Lecture 3: Shallow Neural Networks and Approximation

- Last week we talked about two basic notions of approximation:

① Linear Approximation: Identifying a single finite-dim subspace (or a nested collection) that approximates uniformly well a whole class of inputs.

② Non-linear Approx: we allow the finite-dim model to adapt to the input.

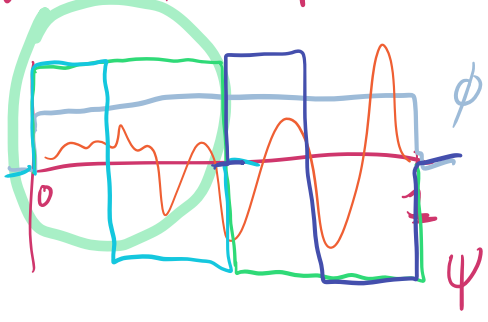
Example: . Fourier representation. (linear).

$$f(t) = \sum_{k=0}^{\infty} a_k e^{i 2\pi k \cdot t}$$

$$\downarrow$$

$$f_N(t) = \sum_{k \le N} a_k e^{i 2\pi k t}.$$

- Wavelet representations. ; Haar wavelets (1920s)

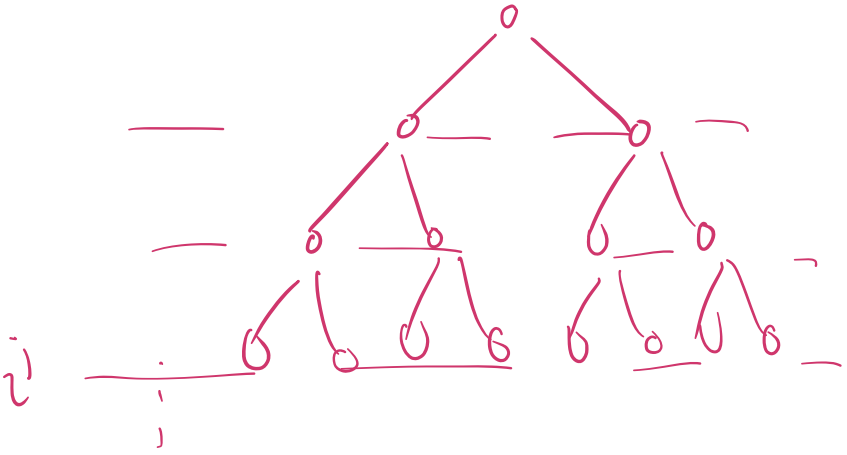


$$\int \phi(t)\,\psi(t)\,dt = 0$$

$$\|\phi\| = \|\psi\| = 1$$

$$\Omega = \{\phi, \psi, \psi, \psi, \ldots, \psi, \ldots\}$$

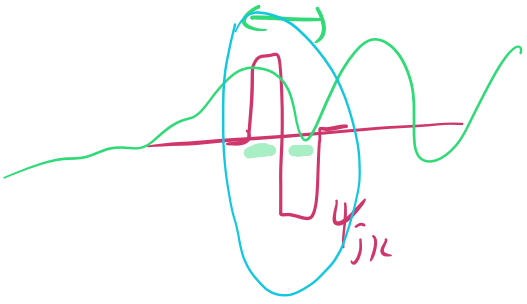
$$\psi_{j,k}(t) = 2^{-j/2}\,\psi_0\left(2^{-j}t - 2^{-j}k\right), \quad k = 0, \dots, 2^{-j}-1$$

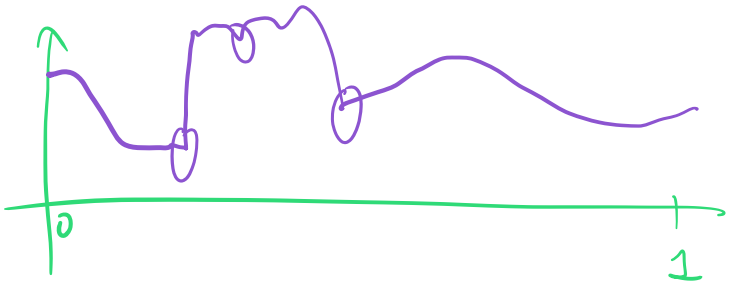


Fact: $\mathcal{D}$ is an orthonormal basis of $L^2([0,1])$.

