Advanced Topics in Deep learning:

One Signal Processing view

Draft.

Edouard Oyallon

March 25, 2023

Contents

Chapter 1

Introduction

The goal of these few notes is to present a (personal) overview of Deep Neural Networks, through the lens of Statistical Signal Processing. Neural Networks are difficult models to analyze because they have many parameters and are quite non-linear, both in term of parametrization and embedding. Indeed, they typically consist of a cascade of linear layers followed by pointwise non-linearities, whose output is a fed to a given loss (e.g., supervised classification loss). Deep learning is a subfield of machine learning that aims at studying this type of models. In general, the number of parameters is substantially larger than the dimensionality of the input and each layer is trained in an end-to-end manner via gradient descent.

In both unsupervised or supervised context, a strength of deep neural networks is their ability to adapt the parameters of a model to the specific bias of a dataset, as well as the simplicity of performing a grid search on a large amount of hyper-parameters. This makes their analysis even more difficult, as there are currently no simple model of complex signals like images, sounds or even a game of Go.

This type of black-box phenomenon is not specific to deep architectures and is actually common to all high-dimensional models, without additional assumptions of structure (e.g., sparsity, linearity, ...). This is usually referred as the curse of dimensionality which occurs when dimension d is large, and we can illustrate it by the proposition below, where $\mathcal{B}(a, \rho) = \{x, \|x - a\| \leq \rho\} \subset \mathbb{R}^d$, and we show that covering the unit cube requires an exponential number of $\epsilon-$balls:

Proposition 1.1. *If* $[0, 1]^d \subset \cup_{n \leq N} \mathcal{B}(a_n, \epsilon)$*, then* $N \geq \mathcal{O}(\epsilon^{-d})$*.*

Proof. Let $K = \lfloor \frac{1}{4\epsilon} \rfloor$ and $\mathbf{n} = (n_1, \ldots, n_d)$, and write $f(n_1, ..., n_d) = \frac{\mathbf{n}}{K} = (\frac{n_1}{K}, ..., \frac{n_d}{K})$. Then, we have: $\forall \mathbf{n}, \exists i(\mathbf{n}), \|f(\mathbf{n}) - a_{i(\mathbf{n})}\| \leq \epsilon$. Furthermore, for $\tilde{\mathbf{n}} \neq \mathbf{n}$, we have $\|a_{i(\mathbf{n})} - a_{i(\tilde{\mathbf{n}})}\| \geq 4\epsilon - 2\epsilon = 2\epsilon$ which implies that $i(\tilde{\mathbf{n}}) \neq i(\mathbf{n})$. Thus, we get $N \geq K^d = \mathcal{O}(\epsilon^{-d})$. $\square$

Remark 1.1. *The previous proposition computed the covering number by re-*

lating the ℓ^2-norm to the ℓ^∞-norm. More generally, we can show that on a d-dimensional normed vector space $(E, \|.\|)$ covering the the unit ball via ϵ-balls requires at least $\mathcal{O}(\epsilon^{-d})$ such balls. This is the notion of covering numbers [38], and also related to the definition of Hausdorff dimension.

During these lectures, I would like to introduce several concepts from the theory of Deep Learning through the scope of symmetries and linearization. This document should not be technical but self-contained, and if not given, a proof will be related to a precise reference.

1.1 Reminders about several Hilbert spaces

Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ be a Hilbert space with corresponding norm given for $x \in \mathcal{H}$ by $\|x\|^2 = \langle x, x \rangle$. We call bounded operator (or simply operator) any linear application $T : \mathcal{H}_1 \to \mathcal{H}_2$ between two Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2$, which satisfies $\sup_{x \in \mathcal{H}_1} \|Tx\| \leq M\|x\|$ for some $M \geq 0$, which is equivalent to saying T is continuous. For $x \in \mathcal{H}$, we will write Tx the images of x by T and the composition will be denoted by T_1T_2. We say it has a finite rank if the image $T\mathcal{H}$ of T has finite dimension. T has an inverse if and only if there exists T' s.t. $TT' = T'T = \mathbf{I}$. If T is bounded, its inverse is too. We write $\mathrm{Sp}(T) = \{\lambda, T - \lambda\mathbf{I} \text{ has no inverse}\} \subset \mathbb{C}$, the spectrum of T. Many proofs of those facts can be found in [32].

Definition 1.1. *An operator T is a compact operator if $\overline{T\mathcal{B}(0,1)}$ is compact.*

One simple caracterization of compact operators is as follows:

Proposition 1.2. *An operator T is compact if and only if T is the limit of a sequence of compact operators.*

In particular, we have the following useful spectral theorem:

Proposition 1.3. *If T is compact, then $Sp(T) = \{\lambda_1, ...\}$ is at most countable and $\mathcal{H} = \bigoplus_{n \geq 0}^{\perp} \ker(T - \lambda_n \mathbf{I})$. Furthermore, the eigen-values can only accumulate at 0 and $\ker(T - \lambda_n \mathbf{I})$ has finite dimension if $\lambda_n \neq 0$.*

We refer to [31] for more elements about the theory of compact operators. We further consider $\mathcal{U}(\mathcal{H}) = \{T, T^*T = \mathbf{I}\}$ the unitary isometries of $\mathcal{H}$, which is a group for the composition.

Integral operators We will also use widely the notion of integral operator. For a reference measure μ on a measurable space $\mathcal{X}$, let:

$$\mathcal{L}^2(\mathcal{X}) = \{f, \int_{\mathcal{X}} |f|^2(u)\, d\mu(u) < \infty\}. \tag{1.1}$$

This space is endowed with a Hilbert structure, and we introduce the notion of *integral operator*: they are operators defined via a parametric integral over a *kernel*:

Definition 1.2. *We call $K : \mathcal{L}^2(\mathcal{X}) \to \mathcal{L}^2(\mathcal{X})$ an integral operator if there exists a measurable function, called a kernel, $k(u,t)$ s.t. $\forall f \in L^2(\mathcal{X})$,*

$$Kf = \int_{\mathcal{X}} k(u,t) f(t)\, dt\,. \tag{1.2}$$

The adjoint of K is given by:

$$K^* f = \int_{\mathcal{X}} f(u) \bar{k}(u,t)\, du. \tag{1.3}$$

Thus, the following identity leads to:

$$\forall f, g \in L^2(\mathcal{X}), \langle Kf, g \rangle = \langle f, K^* g \rangle\,. \tag{1.4}$$

It thus leads to:

$$\forall f \in L^2(\mathcal{X}), \|Kf\|^2 = \int_{u,v} w(u,t) \bar{f}(u) f(t)\, dt\, du, \tag{1.5}$$

where

$$w(u,t) = \int_z \bar{k}(u,z) k(t,z)\, dz\,. \tag{1.6}$$

Note the (expected) hermitian symmetry of the previous kernel: $\bar{w}(u,t) = w(t,u)$. We also recall the following lemma.

