Lecture 2: Approximation Main Ingredients

Goals for today: → Notion linear vs non-linear approximation

→ Get familiar with Fourier represent.

(→ From Fourier to shallow NNs)

- Signals that we measure as inputs to our ML algorithms (images, video, speech, ...) they are inherently defined over continuous domains

- Similarly functions that we aim to learn with ML

$f: X \longrightarrow Y$

X: space of images

Y: label space.

in both cases $X(t)$, f belongs to infinite-dimensional spaces.

Q: How to represent / manipulate signals in infinite-dim spaces?

→ This is what approximation theory tries to answer.

→ General set-up: $\mathcal{F}$: normed vector space of signals

$\mathcal{D} = \{\phi_\lambda\}_{\lambda \in \Lambda}$, $\phi_\lambda \in \mathcal{F}$ Λ: indexing a family/dictionary of functions in $\mathcal{F}$.

eg: $\Lambda = \mathbb{N}, \mathbb{Z}, \mathbb{R}$.

Given $f \in \mathcal{F}$, we want to express it as a linear combination of dictionary elements in $\mathcal{D}$:

$$f = \sum_{\lambda \in \Lambda} \alpha_\lambda \cdot \phi_\lambda \quad (\Lambda \text{ is countable})$$

$$f = \int_\Lambda g(\lambda) \cdot \phi_\lambda \, d\lambda \quad (\text{if } \Lambda \text{ is } \mathbb{R}, \mathbb{R}^n) \ldots$$

The information on f is in the "representation" α_λ / $g(\lambda)$

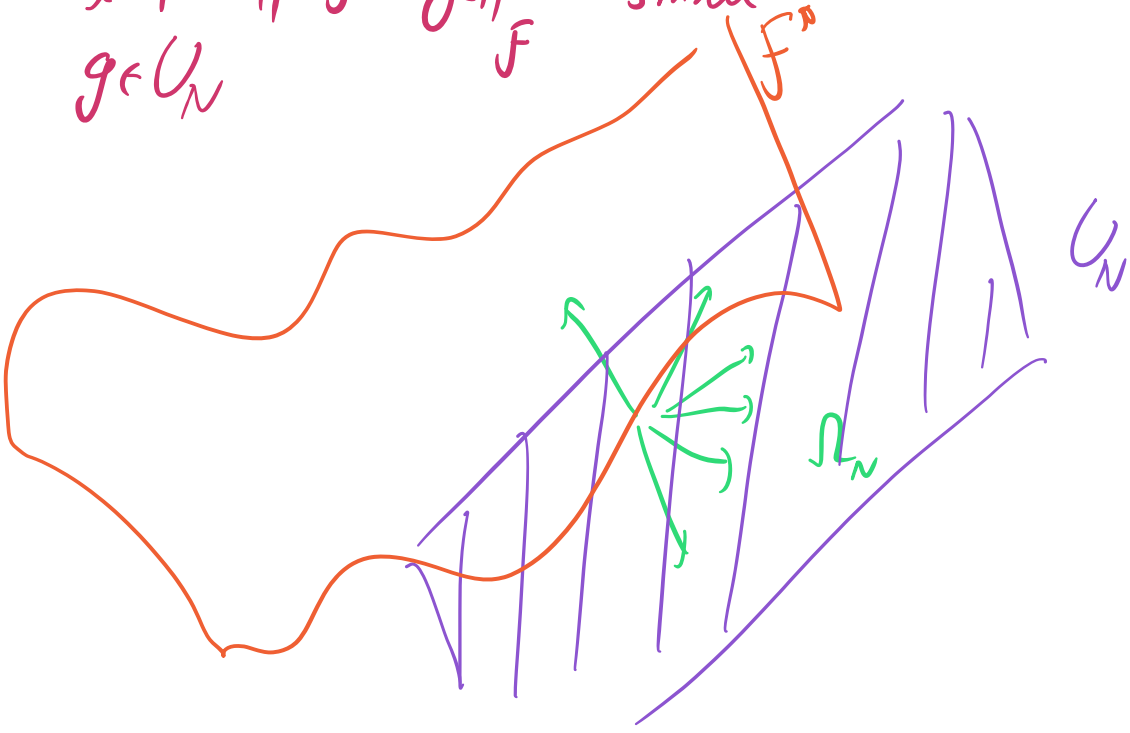
Goal: For a target family of functions $\mathcal{F}^* \subseteq \mathcal{F}$, obtain efficient approximation for any $f \in \mathcal{F}^*$ using the dictionary $\mathcal{D}$.

