Signal Processing and Deep Learning on Graphs

Edouard Oyallon

edouard.oyallon@cnrs.fr

CNRS, ISIR

Graph Laplacian

Adjacency matrix on a Graph

- Let $\mathcal{G} = \{1, ..., n\}$ a graph with n nodes and $\mathcal{E} \subset \mathcal{G} \times \mathcal{G}$ a set of edge. We assume our graph is undirected:

$$(i, j) \in \mathcal{E} \Longleftrightarrow (j, i) \in \mathcal{E}$$

- We will consider weighted graphs: we assume each edge has a weight $\mathcal{A}_{ij}$ which is the adjacency matrix. It satisfies: $\mathcal{A}^T = \mathcal{A}$ and $\mathcal{A}\mathbf{1} = \mathbf{1}$

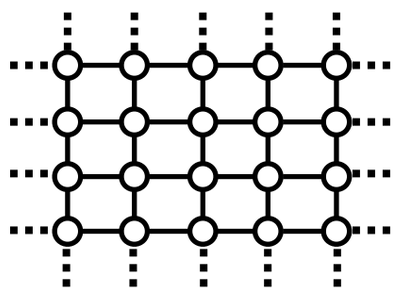
$$\text{and } \mathcal{A}_{ij} > 0 \Longleftrightarrow (i, j) \in \mathcal{E}$$

- Reciprocally, such matrix is the adjacency matrix of a graph.

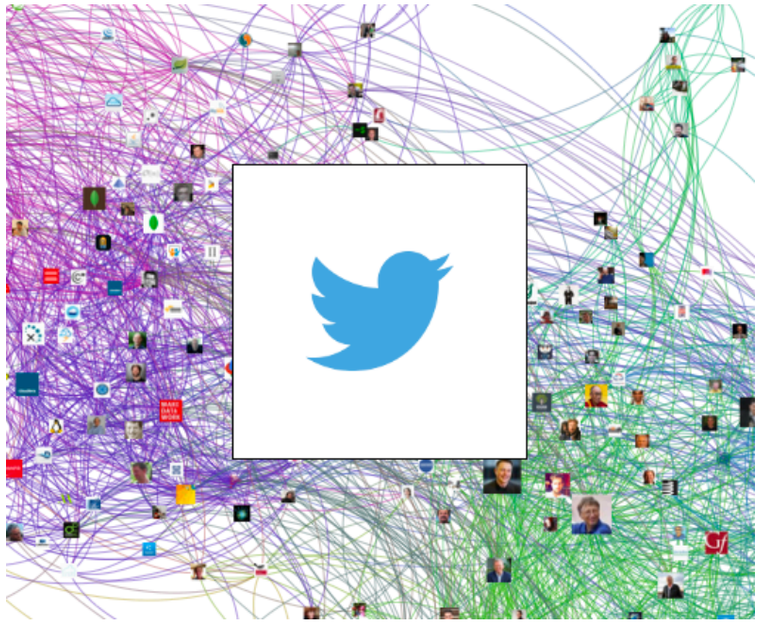
- We write: $L^2(\mathcal{G}) = \{x : \mathcal{G} \to \mathbb{R}, \sum |x(n)|^2 < \infty\}$

Examples of data with graph structure

4 connectivity for images:

