Lec 9

Ann.

* Homework due tomorrow
* Maybe hybrid Thursday

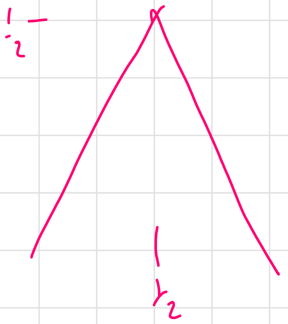
Recap

* Today is final apx lecture

$$\inf_{f \in \mathcal{F}} R(f) \quad \text{vs} \quad \inf_{g \in \text{anything else}} R(g)$$

* Various apx of continuous function (2-hidden & 1-hidden)
* Fourier apx (connections to ntk, non-worst-case estimates.)
* NTK (near initialization, large width network ≈ first order Taylor)

* Depth helps



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

$$N_A\left(\text{ReLU network with } \le L \text{ layers}, \le m \text{ nodes}\right) \le \left(\frac{2m}{L}\right)^L$$

$$\left[2^L \prod_{j \le L} m_j \right]$$

Approximating x^2.

* Why?

"polarization identity"

$$xy = \frac{1}{2}\left((x+y)^2 - x^2 - y^2\right).$$

"smallish" deep networks

efficient $\Delta^2 \Rightarrow$ eff. x^2

$\Rightarrow$ products $\Rightarrow$ multivariate polynomials $\Rightarrow$ Taylor expansions

$\Rightarrow$ "Sobolev space"

$$x^2 = \int_0^1 2b\, \mathbb{1}[x \geq b]\, db$$

$$= \int_0^1 2\, \sigma_r(x-b)\, db.$$

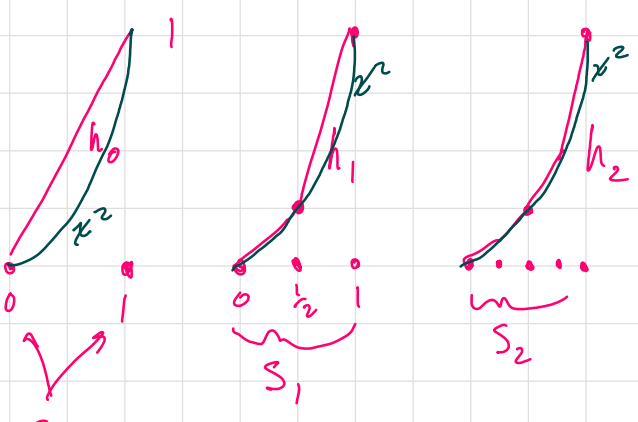
constant; hint can be efficiently approx'd.

Homework: $\left\lceil \frac{2}{\sqrt{\varepsilon}} \right\rceil$ width, depth: $O(\ln 1/\varepsilon)$.

(figured out by Dmitry Yarotsky)

$S_i := \left\{ \frac{0}{2^i}, \frac{1}{2^i}, \dots, \frac{2^i}{2^i} \right\}$ grid $[0,1]$ at scale $1/2^i$

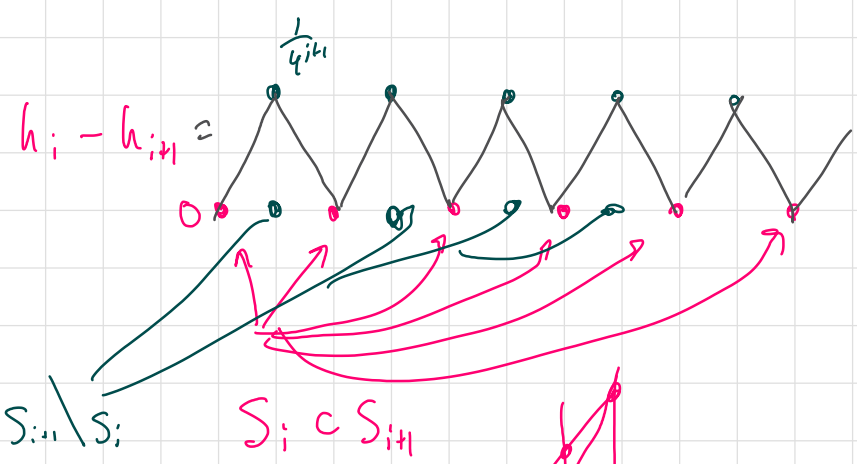
$h_i :=$ secant approximation of x^2 at points in S_i:



