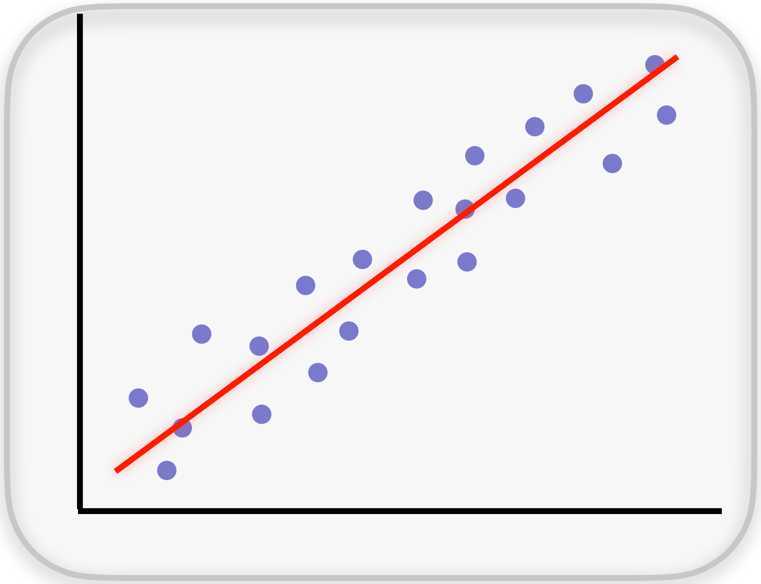
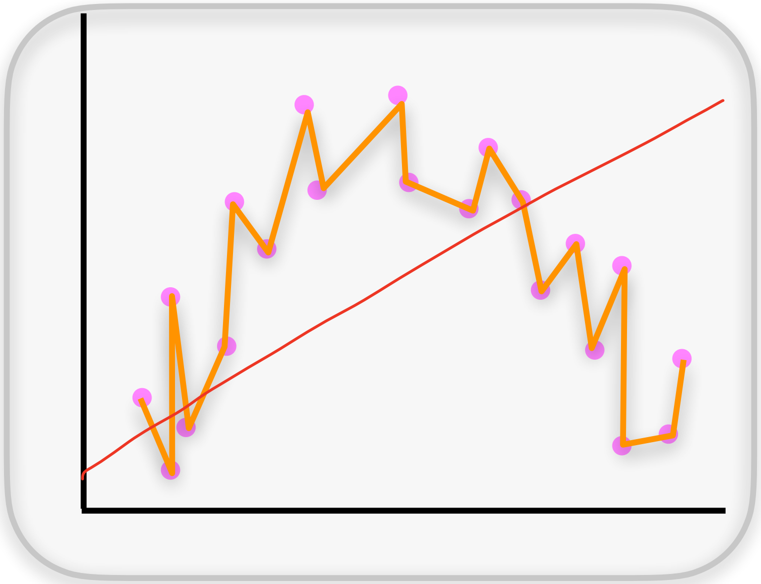
# Approximation Theory