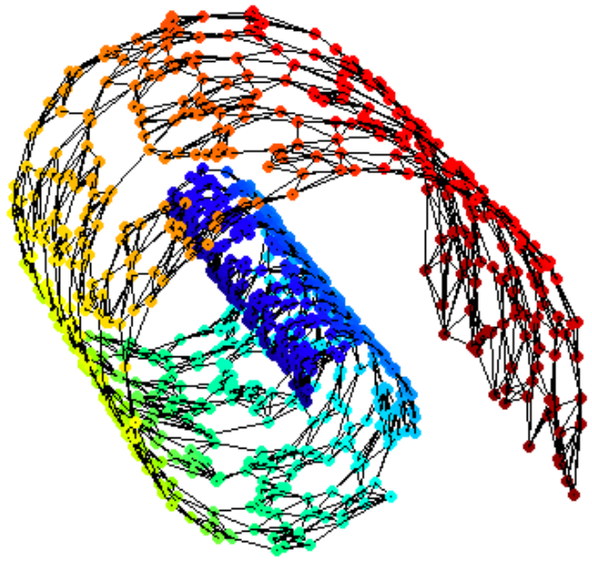


Community graph



$$\mathcal{A}_{i,j} = e^{-\frac{d(x_i,x_j)^2}{2\sigma^2}}$$

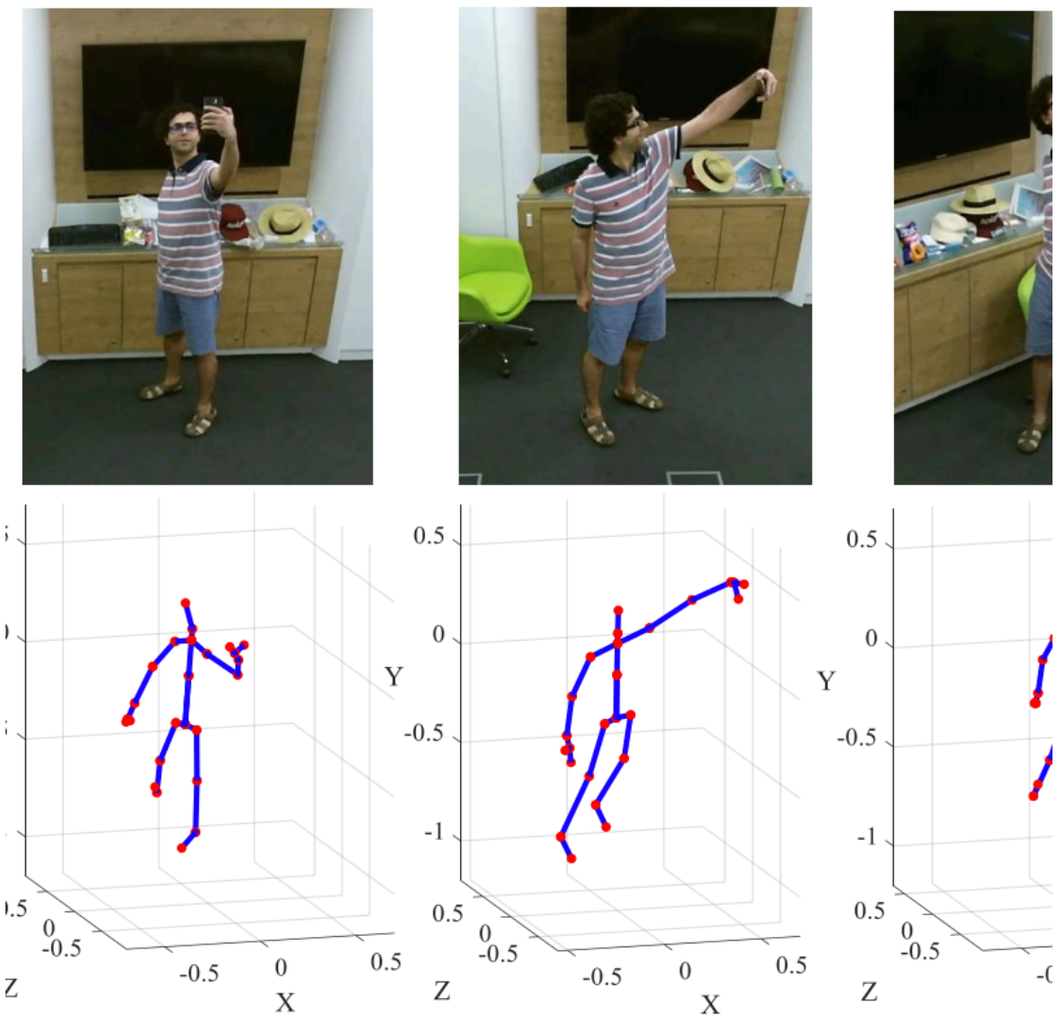
Manifold



Signals with non-Euclidean topology

Objective: finding Φ

$$\forall x \in L^2(\mathcal{G}), \Phi(x) \in \mathbb{R}^d$$



$\mathcal{G}$ is fixed

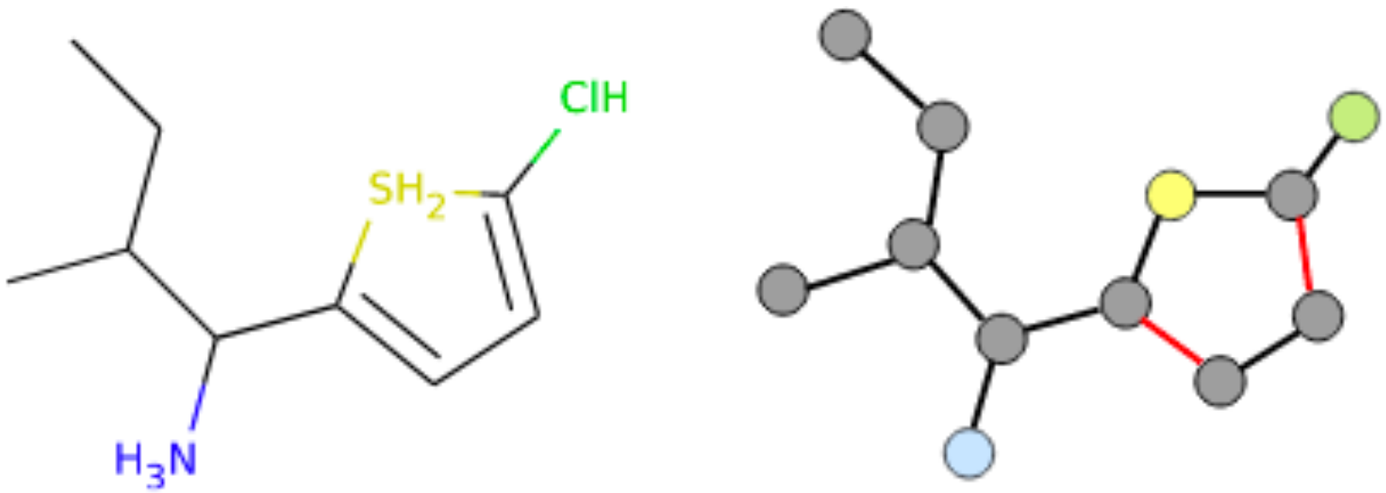
$\{x_1, ..., x_n\}$

"classify movements
from fiducial coordinates"

Graph classification

- Example: molecules

$$x = (\mathcal{A}_1, \mathcal{G}_1, \mathcal{A}_2, \mathcal{G}_2, ..., \mathcal{A}_k, \mathcal{G}_k) \qquad \Phi(x) \in \mathbb{R}^d$$

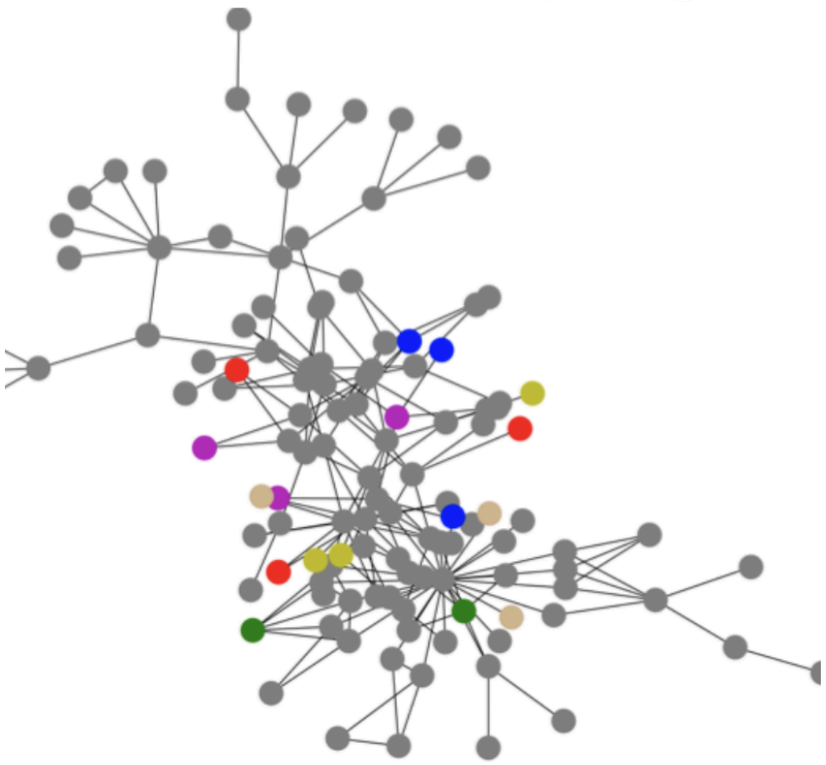


- What is the chemical potential of the given molecule?

Semi-supervised tasks

$$\Phi(X; \mathcal{A}) \in \mathbb{R}^d \qquad \mathcal{A} \in \mathbb{R}^{n \times n} \quad X \in \mathbb{R}^{n \times K}$$

- Only a subset of the nodes have labels Y.
- Yet $\mathcal{A}$ is fully know, X too.
- It is useful for classifying communities for instance.



Graph Filtering

Graph Laplacian

- We write the graph Laplacian as:

$$\Delta = \mathbf{I} - \mathcal{A}$$

It satisfies:

$$\Delta^T = \Delta \qquad \text{and} \qquad \Delta \succcurlyeq 0$$

- Assuming the graph has a single component:

$$\mathrm{Sp}(\Delta) = \{\lambda_1, ..., \lambda_n\} \quad \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$$

$\lambda_1 = 0 \longrightarrow \mathbf{1}$ is our low-pass filter!

λ_2 spectral gap (or fiedler value)

Relation with signal processing

- Consider: $\phi(x) = e^{-\frac{\|x\|^2}{2\sigma^2}}$ then $\phi \geq 0$, $\hat{\phi} \geq 0$

$$\hat{\psi}(\omega) = \hat{\phi}(\omega - \omega_0) \quad \text{and} \quad |\hat{\psi}|^2 + |\hat{\phi}|^2 = 1$$

- Here: $\Phi = \dfrac{\mathbf{I} + A}{2}$ then $\Phi_{ij} \geq 0$ and $\mathrm{Sp}(\Phi) \subset \mathbb{R}_+$

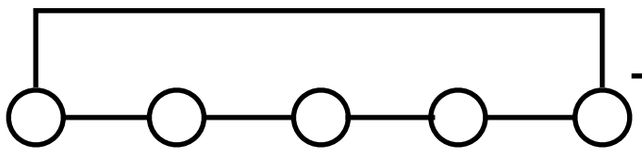
$$\tilde{\Delta} = \frac{\mathbf{I} - \mathcal{A}}{2} \quad \text{and} \quad \tilde{\Delta} + \Phi = \mathbf{I}$$

Spectral methods for graphs

- Performing Harmonic Analysis on graphs?
  - signal processing tools (wavelets, smoothness,. . .)
  - generalising the notion of convolutions to non-'flat group' domains
  - more "principled"

Ref.: Wavelets on graphs via spectral graph theory
DK Hammond, P Vandergheynst, R Gribonval, 2011

- Currently, harmonic analysis relies on an analogy with the 1D setting.



$$\Delta_{\mathcal{G}} = \begin{bmatrix} -2 & 1 & 0 & 0 & 1 \\ 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ 1 & 0 & 0 & 1 & -2 \end{bmatrix} = J + J^T - 2\mathbf{I}$$

with J Toeplitz and:

$$J J^T = \mathbf{I}$$

. . . and an eigendecomposition of J is given via a cosine transform

Convolutions on graphs

similar to a multiplication in Fourier domain

$$f \star g \triangleq \sum_n g[n] \langle f, e_n \rangle e_n$$

Proposition: if f in $L^2(\mathcal{G})$ and g is in $\ell^2(\mathbb{N})$ then $f \star g \in L^2(G)$

Yet, this formulation doesn't take in account the spectrum amplitude thus also:

$$f \tilde{\star} g \triangleq \sum_{\lambda_n \in \mathrm{Sp}(\Delta)} \hat{g}(\lambda_n) \langle f, e_n \rangle e_n$$

Proposition: if f in $L^2(\mathcal{G})$ and g is continuous then $f \star g \in L^2(G)$