Lemma 1.1 (Minkowski integral inequality)**.** *Given two measurable spaces $(X,\mu), (Y,\nu)$ and a measurable function $f : X \times Y \to \mathbb{R}$ we have the inequality:*

$$\left(\int_X | \int_Y f(x,y) d\nu(y)|^2 d\mu(x) \right)^{\frac{1}{2}} \leq \int_Y \left(\int_X |f(x,y)|^2 d\mu(x) \right)^{\frac{1}{2}} d\nu(y)\,. \tag{1.7}$$

and the following lemma allows to compute the norm of T via a bound on its kernel:

Lemma 1.2 (Schur Lemma)**.** *Let K an integral operator with kernel k, then if*

$$\forall v, \int_{\mathbb{R}^d} |k(u,v)|\, du \leq C_1 \text{ and } \forall u, \int_{\mathbb{R}^d} |k(u,v)|\, dv \leq C_2\,, \tag{1.8}$$

we have $\|K\| \leq \sqrt{C_1 C_2}$.

Symmetries Following the approach of [27], fix for instance $f : \mathcal{X} \to \mathbb{R}$. We say that $\mathcal{L} : \mathcal{X} \to \mathcal{X}$ is a symmetry of f if $\mathcal{L}$ is invertible and if:

$$f(\mathcal{L}x) = f(x), \forall x \in \mathcal{X}\,. \tag{1.9}$$

We note that the set $\mathcal{S}ym(f)$ of symmetries $\mathcal{L}$ of f is a group for the composition. We will give a more refined use of the symmetry group for translation invariance in Chapter 2, in order to study invariant signal representations. Symmetries are

potentially difficult to exhibit when there is no clear model of f. Note that if we allow symmetries to permute to x, y, then they characterize the level sets of f, as, by construction:

$$f^{-1}(\{f(x)\}) = \{\mathcal{L}x, \mathcal{L} \in \mathcal{S}ym(f)\}. \tag{1.10}$$

In the case of signals obtain from a physical process and with additional assumptions, symmetries are one of the major analysis tool (e.g., Gauge theory [28]).

Linearization Another mechanism corresponds to the notion of linearization. In particular, one has the following standard result:

Theorem 1.1 (Rademacher theorem)**.** *Let $\mathcal{K} \subset \mathbb{R}^d$ an open set, if $f : \mathcal{K} \to \mathbb{R}$ is locally Lipschitz, then f is almost everywhere differentiable, meaning. In other words, that for almost all $x, h \in \mathcal{K}$, there exists a (bounded) operator L_x, such that*

$$f(x+h) = f(x) + L_x h + o(\|h\|)\,. \tag{1.11}$$

Note that locally Lipschitz on a compact implies globally Lipschitz on this same compact.

In particular, we obtain a local linearization of f. This is of interest, as linear structures are simpler to handle: in particular the case of a low rank L_x allows to substantially reduce the ambient dimension.

1.2 Structure of this document

This document is divided into 5 chapters. Chapter 2 gives a framework to analyze convolutions on groups and introduces the Scattering Transform [26]. Chapter 3 describes Graph Signal Processing techniques [19], and in particular the analogy with a Laplacian: the Laplacian on the sphere is discussed in particular. Then, Chapter 4 proposes several approximation and estimation bounds, following [3], that are easily obtained thanks to several tools that we will have developed during the previous chapters. Chapter 5 discusses training neural networks and a particular regime, the lazy regime [13]. Following the works of Bartlett [6, 5] in the specific case of ReLU Neural Networks, we derive in Chapter 6 several complexity bounds of Deep Neural Networks based on standard complexity measures.

Chapter 2

Group Invariance in Neural Networks

In this section, we will be mainly studying how the formalism of groups can shed light on Neural Network mechanisms. Consider a function $\varphi : \mathcal{X} \to \mathcal{Y}$. We say that it is *equivariant* to $\mathcal{L} : \mathcal{X} \to \mathcal{X}$ if there exists $\mathcal{L}' : \mathcal{Y} \to \mathcal{Y}$ such that

$$\forall x \in \mathcal{X}, \varphi(\mathcal{L}x) = \mathcal{L}'\varphi(x).$$

We say that it is *invariant* if $\mathcal{L}' = \mathbf{I}$. If the spaces $\mathcal{X}$ and $\mathcal{Y}$ resemble each other in a straight-forward way, e.g. $\mathcal{Y} = \mathcal{X}$ or $\mathcal{Y}$ is several copies of $\mathcal{X}$, then we also say that $\mathcal{L}$ is *covariant* if it is not invariant.[1]

Remark that this property is stable under composition. We note that invariance can be easily achieved from covariant operators and via single invariant operator, as we have the following proposition:

Proposition 2.1. *If $\varphi_1, ..., \varphi_{J-1}$ are covariant to $\mathcal{L}$ and φ_J is invariant to $\mathcal{L}$, then $\Phi = \varphi_J \circ ... \circ \varphi_1$ is invariant to $\mathcal{L}$.*

Proof. It follows from the definition, as:

$$\Phi\mathcal{L}x = \varphi_J(\mathcal{L}\varphi_{J-1} \circ ... \circ \varphi_1(x)) = \Phi x\,. \tag{2.1}$$

□

Remark 2.1. *One can notice the analogy with a Convolutional Neural Network [23], that has typically this structure, where we can find $\mathcal{L}$ and $\mathcal{L}'$ which typically perform different numbers of copies of the same operation to move every channel.*

[1] In order to distinguish operators from standard functions, we might often omit the parenthesis for readability reasons.

2.1 Reminder on Groups

At minimum, we consider here a group G with a metric d which is locally compact, which means that each point admits a compact neighborhood. Let U be a space on which G can act. For a function $x \in L^2(U)$, the action $\mathcal{L}_g x$ is defined as $u \mapsto x(g^{-1}u)$. Notably, if $U = G$, We denote $L^p(G) = \{f, \int_G |f|^p < \infty\}$. We write $\mathcal{L}_g x(g') \triangleq x(g^{-1}g')$ the left action of $g \in G$ on $L^2(G)$, and we consider left and right invariant measure μ, which means that for A measurable and any $g \in G$, $\mu(A\mathcal{L}_g) = \mu(\mathcal{L}_g A) = \mu(A)$ (this also referred as unimodularity), see [4] for more details.

2.1.1 Convolution along a Group

Definition 2.1. *For $a \in L^1(G), b \in L^1(U)$, we define the convolution along the group G by:*

$$a \star b(u) \triangleq \int_G a(g)b(g^{-1}u)\, d\mu(g)\,. \tag{2.2}$$

In particular, if $U = G$ and for $a, b \in L^1(G)$, we define the convolution along the group G by:

$$a \star b(g) \triangleq \int_G a(g')b(g'^{-1}g)\, d\mu(g')\,. \tag{2.3}$$

Furthermore, $a \star b \in L^1(G)$.