UNIVERSITY *of* WASHINGTON

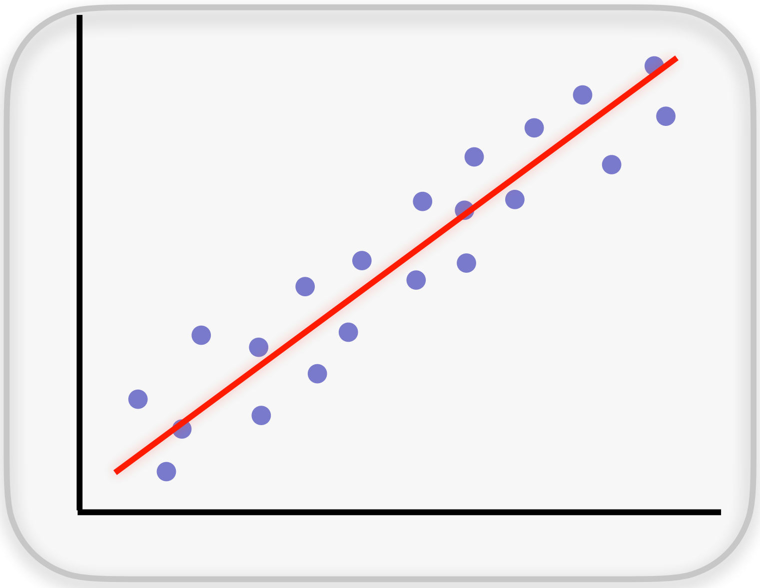
# Expressivity / Representation Power





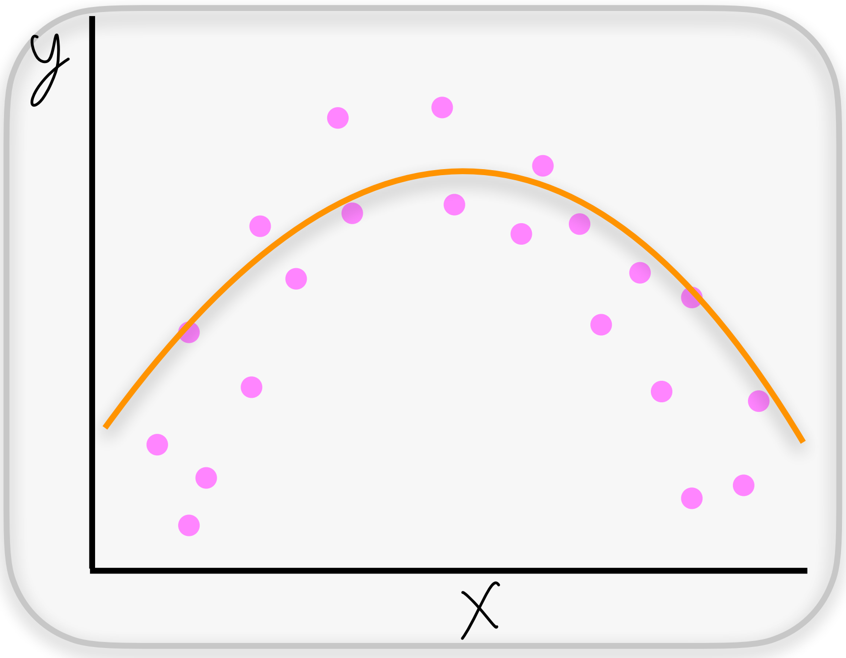
Expressive: Functions in class can represent "complicated" functions.

# Linear Function



best linear fit

# Review: generalized linear regression




Transformed data: $h_1(x) = 1$

$h_2(x) = x$

$h_3(x) = x^2$

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

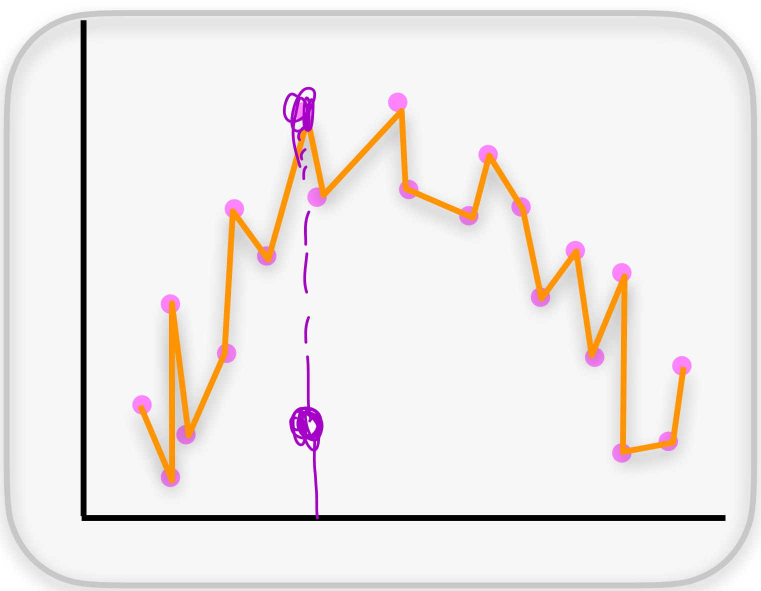
⋮

Hypothesis: linear in h

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

# Review: Polynomial Regression

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



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

$$w = (w_0, \dots, w_p)$$

Lagrange's Interpolation Theorem:

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

$\exists$ polynomial $P$ of degree $n-1$, s.t.

$\forall i$, $y_i = P(x_i)$

(Condition: no $(x_i, y_i)$, $(x_j, y_j)$ s.t. $x_i = x_j$, $y_i \neq y_j$)

# Approximation Theory Setup

- Goal: to show there exists a neural network that has small error on training / $\underbrace{\text{test set}}_{\text{population loss}}$.

- 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) loss $\quad \ell(f(x), y) = \ell(yf(x))$, $\rho$-Lipshitz

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

e.g. hinge loss

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

$\ell$ is 1-Lipshitz

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

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

# Decomposition

$$\mathcal{L}(f) - \mathcal{L}(g)$$

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

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

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

assume $|y| \leq 1$

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

# Specific Setups

- "Average" approximation: given a distribution $\mu$

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

- "Everywhere" approximation

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

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

# Polynomial Approximation

continuous

**Theorem (Stone-Weierstrass)**: for any function $f$, we can approximate it on any compact set $\Omega$ 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!

