Lecture 10: Depth Separation

Recap from last week:

→ Optimizing shallow nets in feature learning regime: positive guarantee, BUT in quantitative

→ Learning lower Bounds under statistical/"gradient" queries: exponential number of orthogonal possible target functions.

Today: Another source of learning hardness: poor approximation using shallow NN

↳ Better approximation using deeper networks?

Depth Separation: Univariate Case

(Motivating question): Let $f: \mathbb{R} \to \mathbb{R}$. All approximation schemes we have seen so far (Fourier, Kernels, Shallow NN) are additive:

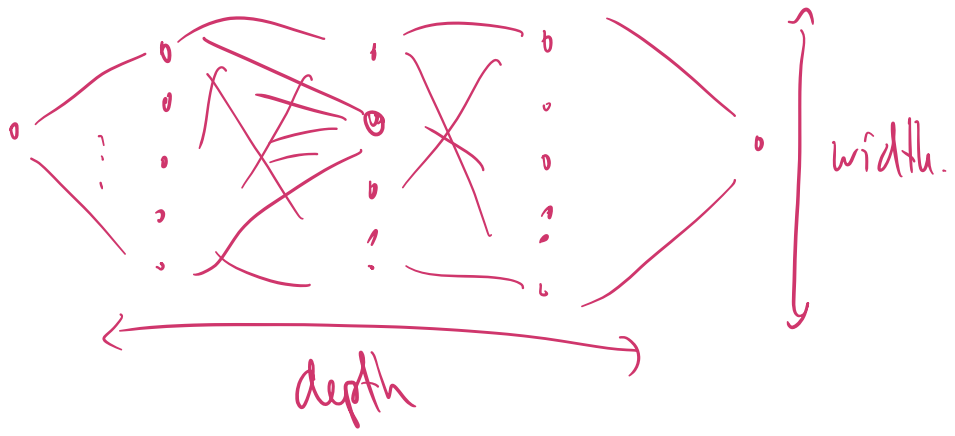
$$f \approx \sum_i \phi_i(x)$$

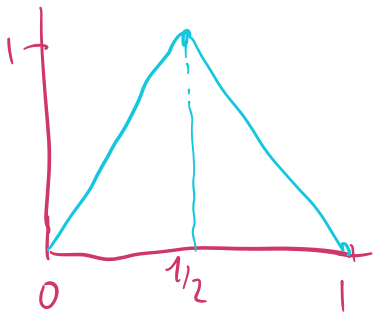
↳ This is in contrast to Approximation by composition

$$f \approx \phi_L \circ \dots \circ \phi_2 \circ \phi_1 \qquad \phi_\ell : \mathbb{R} \to \mathbb{R}$$

Q: Which functions can be easily approximated by composition, yet poorly approximated by addition?

In the context of NNs, we will quantify these options in terms of # of neurons.

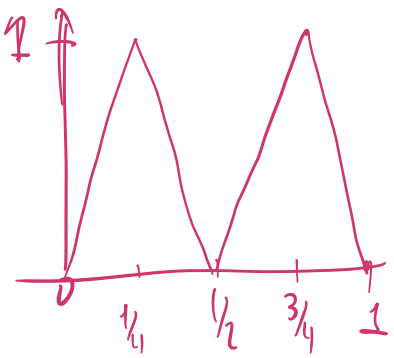




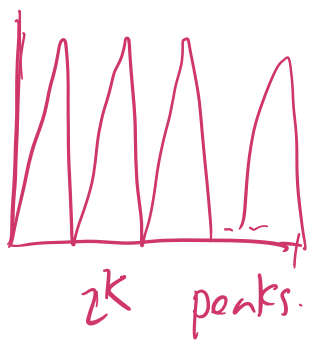
$$\Delta : [0,1] \to [0,1]$$

$$x \longmapsto \begin{cases} 2x & \text{if } x \in [0, 1/2) \\ 2-2x & \text{if } x \in [1/2, 1) \end{cases}$$

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



$$\Delta^k = \underbrace{\Delta \circ \Delta \circ \cdots \circ \Delta}_{k \text{ times}}$$



$f = \Delta^k$

Theorem [Telgarsky '15] : Fix $L \geq 2$, and let $f = \Delta^M$ with $M = L^2 + 2$. Then

(i) f can be expressed as a ReLU net with $3M$ neurons and $2M$ layers. (positive approximation)

(ii) f cannot be approximated using shallow nets: any Relu network g with $\leq 2^L$ neurons and $\leq L$ layers incurs in constant L_1 error: $\|f - g\|_1 \geq \frac{1}{32}$.
(negative approx).

