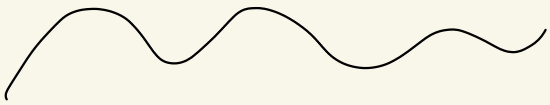
Part 3 Neural Network Approximation

(proof-trivial)

- Approximation (Representation)
- Generalization
- Optimization



Goal: $f: \mathbb{R}^d \to \mathbb{R}$ $\longrightarrow$ neural network, $g: \mathbb{R}^d \to \mathbb{R}$.

population risk $\int \ell(f(x), y)\, dP(x,y) \longleftrightarrow \int \ell(g(x), y)\, dP(x,y)$

upper bound: $\ell(\cdot, y)$ 1-Lipschitz: $\int \big(\ell(g(x),y) - \ell(f(x),y)\big)\, dP(x,y) \le \int |g(x) - f(x)|\, dP(x,y)$

lower bound: w.c. $|g-f|$ large $L_1(P)$, $L_1(\text{Uniform})$

universal uniform

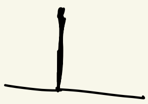
$L_\infty(P)$ approximation

Deep Network: $x \mapsto A_L \sigma_{L-1}(\cdots \sigma_1(A_1 x + b_1) \cdots) + b_L$

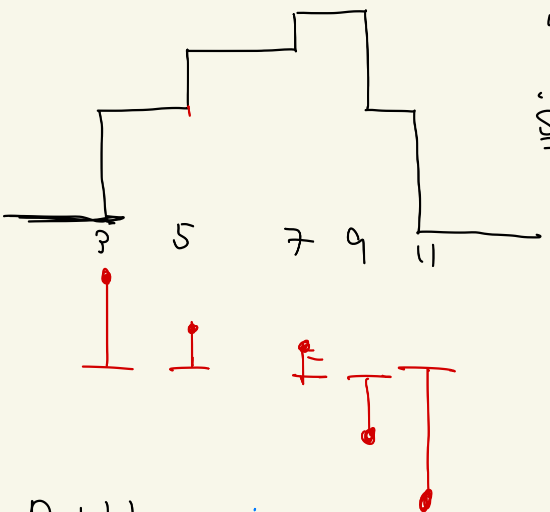
non-linearity/activation: ReLU $z \mapsto \max\{0, z\}$

Univariate

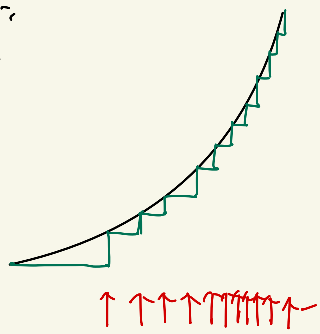
Step Activation



$x \mapsto 2 \cdot \mathbb{1}\{x-3 \ge 0\} + \mathbb{1}\{x-5 \ge 0\} + \mathbb{1}\{x-7 \ge 0\} \cdots$



"smoothe" $\longrightarrow$



Lipschitz L: $\frac{L}{\varepsilon}$

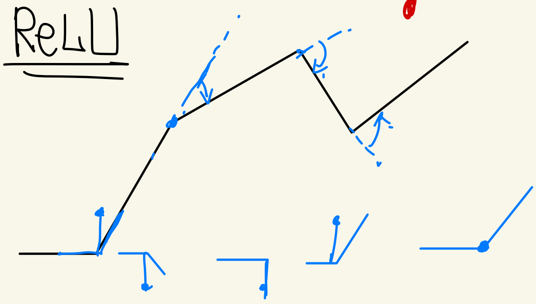
$f(x) = f(0) + \int_0^x f'(b)\, db$

$= f(0) + \int_0^\infty \mathbb{1}\{x - b \ge 0\} f'(b)\, db$

→ Infinite width network

avg L/ε^2.