→ By "efficient", we mean identifying a finite set Ω_N of N elements in $\mathcal{D}$ such that

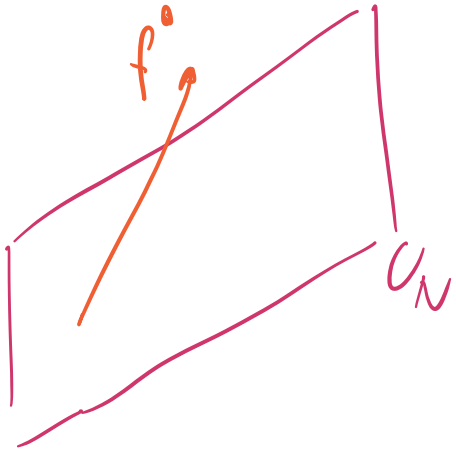
$$U_N = \{\tilde{f} \in \mathcal{F} ; \tilde{f} = \sum_{\lambda \in \Omega_N} \alpha_\lambda \phi_\lambda\} \subseteq \mathcal{F}$$

is "close" to $\mathcal{F}^*$:

$\forall f \in \mathcal{F}$, $\inf_{g \in U_N} \|f - g\|_{\mathcal{F}}$ small



→ Observe that we are selecting Ω_N that needs to work <u>for any</u> $f^* \in \mathcal{F}$. Once we have chosen Ω_N, finding the best approximation of f in U_N, amounts to solving a least-squares problem.



$$\min_{g \in U_N} \|f - g\| = \min_{\alpha \in \mathbb{R}^N} \left\| f - \sum_{k \in \Omega_N} \alpha_k \phi_m \right\|$$

linear approximation.

<u>Example:</u> Suppose $\mathcal{F} = L^2([0,1])$ squared-integrable functions.

Hilbert space.

$$\langle f, g \rangle_{\mathcal{F}} = \int_0^1 f(t)\, g(t)\, dt.$$

$\mathcal{D} = \{\phi_m\}_{m\in\mathbb{N}}$ orthonormal basis of F.

$\{\phi_m\}_{m\in\Omega_N}$ orthonormal basis of $\operatorname{span}\{\phi_m\}_{m\in\Omega_N} = U_N$.

The best approximation

$$\inf_{g\in U_N} \|f-g\|_F^2$$ can be computed explicitly:

if $$f = \sum_{m=0}^{\infty} \langle f, \phi_m\rangle \phi_m$$

$$f_N := P_{U_N} f = \sum_{m\in\Omega_N} \langle f, \phi_m\rangle \phi_m$$

$$f - f_N = \sum_{m\notin\Omega_N} \langle f, \phi_m\rangle \phi_m, \quad \text{with approx error}$$

$$\varepsilon_e(N, f) = \|f - f_N\|^2 = \boxed{\sum_{m\notin\Omega_N} |\langle f, \phi_m\rangle|^2}$$

- Since $\|f\|^2 = \int |f(t)|^2 dt = \boxed{\sum_{m=0}^{\infty} |\langle f, \phi_m\rangle|^2 < +\infty}$

we have $\lim_{N\to\infty} \varepsilon_e(N; f) = 0 \quad \forall f.$

- Suppose w.l.o.g $\Omega_N = \{1, \dots, N\}$.

Decay of $\varepsilon_\ell(N, f)$ as N increases

$\updownarrow$

Decay of $|\langle f, \phi_m \rangle|$ as m increases.