Q: What are the functions that can be well approximated by few (Haar) wavelet coefficients?

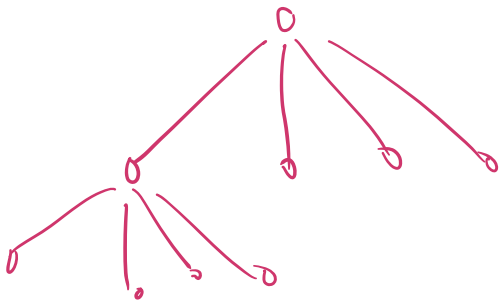
$$(f, \psi_{j,k}) \approx 0$$

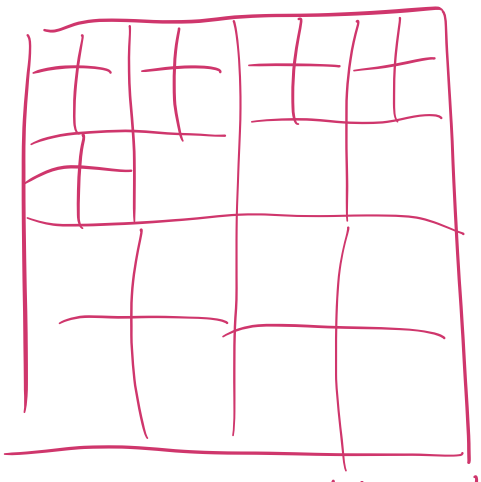




$$\int f(t)\,\psi_{j,k}(t)\,dt = 2^{-j/2}\left[\int_{2^{-j}k}^{2^{-j-1}(k+1)} f(t)\,dt - \int_{2^{-j-1}(k+1)}^{2^{-j}(k+2)} f(t)\,dt\right]$$

→ $f(u,v)$ $(u,v) \in [0,1]^2$





→ In general, in high-dimensions, this wavelet construction increases exponentially in the input dimension!!

→ Alternative?

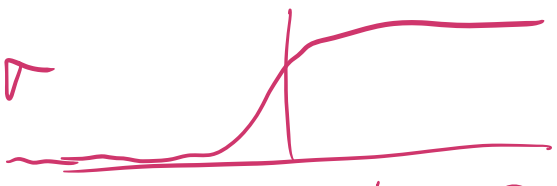
→ Go from "grid"-like dictionaries to "off-the-grid" representations!

Shallow Neural Networks as off-the-grid representations

Consider $\sigma : \mathbb{R} \to \mathbb{R}$ Lipschitz activation function.

$$\phi(x; \theta) = \sigma(\langle x, w \rangle + b) \qquad \theta = (w, b) \in \mathbb{R}^{d+1}$$

- σ: activation.
- $\langle x, w \rangle + b$: affine function from $\mathbb{R}^d \to \mathbb{R}$
- ϕ: Ridge function