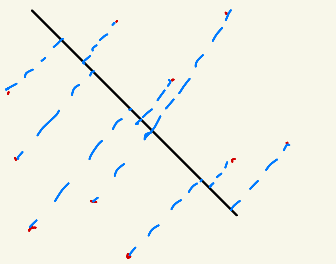
ReLU

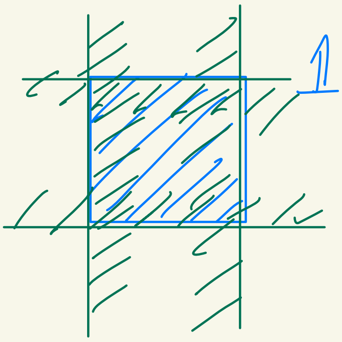


change of slope.

$f(x) = f(0) + r(x) \cdot f'(0) + \int_0^\infty r(x-b) f''(b) \cdot db$, avg $\frac{L}{\varepsilon^2}$

Multivariate box





$\mathbb{1}\{$ inside of the box $\}$

2	3	2
3	4	3
2	3	2

① product ☆☆

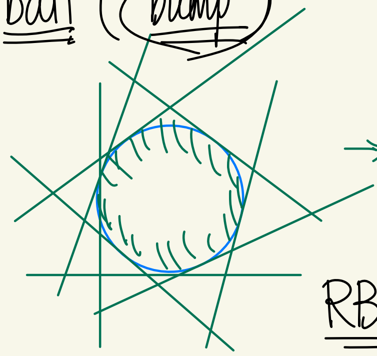
② threshold at 3.5

0	0	0
0	1	0
0	0	0

add a layer

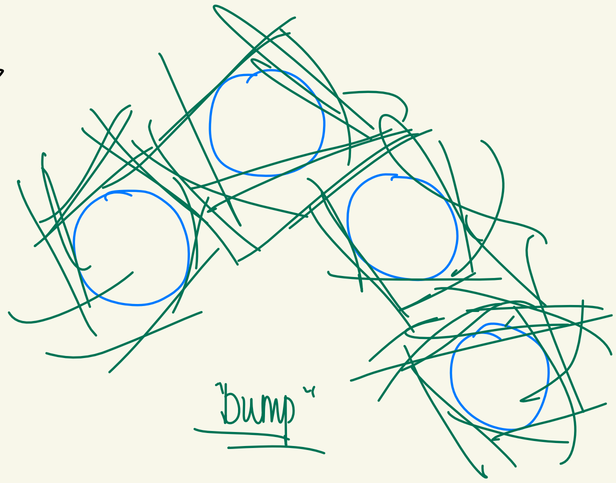
2 Layer

ball ("bump")



→ radial function

2^d nodes



"bump"

RBF: convolve with f

$$\left| f(x) - \int f(z)\cdot p(x-z)\,dz \right|$$

$$= \left| f(x) - \int f(x-z)\cdot p(z)\,dz \right|$$

$$= \left| \int f(x)\,p(z)\,dz - \int f(x-z)\,p(z)\,dz \right| \le \int |f(x) - f(x-z)|\,\underbrace{p(z)}_{\geq 0}\,dz$$

$$\left(\frac{d\cdot L}{\varepsilon}\right)^{O(d)}$$

(Mhaskar-Micchelli '92)

box: 3 layer $(L/\varepsilon)^{O(d)} \longrightarrow \infty$

ball: 2 layer RBF $(d\cdot L/\varepsilon)^{O(d)}$

Univariate bump: $\cos x^p$

$$\mathbb{1}\{\|x\|_\infty \le 1\} = \prod_{i=1}^{d} \mathbb{1}\{|x_i| \le 1\}$$

$$\cos x \cdot \cos x = \tfrac{1}{2}\cos 2x + 1$$

$$2\cos x_1 \cdot \cos x_2 = \cos(x_1+x_2) + \cos(x_1-x_2)$$

↓

Polynomial

1885

Weierstrass Approximation Thm : polynomials can uniformly approximate continuous functions over compact sets.