We recall also the norm inequalities for convolution, given by:

Lemma 2.1 (Young's inequality for convolution)**.** *Let $p, q, r \geq 1$ and $a \in L^p(G, \mu), b \in L^q(G, \mu)$ s.t. $\frac{1}{r} + 1 = \frac{1}{q} + \frac{1}{p}$, then:*

$$\|a \star b\|_r \leq \|a\|_p \|b\|_q\,.$$

Remark 2.2. *If $a \in L^2(G)$ and $b \in L^1(G)$, then by Young's inequality, we have $a \star b \in L^2(G)$.*

We will sometimes use the following lemma, which allows to obtain an "identity" element for the convolution:

Lemma 2.2 (Approximation to the identity)**.** *There exists a sequence $\delta_n \in L^1(G)$, compactly supported, $\int \delta_n d\mu = 1$, $\delta_n \geq 0$ such that $\delta_n \star a \to a, \forall a \in L^2(G)$.*
Alternatively, we also have $\delta_n \star a \to a, \forall a \in L^1(G)$.

Proof. For a neighborhood $\mathcal{V}_e$ of size ϵ_n arbitrary small around e, we consider

$\delta_n(g) = \frac{1_{\mathcal{V}_e}(g)}{\mu(\mathcal{V}_e)}$. Then, $\int \delta_n d\mu = 1$ and using the Lemma 1.1:

$$\begin{aligned}
\int_G |\delta_n \star a(g) - a(g)|^2 d\mu(g) &= \int_G |\int_G \delta_n(g')(a(g) - a(g'^{-1}g))d\mu(g')|^2 d\mu(g) \\
&\leq \int_G \sqrt{\int_G \delta_n(g')^2 |a(g) - a(g'^{-1}g)|^2 d\mu(g) d\mu(g')} \\
&= \int_G \delta_n(g') \|a - L_{g'}a\| d\mu(g')
\end{aligned}$$

Now since $a \in L^2, \|a - L_{g'}a\| < \epsilon$ (for instance via density and dominated convergence theorem) for $g' \in \mathcal{V}'_e$. We pick ϵ_n such that $\mathcal{V}_e \subset \mathcal{V}'_e$, and obtain the conclusion. □

Remark 2.3. *We note that in the case where $G = \mathbb{R}^d$ with an additive law, we can pick $\hat{\delta}_n(\omega) = e^{-\frac{\|\omega\|^2}{2n^2}}$, where the Fourier Transform is discussed in Sec. 2.1.2, which satisfies $\int \delta_n d\mu = 1$ and $a \star \delta_n \to a$ in $L^2(\mathbb{R}^d)$ thanks to Parseval's theorem.*

We then can characterize the operators that commute with the action of translation: they correspond roughly to convolutions, as (and many variants of this proposition can be obtained, in weaker sens):

Proposition 2.2. *Let $W : L^1(G) \to L^1(G)$, be a continuous operator. Then one has:*

1. $\forall g, \mathcal{L}_g W = W \mathcal{L}_g$,
2. $\exists f \in L^1(G) : W\delta_n \to f \in L^1$,

if and only if $\exists f \in L^1(G) : \forall x \in L^1(G), Wx = x \star f$.

Proof. Convolutions clearly satisfy (2). For the other direction, note that, by linearity and continuity, for $x \in L^1(G)$, we get:

$$W(x \star \delta_n) = W\left(\int_G x(g')\mathcal{L}_{g'}\delta_n d\mu(g')\right) \tag{2.4}$$

$$= \int_G x(g') W \mathcal{L}_{g'} \delta_n d\mu(g') \tag{2.5}$$

$$= \int_G x(g') \mathcal{L}_{g'} W \delta_n d\mu(g') \tag{2.6}$$

$$= x \star W\delta_n \,. \tag{2.7}$$

Then, from Young's inequality, we get:

$$\|x \star W\delta_n - x \star f\|_1 \leq \|x\|_1 \|W\delta_n - f\|_1 \tag{2.8}$$

leading to the conclusion. □

Remark 2.4. *Those concepts can be extended to L^1 and L^2 via distributions and duality in $C^0(G)$, see e.g. [20, 11].*

Remark 2.5. *Note an application to this Lemma to signals $L^2(\mathbb{R}^d) \times \Lambda$, where the translation is given by:*

$$\forall \lambda \in \Lambda, \mathcal{L}_a x(u, \lambda) \triangleq x(u - a, \lambda) \,. \tag{2.9}$$

This can be extended to standard Convolutional linear layers for Neural Networks. If $W : L^2(\mathbb{R}^d) \times \Lambda \to L^2(\mathbb{R}^d) \times \Lambda'$ and $W\mathcal{L}_a = \mathcal{L}_a W$, the previous Lemma gives the existence of $k_{\lambda,\lambda'}$ s.t. $Wx = \{x \star \sum_{\lambda \in \Lambda} k_{\lambda,\lambda'}\}_{\lambda' \in \Lambda'}$.

2.1.2 Fourier analysis on groups

In this subsection, we develop some tools that allow to characterize and better manipulate the convolution operators on a group. To do so, we consider the action of G on the isometries of $L^2(G)$, which are called (unitary) representation. Non-unitary representation is a wider theory, yet this is beyond the scope of of this class. Ideally, our objective is to decompose $L^2(G)$ in subspaces which will be stable by convolutional operators. In the following, all representations will be unitary.

Definition 2.2. *Let $\mathcal{H}$ be a Hilbert space and a group G. We call a* ***unitary representation*** *any continuous group morphism $\rho : G \to \mathcal{U}(\mathcal{H})$.*

We only consider groups for which such a representation exists, and [18] shows its existence for compact groups. We will be mainly interested in:

$$\begin{aligned} G &\to \mathcal{U}(L^2(G)) \\ g &\to \mathcal{L}_g \,, \end{aligned} \tag{2.10}$$

where $\mathcal{L}_g : L^2(G) \to L^2(G)$ is the translation of a signal by g. Note that the direct sum $\rho_1 \oplus \rho_2 : G \to \mathcal{U}(\mathcal{H}_1 \oplus \mathcal{H}_2)$ of two representations $\rho_1 : G \to \mathcal{U}(\mathcal{H}_1), \rho_2 : G \to \mathcal{U}(\mathcal{H}_2)$ is itself a representation. Note that a representation on $\mathcal{H}$ induces canonically a representation on a subspace of $\mathcal{H}' \subset \mathcal{H}$, by restriction of $\rho(g)_{|\mathcal{H}'}$. The group quotient of representations is also a representation, yet we do not introduce this notion because we will not use it.