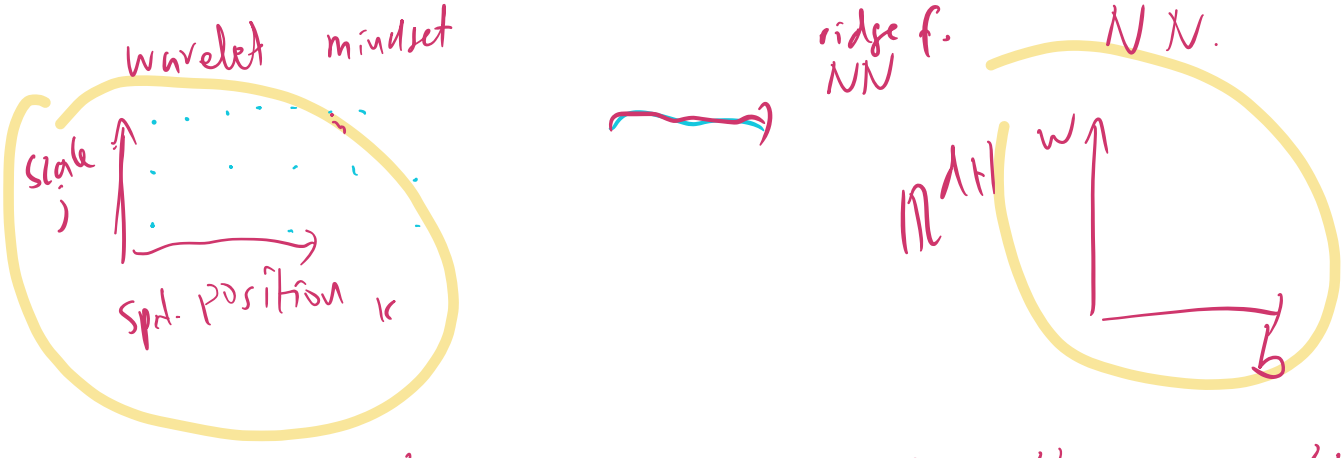


$\phi : \mathbb{R}^d \to \mathbb{R}$

$\rightarrow f_K(x;\theta) = \sum_{k=1}^{K} \phi(x;\theta_k)$ linear combination of ridge functions.

$\theta = \{\theta_k\}_k$.

$$H_\sigma = \{ f_K ; K \in \mathbb{N} \}$$: space of functions represented as shallow



Q: How expressive is the set H_σ?

Universal Approximation Theorems

→ We want to understand whether H_σ can approximate arbitrarily well continuous functions $f: \mathbb{R}^d \to \mathbb{R}$.

For a given error metric d, $\forall f \in C(\mathbb{R}^d), \forall \varepsilon > 0$ $\exists \tilde{f} \in H_\sigma$ with $d(f, \tilde{f}) \leq \varepsilon$.

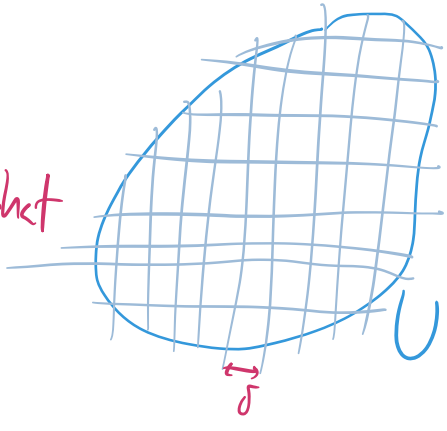
Let's first establish UAT using 3-layer NN.

Theorem: let $\sigma : \mathbb{R}^d \to \mathbb{R}$ continuous, $\varepsilon > 0$.

Theorem: Let $g : [0,1]^d \to \mathbb{R}$ continuous, Let $\varepsilon > 0$ so that $\forall x, x' \in [0,1]^d$, $\|x - x'\|_\infty \leq \delta$ then $|g(x) - g(x')| \leq \varepsilon$. Then there exists a 3-layer neural Network f with ReLu activation and $\Omega(\delta^{-d})$ such that $\|f - g\|_1 \leq 2\varepsilon$.

Proof: [Telgarski '20]

Lemma: Let any $U \subset \mathbb{R}^d$ bounded, along with a partition P of U into rectangles of diameter δ $P = \{R_1 \dots R_N\}$. Then there exists scalars $\alpha_1 \dots \alpha_N$ such that

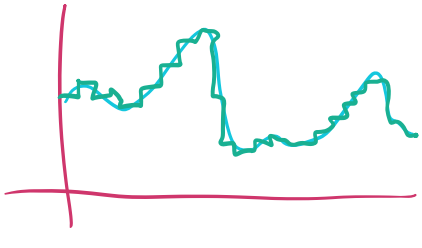


$$\sup_{x \in U} |g(x) - h(x)| \leq \varepsilon,$$

$$h(x) = \sum_i \alpha_i \mathbb{1}_{R_i}(x).$$

in other words, we are building a piece-wise constant approximation of a continuous function.