Proof:

(i) Track the number of affine regions, $N_A(f)$, for each ReLU network.

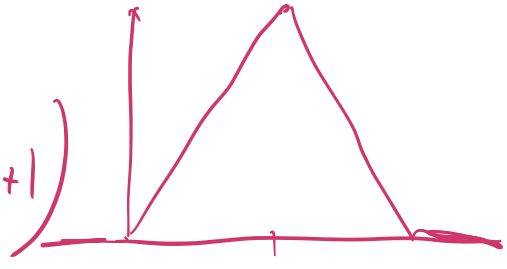
Lemma: Let $f: \mathbb{R} \to \mathbb{R}$ be a Relu net with L layers and widths $(m_1 \dots m_2)$, with $m = \sum_i m_i$. Then

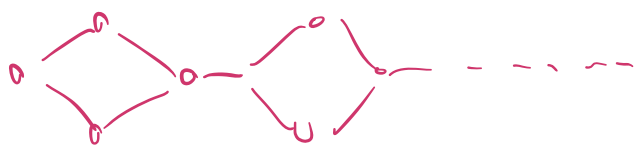
$$N_A(f) \leq \left(\frac{2m}{L}\right)^L$$

- Proof of this Lemma (Telgarski '15): uses induction over depth.
- $N_A(f+g) \leq N_A(f) + N_A(g)$
- $N_A(f \circ g) \leq N_A(f)\, N_A(g)$ ← This is tight for $f = g = \Delta$.

(ii) Observe that $\Delta(x)$ can be implemented with a simple ReLU net: $\sigma(t) = \max(0, t)$

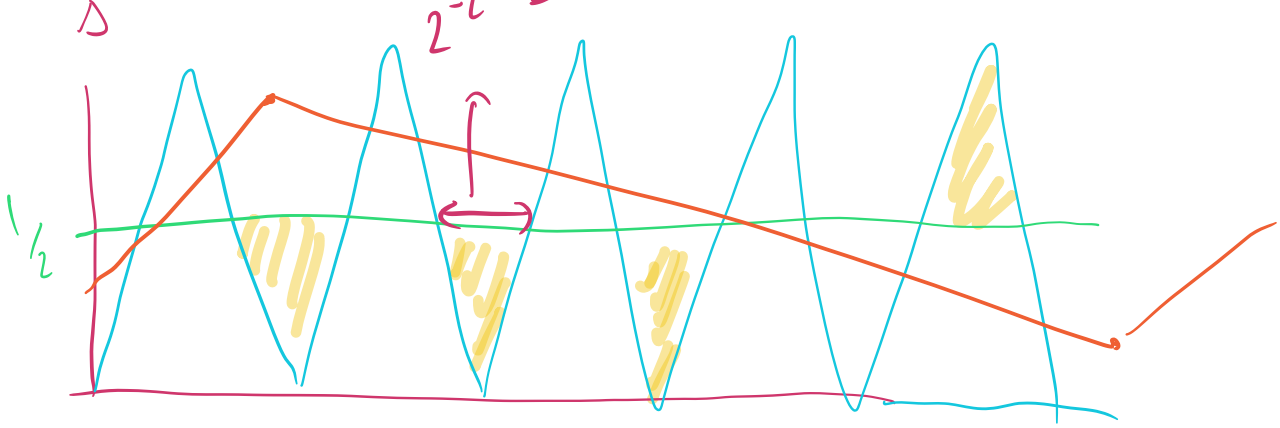
$$\Delta(x) = \sigma\left(-2\sigma\left(x - \tfrac{1}{2}\right) - 2\sigma\left(-x + \tfrac{1}{2}\right) + 1\right)$$





Δ^M.