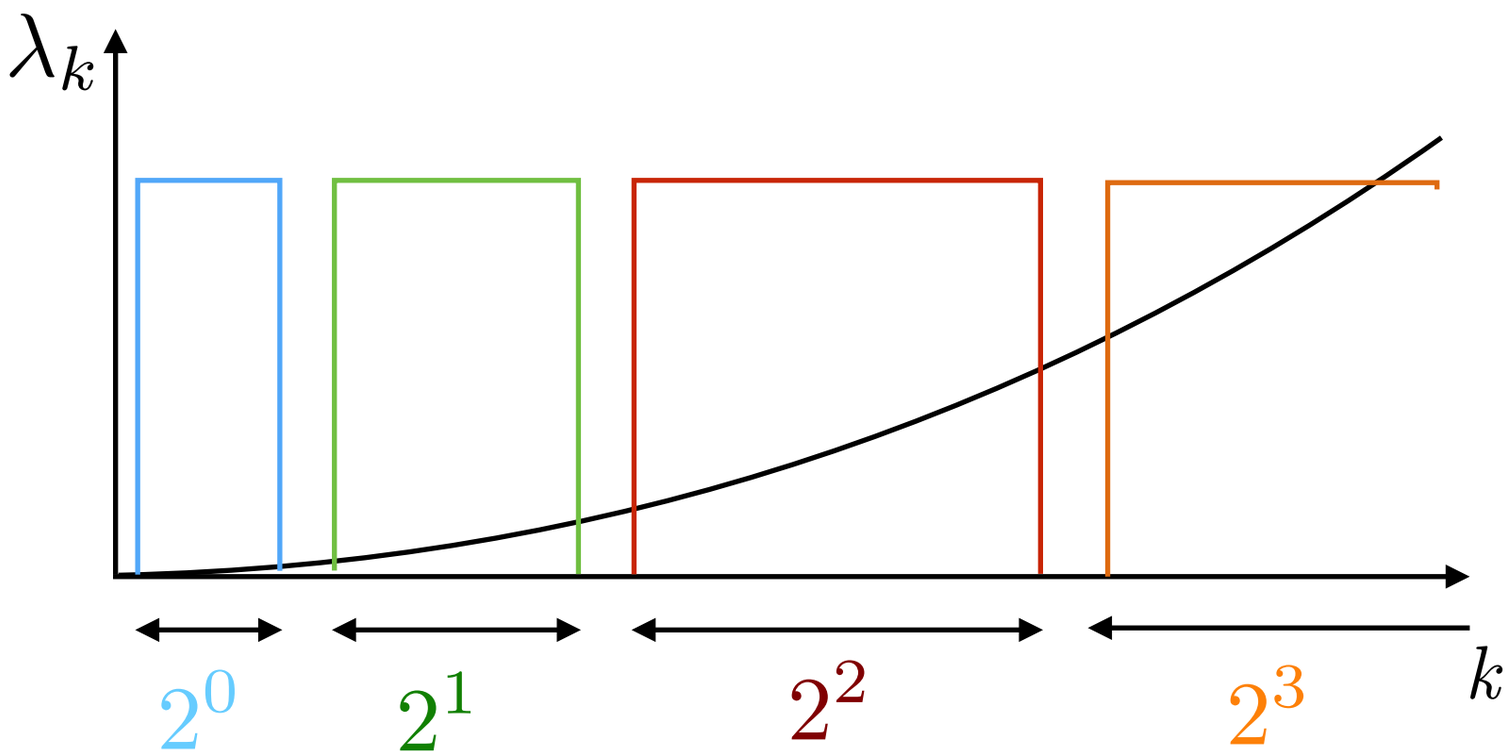
Proof:

$$\mathrm{Sp}(\Delta) \subset [0, \|\Delta\|]$$

Wavelets on graphs?

Get the spectrum: $\Delta e_k = \lambda_k e_k$

Ref.: Wavelets on graphs via spectral graph theory, David K. Hammond, Pierre Vandergheynst, Rémi Gribonval



The same type of ideas apply

Guarantees of this transform are task specific and often bound to potential analogy with an Euclidean grid. *Let's follow this idea!*

Diffusion wavelets on graphs

- A specific class wavelets transform, which is obtained by analogy with DoG:

$$\psi_j \triangleq \mathcal{A}^{j+1} - \mathcal{A}^j$$

$$\phi = \mathcal{A} + \mathbf{I}$$

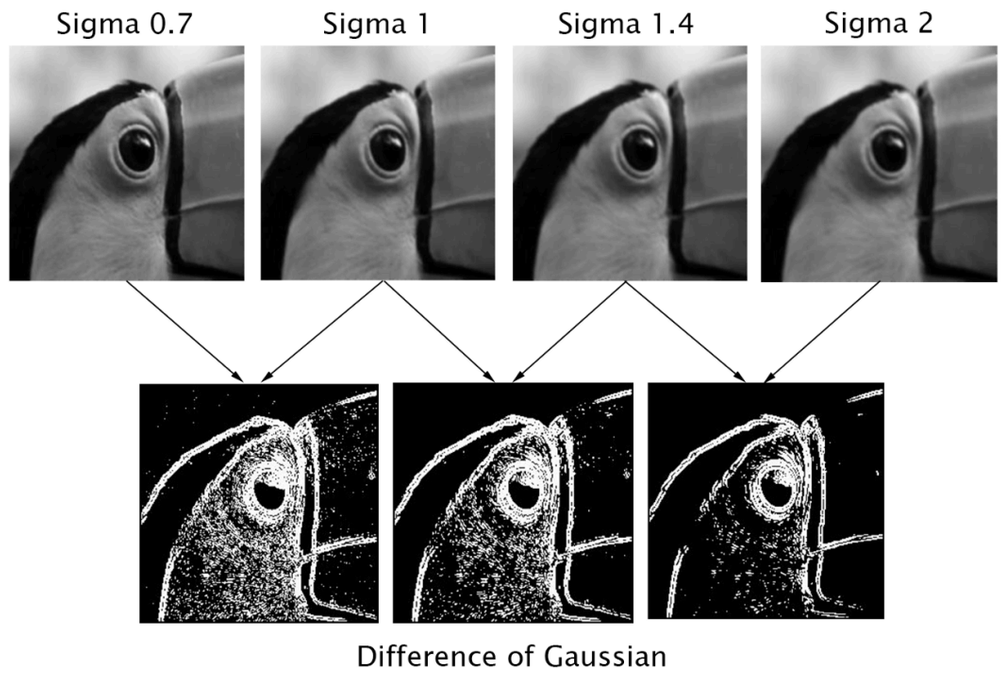
$$\|Wx\|^2 = \|\phi x\|^2 + \sum_j \|\psi_j x\|^2$$

$$= \sum_j x^T \mathcal{A}^{2j} (\mathcal{A} - \mathbf{I})^2 x + x^T (\mathcal{A} + \mathbf{I})^2 x$$

- Lemma: $0 \preccurlyeq (\mathcal{A} - \mathbf{I})^2 \preccurlyeq \mathbf{I} - \mathcal{A}^2$

- Prop.: W is a frame, ie:

$$\|Wx\|^2 \leq 5\|x\|^2$$

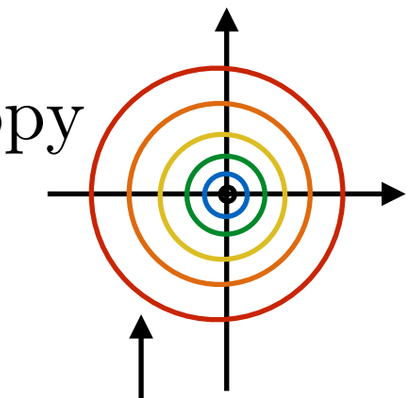


(Unfortunate) Isotropy of the Laplacian

Diagonalisable in $\mathbb{C}$ ↑ ∂x_i ⟶ $\Delta = \sum_i \frac{\partial^2}{\partial x_i^2}$ ⟶ $\Delta(e^{i\omega t}) = -\|\omega\|^2 e^{i\omega t}$

Diagonalisable in $\mathbb{R}$ ↑

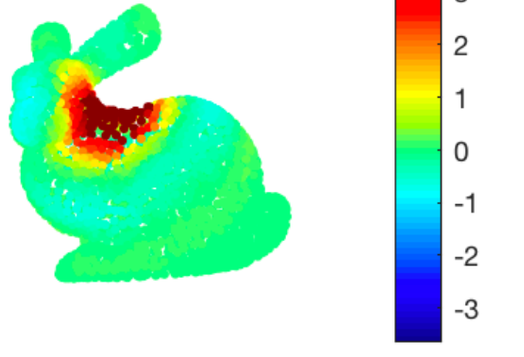
Isotropy ↑



Phase-shift: $\mathcal{L}_a x(u) = \sum_n \frac{x^{(n)}(u)}{n!} a^n \Rightarrow \widehat{\mathcal{L}_a x}(\omega) = \sum_n \frac{w^n \hat{x}(\omega)}{n!} a^n = e^{i\omega a}\hat{x}(\omega)$

