Lec 8

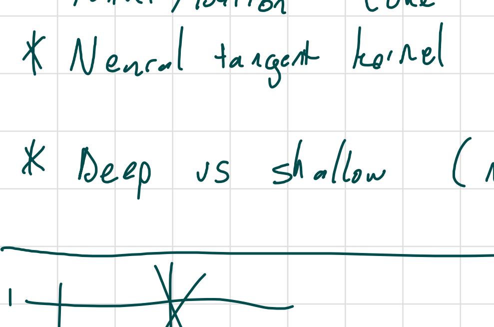
* hwk 1 due Wednesday. (No late hwk.)

Recap Apx:

* Elementary/constructive approximations
* "universal approximation" (Stone-Weierstrass) (one hidden layer)
* Fourier/Barron (one hidden layer)
* Neural tangent kernel

} near initialization

* Deep vs shallow (non-algorithmic)



$$\Delta(x) := 2\sigma(x) - 4\sigma(x-\tfrac{1}{2}) + 2\sigma(x-1) = \begin{cases} 2x & x\in[0,1/2) \\ 2(1-x) & x \in [1/2, 1) \\ 0 & \text{o.w.} \end{cases}$$

$$\underbrace{\Delta \circ \cdots \circ \Delta}_{L}(x) = \Delta^L(x)$$

(2^{L-1} regular copies of Δ)

Proposition. $\Delta^L(x) = \Delta(\langle 2^{L-1}x\rangle) = \Delta(2^{L-1}x - \lfloor 2^{L-1}x \rfloor)$ $\forall x\in[0,1]$

Pf. By induction on depth; peel one off every iteration; details in notes.

Theorem (Telgarsky '15).
$\forall$ depth $L \geq 2$
$f(x) := \Delta^{L^2+2}(x)$ use $3(L^2+2)$ ReLUs total in $2(L^2+2)$ layers

$\forall g: \mathbb{R}\to\mathbb{R}$ $\leq 2^L$ ReLUs in $\leq L$ layers.

$$\int_0^1 |g(x) - f(x)| \geq \frac{1}{32}.$$

Remark (Why L^2?).
Seems to imply existence of naturalish prediction problems where risk is large.

Proof plan:

① Characterize $f(x) := \Delta^{L^2+2}(x) = \Delta(\langle 2^{L^2+1}(x)\rangle)$ ✓
② Define & bound "complexity" of g.
③ 1 & 2 together imply theorem

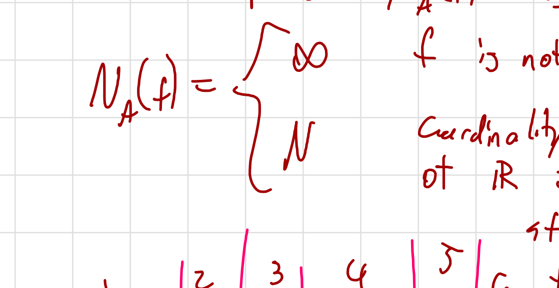
Remark Effective justifications of depth & architectural choices are all open.

---

② Define & bound "complexity"

Definition. Given $f: \mathbb{R}\to\mathbb{R}$, define "number of affine pieces" $N_A(f)$ as

$$N_A(f) = \begin{cases} \infty & f \text{ is not piecewise affine} \\ N & \text{Cardinality of smallest partition of } \mathbb{R} \text{ (into intervals) such that } f \text{ is affine restricted to any partition element} \end{cases}$$



Remark. $\exists$ multivariate version; similar reasoning to VC dimension in later lectures.

Lemma: Given $f:\mathbb{R}\to\mathbb{R}$, ReLU network with m ReLUs, L layers, $(m_1, \dots, m_L)$ widths, $\sum_{i=1}^L m_i = m$.

* $N_A(f) \leq \left(\frac{2m}{L}\right)^L$.

* For any node in layer i, letting $g:\mathbb{R}\to\mathbb{R}$ denote its output as a function of whole network input,
$$N_A(g) \leq 2^i \prod_{j<i} m_j.$$

Remark. "Complexity" N_A grows exponentially in L, polynomially in m.

Remark Eldan & Shamir '15
$\exists f: \mathbb{R}^d \to \mathbb{R}$ with 2 hidden layers (poly(d) nodes), s.t. $\forall g: \mathbb{R}^d\to\mathbb{R}$ with 1 hidden layer and $o(2^d)$ nodes s.t.
$$\|f-g\|_u \geq \text{constant}.$$

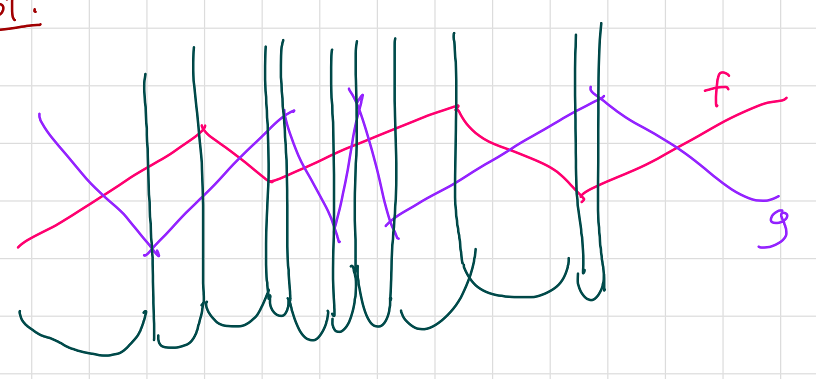
Lemma. Let ($k\geq 1$) $\underbrace{g_1 \cdots g_k, f}_{\text{functions}}$, $\underbrace{a_1 \cdots a_k, b}_{\text{scalars}}$ be given.