(iii)



↳ we have 2^{L^2+1} copies of $\Delta \rightsquigarrow 2^{L^2+2} - 1$ half-triangles. Each halftriangle has area $\frac{1}{4} 2^{-L^2-2} = 2^{-L^2-4}$

↳ g has $N_A(g) \le \left(2 \cdot \frac{2^L}{L}\right)^L \le 2^{L^2}$ linear regions.

$$\int_0^1 |f - g| \, dx \ge (\#\text{ untouched triangles}) \cdot (\text{triangle area})$$

$$\ge \frac{1}{2}\left(\text{Total triangles} - 2 \cdot N_A(g)\right] \cdot 2^{-L^2-4}$$

$$\ge \frac{1}{2}\left(2^{L^2+2} - 1 - 2 \cdot 2^{L^2}\right) \cdot 2^{-L^2-4} \ge 2^{-5} \quad \square$$

Remarks:
- → Separates $\sim L^2$ depth from $\sim L$ depth.
- → Very tailored to ReLu piece-wise linear structures ⇝ NOT too "intrinsic".

Q: High-dimensional perspective?

Depth Separation in the High-Dimensional Regime

→ From UAT, shallow is sufficient in a qualitative sense.
Now we will be interested in finite guarantees.

→ Focus on simplest instance of Depth separation: shallow vs (1 hidden layer) 2 hidden layer.

⇒ Problem setup: Let $\Omega \subseteq \mathbb{R}^d$ input domain; Let ν be a probability distribution (data) on Ω.

Let $$\mathcal{F}_2(m) := \left\{ f: \Omega \to \mathbb{R} \;;\; f(x) = \sum_{j=1}^{m} \phi_j(\langle x, w_j \rangle) \right.$$
$$\left. \phi_j \text{ is Lipschitz } \right\}$$
$\phi_j : \mathbb{R} \to \mathbb{R}$

(ex $\phi_j(t) = a_j \cdot \sigma(t - b_j)$ recovers 2NN "classic".

$$\mathcal{F}_3(m) = \left\{ f: \text{3 layer NN of widths } m_1, m_2 \right\}, \quad m = m_1 + m_2$$

Fix $\varepsilon > 0$, and consider a target function $f^\circ : \Omega \to \mathbb{R}$

The rate of approximation of f° on each $\mathcal{F}_2(m)$ $\mathcal{F}_3(m)$

$$m_2(\varepsilon, d) = \inf \left\{ m \;;\; \inf_{f \in \mathcal{F}_2(m)} \|f^\circ - f\|_\nu^2 \le \varepsilon \right\}$$

↳ In words, width required to approximate f° to error ε in dimension d.

↳ $\mathbb{E}_{x \sim \nu} (f^\circ(x) - f(x))^2$

Punch line: We can find f° such that

$m_3(\varepsilon, d) = \text{poly}(d, 1/\varepsilon)$, but

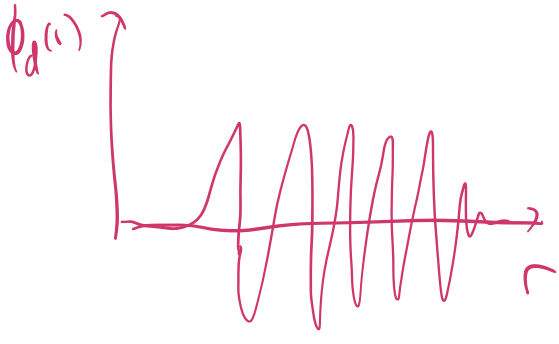
$m_2(\varepsilon, d) = \exp(d)$ (for any fixed ε)

Theorem [Eldan & Shamir '16] Suppose σ activation function satisfies mild growth conditions and $\neq$ polynomial. Then for each $d \in \mathbb{N}$, there exists probability measure ν in $\mathbb{R}^d$ and $f^\circ : \mathbb{R}^d \to \mathbb{R}$ such that $m_3(\varepsilon, d) = \text{poly}(d, \varepsilon^{-1})$, but f° cannot be ε-approximated for $\varepsilon = \Omega(1)$ by any shallow NN of width $\exp(o(d))$.