Proof of lemma: exercise.



Let P be a partition of $[0, 1+c)^d$ into rectangles $\prod_j [a_j, b_j]$ with diameter $\leq \delta$. $b_j - a_j \leq \delta$.

→ Let $h = \sum_i \alpha_i \mathbb{1}_{R_i}$ from Lemma, so we know

$\| g - h \|_1 \leq \varepsilon$.

→ We will build $f(x) = \sum_i \alpha_i g_i(x)$

where each g_i is a ReLu net that approximates $\mathbb{1}_{R_i}$ with $O(d)$ nodes.

Since $|P| = \Theta(\delta^{-d})$, then

$O(d) \cdot \Theta(\delta^{-d}) = \Omega(\delta^{-d})$.

· $\| g - f \|_1 \leq \| f - h \|_1 + \| h - g \|_1$

$\leq \| \sum_i \alpha_i (\mathbb{1}_{R_i} - g_i) \|_1 + \varepsilon$

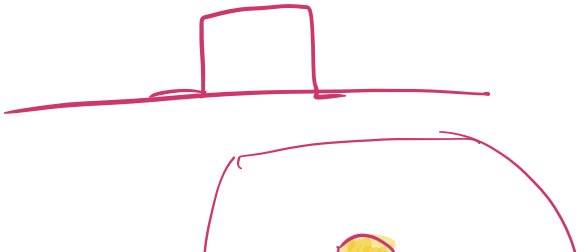
$\leq \sum_i |\alpha_i| \, \| \mathbb{1}_{R_i} - g_i \|_1 + \varepsilon$

→ If we can build g_i so that

$\| \mathbb{1}_{R_i} - g_i \|_1 \leq \frac{\varepsilon}{\|\alpha\|_1}$, then

$\| f - g \|_1 \leq 2\varepsilon$.

· Fix i, and let

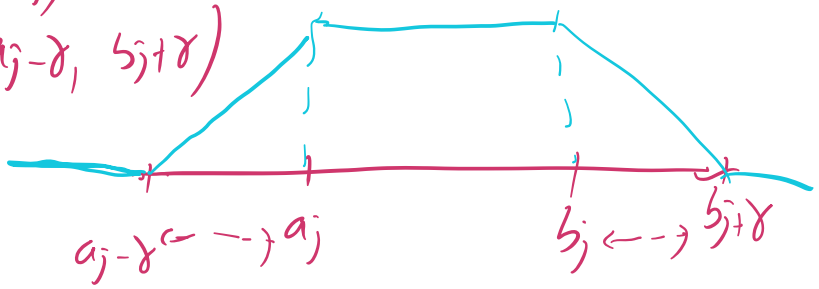


$$R_i = \bigotimes_{j=1}^{d} [a_{i,j}, b_{i,j})$$

For $\gamma > 0$, for each $j \in \{1..d\}$ we define

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

$$g_{j,\gamma}(t) = \begin{cases} 1 & \text{if } t \in [a_j, b_j) \\ 0 & \text{if } t \notin [a_j-\gamma, b_j+\gamma) \\ [0,1] & \text{otherwise.} \end{cases}$$