Since $\mathcal{L}_g$ commutes with the convolutions, analyzing the invariant subspaces of a unitary representation allows to obtain the invariant subspaces of a convolution: those spaces are linked to a notion of frequencies. Let us describe them further.

Definition 2.3. *A closed subspace $F \subset \mathcal{H}$ is invariant w.r.t. ρ, if:*

$$\forall g, \rho(g) F \subset F \,. \tag{2.11}$$

Proposition 2.3. *Assume ρ is a unitary representation. Then, F is an invariant subspace of ρ if and only if $F^\perp$ is an invariant subspace of ρ.*

Proof. Indeed, for $g \in G$, if F is stable by $\rho(g)$, then, as $\rho(g)$ is adjoint, then $F^\perp$ is stable by $\rho(g)$. This is true for all $g \in G$, $F^\perp$ is invariant. □

For $\mathcal{K} \subset \mathcal{H}$ an invariant subspace of ρ, we write $\forall k \in \mathcal{K}, \rho_{|\mathcal{K}}(g)k = \rho(g)k$ the restriction of $\rho(g)$ to $\mathcal{K}$. We will also need:

Definition 2.4 (Irreducible representation)**.** *A representation $\rho : G \to \mathcal{U}(\mathcal{H})$ is irreducible if its only invariant subspaces are $\mathcal{H}$ and $\{0\}$.*
A representation ρ is said to be completely reducible, if $\mathcal{H} = \bigoplus_{n\in\mathbb{N}}^{\perp} \mathcal{H}_n$ and $\rho_{|\mathcal{H}_n}$ is irreducible for all n.

Proposition 2.4 (Irreducible representation in finite dimension)**.** *If ρ is a finite dimensional representation, then ρ is completely reducible.*

Proof. The proof can be obtained by induction using Prop. 2.3. □

This applies in particular on representations $G \to \mathcal{U}_n$, the group of the unitary matrix on $\mathbb{C}^n$, and for instance, if $G = \mathbb{R}$, this leads to the concept of one parameter group of $GL_n(\mathbb{C})$, generated by $\mathfrak{h} \in M_n(\mathbb{C})$, acting on $x \in \mathbb{C}^n$ via

$$t \to e^{t\mathfrak{h}} x \,. \tag{2.12}$$

If $e^{t\mathfrak{h}}$ is unitary, then it is a normal operator and its spectrum are elements of modulus 1: $\mathfrak{h}$ has its spectrum in $\mathbf{j}\mathbb{C}$ and thus $\mathfrak{h}^* = -\mathfrak{h}$.

Representations are of high interest when studying convolutions, because if $Wx = x \star \psi$, $\psi \in L^1(G)$ and ρ is a representation that acts on $\mathcal{U}(L^2(G))$ via translation as in Eq. (2.10), we note that a representation and a convolution commutes:

$$\forall g, \rho(g)W = W\rho(g) \,. \tag{2.13}$$

Consequently, if F is a characteristic space of W (which is stable by W), then it is stable by $\rho(g)$. **Thus, we can study convolutions on potentially smaller subspaces.** An ideal setting is of course when those spaces are of dimension 1: they are then the eigen-spaces of W. We now describe those subspaces for various type of groups.

The Euclidean case, $G = \mathbb{R}^n$

We consider the standard Fourier Transform $\mathcal{F} : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$, defined by:

$$\mathcal{F}x(\omega) \triangleq \int_{\mathbb{R}^n} x(u) e^{-\mathbf{j}\omega^T u} \, du \,. \tag{2.14}$$

We also write a character (which is a complex morphism $G \to \mathbb{C}$):

$$\chi_\omega(u) \triangleq e^{\mathbf{j}\omega^T u} \,. \tag{2.15}$$

Note that $\chi_\omega \notin L^2(\mathbb{R}^n)$. It is well known that the Fourier Transform is an isometry of $L^2(\mathbb{R}^n)$ [35]. We observe that, for all $a \in \mathbb{R}^n$, Fourier transforms the translation into a phase multiplication, as:

$$\mathcal{F}\mathcal{L}_a x(\omega) = e^{-\mathbf{j}a^T\omega} \mathcal{F}x(\omega) = \chi_a \mathcal{F}x(\omega). \tag{2.16}$$

In the following, we write $E_\omega = \text{vect}(\chi_\omega)$. Note that a Fourier transform is neither irreducible, neither contains an invariant subspace, as:

Proposition 2.5. *$\rho : a \to (u \to \chi_a(u)x(u))$ is a unitary representation, unitarily equivalent to $a \to \mathcal{L}_a$, which is neither irreducible nor has any invariant subspaces.*

Proof. Following the Remark 2.8 (which is proven in [35]), as $\mathbb{R}^d$ is a commutative group, such an invariant subspace has to be of dimension 1. Given that $\dim L^2(\mathbb{R}^n) = \infty$ and $\mathbb{R}^n$ is commutative, then, ρ is not irreducible. Assume that one can obtain an invariant subspace F, then it is of dimension 1 by commutativity, and $\exists x \in L^2(\mathbb{R}^n), x \neq 0, \forall a, x(\omega)\chi_a(\omega) = \lambda x(\omega)$. Then, $|\lambda| = 1$, and applying $\mathcal{F}^*$ gives $|\hat{x}(u+a)| = |\hat{x}(u)|, \forall a$, thus $|\hat{x}|$ is constant which is not in $L^2(\mathbb{R}^n)$. □

In order to get a representation which is an "uncountable" sum of irreducible representations, we would need to use the notion of direct integral, which can be understood as a squared integrable section of $\cup_\omega E_\omega$, that writes:

$$L^2(\mathbb{R}^n) = \int_x \bigoplus_\omega E_\omega(x)\,.$$

For the sake of simplicity and concision, we do not introduce those tools here but we refer the reader to [35].

Remark 2.6. *It is possible to consider only signals with same compact support (without losing in generality, $[0,T]^d$), and to use Shannon theorem to embed them on the Torus: indeed, if so, to a compactly supported signal, we can associate a T-periodic signal along each of the d dimensions. See [25].*

The Abelian case, finite dimensional case.

In this subsection, we assume that $\dim \mathcal{H} < \infty$. This assumption is due to Prop. 2.5, which shows unitary representations of infinite dimensional space can not be always semi-simple. Again, our goal will be to decompose $\mathcal{H}$ in a sum of smaller subspaces which are invariant. In the Abelian case (i.e., commutative case), invariant subspaces of unitary representations are of dimension 1. In the finite dimensional case (i.e., $\dim \mathcal{H} < \infty$), it implies that any unitary representation is diagonalizable in an orthogonal basis. We remind the Schur lemma, which is an important tool of representation theory:

Lemma 2.3 (Schur lemma)**.** *Let $\rho : G \to \mathcal{U}(\mathcal{H})$ be a finite dimensional irreducible representation. Consider L an operator of $\mathcal{H}$ such that:*

$$\forall g \in G, \rho(g)L = L\rho(g)\,, \tag{2.17}$$

then, $\exists \lambda \in \mathbb{C}$ such that $L = \lambda \mathbf{I}$.