Q: What is the nature of f°?

f° is chosen as a radial function:

$f^\circ(x) = \phi_d(\|x\|)$, for $\phi_d : \mathbb{R} \to \mathbb{R}$, and when ϕ_d will oscillate, $\text{Lip}(\phi_d) = \text{poly}(d)$



$\rightarrow$ $f(x) = \phi_d(\|x\|)$

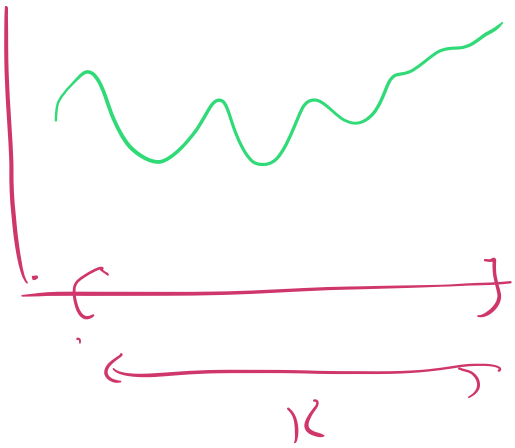
(-) positive result: $\left(F_1 : \begin{matrix} \mathbb{R}^d \to \mathbb{R} \\ x \mapsto \|x\|^2 \end{matrix}\right)$ $\|x\|^2 = \sum_{i=1}^{d} x_i^2$

$\left(F_2 : \begin{matrix} t \mapsto \phi_d(\sqrt{t}) \\ \mathbb{R} \to \mathbb{R} \end{matrix}\right)$ $\leftarrow$ $f^\circ = F_2 \circ F_1$

↳ Idea: "spend" 1st hidden layer to approximate F_1.

"spend" 2nd hidden layer to approximate F_2.

↳ Given a compact domain $K \subset \mathbb{R}^n$ and $F: K \to \mathbb{R}$ with $\mathrm{Lip}(F) \le \beta$, we

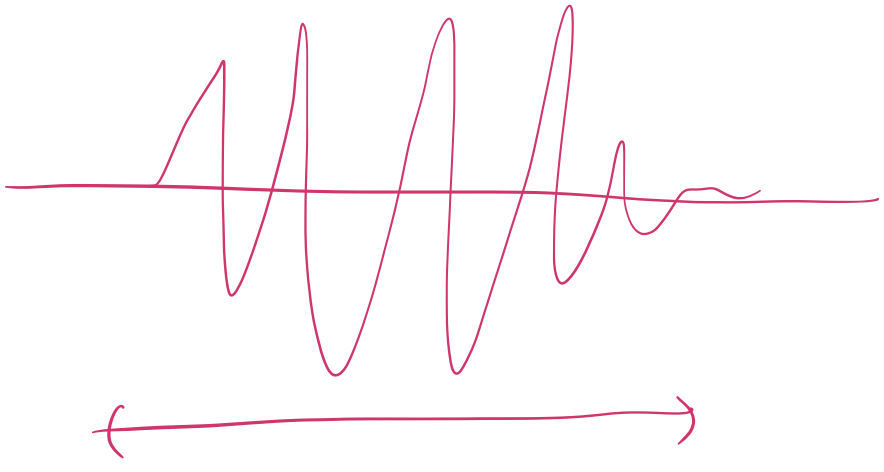


can approximate F as

$$\tilde{F}(t) = \sum_{i=1}^{N} a_i \sigma(t - t_i)$$

$|K|, \beta, \varepsilon^{-1}$ poly(d)