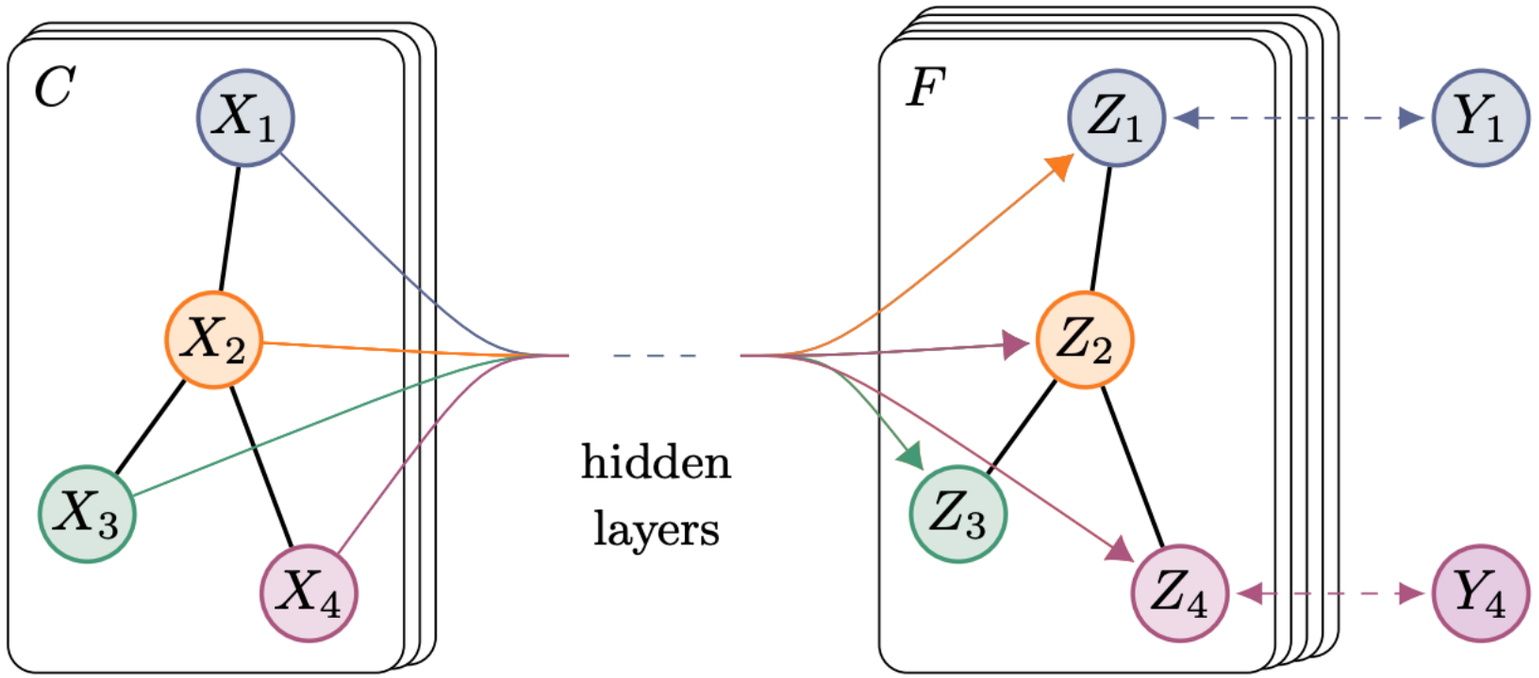
?? ⟶ $\Delta_{\mathcal{G}}$ ⟶

No Phase

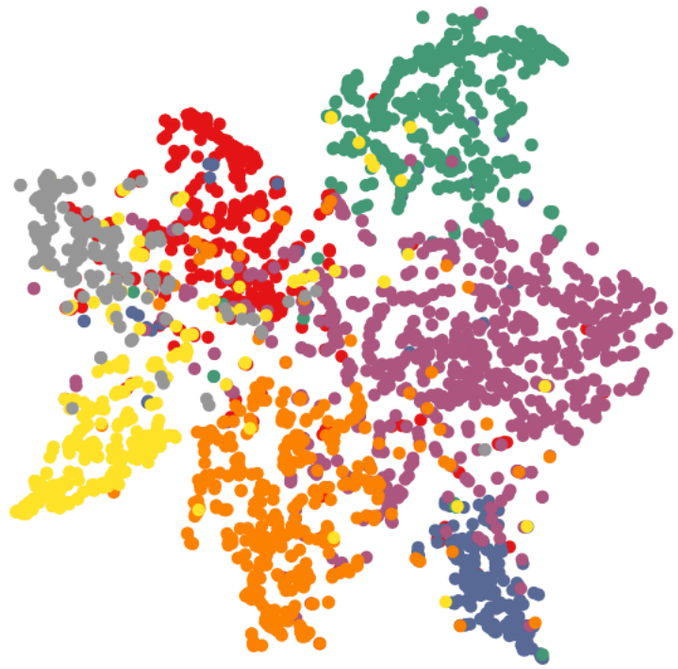


A short Review of Deep Learning models

GCN "a la" Kipf



(a) Graph Convolutional Network



(b) Hidden layer activations

Issue: over smoothing!

$$H^{(l+1)} = \sigma(AH^{(l)}W^{(l)})$$

($A = I$ leads to a MLP)

Non-linearity

Averaging (not learned)

Weights (learned)

A few Spectral free variants

- Chebynet: Learn filters of type:

$$g_\theta(\Delta) = \sum_{i=1}^{n} \theta_i T_i(\Delta)$$

Chebychev polynomial (fast to compute)

- Scattering-Net: $\psi_j \triangleq \mathcal{A}^{j+1} - \mathcal{A}^j$

$$\phi = \mathcal{A} + \mathbf{I}$$

- Many variants that combine "pooling". Most of them suffer from the oversmoothing phenomenon (depth is roughly 2)

Universal Approximation via Shallow Neural Networks

Edouard Oyallon

edouard.oyallon@cnrs.fr

CNRS, ISIR

Motivation

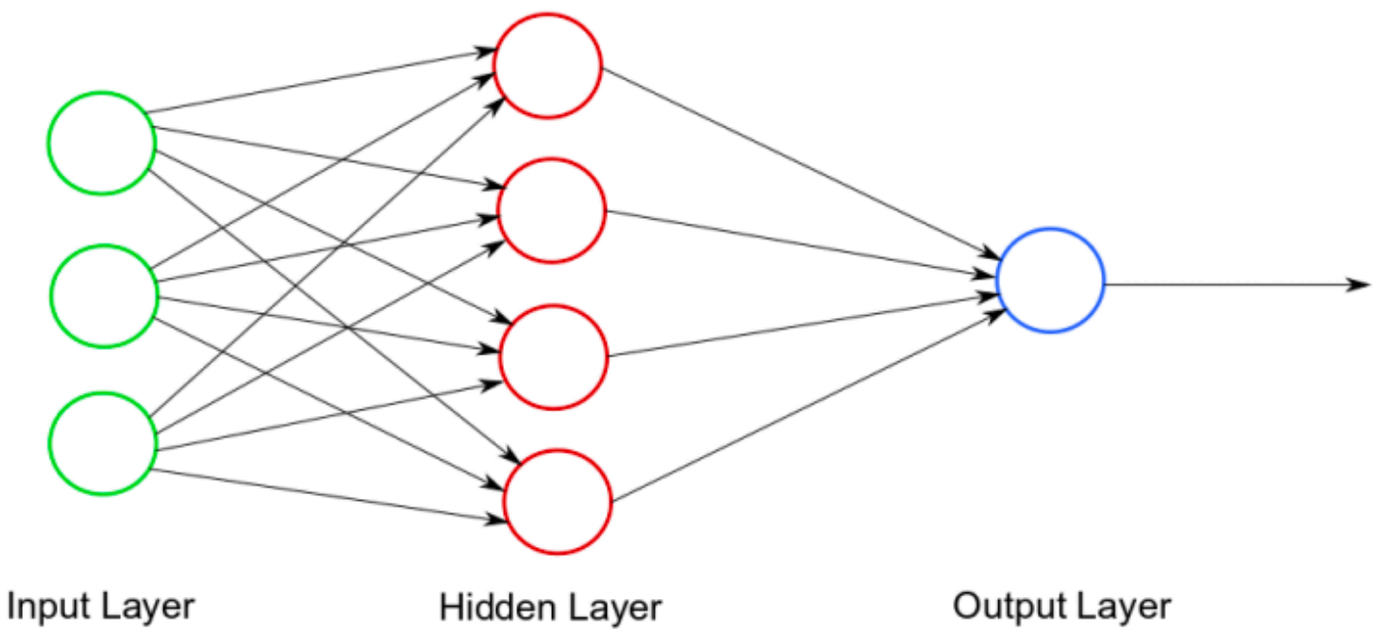
Definition

- We will consider real-valued 1-hidden layer neural network with width N, which are defined via:

$$f(x) = \sum_{i=1}^{N} \alpha_i \rho(\langle \beta_i, x \rangle)$$

- ρ will be a ReLU if no ad-hoc explanations.

- The data are assumed to be bounded.



Fourier as a NN

- A very simple example of a one hidden layer Neural Network: for f in $L^2(\mathbb{R}^d)$, we have:

$$f(u) = C_d \int_{\mathbb{R}^d} \hat{f}(\omega) e^{i\omega^T u} \, d\omega$$

- The sum can be understood as a linear operator, and the complex exponential as a non-linearity.