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

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

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

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

# Kernel Method

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

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

**Polynomial kernel**

d-dim input, choose degree $p$

$\phi(x) = (1, x_1, x_2, \ldots, x_d, x_1^2, x_1 x_2, x_1 x_3, \ldots, \ldots x_d^p)$

**Gaussian Kernel**

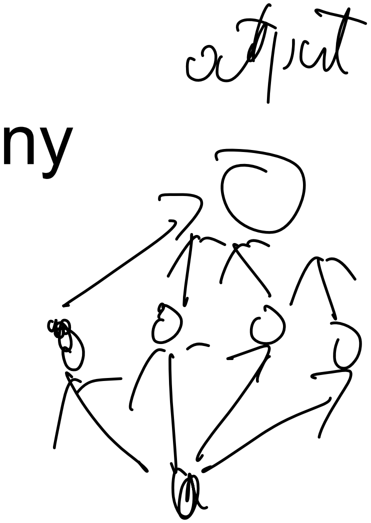
1-dim, $K(x, x') = \exp\left(-\frac{(x-x')^2}{2\sigma^2}\right)$

$\phi(x) = e^{-\frac{x^2}{2\sigma^2}} \cdot \left(1, \sqrt{\frac{1}{1!}}\frac{x}{\sigma}, \sqrt{\frac{1}{2!}}\left(\frac{x}{\sigma}\right)^2, \sqrt{\frac{1}{3!}}\left(\frac{x}{\sigma}\right)^3, \ldots\right)$

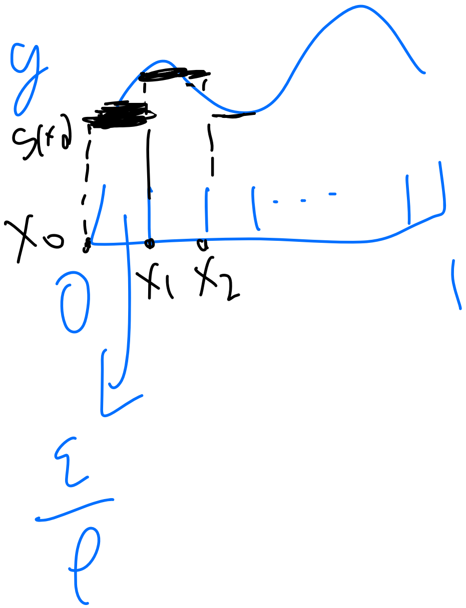
# 1D Approximation

**Theorem**: Let $g : [0,1] \to R$, and $\rho$-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 $m \triangleq \lceil \frac{\rho}{\varepsilon} \rceil$, $x_i = \frac{(i-1)\varepsilon}{\rho}$, $i=1,\dots,m$

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

$a_1 = g(0)$, $a_i = g(x_i) - g(x_{i-1})$, $i=1,\dots,m$

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

  $f(x) = g(x_0)$

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

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

\- \- \-

$|g(x) - f(x)| = |g(x) - f(x_i)|$, say $x \geq x_i$, $x < x_{i+1}$

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

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

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

# Multivariate Approximation

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

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

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

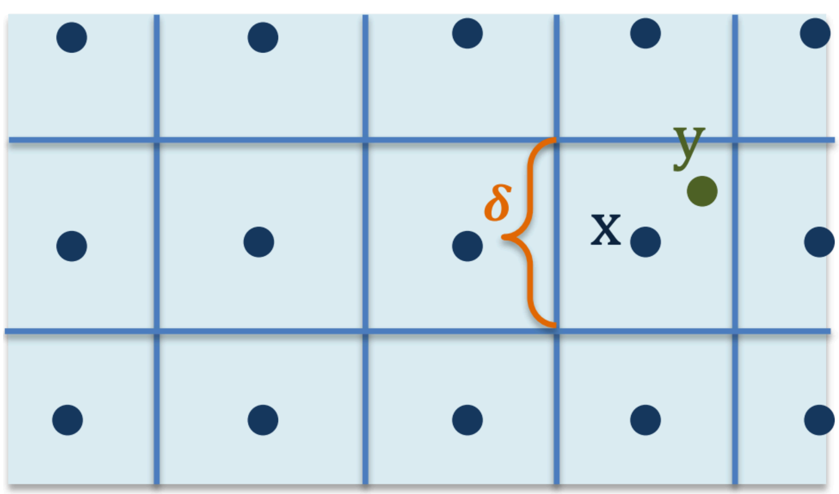
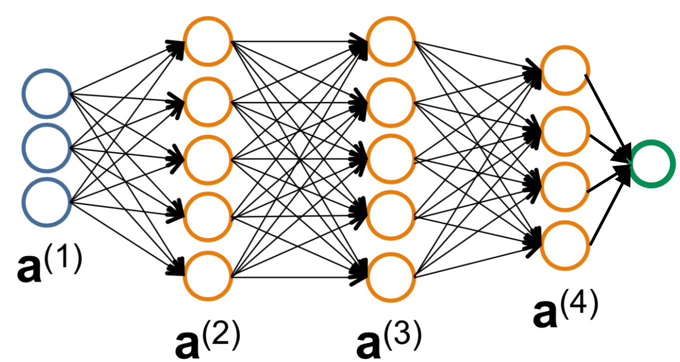
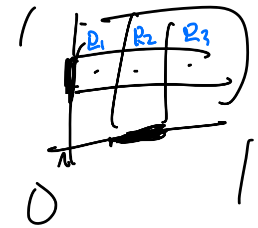