Proof. Here, let λ an eigenspace of L, $\dim \ker(L - \lambda\mathbf{I}) > 0$, which is stable by $\rho(g)$ for all g, thus $L = \lambda\mathbf{I}$. □

Remark 2.7. *As noted in [36], if we allow dim $\mathcal{H} = \infty$, we will get that $L = \lambda W, W \in \mathcal{U}(\mathcal{H})$.*

Proposition 2.6. *Assume again ρ is finite dimensional. If G is commutative and $\rho : G \to \mathcal{U}(\mathcal{H})$ is an unitary irreducible representation, then dim $\mathcal{H} \leq 1$.*

Proof. Indeed, if dim $\mathcal{H} > 0$, then, let $g' \in G$ then $\forall g, \rho(g')\rho(g) = \rho(g)\rho(g')$. From the Lemma above, $\rho(g') = \lambda_{g'}\mathbf{I}$. Then, clearly every subspace of dimension larger than 1 is stable by $\rho(g)$ for any g, thus dim $\mathcal{H} = 1$. □

We can now conclude:

Proposition 2.7. *Assume every invariant subspace of $\rho : G \to \mathcal{U}(\mathcal{H})$ is finite dimensional and that G is commutative, then $\mathcal{H} = \bigoplus_{n\in\mathbb{N}} \mathcal{H}_n$, where $\mathcal{H}_n$ is of dimension 1.*

Proof. This a direct conclusion of Prop. 2.6 and Prop. 2.4. □

Remark 2.8. *In infinite dimension, the existence of invariant subspaces is not guarantee, however, we have the following: let F an invariant and irreducible subspace of a representation $\rho : G \to \mathcal{U}(\mathcal{H})$ (without any assumptions on the dimension of $\mathcal{H}$) over an abelian group G. Then, dim $F = 1$. The proof requires the Spectral theorem for self-adjoint operators.*

Proof. If we fix $g_0 \in G$, then F is stable by $\rho(g_0)$ and $F^\perp$ is stable by $\rho(g_0)$. □

Example 2.1. *Consider $\mathcal{S}^1 = [0, 2\pi]$. Then, for any $\epsilon > 0$, it is easy to build a cyclic sub-group G such that $d(\mathcal{S}^1, G) < \epsilon$, as such a sub-group is isomorph to $\mathbb{Z}/n\mathbb{Z}$. Note that this construction is not true anymore for $\mathcal{S}^n, n \geq 2$ []: there exists no finite sub-groups that approximate the sphere with precision ϵ. Otherwise, it is possible to note that any $f \in L^2(\mathcal{S}^1)$ can be written:*

$$f(\theta) = \sum_n c_n e^{in\theta} \,. \tag{2.18}$$

and in this case we have the invariant (irreducible) subspaces $\mathcal{H}_n = Vect(\theta \to e^{in\theta})$.

The compact case

Now, we will briefly discuss the case of compact groups. In particular, this englobles finite groups, commutative or not, and the case of compact groups.

Theorem 2.1 (Peter-Weyl theorem)**.** *Let G be a compact group and $\rho : G \to \mathcal{U}(\mathcal{H})$ is a unitary representation, then $\mathcal{H} = \bigoplus_{n>0} \mathcal{H}_n$, where each $\mathcal{H}_n$ is a finite dimensional irreducible subspace of ρ.*

Proof. Here, the difficult part is to show the finite dimensional aspect. Fix $\|w\| = 1$ and:

$$Th = \int_G \langle \rho(g)w, h\rangle \rho(g)w \, d\mu(g) \tag{2.19}$$

Then, $T^* = T$ (by unimodularity), $\rho(g)$ and T commute for all g (by construction), $\forall h, \langle h, Th\rangle \geq 0$. The difficult part is to show that T is compact, and for this, we will use this Lemma:

Lemma 2.4. *For $\delta > 0$, there there exists $g_1, ..., g_n$ s.t. G is a disjoint union of S_i, $g_i \in S_i$ and $d(g_i, g) < \delta, \forall g \in S_i$.*

Proof. By compacity, consider a finite covering $K_1, ..., K_n$ of diameter at most ϵ, and we consider $S_i = K_i \setminus \cup_{j \leq i} K_j$. □

Now, we fix $h \in \mathcal{H}, \|h\| \leq 1$, and $\epsilon > 0$. Then, ρ is continuous on a G, thus ρ is unformly continuous on G, thus $\|\rho(g) - \rho(g')\| \leq \epsilon$ if $d(g, g') \leq \delta$. Then, we have:

$$\begin{aligned}
&\|\sum_{i \leq n} \mu(S_i)\langle \rho(g_i)w, h\rangle \rho(g_i)w - Th\| \\
&= \|\sum_{i \leq n} \int_G 1_{g \in S_i} \left(\langle \rho(g)w, h\rangle \rho(g)w - \langle \rho(g_i)w, h\rangle \rho(g_i)w \right) d\mu(g)\| .
\end{aligned}$$

Now, we note that:

$$\begin{aligned}
&\|\langle \rho(g)w, h\rangle \rho(g)w - \langle \rho(g_i)w, h\rangle \rho(g_i)w\| \\
&\leq |\langle \rho(g)w - \rho(g_i)w, h\rangle| \|\rho(g)w\| + |\langle \rho(g_i)w, h\rangle| \|\rho(g)w - \rho(g_i)w\| \\
&\leq 2\epsilon
\end{aligned}$$

Observe then that $h \to \sum_{i \leq n} \mu(S_i)\langle \rho(g_i)w, h\rangle \rho(g_i)w$ is of finite rank. Thus, $\tilde{T}$ is compact as a limit of finite rank operator. Then, if $Th = 0$, then $\langle h, Th\rangle = 0, \forall i \leq k$ and $\int_G |\langle \rho(g)w, h\rangle|^2 = 0$, thus $\langle \rho(g)w, h\rangle = 0$ (by continuity), thus $\langle w, h\rangle = 0$. Consequently, if $\{w_n\}_{n \in \mathbb{N}}$ is a Hilbert basis, set $T = \sum_{n \in \mathbb{N}} \frac{1}{n^2} T_n$, where $T_n = \int_G \langle \rho(g)w_n, h\rangle \rho(g)w_n \, d\mu(g)$. As a limit of compact operators, this operator is compact. Furthermore, $Th = 0$ by positivity implies that $\forall n, T_n h = 0$ and thus $\langle w_n, h\rangle = 0$. Thus, $h = 0$, by definition of a Hilbert basis.