- Can we pick other non-linearities?

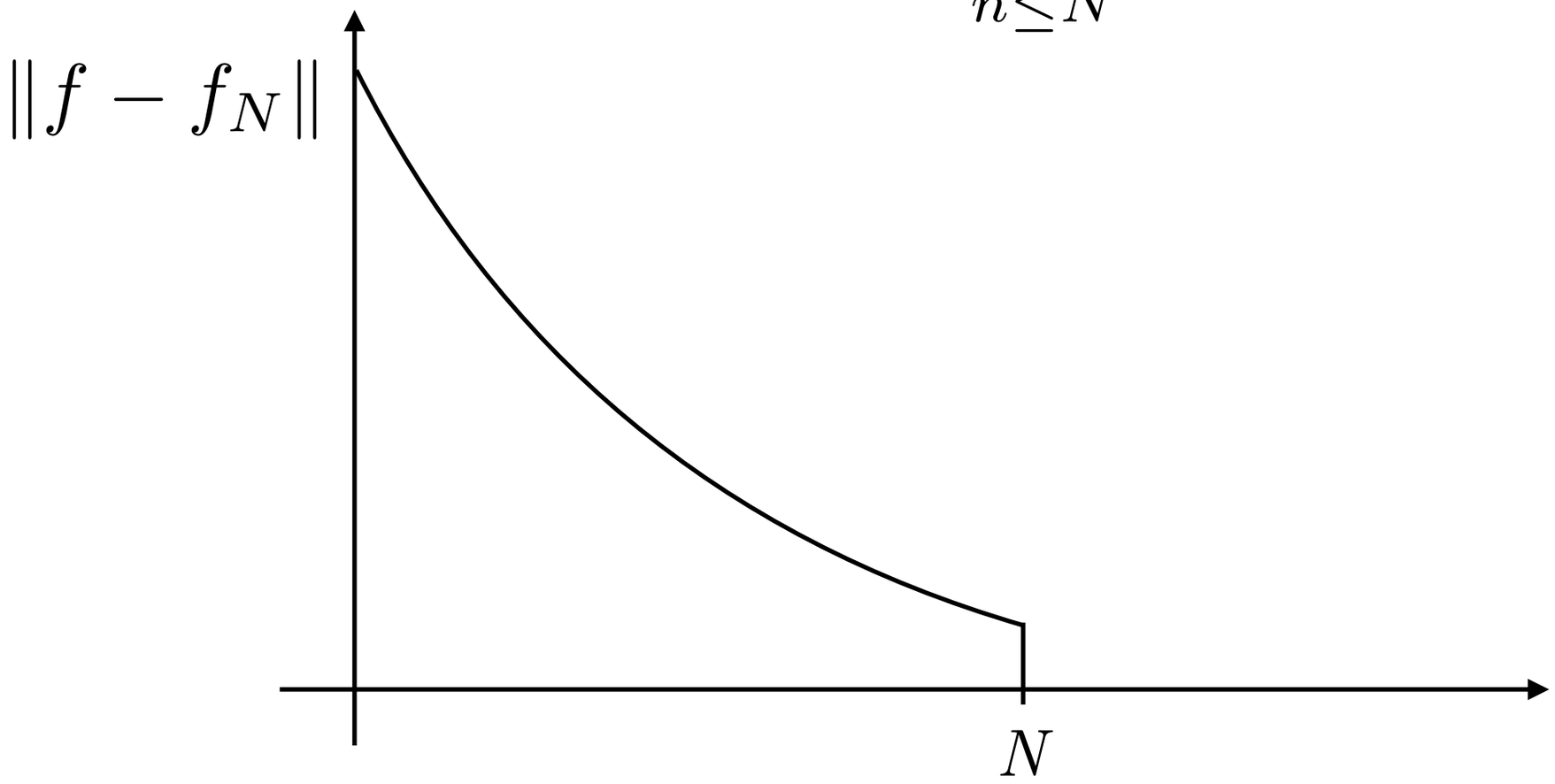
Cybenko's theorem

- Universal approximation theorem (Cybenko, 1989): Fix a compact K and $m \in \mathbb{N}^*$, then for any $f : K \to \mathbb{R}^m$ continuous, if $\rho : \mathbb{R} \to \mathbb{R}$ is not polynomial, then for any$\epsilon > 0$ there exists W_1, W_2, b, s.t.

$$\sup_{x \in K} \|f(x) - W_2 \rho(W_1 x + b)\| < \epsilon$$

Approximation rates?

$$f \approx \sum_{n \leq N} \lambda_n g_n = f_N$$



Examples: Fourier, wavelets, ...

Spherical symmetry

$$\Phi(x;w) = \sum_{k \leq K} \alpha_k \rho(\langle x, w_k \rangle) \quad \text{and let:} \quad w_k = \|w_k\| \theta_k$$

- A neural network of 1-layer writes, for $x \in \mathcal{B}(0, \frac{1}{2})$:

$$\Phi(x;w) = \sum_{k \leq K} \alpha_k \rho(\langle x, w_k \rangle)$$

$$= \sum_{k \leq K} \frac{\alpha_k}{\|w_k\|} \rho(\langle x, \theta_k \rangle)$$

$$= \int_{\mathcal{S}^{d-1}} \rho(\langle x, \theta \rangle) \, d\mu(\theta) \text{ where: } \mu = \sum_{k \leq K} \frac{\alpha_k}{\|w_k\|} \delta_{\theta_k}$$

- Then let $|t| = \sqrt{1 - \|x\|^2}$ and let:

$\tilde{\Phi}((x,t)) \triangleq |t| \Phi(\frac{x}{t}) = \Phi(x)$ if $t > 0$

if Φ is lipschitz/continuous
so is $\tilde{\Phi}$

($\tilde{\Phi}$ is odd)

Fourier tools on the Sphere

The "convolution":

$$f \circledast g(x) = \int_{\mathcal{S}^{d-1}} f(y) g(\langle x, y \rangle)\, d\sigma(y)$$

on the sphere

a real valued function

Reminders about the Sphere

- In hyper spherical coordinate, we have:

$$\langle f, g \rangle_{\mathcal{S}^{d-1}} = \int_{x \in \mathcal{S}^{d-1}} f(x) g(x) \, d\sigma(x)$$

Sphere as a Manifold

What are the eigenvalues of a Laplacian?

$$\Delta e_\omega = -\|\omega\|^2 e_\omega, \text{where } e_\omega(u) = e^{i\omega^T u}$$

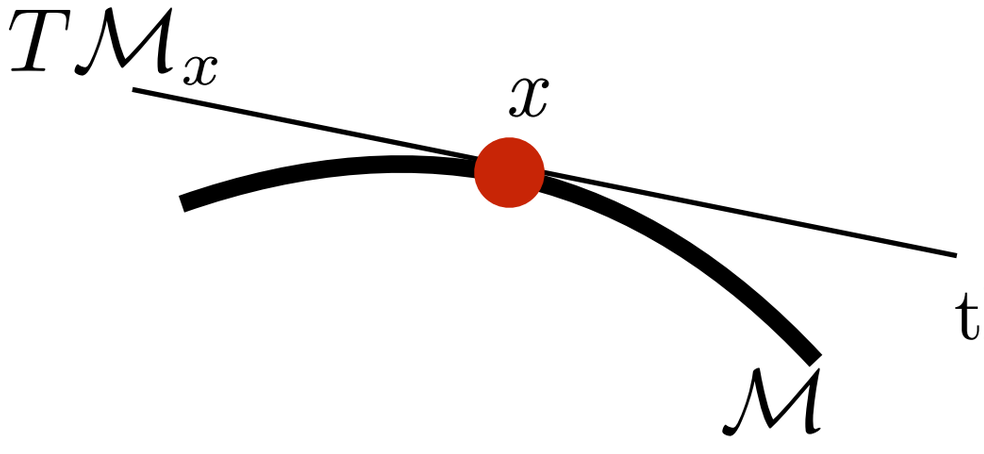
Manifold, tangent space

- By now, we consider a manifold $\mathcal{M}$ which is a space locally diffeomorph to $\mathbb{R}^d$.

- Each point x has thus a tangent bundle $T\mathcal{M}_x$ (a collection of tangent vector) equipped with a local scalar product.

