f(x)| = |g(x) - f(x_i)|, \quad x_i \leq x, \text{ closest}$$

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

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

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

$\square$

Multivariate Approximation

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

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

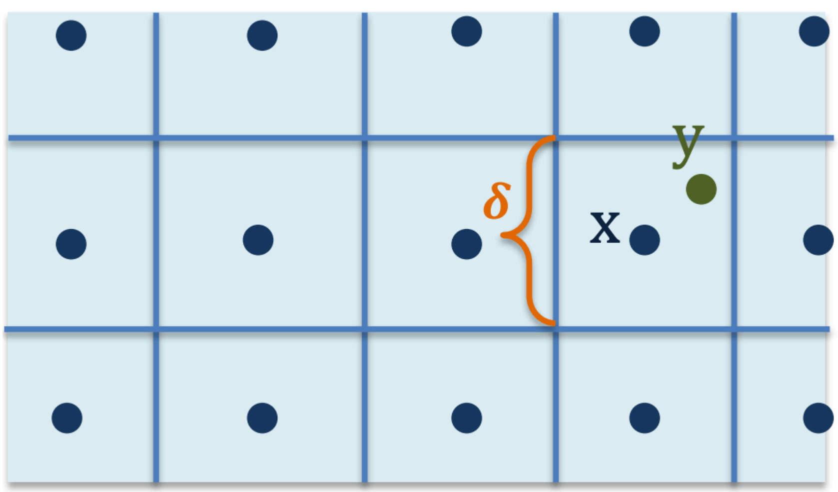
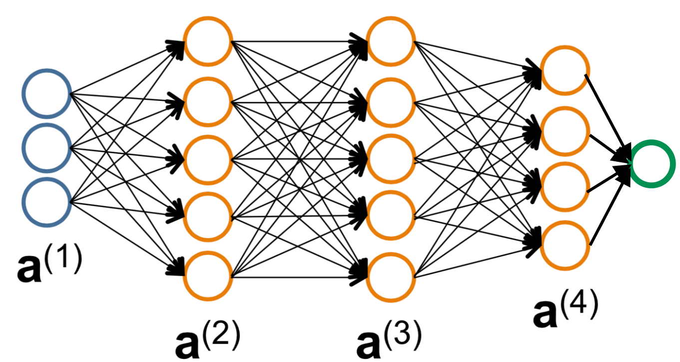


Figure credit to Andrej Risteski



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

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

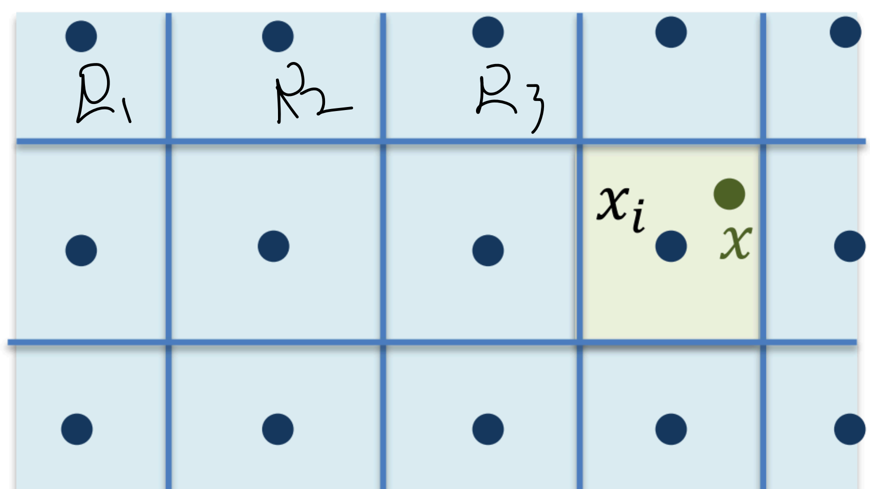


Figure credit to Andrej Risteski

Proof of Partition Lemma

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

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

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

$$\le \varepsilon \qquad \square$$

Proof of Multivariate Approximation Theorem

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

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

2) find a linear combination to represent h

$\Rightarrow \|f-g\|_1 \leq \|f-h\|_1 + \|h-g\|_1$

Let $f = \sum_{i=1}^{N} \alpha_i f_i$, $f_i \approx \mathbb{1}_{R_i}(x)$

$\alpha_i \triangleq g(x_i)$

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

$$\leq \sum_i |\alpha_i| \, \| \mathbb{1}_{R_i} - f_i \|_1$$

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

$\Rightarrow \|f - g\|_1 \leq \varepsilon$

if $\sum_{i=1}^{N} |\alpha_i| = 0 \Rightarrow g(x_i) = 0$

$\Rightarrow |g(x)| \leq \varepsilon$

use 0-network

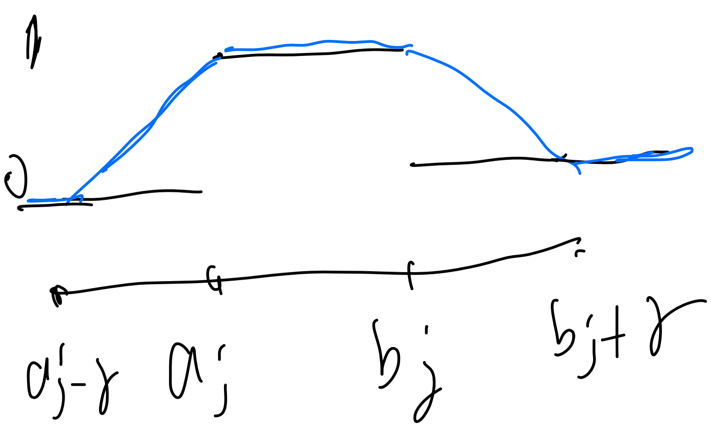
Proof of Multivariate Approximation Theorem

(1) bump function / smooth approximating

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

Given $r > 0$, define (σ: ReLU)

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



$$\begin{cases} \text{if } z \in [a_j, b_j] \Rightarrow g_{r,j}(z) = 1 \\ z \notin (a_j - r, b_j + r) \Rightarrow g_{r,j}(z) = 0 \end{cases}$$

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

Proof of Multivariate Approximation Theorem

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

$x = \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^d \end{pmatrix}$ $\qquad g_r(x) = \begin{cases} 1 & \text{if } x \in R_i \\ 0 & \text{if } x \notin [a_1 - r, b_1 + r] \times [a_2 - r, b_2 + r] \cdots \\ [0,1] & \text{o.w.} \end{cases}$

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

$\Rightarrow g_r \to \mathbb{1}_{R_i}$

choose $f_i = g_r$

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