Now to conclude, we observe that the characteristic subspaces (of finite dimension) of T previously exhibited are stable by $\rho(g)$ for all g. Thus, they contain the irreducible subspaces of $\rho(g)$, and 0 is not an eigen value: we have the conclusion. □

Example 2.2. *Consider $SO_3(\mathbb{R})$, and we consider $\rho : SO_3(\mathbb{R}) \to L^2(\mathcal{S}^2)$, the representation of rotation on the sphere. For such groups, we will see in the next chapter that any $L^2(\mathcal{S}^2) = \bigoplus_{n \geq 0} \mathcal{H}_n^d$ where $\mathcal{H}_n^d$ is the homogeneous harmonic polynomials of degree d, and it can be shown that those subspaces are irreducible [1].*

Product and semi-direct products: an example, the roto-translation

Via $\rho((g,g')) = (\rho(g), \rho(g'))$, it is clear that the representations $\rho_1 : G \to \mathcal{U}$ and $\rho_2 : G' \to \mathcal{U}'$ induce some representations $G \times G'$ on $\mathcal{U} \times \mathcal{U}'$. We remind here the notion of semi-product of group. Let G, G' two subgroups of a same group H, we say that $G \times G'$ has a semi-direct product structure (and we write $G \ltimes G'$) if it can be seen as a subgroup of H and that G' acts by conjugation on G via the inner operation:

$$\forall g_1, g_2 \in G, g_1', g_2' \in G' (g_1, g_1')(g_2, g_2') = (g_1(g_1' g_2 g_1'^{-1}), g_1' g_2') \,. \tag{2.20}$$

This law is in general not commutative. In this section, we will mainly focus on the group of roto-translation (or rigid-motion), given by $SL(E) = SO_2 \ltimes \mathbb{R}^2$, where we have, for $g = (u, \theta) = \mathcal{L}_u r_\theta, g' = (v, \varphi) = \mathcal{L}_v r_\varphi$:

$$g.g' = \mathcal{L}_u r_\theta \mathcal{L}_v r_\varphi = \mathcal{L}_u r_\theta \mathcal{L}_v r_{-\theta} r_\theta r_\varphi = (u + r_\theta v, \theta + \varphi) \,. \tag{2.21}$$

(because by conjugation, $r_\theta \mathcal{L}_v r_{-\theta} = \mathcal{L}_{r_\theta v}$). Then, we note there that for $x \in L^2(SL(E))$, defining the separable Fourier transform:

$$\tilde{\mathcal{F}} x(\omega, n) = \int_{[0,2\pi]} \int_{\mathbb{R}^2} x(u, \theta) e^{-i\omega^T u - in\theta} \, du d\theta \,, \tag{2.22}$$

and we see that $\tilde{\mathcal{F}} \mathcal{L}_g = \mathcal{L}_g \tilde{\mathcal{F}}$ commutes with the action of $SL(E)$. Now, we also have:

$$x(u, \theta) = \sum_{n \in \mathbb{Z}} \Big(\int_{[0,2\pi]} x(u, \theta') e^{-in\theta'} \, d\theta' \Big) e^{in\theta} \tag{2.23}$$

$$= \sum_{n \in \mathbb{Z}} \int_{\mathbb{R}^2} \tilde{\mathcal{F}} x(\omega, n) e^{i\omega^T u + in\theta} \, d\omega \,. \tag{2.24}$$

Applying $g' = (u', \theta')$ on x leads to:

$$\mathcal{L}_{g'} x(u, \theta) = \sum_{n \in \mathbb{Z}} \int_{\mathbb{R}^2} \tilde{\mathcal{F}} x(\omega, n) e^{i\omega^T r_{\theta'} u - i\omega^T u' + in\theta - in\theta'} \, d\omega \tag{2.25}$$

$$= \sum_{n \in \mathbb{Z}} \int_0^{+\infty} \int_{v \in \mathcal{S}^1} \tilde{\mathcal{F}} x(\rho v, n_\theta) e^{i\rho v^T r_{\theta'} u - i\rho v^T u' + in\theta - in\theta'} \, \rho \, d\rho dv \,. \tag{2.26}$$

We note that the right term $\mathcal{H}_{\rho,n} = \{(u, \theta) \to e^{i\rho v^T u + in\theta}, v \in \mathcal{S}^1\}$ is an invariant subspace by the action of $SL(E)$. As noted in Sec. 2.1.2, we would need the concept of section to formalize better this concept, and we leave those technical details to the reader and refer him to [35].

2.1.3 Invariance along $\mathbb{R}^n$

Linear invariance

In this subsection, we will focus on linear invariance in $\mathbb{R}^n$.

Proposition 2.8. *A continuous linear operator $A : L^1(\mathbb{R}^n) \to \mathbb{R}$ is invariant to the group of translations on $\mathbb{R}^n$ if and only if $\exists \lambda, Ax = \lambda \int x$.*

Proof. From Lemma 2.2, let $\delta_n \geq 0, \int \delta_n = 1$ approximating the Dirac distribution as in Lemma 2.2, then, by continuity:

$$A(\delta_n \star x) = A(\int \delta_n(. - u)x(u)du) \tag{2.27}$$

$$= \int x(u)A\mathcal{L}_u\delta_n du \tag{2.28}$$

$$= A\delta_n \int x(u)du\,, \tag{2.29}$$

and by continuity, the left term converges to Ax and it implies that $A\delta_n$ converges to some real value, as this is true for any $x \in L^1(\mathbb{R}^n)$. □

Non-linear demodulation via complex wavelets

We now discuss the demodulation effect linked to the complex envelop of an analytic signal. Analytic signals are by definition the signals which have their Fourier transform equal to 0 on half of the frequency space, i.e. $\exists e : \forall \omega, \omega^T e \leq 0, \hat{\psi}(\omega) = 0$. It can be proved that they help to demodulate signals [29]. We discuss a strategy to obtain an approximativaly analytic signal. Consider a well localized low-pass filter θ, real, symmetric around 0 and with most of its energy concentrated in this neighborhood. Write $\hat{\psi}(\omega) = \hat{\theta}(\omega - \omega_0)$ for some frequency ω_0. An informal computation shows that:

$$\begin{aligned}
\widehat{(\mathcal{L}_a x) \star \psi}(\omega) &= e^{i\omega^T a}\hat{x}(\omega)\hat{\psi}(\omega) \\
&= \sum_{n\geq 0} \hat{x}(\omega)\frac{(i\omega^T a)^n}{n!}\hat{\psi}(\omega) \\
&\approx \sum_{n\geq 0} \hat{x}(\omega)\frac{(i\omega_0^T a)^n}{n!}\hat{\psi}(\omega) \\
&= e^{i\omega_0^T a}\widehat{x \star \psi}(\omega)\,.
\end{aligned}$$

We note that we implicitly used the fact that the infinitesimal generator of the translation is the derivation. Applying a modulus thus allows to obtain smoother coefficients. This demodulation strategy can be actually quantified through the following Lemma, that can be found in [26]:

Lemma 2.5 (Demodulation)**.** *Assume $\phi \geq 0$, then for any $x \in L^2(\mathbb{R}^d)$ and $u \in \mathbb{R}^d$:*

$$|x \star \psi| \star \phi(u) \geq \sup_{\eta \in \mathbb{R}^d} |x \star \psi \star \phi^{\langle \eta}|(u)\,, \tag{2.30}$$

where $\phi^{\langle\eta}(u) = e^{iu^T\eta}\phi(u)$.