- We can define a gradient for smooth $f : \mathcal{M} \to \mathbb{R}$

$$\forall x \in \mathcal{M}, \langle \nabla_{\mathcal{M}} f(x), y \rangle_{T\mathcal{M}_x} \triangleq df(x).y$$



$f : \mathcal{M} \to \mathbb{R}$ smooth

iff for any $\phi : \mathcal{U} \subset \mathbb{R}^d \to \phi(\mathcal{U}) \subset \mathcal{M}$ smooth

then $f \circ \phi$ smooth

also: $\nabla(f \circ \phi) = d\phi^T \nabla_{\mathcal{M}} f$

Divergence theorem

- Divergence Theorem: Consider a regular open set Ω with a smooth boundary $\partial\Omega$, then for u, v smooth we have:

Green identity (1)
$$\int_\Omega u\Delta v + \nabla u.\nabla v \, d\lambda = \int_{\partial\Omega} u(\nabla v.\mathbf{n}) \, d\sigma$$

Green identity (2)
$$\int_\Omega u\Delta v \, d\lambda - \int_\Omega v\Delta u \, d\lambda = \int_{\partial\Omega} u(\nabla v.\mathbf{n}) \, d\sigma - \int_{\partial\Omega} v(\nabla u.\mathbf{n}) \, d\sigma$$

Note: the sphere has no (manifold) boundary!!

A symmetric operator

- Consider a manifold with empty boundary. Thanks to the previous theorem, we formally define the Laplacian on a manifold as the unique operator satisfying:

$$\int_{\mathcal{M}} u\Delta v \, d\lambda = -\int_{\mathcal{M}} \langle \nabla u, \nabla v \rangle_{T\mathcal{M}} \, d\lambda$$

$$\text{with} \qquad \Delta_{\mathcal{M}} : L^2(\mathcal{M}) \to L^2(\mathcal{M})$$

- Proposition: it is symmetric and non positive

- In fact, it is also compact... (much more work)

Spectral theorem and convolutions

- A given Laplacian has some eigen-couples: $\{e_n, \lambda_n\}_{n\in\mathbb{N}}$.
- Similarly to a graph, we can consider for $g \in \ell^2(\mathbb{N})$:

$$f \star g \triangleq \sum_n g[n]\langle f, e_n\rangle e_n \in L^2(\mathcal{M})$$

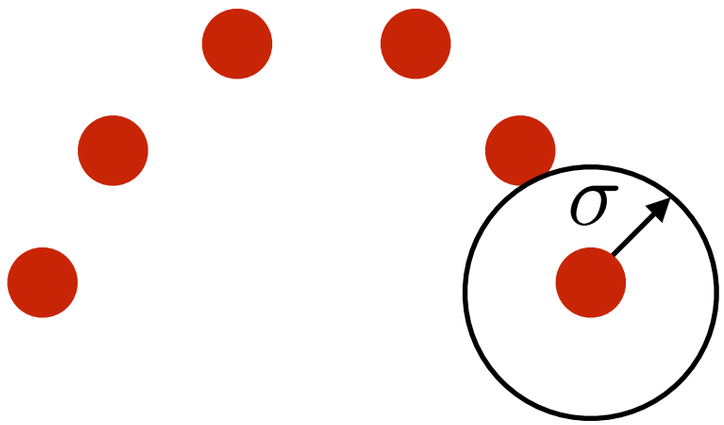
and

$$\|f \star g\| \leq \|g\|_{\ell^2(\mathbb{N})}\|f\|$$

- Is it consistent with graphs?

From Manifold to graph Laplacian...

N points



$$\Delta_{\mathcal{G}} f(x) = \frac{\sum_{i=1}^{N} K_\sigma(x_i, x) f(x_j)}{\sum_{i=1}^{N} K_\sigma(x_i, x)} - f(x)$$

where $K_\sigma(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$

As $\sigma \to 0, N \to \infty$ then $\Delta_{\mathcal{G}} \to \Delta_{\mathcal{M}}$

(if only N grows, there is a bias and we need to take in account the local connectivity)

Manifold on the sphere

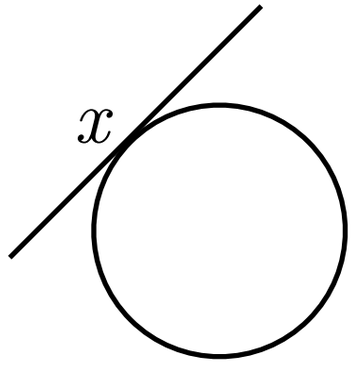
- Now, we consider a specific example of high interest: the sphere. Our first observation is that if $f : \mathcal{S}^{d-1} \to \mathbb{R}$

$$\forall \|x\| = 1, \Delta_{\mathbb{R}^d}(f \circ i)(x) = \Delta_{\mathcal{S}^{d-1}} f(x) \text{ where } i(x) = \frac{x}{\|x\|}$$

- Sketch of the Proof:

If $\|x\| = 1$ then:

$$i(x+\epsilon) = \frac{x+\epsilon}{\|x+\epsilon\|} = \frac{1}{\|x\|}(x + \epsilon - x\langle x, \epsilon\rangle) + o(\epsilon)$$



and: $\nabla(f \circ i)(x) = di(x)^T \nabla_{\mathcal{S}^{d-1}} f(i(x))$

which leads to:

$$\nabla(f \circ i)(x)^T \nabla(g \circ i)(x) = \frac{1}{\|x\|^2} \langle \nabla_{\mathcal{S}^{d-1}} f, \nabla_{\mathcal{S}^{d-1}} g \rangle_{T\mathcal{S}^{d-1}_x}$$

- Now the left part is invariant by rescaling and we can integrate on the unit ball, and apply Green formula.

Fourier on the sphere

How can we Formally define a Fourier notion on the Sphere?
A class of polynomials will have similarity with complex exponentials.

1D complex exponential

- Any functions 2π periodic, with enough regularity can be decomposed as:

$$f(x) = \sum_n c_n e^{inx}$$

- Here, $P_N(x) = \sum_{|n| \leq N} c_n e^{inx}$ is a trigonometric polynomial.

- Trigonometric polynomials are dense in $L^2[0, 2\pi]$, and in other words, we can build a sequence such that:

$$\|f - P_N\| \to 0$$

Spherical harmonics

- We say that a polynomial is a spherical harmonic (of degree k) if it is homogeneous (of degree k) and harmonic:

$$\exists k \in \mathbb{N}, \forall x \in \mathbb{R}^d, \lambda \in \mathbb{R} : P(\lambda x) = \lambda^k P(x)$$

$$\text{and} \quad \Delta P = 0$$

- It defines a subspace of $\mathbb{K}[X_1, ..., X_d]$ and we write it $\mathcal{H}_k^d$
- Spherical harmonics are dense in $L^2(\mathcal{S}^{d-1})$: $\|f - P_N\| \to 0$
- They are solution of: $\forall P \in \mathcal{H}_k^d, \Delta_{\mathcal{S}^{d-1}} P = -k(k+d-2)P$
- They are exactly the eigen-vector of the spherical Laplacian and invariant subspaces of rotations.

Approximating periodic functions with trigonometric polynomials

- It is well known that one can introduce the Fourier coefficient:

$$c_n(f) = \frac{1}{2\pi}\int_0^{2\pi} f(t)e^{-int}dt$$

- They satisfy $c_n(e_m) = \delta_{n=m}$ with $e_m(t) = e^{imt}$

- And we have the Perseval identify $\|f\| = \sum_n |c_n(f)|^2$

- And the identity in L^2:

$$f = \sum_n c_n e_n$$

Zonal polynomials

- Here, we let: $P_n(t) = (-1)^n \frac{\Gamma(\frac{d-1}{2})}{2^n \Gamma(n+\frac{d-1}{2})} (1-t^2)^{\frac{3-d}{2}} \frac{d}{dt^n}\left((1-t)^{n+\frac{d-3}{2}}\right)$

and $$K_n(f)(x) = \int_{\mathcal{S}^{d-1}} P_n(\langle x, y\rangle) f(y)\, d\sigma(y)$$

I will explain soon why

Then:

$$\forall P \in \mathcal{H}_n^d, P = K_n(P) \quad \text{and} \quad K_n(P) = 0$$

where $P \in \mathcal{H}_m^d, m \neq n$

and:

$$\boxed{\forall f \in L^2(\mathcal{S}^{d-1}), f = \sum_{n=0}^{\infty} K_n(f)}$$

and:

$$\forall f \in L^2(\mathcal{S}^{d-1}), \|f\|^2 = \sum_{n=0}^{\infty} \|K_n(f)\|^2$$

Differentiability

- Proposition: if f is $\mathcal{C}^{2p}(\mathcal{S}^1)$, then: $|c_n(f)| = o(\frac{1}{n^{2p}})$

- Proposition: if f is $\mathcal{C}^{2p}(\mathcal{S}^{d-1})$, then: $\|K_n(f)\| = o(\frac{1}{n^{2p}})$