Figure credit to Andrej Risteski

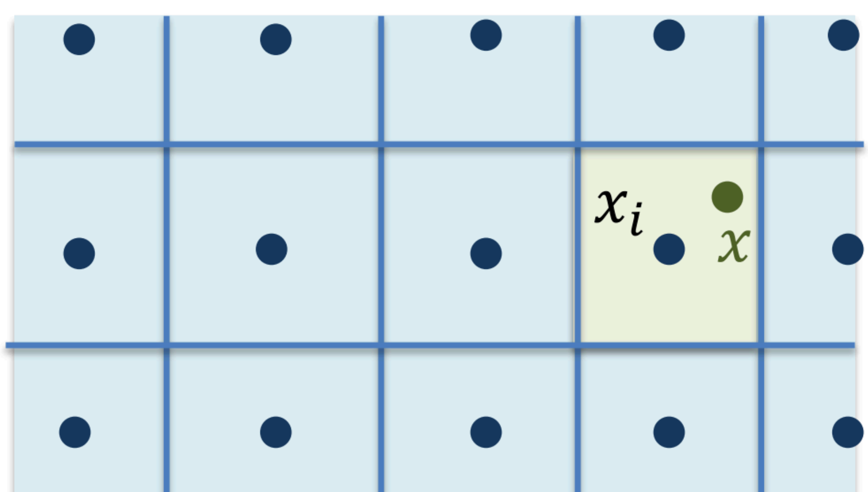





# Partition Lemma

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

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



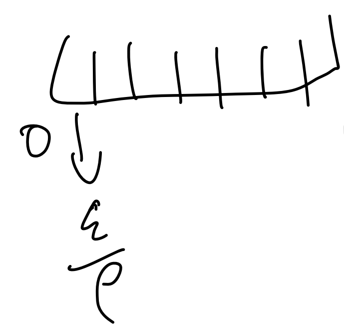



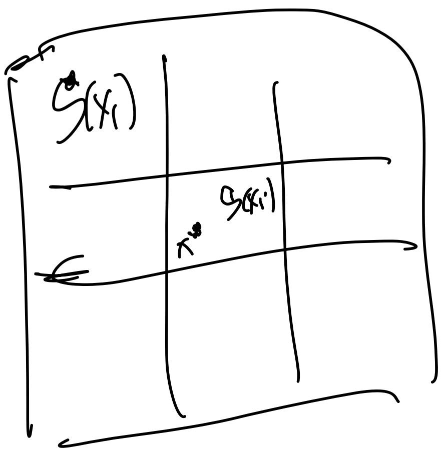

Figure credit to Andrej Risteski

2D

# Proof of Partition Lemma

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

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




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

(we know $\|x-x_i\|_\infty \le \delta$)

$$\le \sup_{i\in\{1,\dots,N\}} \sup_{x\in Q_i} \varepsilon$$

$$= \varepsilon$$

$\square$

# Proof of Multivariate Approximation Theorem

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

function $f$ $\begin{cases} 1) \text{ use 2-layer NN to approximate } x \mapsto \mathbb{1}_{R_i}(x) \\ 2) \text{ find a linear combination to represent } h \end{cases}$

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

For 2) say we have $f_i$ that approximates $\mathbb{1}_{R_i}$

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

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

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

if from 1), we show $\|\mathbb{1}_{R_i} - f_i\|_1 \le \dfrac{\varepsilon}{\sum_{j=1}^{N} |\alpha_j|}$, for $\forall i$ $\Rightarrow \|f-h\|_1 \le \varepsilon$

what if $\sum_{j=1}^{N} |\alpha_j| = 0$? $\Rightarrow g(x_i) = 0, \forall i$, $|g(x)| \le \varepsilon, \forall x \Rightarrow f(x) = 0$

# Proof of Multivariate Approximation Theorem

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

$$\mathbb{1}_{R_i} = \mathbb{1}_{[a_1,b_1]}\cdot\mathbb{1}_{[a_2,b_2]}\cdots\cdot\mathbb{1}_{[a_d,b_d]}$$

☆ bump function

Given $r>0$, define ($\sigma$: ReLU)

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


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

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

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

# Proof of Multivariate Approximation Theorem

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

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

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

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

choose $f_i = g_r$ can approximate $\mathbb{1}_{R_i}$

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

$\square$