Fact: For any $s > 1/2$, there exists $A, B > 0$ such that if

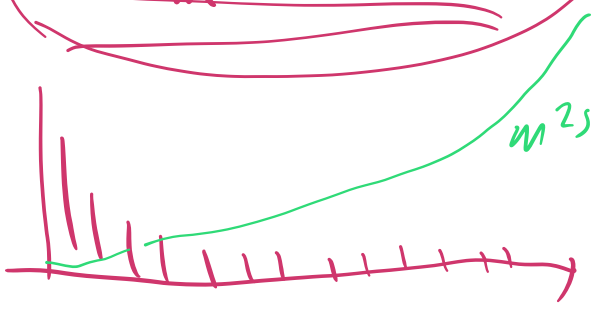
$$\sum_{m=0}^{\infty} |m|^{2s} \underbrace{|\langle f, \phi_m \rangle|^2}_{\beta_m} < +\infty$$

then

$$A \sum_m m^{2s} |\langle f, \phi_m \rangle|^2 \leq \sum_{N=0}^{\infty} N^{2s-1} \varepsilon_\ell(N, f) \leq$$

$$\leq B \sum_m m^{2s} |\langle f, \phi_m \rangle|^2$$

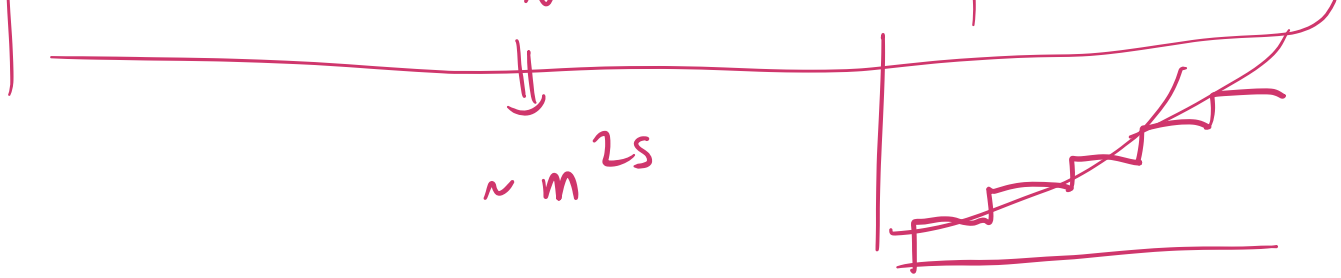
Thus $\varepsilon_\ell(N, f) = O(N^{-2s})$.



Proof: Apply definition $\varepsilon_\ell(N, f) = \sum_{m=N}^{\infty} \underbrace{|\langle f, \phi_m \rangle|^2}_{\beta_m}$

$$\sum_{N=0}^{\infty} N^{2s-1} \sum_{m=N}^{\infty} \beta_m^2 = \sum_{m=0}^{\infty} \beta_m^2 \sum_{N=0}^{m} N^{2s-1}$$

$$\int_0^m x^{2s-1}\, dx \leq \sum_{N=0}^{m} N^{2s-1} \leq \int_1^{m+1} x^{2s-1}\, dx$$



Let's verify $\varepsilon_\ell(N, f) = O(N^{-2s})$.

Observe that $\varepsilon_\ell(N, f) \le \varepsilon_\ell(m, f) \quad \forall\, m \le N$

$$\varepsilon_\ell(N, f) \sum_{m=N/2}^{N} m^{2s-1} \le \sum_{m=N/2}^{N} \varepsilon_\ell(m, f)\, m^{2s-1}$$

Since $\sum_{m=1}^{\infty} \varepsilon_\ell(m, f)\, m^{2s-1} < +\infty$, it follows that $\lim_{N\to\infty} \sum_{m=N/2}^{N} \varepsilon_\ell(m, f)\, m^{2s-1} = 0.$

Moreover, since $\exists\ C > 0$ $\sum_{m=N/2}^{N-1} m^{2s-1} \ge C \cdot N^{2s}$

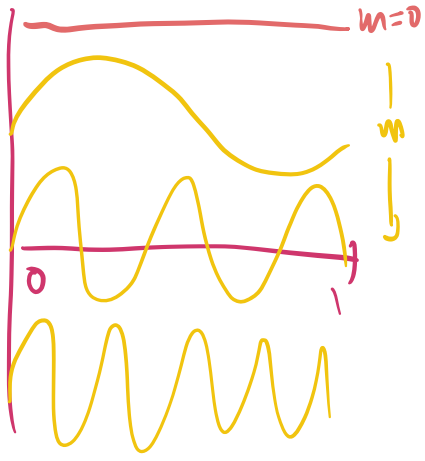
$\lim_{N\to\infty} \varepsilon_\ell(N, f) \cdot N^{2s} = 0.$ □