- Proof: We observe, using the Laplacian's relations:
  $\forall P \in \mathcal{H}_n^d, -n(n+d-2)P = \Delta_{\mathcal{S}^{d-1}} P = K_n(\Delta_{\mathcal{S}^{d-1}} P)$

  and:

$$\Delta^p_{\mathcal{S}^{d-1}} f = \sum_n K_n(\Delta^p_{\mathcal{S}^{d-1}} f) = \sum_n \Delta^p_{\mathcal{S}^{d-1}} K_n(f) = \sum_n (-n(n+d-2))^p K_n(f)$$

Convolution in 1D

- Define the convolution between $f, g \in L^2(\mathcal{S}^1)$:

$$f \star g = \int_0^{2\pi} f(\theta - \theta') g(\theta') \, d\theta'$$

- Then, it is well-known that:

$$c_n(f \star g) = c_n(f) c_n(g)$$

Convolution on the sphere

- Here, we introduce for:

$$f \in L^2(\mathcal{S}^{d-1}), g \in L^1([-1,1])$$

the following convolution:

$$f \circledast g(x) \triangleq \int_{y \in \mathcal{S}^{d-1}} f(y) g(\langle x, y \rangle) d\sigma(y)$$

Approximation on the Sphere

Poisson kernel in 1D

$$\theta \in \mathbb{R}, 0 \leq r < 1 \quad P_r(\theta) = \sum_{n \in \mathbb{Z}} r^{|n|} e^{in\theta} = \frac{1 - r^2}{1 - 2r\cos\theta + r^2}$$

Here: $\int_{[0,2\pi]} P_r(\theta)\, d\theta = 1$ and $P_r(\theta) \geq 0$

- If $f \in L^2([0, 2\pi])$ then:

$$(f \star P_r)(\theta) = \int_0^{2\pi} f(\theta - \theta') P_r(\theta')\, d\theta' = \sum_{n \in \mathbb{Z}} \hat{f}[n] r^{|n|} e^{in\theta}$$

- And it acts as a dirac:

$$f \star P_r \overset{r \to 1}{\to} f$$

- If f is continuous:

$$\| f \star P_r - f \|_\infty \to 0$$

Poisson Kernel in n-D

Poisson kernel: $\forall x, y \in \mathcal{S}^{d-1}$ $0 \leq r < 1$ $\quad P_r(\langle x, y \rangle) = \dfrac{1 - r^2}{(1 - 2r\langle x, y \rangle + r^2)^{d/2}}$,

- Then: $P_r(\langle x, y \rangle) \geq 0$ and $\displaystyle\int_{\mathcal{S}^{d-1}} P_r(\langle x, y \rangle) d\sigma(y) = 1$

- Additionally:

$$P_r(\langle x, y \rangle) = \sum_{n=0}^{\infty} P_n(\langle x, y \rangle) r^n$$

(Zonal polynomials are generated by this analytic sum)

and:

$$K_n(f \circledast P_r) = \sum_{n=0}^{\infty} K_n(f) r^n P_n(\langle x, y \rangle)$$

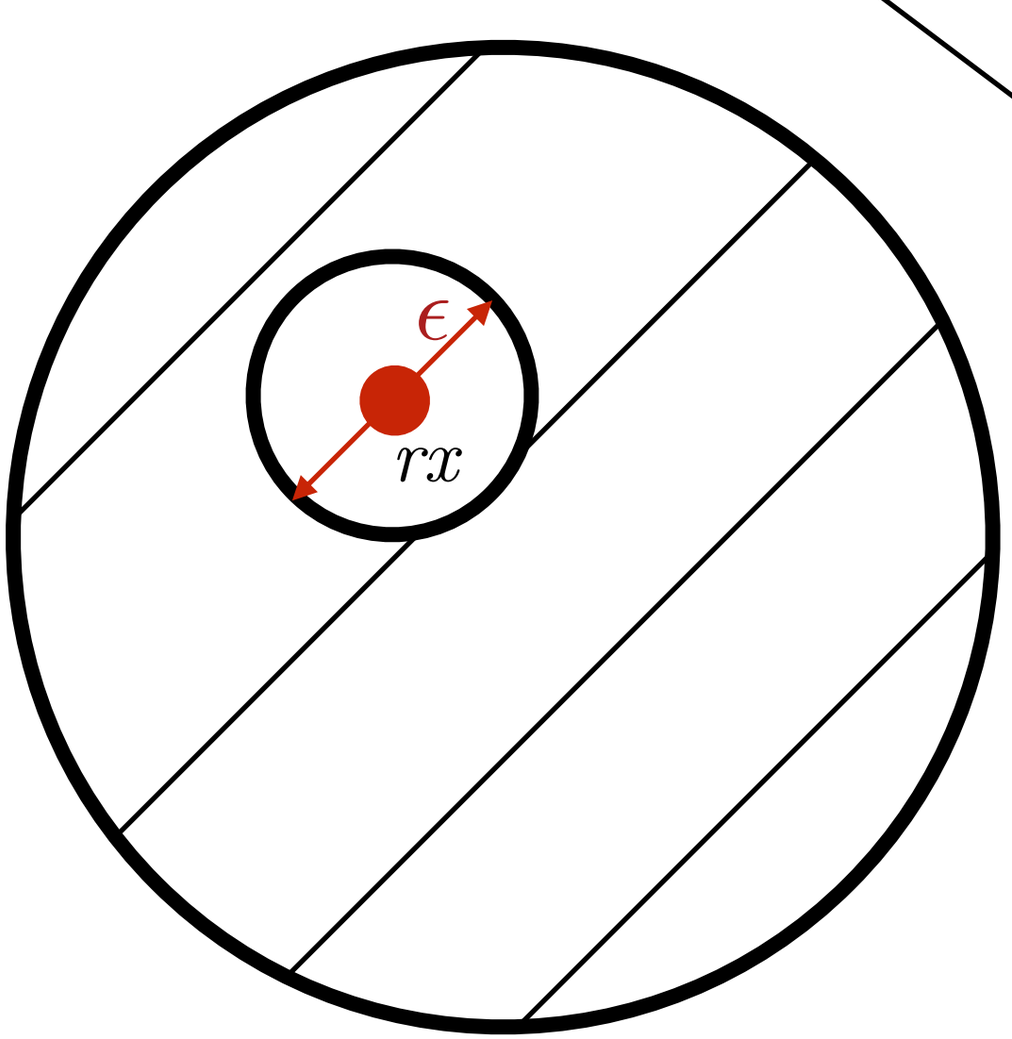
Proof's sketch

- Theorem: If P satisfies $\Delta P = 0$ then for $0 \leq r < 1, \|x\| = 1$

$$P(rx) = \int_{\mathcal{S}^{d-1}} P(y) \frac{1 - r^2}{\|rx - y\|^d} \, d\sigma(y)$$

Proof: Pick a good function u and apply the divergence theorem:

$$\int_{\Omega_\epsilon} u\Delta P + P\Delta u \, d\lambda = \int_{\partial\Omega_\epsilon} P\frac{\partial u}{\partial n} d\sigma - \int_{\partial\Omega_\epsilon} u\frac{\partial P}{\partial n} d\sigma$$



Must cancel

with:

$$u(x, y) = \frac{1}{\|rx - y\|^{d-2}} - \frac{1}{\|ry - x\|^{d-2}}$$

and

$$\Omega_\epsilon = \mathcal{B}(0, 1) \backslash \mathcal{B}(rx, \epsilon)$$

Unit approximation

- If f is continuous, then $\|f \circledast P_r - f\|_\infty \to 0 \quad \text{as} \quad r \to 1$

$$\left|\int_{\mathcal{S}^{d-1}} P_r(x,y)(f(x)-f(y))d\sigma(y)\right| \leq \int_{\mathcal{S}^{d-1}} P_r(x,y)|f(x)-f(y)|d\sigma(y)$$

$$\leq \int_{\|x-y\|\leq\delta} P_r(x,y)|f(x)-f(y)|d\sigma(y)$$

$$+ \int_{\|x-y\|>\delta} P_r(x,y)|f(x)-f(y)|d\sigma(y)$$

Everything is consistent!

- Our convolutions with the Laplacian on a manifold correspond to the convolution with the sphere.

- $SO_d(\mathbb{R})$ isn't commutative as it's invariant subspaces aren't of dimension 1

- "Fourier on the sphere" allows to study the regularity of a given function.