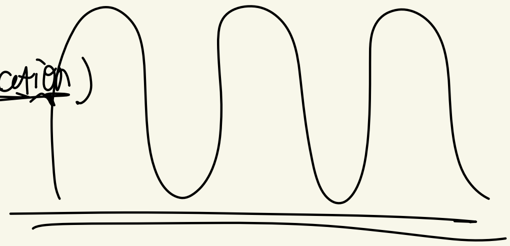
Bernstein's proof

$$\sum \binom{n}{x} p^x (1-p)^{n-x}$$

$\Downarrow$

Stone-Weierstrass thm : polynomial-like functions

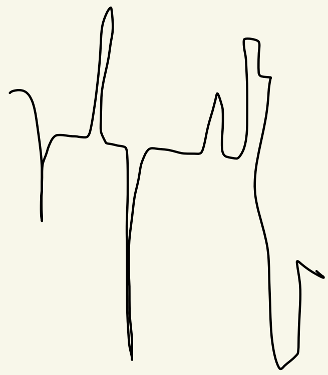
approximate ctx functions (closed under multiplication)



Thm (Hornik-Stinchcombe-White '89)

$\sigma : \mathbb{R} \to \mathbb{R}$, $\lim_{z \to -\infty} \sigma(z) = 0$, $\lim_{z \to \infty} \sigma(z) = 1$

$$H_\sigma = \{ x \mapsto \sigma(a^T x - b) : (a,b) \in \mathbb{R}^{d+1} \}$$



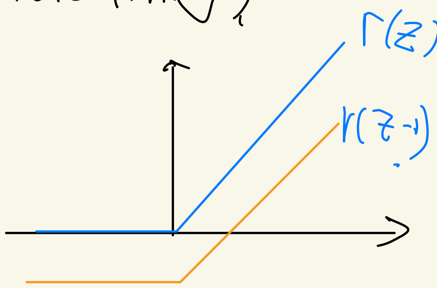
span$\{H_\sigma\}$ uniformly approximates ctx functions on compact sets

Proof sketch: $H_{\cos}$ is closed $\quad 2\cos a \cos b = \cos(a+b) + \cos(a-b)$

$H_{\cos}$ with span$\{H_\sigma\}$

uniformly approximate (univariate fitting)

* $H_{\exp}$: $e^a \cdot e^b = e^{a+b}$ $\quad \square$



* Leshno-Lin-Pinkus-Schocken '93 :

HSW holds iff σ is not a polynomial

$\sigma(z) \triangleq r(z) - r(z-1)$

(Barron '93) $\quad x \mapsto \exp(i a^T x) \Rightarrow \int \exp(i a^T x) \tilde{f}(a)\, da$

Depths $\quad$ Radial functions $\quad$ 2 layer ReLU

$f(\|x\|^2)$ with Lipschitz constant L.

- $h(x) \approx_\varepsilon \|x\|_2^2 = \sum x_i^2$ $\quad \to$ Layer 1 $\to$ $d \cdot L/\varepsilon$. ReLU
- $g \approx_\varepsilon f$ $\quad \to$ Layer 2 $\to$ L/ε ReLU

$$|f(\|x\|^2) - g(h(x))| \le |f(\|x\|^2) - f(h(x))| + |f(h(x)) - g(h(x))|$$

$$\le L \cdot \left| \|x\|^2 - h(x) \right| + \varepsilon \le O(\varepsilon)$$

$g \circ h \to$ size: poly$(L, d, \frac{1}{\varepsilon})$

However, Thm $\exists$ radial function f

(Eldan-Shamir '15) expressible with two layer ReLU of width poly(d)

s.t. every g with a single ReLU layer of width $\boxed{2^{O(d)}}$

satisfies $\int (f(x)-g(x))^2 \, dP(x) \geq \Omega(1)$

A probability measure P

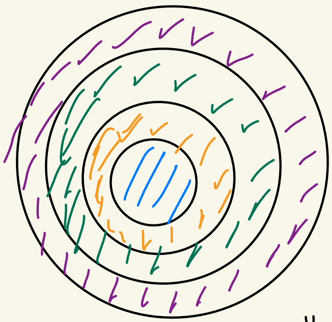
Thm (Daniely, 17') $(x, x') \sim P = \text{Uniform}(S^2)$