(ii)



$$\mathbb{E}_{x \sim \nu} |F_1(x) - \tilde{F}_1(x)|^2 \le \varepsilon$$

(ii) negative result: Suppose data distribution ν admits a density with respect to Lebesgue: $\nu(dx) = \varphi^2(x)\,dx$. (↳ since density is nonneg.) Then

$$\|f - g\|_\nu^2 = \int |f(x) - g(x)|^2 \varphi^2(x)\,dx = \int (f\varphi - g\varphi)^2 dx = \|f\varphi - g\varphi\|_2^2$$

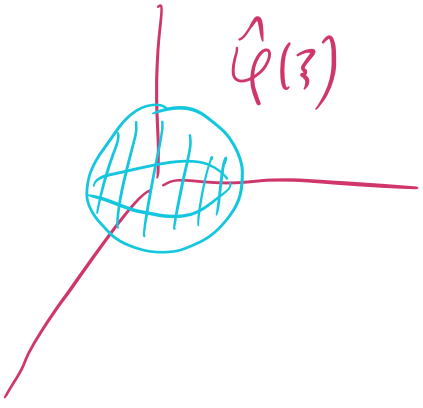
- By Parseval identity, we can express this norm in Fourier:

$$\|f\varphi - g\varphi\|_2^2 = \|\widehat{f\varphi} - \widehat{g\varphi}\|_2^2$$

- Recall $\widehat{f\varphi} = \hat{f} * \hat{\varphi}$. We construct data distribution $\nu(x) = \varphi^2(x)\,dx$ such that $\hat{\varphi}$ is well localised in freq;
$\hat{\varphi}(\xi) = \underbrace{\mathbb{1}\{\|\xi\| \le 1\}}_{B}$ (indicator of unit ball in d-dimension)

↳ we verify that $\varphi^2(x)$ is indeed a density. $\varphi^2(x) \ge 0$

$$\int \varphi^2(x)\,dx = \int |\hat{\varphi}|^2(\xi)\,d\xi = 1.$$



→ Now we compute fourier transform of

$$f(x) = \sum_{j=1}^{m} \phi_j(\langle x, w_j \rangle).$$

"ridge" function.

↳ Focus on a single term $f_0(x) = \phi(\langle x, w \rangle)$

$$\hat{f}_0(\xi) = \int f_0(x)\, e^{-i\langle x, \xi \rangle}\,dx$$

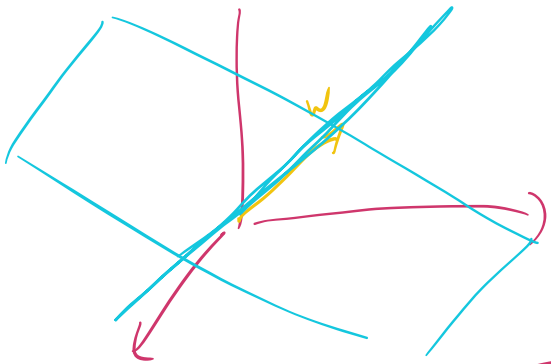
; if is well-defined in the sense of tempered distributions, thanks to polynomial growth assumptions on ϕ.

$$\hat{f}_0(\xi) = \int \phi(\langle x, w \rangle)\, e^{-i\langle x, \xi \rangle}\,dx$$

change of variables $w \to \frac{w}{\|w\|}$

$$= \int \phi(x_1)\, e^{-i x_1 \xi_1} \cdot e^{-i\langle x_{-1}, \xi_{-1}\rangle}\, dx_1 \cdot dx_{-1}$$

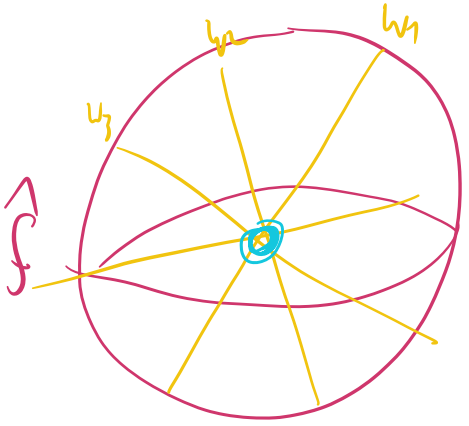
$$= \underbrace{\left(\int_{\mathbb{R}} \phi(x_1) e^{-i x_1 \xi_1}\, dx_1\right)}_{\text{univariate F. Transform of } \phi \;\; \hat{\phi}(\xi_1)} \cdot \underbrace{\left(\int_{\mathbb{R}^{d-1}} e^{-i\langle x_{-1}, \xi_{-1}\rangle}\, dx_{-1}\right)}_{\text{F Transform of constant } f \sim \delta(\xi_{-1})}$$