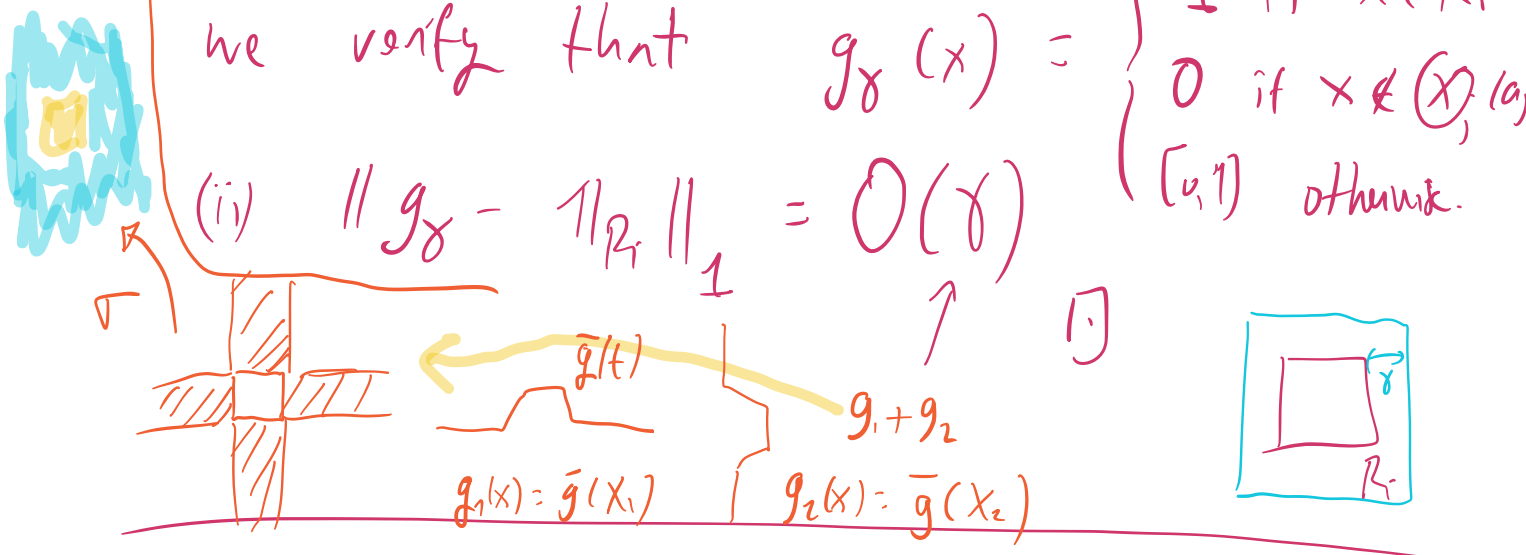


$$g_\gamma(x) = \sigma\left(\frac{1}{d}\sum_{j=1}^{d} g_{\gamma_{ij}}(x_j) - (d-1)\right)$$

chat!

we verify that $g_\gamma(x) = \begin{cases} 1 & \text{if } x \in R_i \\ 0 & \text{if } x \notin \bigotimes_j (a_j-\gamma, b_j+\gamma) \\ [0,1] & \text{otherwise.} \end{cases}$

(ii) $\| g_\gamma - 1_{R_i} \|_1 = O(\gamma)$



Q: Are there more powerful characterizations of universal approximation?

Thm: [Stone–Weierstrass] Let $\Omega = [0,1]^d$. Let $\mathcal{F}$ function class with:

(i) each $f \in \mathcal{F}$ is continuous.

(ii) $\forall x \in \Omega$, $\exists f \in \mathcal{F}$ with $f(x) \neq 0$.

(iii) $\forall x \neq x'$, $\exists f \in \mathcal{F}$ with $f(x) \neq f(x')$

(iv) $\mathcal{F}$ is closed under multiplication and a vector space operations (ie $\mathcal{F}$ is an algebra).

Then $\mathcal{F}$ is a universal approximator: For any $g : \mathbb{R}^d \to \mathbb{R}$ continuous and $\varepsilon > 0$, $\exists f \in \mathcal{F}$ with

$$\sup_{x \in \Omega} |f(x) - g(x)| \leq \varepsilon.$$

- σ a "sigmoid" $\lim_{t \to -\infty} \sigma(t) = 0$, $\lim_{t \to +\infty} \sigma(t) = 1$ [Hornik, Stinchombe, White '89]
- $\sigma \neq$ polynomial [Leshno '93]

Both results use S-W.

Q: Is this UAT "surprising" / different from last week?

The Fourier Perspective

→ Let $f \in C(\mathbb{R}^d)$ and consider its restriction

in a compact set Ω. ($\Omega = [0,1]^q$).