① $N_A(f+g_1) \leq N_A(f) + N_A(g_1)$
② $N_A(\sum_i a_i g_i + b) \leq \sum_i N_A(g_i)$
③ $N_A(f\circ g) \leq N_A(f) N_A(g)$
④ $N_A(x \mapsto \underbrace{f(\sum a_i g_i + b)}_{\text{e.g., one node}}) \leq N_A(f)\sum_{i=1}^k N_A(g_i)$.

[width additively increases complexity]
[depth multiplicatively increases complexity]

Proof.

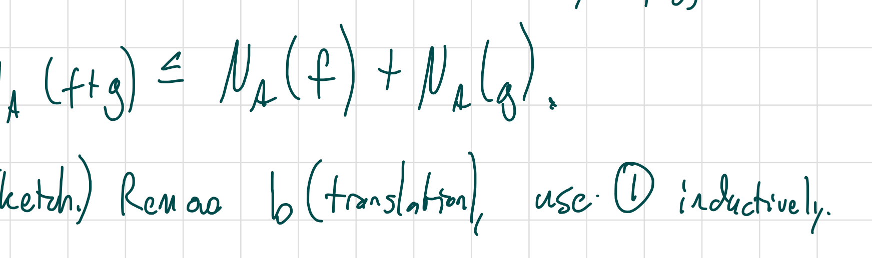
①



Sort the changepoints of the two partitions; within any adjacent pair, both f & g are affine, $f+g$ also affine in that interval, thus
$$N_A(f+g) \leq N_A(f) + N_A(g).$$

② (Sketch) Remove b (translation), use ① inductively.

③ $N_A(f\circ g)$.



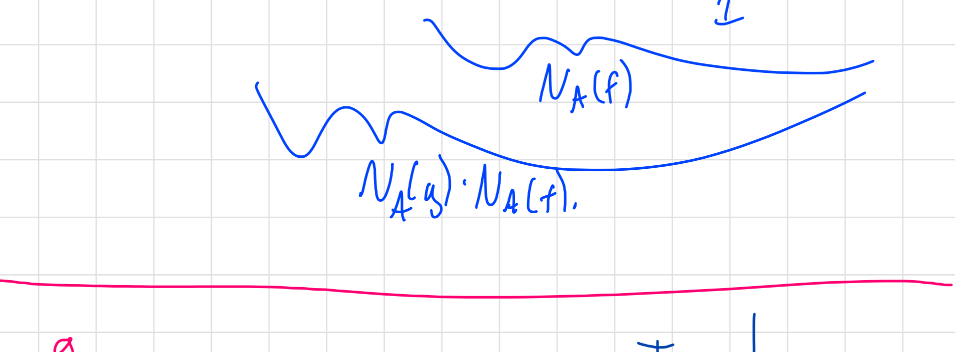
Let $P_A(g)$ denote the corresponding partition. For any $U \in P_A(g)$, $g(U)$ is an interval, thus $N_A(f_{|g(U)}) \leq N_A(f)$.

Given any $T \in N_A(f_{|g(U)})$,

then $f_{|g(U)}$ is affine in T,

and g is affine in $U \cap g^{-1}(T)$.

$$N_A(f\circ g) \leq \sum_{U\in P_A(G)} N_A((f\circ g)_{|U}) \leq \underbrace{\sum_{U\in P_A(G)} \underbrace{\sum_{T\in P_A(f_{|g(U)})} \underbrace{N_A((f\circ g)_{U\cap g^{-1}(T)})}_{1}}_{N_A(f)}}_{N_A(g)\cdot N_A(f)}.$$



Remark. For images of $P_A(g)$ to hit all pieces of $P_A(f)$ is delicate; $\Delta\circ\Delta$ approximately meets this bound.

To prove
$$N_A(g) \leq 2^i \cdot \prod_{j<i} m_j$$

$$N_A(g) = N_A\left(x \mapsto \sigma\left(\sum_{i=1}^{m_{i-1}} a_i h_i(x) + b\right)\right)$$
(using ④)
$$\leq 2\cdot \sum_{i=1}^{m_{i-1}} N_A(h_i)$$
$$\leq 2\cdot m_{i-1} 2^{i-1}\prod_{j<i-1} m_j = 2^i \prod_{j<i} m_j.$$

To prove full theorem