The support of $\hat{f}_0$ is the ray span $\{w\}$

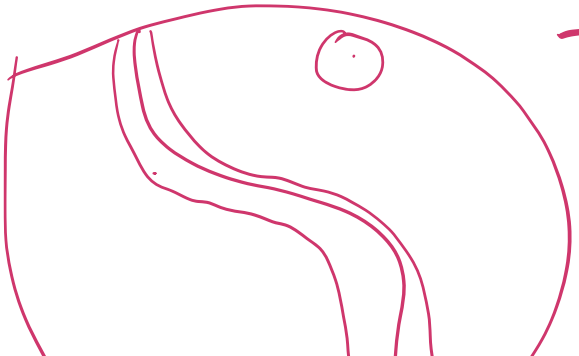
(by linearity of F Transform)

→ The support of the Fourier transform of $f(x) = \sum_{j=1}^{m} \phi_j(\langle x, w_j\rangle)$ is the union of rays $\bigcup_{j=1}^{m} \text{span}\{w_j\}$

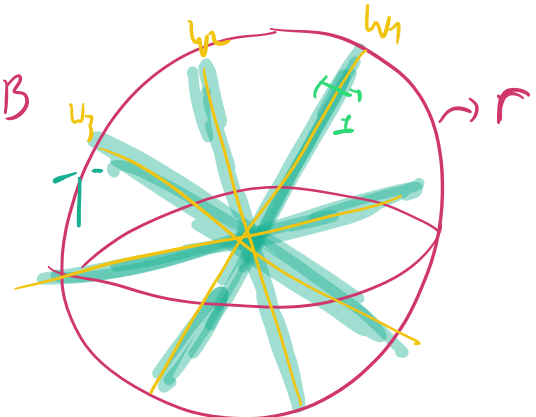


→ The support of $\hat{f} * \hat{\varphi}$?

Recall that $\hat{\varphi} = \mathbb{1}(\xi \in B)$



$$T = \bigcup_{j=1}^{m} \text{span}\{w_j\} + B$$



→ What fraction of the surface of the d-dim sphere of radius r intersects T?

$$\frac{\mathrm{Vol}_{d-1}(T \cap r \cdot S^{d-1})}{\mathrm{Vol}_{d-1}(r S^{d-1})} \leq m \cdot e^{-d}$$

when r is large wrt radius of B.

→ Can we find g (target) such that $\hat{g} * \varphi$ has energy "spread" out in $r S^{d-1}$?

→ A: Radial Function! g such that

$$g(\theta x) = g(x) \quad \forall x \quad \forall \theta^T \theta = Id.$$

$$\hat{g}(\theta \xi) = \int g(x) e^{-i\langle x, \theta \xi \rangle} dx$$

$$= \int g(x) e^{-i\langle \theta^T x, \xi \rangle} dx$$

$$= \int g(\theta x) e^{-i\langle x, \xi \rangle} dx$$

$$= \hat{g}(\xi)$$

↳ So we need to construct g such that:

(i) g is radial
$g(x) = \phi_d(\|x\|)$ $\left(\Rightarrow g\varphi \text{ radial} \Rightarrow \hat{g}+\hat{\varphi} \text{ radial}\right)$

(ii) ϕ_d to <u>oscillate</u> . ; energy in the high-frequencies.

<u>Remarks</u>

→ The data distribution is heavy-tailed ; what happens when we replace it for ν more concentrated; $\nu = \text{Unif}(S^{d-1})$
[Daniely '17] gets also exponential lower bounds (but conditioned on weights not too large).

→ Extension beyond radial functions [VBJO '22] using similar proof technique.