In summary, the linear approximation error of f in $\mathcal{D}$ decays at least as N^{-2s} if f belongs to the set

$$W_{\mathcal{D},s} := \left\{ f \in \mathcal{F};\ \sum m^{2s} |\langle f, \phi_m \rangle|^2 < +\infty \right\}$$

Q: What is a "canonical" example of orthonormal basis of $L^2([0,1])$ where this linear approx. scheme is appropriate?

The Fourier Basis: $\{e^{i2\pi mt}\}_{m\in\mathbb{Z}}$ is an orthonormal basis of $L^2([0,1])$. ($e^{i2\pi mt} = \phi_m(t)$)



$$f(t) = \sum_{m=-\infty}^{\infty} \underbrace{\langle f, \phi_m \rangle}_{\text{Fourier coefficients of } f} \phi_m \quad \leftarrow (*)$$

$\hat{f}_m$.

Next week: More flavors of Fourier transform).

Recall:

> Decay of Fourier coefficients $(\hat{f}_m)_m$
> $\updownarrow$
> smoothness of f

- Linear approximation of f in Fourier: Keep the N lowest frequencies and discard the rest.

$$f_N(t) = \sum_{|m|\le N/2} \hat{f}_m \phi_m$$

low-pass filter.

N: sampling rate.

- Recall: The Fourier transform of f' (derivative of f)

is $\longrightarrow$ $i(w)\hat{f}(w)$ (where $\hat{f}(w)$ is the F.T. of f).

so $\sum_m |m|^{2s} |\hat{f}_m|^2 < +\infty$ roughly measures

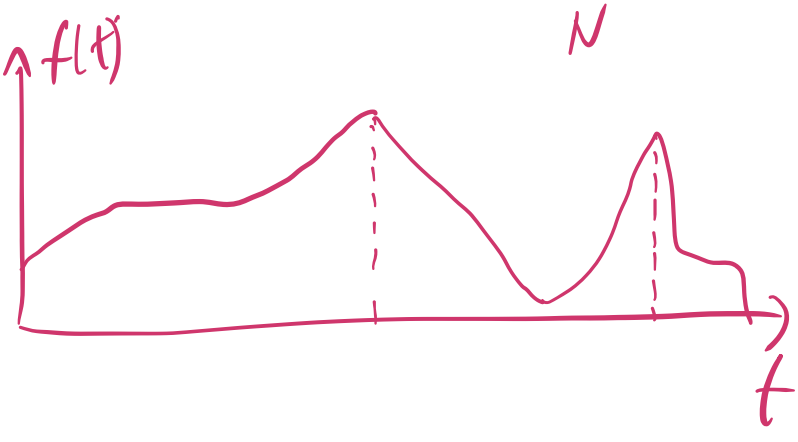
if f admits s derivatives (Sobolev Space $W^s([0,1])$)

↳ $\|f\|_{L^2}^2 = \sum_m (\hat{f}_m)^2$ (Parseval Identity).

$(s=1) \rightsquigarrow \|f'\|_{L^1}^2 = \sum_m |m|^2 |\hat{f}(m)|^2$

In conclusion, we can characterise $f \in W^s([0,1])$

if and only if $\sum_N N^{2s-1} \varepsilon_e(N, f) < +\infty$



↳ When f is discontinuous, (or f' discontinuous) rates of approximation cannot leverage local regularity of f.

→ Linear approximation exhibits <u>no adaptivity</u>.

The subspace $(V_N \subset L^2([0,1]))$ is chosen once and

for all target functions in $\mathcal{F}^*$.

Q: Can we do better?