"h_i is affine interpolation of x^2 along S_i"

$N_A(h_i) = 2^i$, seems need 2^i ReLUs

$h_i - h_{i+1} =$



$S_i \subset S_{i+1}$

$x \in S_{i+1}$ $\quad \varepsilon := \frac{1}{2^{i+1}}$

$$h_i(x+\varepsilon) - h_{i+1}(x+\varepsilon) = \frac{1}{2}\left(x^2 + (x+2\varepsilon)^2\right) - (x+\varepsilon)^2$$

$$= \frac{1}{2}\left(x^2 + x^2 + 4x\varepsilon + 4\varepsilon^2\right) - \left(x^2 + 2x\varepsilon + \varepsilon^2\right)$$

$$= \varepsilon^2.$$

By induction

$$h_i(x) = h_0(x) + \sum_{j<i} \left(h_{j+1}(x) - h_j(x)\right)$$

$$= x - \sum_{j=1}^{i} \frac{\Delta^j(x)}{4^j}.$$

Theorem (Yarotsky).

Let h_i be as above.

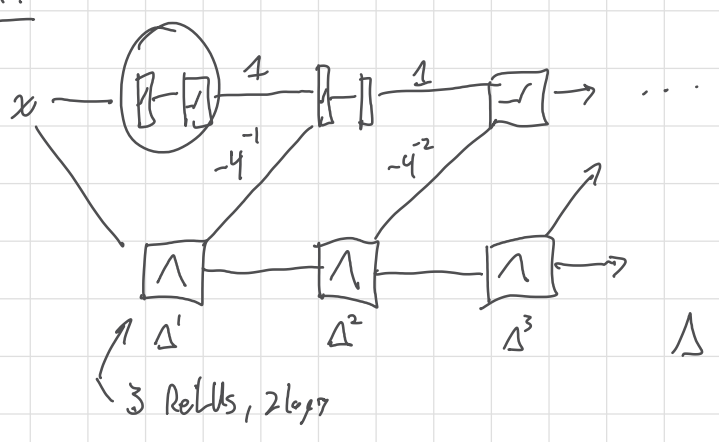
① h_i can be written with $\leq 4i$ ReLUs in $\leq 2i$ layers

② $\sup_{x \in [0,1]} |h_i(x) - x^2| \leq \frac{1}{4^{i+1}}$.

③ Any ReLU network $g: \mathbb{R} \to \mathbb{R}$ with $\leq L$ layers, $\leq M$ nodes

satisfies $\int_0^1 (g(x) - x^2)^2\, dx \geq \frac{1}{5760}\left(\frac{2M}{L}\right)^{4L} \geq N_\Delta(g)$

Proof.

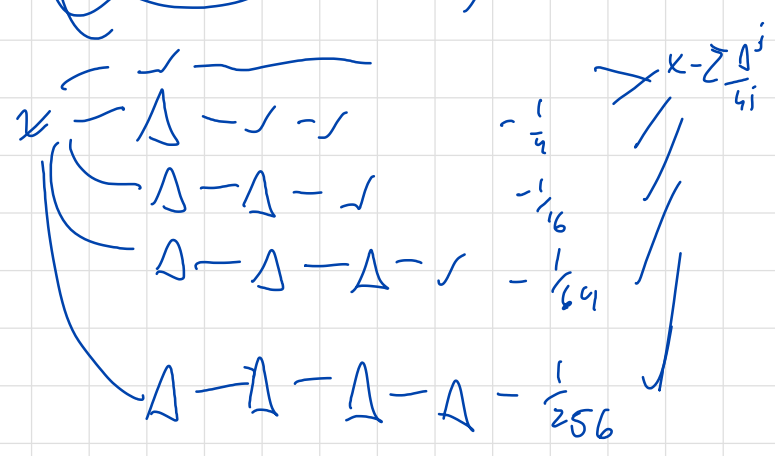
① 

5i ReLU, 2i layers.

$$h_i(x) = x - \sum_{j=1}^{i} \frac{\Delta^j(x)}{4^j}$$

$O(i)$

$O(i^2)$



tooth. max. Relu(tooth. max. Relu(x))