Back to One layer Neural Networks

Shallow Neural Networks

Spherical symmetry

$$\Phi(x;w) = \sum_{k \leq K} \alpha_k \rho(\langle x, w_k \rangle) \quad \text{and let:} \quad w_k = \|w_k\| \theta_k$$

- A neural network of 1-layer writes, for $x \in \mathcal{B}(0, \frac{1}{2})$:

$$\Phi(x;w) = \sum_{k \leq K} \alpha_k \rho(\langle x, w_k \rangle)$$

$$= \sum_{k \leq K} \frac{\alpha_k}{\|w_k\|} \rho(\langle x, \theta_k \rangle)$$

$$= \int_{\mathcal{S}^{d-1}} \rho(\langle x, \theta \rangle) \, d\mu(\theta) \text{ where: } \mu = \sum_{k \leq K} \frac{\alpha_k}{\|w_k\|} \delta_{\theta_k}$$

- Then let $|t| = \sqrt{1 - \|x\|^2}$ and let:

$\tilde{\Phi}((x,t)) \triangleq |t| \Phi(\frac{x}{t}) = \Phi(x)$ if $t > 0$

if Φ is lipschitz/continuous
so is $\tilde{\Phi}$

($\tilde{\Phi}$ is odd)

From $\mathcal{M}$ to L^2

- Instead to consider finite measures $\mu \in \mathcal{M}(\mathcal{S}^{d-1})$, consider as a reference measure the uniform measure and:

$$L^2(\mathcal{S}^{d-1}) = \{ \int_{\mathcal{S}^{d-1}} |p|^2 \, d\sigma < \infty \}$$

and:

$$d\mu = p \, d\sigma$$

- And we study approximation of the type;

$$\forall x \in \mathcal{S}^{d-1}, \quad f(x) \approx \int_{\mathcal{S}^{d-1}} \rho(\langle x, \theta \rangle) p(\theta) \, d\sigma = \rho \circledast p(x)$$

regularity to define

well defined because ρ is smooth bounded

Decomposition of ρ

Remind that: $\rho = \sum_n \lambda_n(\rho) P_n$

- A bit of calculus allows to show that...

$$\lambda_n(\rho) = \frac{\Lambda_{d-1}}{\Lambda_d \dim \mathcal{H}_n^d} \int_{-1}^{1} P_n(t)(1-t^2)^{\frac{d-1}{2}} \rho(t)\, dt$$

- Some algebra (/!\ harsh) show that:

$$\lambda_{2n+3} = 0 \text{ , } \lambda_1 \neq 0, \lambda_2 \neq 0$$

and:

$$\lambda_{2n} \sim C(d)(-1)^{n+1}(2n)^{-\frac{d+3}{2}}$$

- Alose, we remind that for any p integrable:

$$K_n(p \circledast \rho) = K_n(p)\lambda_n(g)$$

$\mathcal{F}_2$ as a vector space

Let $Tp = \rho \circledast p$ which defines an integral operator.

Proposition: $\mathcal{F}_2 = \{\rho \circledast p, p \in L^2(\mathcal{S}^{d-1})\}$ is a vector space with norm

$$\|f\|_{\mathcal{F}}^2 \triangleq \sum_{\lambda_n(\rho)\neq 0} \frac{\|K_n(f)\|^2}{\lambda_n(\rho)^2}$$

This norm is such that: $\|Tp\|_{\mathcal{F}} = \|p\|$ and $f \in \mathcal{F} \Longleftrightarrow \|f\|_{\mathcal{F}} < \infty$

- Example: f odd and $\mathcal{C}^{2k}$ with: $2k \geq \frac{d+5}{2}$.

Approximation guarantees

- Proposition: there is $\delta > 0$ s.t. if f *is* odd, η-Lipschitz, $f(0) = 0$ then there exists $\|g\| \leq \delta$ and $\|g \circledast p - f\|_\infty = \mathcal{O}(\eta \log(\frac{\delta}{\eta})(\frac{\delta}{\eta})^{-\frac{2}{d-3}})$

Proof(2)

Lemma: $0 < r < 1, \alpha > 0$ then: $\sup_{x>0} x^\alpha r^x = \mathcal{O}((1-r)^{-\alpha})$

On the other hand, the norm in $\mathcal{F}_2$ is:

$$\sum_{n\geq 0} \lambda_{2n}^{-2} r^{2n} \|K_n(f)\|^2 \leq \sup_n \lambda_{2n}^{-2} r^{2n} \eta^2$$

$$= \mathcal{O}(\sup_n (2n)^{d+3} r^{2n}) \eta^2$$

$$= \mathcal{O}\left((1-r)^{-d-3} \eta^2\right),$$

Thus if: $(1-r)^{-\frac{d-3}{2}} \eta = \delta$ then, we get the conclusion!

How can we compute f? Random sampling

Proposition:

- Let $f \in L^2(\mathcal{S}^{d-1})$ then there are $v_1, ..., v_n \in \mathcal{S}^{d-1}$ s.t.:

$$\|f \circledast \rho - \sum_{i=1}^{n} \rho(v_i^T .) f(v_i)\| \leq \sqrt{\frac{8\pi d}{n}} \|f\| \sim \sqrt{\frac{8\pi d}{n}} \|f \circledast \rho\|_{\mathcal{F}_2}$$

Proof: random sampling

High norm <-> difficult to sample.

Limits: from $\mathcal{F}_2$ to $\mathcal{F}_1$

- Unfortunately, our bounds are only for the L^2-norm.
- Using finite measures (meaning that $|\mu(\mathcal{S}^{d-1})| < \infty$):

$$|g(z) - \int_{\mathcal{S}^{d-1}} \rho(\langle v, z\rangle)\, d\mu(z)| \leq \epsilon \gamma_1(g)$$
$$\text{with } \#\text{supp}(\mu) = \mathcal{O}(\epsilon^{-\frac{2d}{d+3}})$$

and a norm is given by:

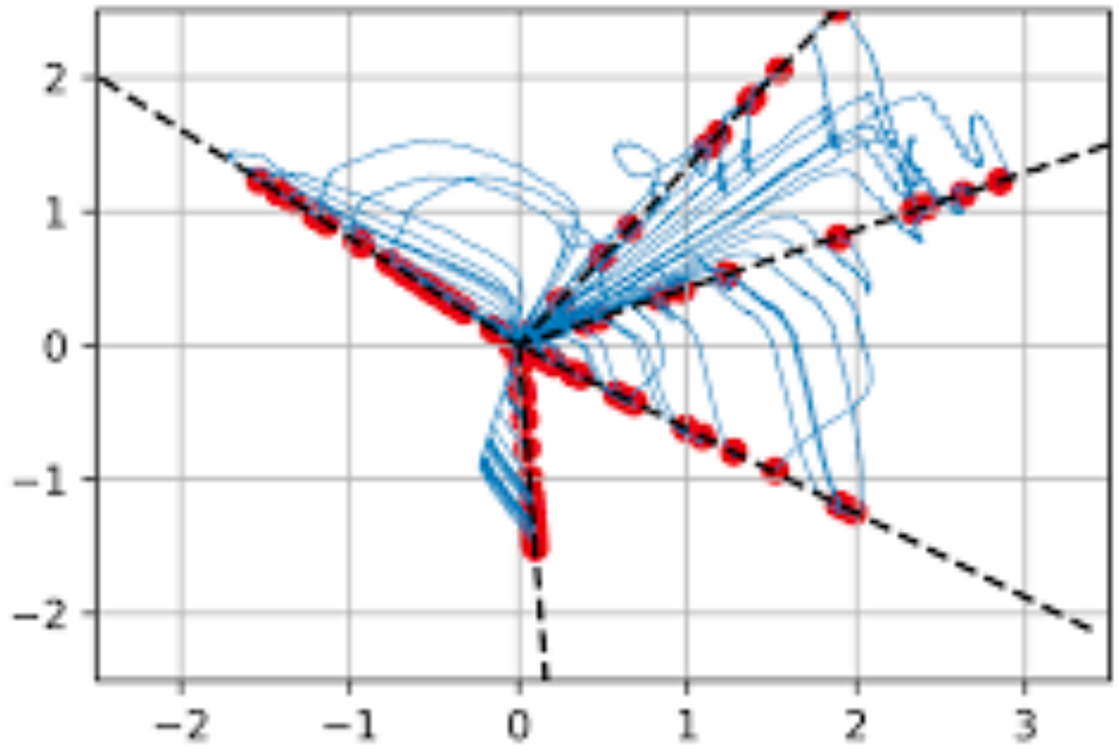
$$\gamma_1(g) = \inf_{g(z)=\int_v \rho(\langle v,z\rangle) d\mu(v)} |\mu(\mathcal{S}^{d-1})|$$

One can show adaptativity to linear structure, convex relaxation (difficult to optimise...)

Optimization of 1-hidden layer NN

- In the mean field limit, one can get convergence guarantees on the flow of an infinite width NN.

pde given by: $\partial_t \mu_t = -\mathrm{div}(v_t \mu_t)$ where v_t depends on the cost function.



Ref.: On the global convergence of gradient descent for over-parameterized models using optimal transport
Chizat and Bach

- Those guarantees are purely asymptotic and seem difficult to extend to deeper NNs.