Proof. Indeed, for $\eta \in \mathbb{R}^d$, we get for a given u:

$$\begin{aligned}
&|\iint x(v)\psi(w-v)\phi(u-w)e^{i(u-w)^T\eta}\,dwdv| \\
&\leq \int |\int x(v)\psi(w-v)dv|\phi(u-w)\,dw\,.
\end{aligned}$$

□

Stability to deformations

Here, we introduce for $\tau \in \mathcal{C}^\infty(\mathbb{R}^d)$ the (linear) deformation operator for $x \in L^2(\mathbb{R}^d)$:

$$\mathcal{L}_\tau x(u) \triangleq x(u-\tau(u))\,. \tag{2.31}$$

We explicitely write $\nabla\tau$ the **Jacobian** of τ as in the standard litterature of the Scattering Transform [26]. From the global inversion theorem, we see that if $\|\nabla\tau\|_\infty = \sup_{u\in\mathbb{R}^d}\|\nabla\tau(u)\| < 1$, then, $\mathbf{I}-\tau$ is invertible, the differential is clearly invertible by standard consideration and the function is injective, thus τ is a diffeomorphism. We say it is smooth if it is $\mathcal{C}^k$ for any k. In the following, we will show that the stability to deformation is a delicate property to obtain. We remind the global inversion theorem [30, Corollary 4.3]:

Theorem 2.2 (Global diffeomorphism)**.** *Let* $\phi \in \mathcal{C}^\infty(\mathbb{R}^d)$*,* ϕ *is a* $\mathcal{C}^\infty$*-diffeomorphism with* $\phi(\mathbb{R}^d) = \mathbb{R}^d$ *if and only if:*

$$\det(\nabla\phi(u)) \neq 0, \forall u \in \mathbb{R}^d$$

and

$$\lim_{\|u\|\to+\infty}\|\phi(u)\| = +\infty\,.$$

Another useful version of this theorem:

Theorem 2.3 (Global diffeomorphism)**.** *Let* $\Omega \subset \mathbb{R}^d$ *an openset, and* ϕ *a smooth injective function. Then:*

$$\det(\phi(u)) \neq 0$$

if and only if $\phi(\Omega)$ *is an openset and* ϕ *is a smooth diffeomorphism from* Ω *to* $\phi(\Omega)$*.*

We in particular demonstrate that:

Lemma 2.6. *Let* $\tau \in \mathcal{C}^\infty(\mathbb{R}^d)$*. Then, if* $\|\nabla\tau\|_\infty \triangleq \sup_u\|\nabla\tau(u)\| < 1$*, then* $\mathbf{I}-\tau$ *is a smooth diffeomorphism. Furthermore:*

$$(1-\|\nabla\tau\|_\infty)^d \leq \det(\mathbf{I}-\nabla\tau(u)) \leq (1+\|\nabla\tau\|_\infty)^d$$

Proof. Fix $u \in \mathbb{R}^d$ and let λ, e an eigen-couple of $\mathbf{I} - \nabla\tau(u)$. Then, $e - \nabla\tau(u).e = \lambda e$ and $\nabla\tau(u).e = (1-\lambda)e$. Thus, $|1-\lambda| \leq \|\nabla\tau(u)\|_\infty$ and $0 < 1 - \|\nabla\tau(u)\|_\infty \leq \lambda \leq 1 + \|\nabla\tau(u)\|_\infty < 2$. So we can lower and upper bound the determinant. Next for u large enough:

$$\|u - \tau(u)\| \geq \|u\| - \|\tau(u)\| \geq \|u\|(1 - \|\nabla\tau\|_\infty) - \|\tau(0)\| .$$

Thus, $\mathbf{I} - \tau$ is a diffeomorphism. □

In particular, this implies that $L\tau$ is bounded:

Corollary 2.1. *If $\|\nabla\tau\|_\infty < 1$, then $(1+\|\nabla\tau\|_\infty)^{-d} \leq \|L_\tau\| \leq (1-\|\nabla\tau\|_\infty)^{-d}$.*

Proof. Indeed, with $u' = u - \tau(u)$

$$\|L_\tau x\|^2 = \int_u |x(u - \tau(u))|^2 \, du = \int_u \frac{1}{\det(\mathbf{I} - \nabla\tau)((\mathbf{I} - \tau)^{-1}u')} |x(u')|^2 \, du$$

□

In the following, we will be mainly interested in the action of small deformations: we will measure the distance between the deformation field $\mathbf{I} - \tau$ and the identity via the following quantity:

$$\sup_x \|\tau(x)\| + \sup_x \|\nabla\tau(x)\| + \sup_{x,y} \|\tau(x) - \tau(y)\| .$$

We will use the notation $\|\Delta\tau\|_\infty = \sup_{x,y} \|\tau(x) - \tau(y)\|$, which is a quantity small in practice.

2.2 Scattering Transform on Euclidean groups

2.2.1 Wavelet Transform on $\mathbb{R}^d$

We now introduce the notion of mother wavelet: a wavelet family is obtained by the dilation of a mother filter, and allows to obtain a dilated basis of $L^2(\mathbb{R}^d)$. In the case of signals such as sounds or images, they allow to obtain a sparse representation with a predefined basis. The simplest example of wavelets is given by the Haar wavelets. See [25] for an exaustive description of wavelets en their properties. In the context of the Scattering Transform, we will discuss rather different properties of wavelets, such as their stability to deformations.