$h(x,x') = \sin(\pi d^3 x^T x')$ $\forall g$ with a single ReLU layer of width

$d^{O(d)}$ and weight magnitude $O(2^d)$

$\int (h(x,x') - g(x,x'))^2 \, dP(x,x') \geq \Omega(1)$

h can be approximated to accuracy ε by f with 2 layer ReLU of size poly$(d, \frac{1}{\varepsilon})$

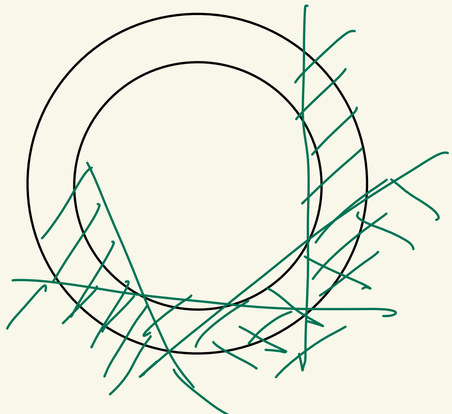


"shell"

each shell approximation

$\downarrow$

overall function

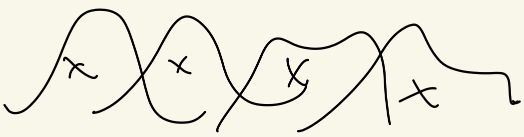


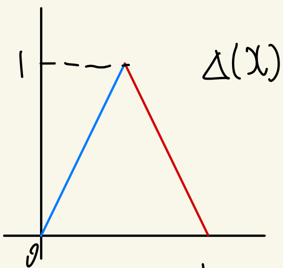
$x \mapsto \mathbb{1}\{\|x\| \in [1-\frac{1}{d}, 1]\}^2$

fraction

Depth < Benefits of depth.

What do shallow representations do exceptionally badly?





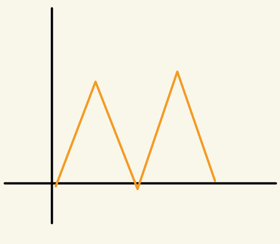
Δ

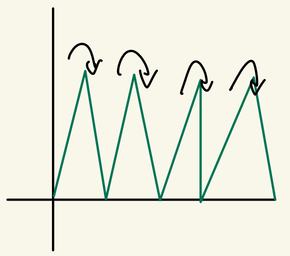
$$\Delta(x) = r(2x) - r(4x-2)$$

$$= \begin{cases} 2x, & x \in [0, \frac{1}{2}] \\ 2(1-x), & x \in [\frac{1}{2}, 1] \end{cases}$$

Composition:

$$f(\Delta(x)) = \begin{cases} x \in [0, \frac{1}{2}] \Rightarrow f(2x) = f \text{ squeezed into } [0, \frac{1}{2}] \\ x \in [\frac{1}{2}, 1] \Rightarrow f(2(1-x)) = f \text{ reversed, squeeze} \end{cases}$$





$\Delta^2 = \Delta \circ \Delta$

Δ^K: $O(K)$ layer & nodes, $O(2^K)$ bumps (oscillation)

Thm (Telgarsky '15) K: # of layer.