② Consider any $z \in [0,1]$, write $z = x + \delta$ ($x \in S_i$, $\delta \in [0, \varepsilon]$, $\varepsilon \in [0, \frac{1}{2^i}]$)

$$h_i(x+\delta) - (x+\delta)^2 = \frac{\varepsilon - \delta}{\varepsilon} x^2 + \frac{\delta}{\varepsilon}(x+\varepsilon)^2 - (x+\delta)^2$$

$$= x^2 - \frac{\delta}{\varepsilon} x^2 + \frac{\delta}{\varepsilon}\left(\varepsilon^2 + 2x\varepsilon + x^2\right) - \left(x^2 + 2\delta x + \delta^2\right)$$

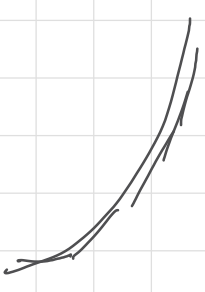
$$= x^2\left(1 - \frac{\delta}{\varepsilon} + \frac{\delta}{\varepsilon} - 1\right) + x\left(2\delta - 2\delta\right) + \left(\delta\varepsilon - \delta^2\right)$$

$$= \delta(\varepsilon - \delta).$$

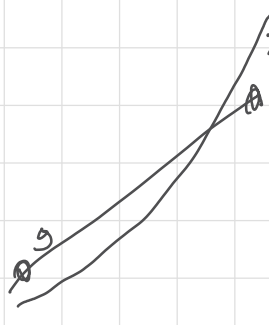
$$\leq \sup_{\delta \in [0,\varepsilon]} \delta(\varepsilon - \delta) = \frac{\varepsilon}{2}\left(\varepsilon - \frac{\varepsilon}{2}\right)$$

$$= \frac{\varepsilon^2}{4} = \frac{1}{4^{i+1}}.$$

③ $N_\Delta(g) \leq \left(\frac{2M}{L}\right)^L$



Constant fraction of affine pieces have length $\frac{1}{2N_\Delta(g)}$. Within each piece:



Many long intervals with large error

$\Rightarrow$ large error over $[0,1]$. //

Approximating functions with "nice Taylor expansions"

(Sobolev space)

Squaring $\Rightarrow$ products

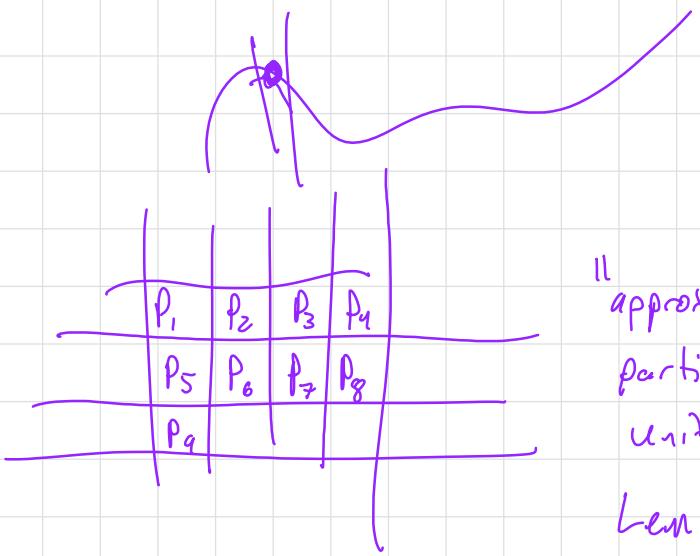
$$xy = \frac{1}{2}\left((x+y)^2 - x^2 - y^2\right)$$

$$\frac{1}{2}\left(h_i(x+y) - h_i(x) - h_i(y)\right)$$

$$4h_i\left(\frac{x+y}{2}\right)$$

$xy \to \prod_{i \in S} x_i$ Monomials

$\to$ Polynomial



"approximate partition of unity"

Lemma 5.5

Office hours

defn "rotational invariance"

M rot matrix

$$M^T M = I$$

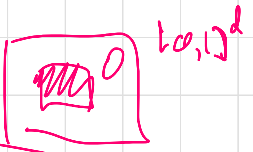
$$Mx \overset{d.}{=} x$$

↑ "equal in dist"

$$\mathbb{E} f(x) = \mathbb{E} f(Mx)$$

$$\int \ell(y f(x)) \, d\rho(x,y)$$

$$\int_{[0,1]^d} \ell(f(x)) \, h(x) \, dx$$



density

$$h \geq 0 \qquad \int_{[0,1]^d} h = 1$$

$$\int \ell(y f(x)) \, d\rho(x,y) = \mathbb{E}_{x,y} \, \ell(y(f(x)))$$