Fourier decomposition: $\forall \xi \in \mathbb{Z}^d$, we consider

$$\hat{f}(\xi) = \left\langle f, e^{i 2\pi \langle x, \xi \rangle} \right\rangle_{L^2(\Omega)} =$$

$$= \int_\Omega f(x) e^{-i 2\pi \langle x, \xi \rangle} dx.$$

$f \simeq \sum \alpha_i \mathbb{1}_{R_i}$ ← Before.

$f \simeq \sum \alpha_i \sigma(\langle x, w_i \rangle)$ ← now

Fourier Inversion lemma: If we consider the partial sums

$$f_{M(R)}(x) := \sum_{\xi;\ \|\xi\| \le R} \hat{f}(\xi) e^{2\pi i \langle x, \xi \rangle}$$

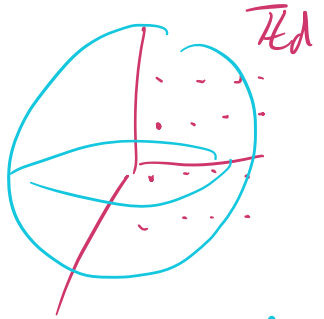
$R \to \infty \downarrow$

f

$\sigma(\langle x, \xi \rangle)$ with $\sigma(t) = e^{i\pi t}$

M: number of integer points in $\{\|x\| \le R\}$.

$\sigma : \mathbb{R} \to \mathbb{C}$



This Fourier Inversion $\hookrightarrow$ UAT.

Moreover, we have tight control over approximation rates.

Smoothness of f $\hookleftarrow$ decay of $\hat{f}(\xi)$.

$f \in H^s \Rightarrow \|f - f_M\| = O(M^{-s/d})$

Sobolev class with s derivatives

with s dimensions

↳ curse of dimensionality.

→ Q: How to overcome this curse?

A: look for another class of functions with "sparse" fourier decomposition.

Idea: Suppose that our target function

$$f(x) = \int \tau(\langle x, \theta \rangle) \cdot g(\theta)\, d\theta$$

with $\rightarrow \int |g(\theta)|\, d\theta < +\infty \quad (\Leftrightarrow g \in L^1)$.

!! $\quad g(\theta) = \operatorname{sign}(g(\theta)) \cdot \|g\|_1 \cdot \frac{|g(\theta)|}{\|g\|_1} := q(\theta)$

$q \geq 0 \quad \int q(\theta)\, d\theta = 1.$

$$f(x) = \int \tau(\langle x, \theta \rangle) \operatorname{sign}(g(\theta)) \cdot \|g\|_1 \cdot q(\theta)\, d\theta.$$

$$= \|g\|_1 \cdot \mathbb{E}_{\theta \sim q}\big[\underbrace{\tau(\langle x, \theta \rangle \cdot \operatorname{sign}(g(\theta))}_{\phi(x,\theta)}\big].$$

So we can consider a Monte-Carlo approximation!

$$f_M(x) = \frac{1}{M} \sum_{i=1}^{M} \phi(x, \theta_i), \quad \theta_i \underset{iid}{\sim} q.$$

$\mathbb{E}\, \phi(x, \theta_i) = f(x)$

$$\mathbb{E}\,\|f - f_M\|^2 = \frac{1}{M}\mathrm{Var}(\phi(x,\theta)) \qquad \pi\,\phi(x,\theta_1) = f(x)$$

$$\leq \frac{1}{M}\,\mathbb{E}\,\|\phi(x,\theta)\|^2 \leq \left(\frac{1}{M}\right)\sup_\theta \mathbb{E}_x\,\phi^2\;\|g\|_1^2$$

There is no curse! (Statistical rate of approximation).

Theorem: [Barron '93] Suppose

$$C = \int \|\widehat{\nabla f(\xi)}\|\,d\xi < +\infty \qquad \widehat{\nabla f(\xi)} = \xi \cdot \hat{f}(\xi)$$

(a weighted L_1 norm in Fourier).

with $f, \hat{f} \in L_1$. Then we can approximate f to accuracy ε with a Relu/sigmoid shallow NN with $\sim C/\varepsilon^2$ hidden units.

(no curse in this function class) "Barron space".