$\exists$ ReLU network $f: [0,1] \to [0,1]$ with $\begin{cases} 4 \text{ distinct parameters} \\ 3k^2+9 \text{ nodes} \\ 2k^2+9 \text{ layers} \end{cases}$

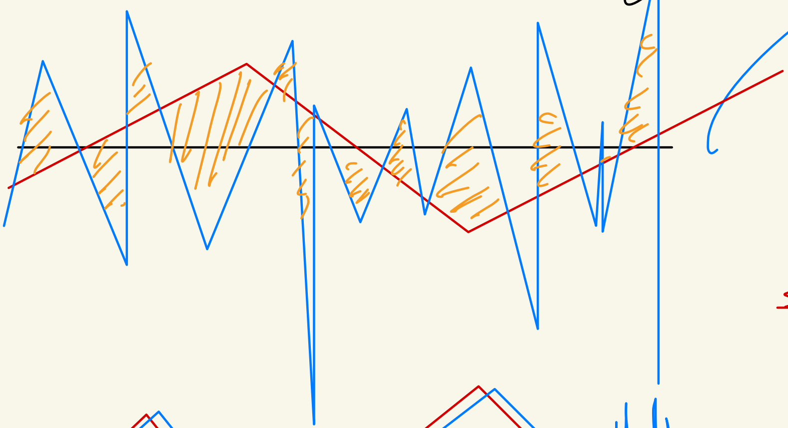
s.t. $\forall$ ReLU network $g: \mathbb{R}^d \to \mathbb{R}$

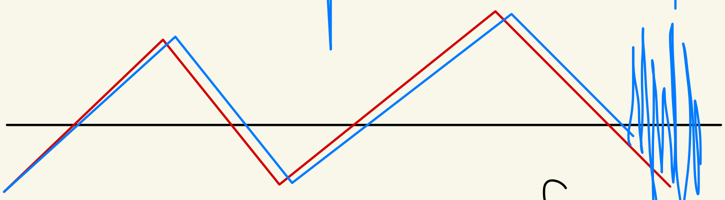
with $\leq k$ layer, $\leq 2^k$ nodes

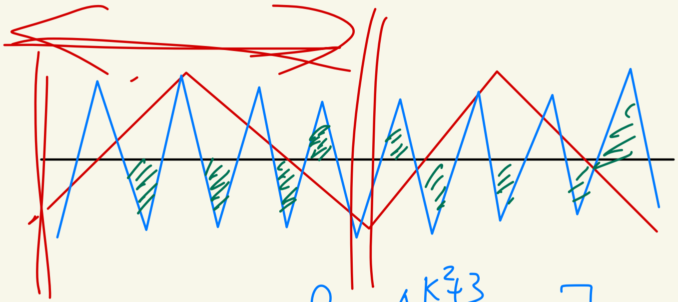
$$\Rightarrow \int_{[0,1]} |f(x) - g(x)|\, dx \geq \frac{1}{32}$$

Proof:

1. g with few oscillations ×
2. $f = \Delta^{k^2+3}$: regular oscillatory f
3. width m, depth L $\Big\}$ $< O(m^L)$







$f = \Delta^{k^2+3}$ $\square$

$\int_{[0,1]} |f-g|$ ×

L_∞ ✓

L_1 ✓

Δ^k: 2^k-Lipschitz

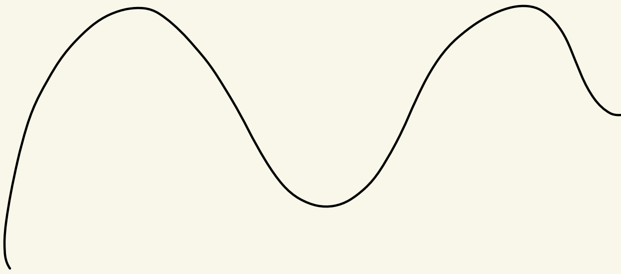
non-realistic

Depth / Radial function
 \ Δ^{k^2+3} function

$O(k^2)$ $O(1)$

if $O(k)$ $O(2^k)$

depth width



Sobolev Spaces