Non-linear approximation

In the previous linear approx, we can characterize the best approximate N-dim subspace as

$$\inf_{U;\dim(U)\le N} \sup_{f\in\mathcal{F}^*} \|f - P_{U_N} f\|^2 = \varepsilon_e(N;\mathcal{F}^*)$$

(N-width of class)

$$\sup_{f\in\mathcal{F}^*} \inf_{U;\dim(U)\le N} \|f - P_{U_N} f\|^2 = \varepsilon_n(N;\mathcal{F}^*)$$

Non-linear approx

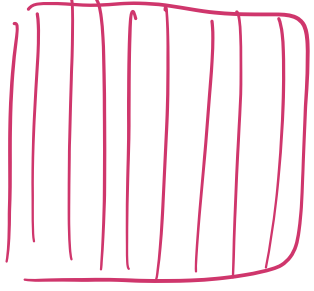
$$\inf_x \sup_y F(x,y) \quad \text{vs} \quad \sup_y \inf_x F(x,y)$$

Fix (x,y)

$$\underbrace{\inf_{x'} F(x',y)}_{g(y)} \le F(x,y) \le \underbrace{\sup_{y'} F(x,y')}_{H(x)}$$

$\forall (x,y)$ $\quad g(y) \le H(x)$

$\Rightarrow \quad \sup_y g(y) \leq \inf_x H(x)$



→ First instance of a fundamental tradeoff that we will encounter in DL / ML.

Better mathematical / Statistical performance

vs

More computationally challenging.

→ How does non-linear approximation work when $\mathcal{D}$ is still an orthonormal basis?

→ $f \rightsquigarrow \beta_m = \langle f, \phi_m \rangle_m$, $m \in \mathbb{N}$.

→ Let us sort the coefficients $|\beta_m| = |\langle f, \phi_m \rangle|$

$\beta_{s(m)} = \bar{\beta}_m$ with $\bar{\beta}_{m+1} \geq \bar{\beta}_m \quad \forall m$.

→ The best N-term non-linear approximation is

$$f_N = \sum_{m=1}^{N} \beta_{s(m)} \phi_{s(m)}, \text{ with error}$$

$$\varepsilon_n(N, f) = \sum_{m=N+1}^{\infty} |\beta_{s(n)}|^2$$

Similarly as before, the rate of approximation $\varepsilon_n(N,f)$ is driven by the decay of β_m.

Q: When is non-linear approximation provably better than linear approximation?

A: If we work with Fourier representation, gains are not relevant for "natural" classes of signals.

lacunary series

However, if we go beyond Fourier dictionaries, and into wavelet representations, then we obtain substantial separation between linear and non-linear approximation.

Theorem: If f has K discontinuities in $[0,1]$ and is uniformly Lipschitz α between these discontinuities, then

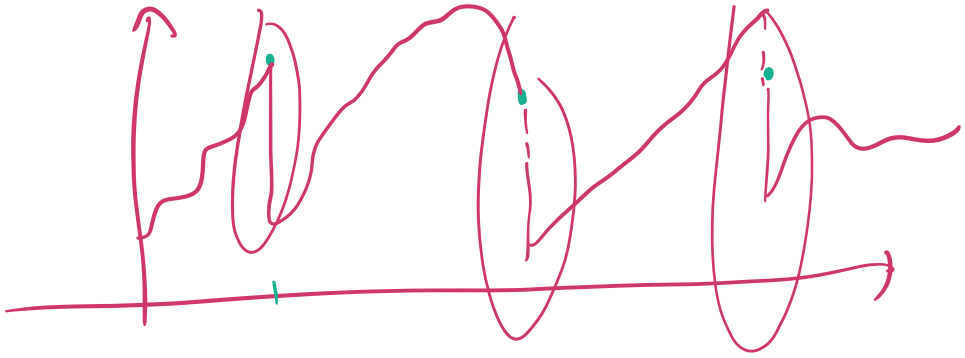
↳ $|f(x)-f(x')| \leq C \cdot |x-x'|^\alpha$

$$\varepsilon_L(N,f) = \Theta\left(K \cdot \|f\|_{C^\alpha}^2 \, N^{-1}\right) \quad \text{vs non-adapt.}$$

$$\varepsilon_n(N,f) = \Theta\left(\|f\|^2 \cdot N^{-2\alpha}\right)$$

adaptive to the local regularity of f

(using an appropriate dictionary of wavelets)



Next week: Take the two concepts from today to shallow NNs.