Definition 2.5. *ψ is a mother wavelet if, $\psi \in L^2(\mathbb{R}^d)$ and $\int_{\mathbb{R}^d} \psi(u) \, du = 0$.*

This function is systematically associated to a real-valued low-pass filter $\phi \in L^2(\mathbb{R}^d)$. We now consider the rotation and dilation of a given mother wavelet, leading to:

$$\psi_{j,\theta}(u) \triangleq \frac{1}{2^{jd}} \psi(\frac{r_{-\theta}u}{2^j}) . \tag{2.32}$$

Considering a family of such filters, we ask them to be close to a tight-frame (see [25]), via:

Definition 2.6. *A discrete set of indexes $\Lambda = \{(j,\theta)\} \subset \{1,...,J\} \times SO_d(\mathbb{R})$ is said to be admissible for $\epsilon > 0$ if:*

$$1 - \epsilon \leq \sum_{\lambda} |\widehat{\psi_\lambda}(\omega)|^2 + |\widehat{\phi}_J(\omega)|^2 \leq 1 \tag{2.33}$$

Example 2.3. *In the case of images, the wavelets $\psi_{j,\theta}$ will be parametrized by a discrete angular parameter $\theta \in 2\pi\mathbf{Z}/K\mathbf{Z}$, which implies that $x \star \psi_{j,\theta}$ will be covariant to the action of rotations of $2\pi\mathbf{Z}/K\mathbf{Z}$ for some $K \in \mathbb{N}$.*

Proposition 2.9. *Assume that:*

$$\sum_{\lambda} |\hat{\psi}_\lambda|^2(\omega) \leq 1 \tag{2.34}$$

is admissible, and, around 0, $|\psi(\omega)| = o(\omega)$ then there exists ρ, continuous, $\rho \geq 0$, $\rho \in L^1$, $|\hat{\rho}(\omega)|^2 \leq 1 - \sum_\lambda |\hat{\psi}_\lambda(\omega)|^2, \forall\omega$, $\hat{\rho}(0) = 1$.

Proof. We follow a similar proof to [39, Lemma 6.1], with a tiny modification to adapt it to any dimension. We note that: $\sum_\lambda |\hat{\psi}_\lambda(\omega)|^2 = o(\|\omega\|^2)$ (by dilation and rotation). If ρ is positive, then it is the square of a positive function, thus we can instead consider ρ^2, which has Fourier transform equal to $\hat{\rho} \star \hat{\rho}$. If $\hat{\rho}_\epsilon(\rho,\theta) = f(\frac{\rho}{\epsilon})$, $f : \mathbb{R}^+ \to \mathbb{R}^+$ smooth with compact support in $\mathcal{B}(0,\epsilon)$ and a maximum in 0 equal to 1, then the support of $\hat{\rho}_\epsilon \star \hat{\rho}_\epsilon$ is in $\mathcal{B}(0,2\epsilon)$ and the maximum of this function is reached in 0 (by symmetry). Also by symmetry, it is clear that this function is radial, and thus for some $\alpha > 0$(because we have a maximum in 0), $\hat{\rho}_\epsilon \star \hat{\rho}_\epsilon = 1 - \alpha\|\omega\|^2 + o(\|\omega\|^2)$. By compacity, we thus have $|\hat{\rho}_\epsilon \star \hat{\rho}_\epsilon| \leq 1 - \beta\|\omega\|^2$ and $\sum_\lambda |\hat{\psi}_\lambda(\omega)|^2 \leq \gamma\|\omega\|^2$. Dilating $\hat{\rho}_\epsilon \star \hat{\rho}_\epsilon$ allows to conclude as in [39]. □

This implies that it is always possible to find a low-pass filter (potentially dilated by a factor J), which is positive and that allows to "fill" the holes in Fourier.

2.2.2 Scattering Transform on $\mathbb{R}^d$

We consider an initial signal given by $U_0 x \triangleq x$. We then write $V_J x = \{x \star \psi_\lambda\}_{\lambda\in\Lambda}$ and $A_J x \triangleq x \star \phi_J$. In general, $V_J x$ corresponds to the high-frequencies of x whereas $A_J x$ corresponds to the lower-frequencies, and is thus more invariant to translations. If the wavelets of V_J are chosen analytic, then its (point-wise) modulus $|V_J x|$ will tend to be smoother. In the following, we assume that $\|W_J\| \leq 1$. The idea will be to cascade progressively several operators $|V_j|$ leading to U_n as follow:

$$U_{n+1}x = |V_J U_n x| \,. \tag{2.35}$$

From this, we define the Scattering coefficients as follow:

Definition 2.7. *For a signal $x \in L^2(\mathbb{R}^d)$ scattering of order n is defined by:*

$$S_n^J x = \{A_J x, ..., A_J U_n x\} \,. \tag{2.36}$$

Theorem 2.4 (Non-expansivity)**.** $\forall x, y \in L^2(\mathbb{R}^d), \forall n, \|S_n^J x - S_n^J y\| \leq \|x - y\|$.

Proof. We note that by $1-$Lipschitz continuity of $|.|$ and the approximative isometry property of the wavelet transform:

$$\|A_J U_n x - A_J U_n y\|^2 + \|U_{n+1}x - U_{n+1}y\|^2 \tag{2.37}$$
$$= \|A_J U_n x - A_J U_n x\|^2 + \||V_J|U_n x - |V_J|U_n y\|^2 \tag{2.38}$$
$$\leq \|A_J U_n x - A_J U_n y\|^2 + \|V_J U_n x - V_J U_n y\|^2 \tag{2.39}$$
$$\leq \|U_n x - U_n y\|^2 \tag{2.40}$$
$$\tag{2.41}$$

Summing those inequality, we get:

$$\|S_n^J x - S_n^J y\|^2 \leq \|U_0 x - U_0 y\|^2 - \|U_{n+1}x - U_{n+1}y\|^2 \leq \|x - y\|^2 . \tag{2.42}$$

□

Lemma 2.7. *Assume that $x \in L^2(\mathbb{R}^d)$ has a compact support in Fourier, i.e., $\hat{x}(\omega) = 0, \forall \|\omega\| > A$. Also assume that $\hat{\psi}(\omega) = 0$ and that $\hat{\psi}$ is continuous and is ϵ-admissible as in Def. 2.6. Then:*

$$\lim_n \|U_n x\| = 0 . \tag{2.43}$$

Proof. See [39]. □

Theorem 2.5 (Energy preservation)**.** *For $x \in L^2(\mathbb{R}^d)$, and ψ as in Lemma 2.7. Assume first that x has a compact support in Fourier, we then have:*

$$(1 - \mathcal{O}(\epsilon))\|x\| \leq \lim_{n\to\infty} \|S_n^J x\| \leq \|x\| \tag{2.44}$$

Furthermore, for $x \in L^2(\mathbb{R}^d)$, if $\epsilon = 0$ in Def. 2.6, then:

$$\lim_{n\to\infty} \|S_n^J x\| = \|x\|$$

Proof. The right inequality is proved with $y = 0$ from supra. For the left side, we have:

$$\begin{aligned}(1-\epsilon)\|U_n x\|^2 &\leq \|A_J U_n x\|^2 + \|V_J U_n x\|^2 \\ &= \|A_J U_n x\|^2 + \||V_J|U_n x\|^2 \\ &= \|A_J U_n x\|^2 + \|U_{n+1}x\|^2\end{aligned}$$

This leads again to:

$$\|x\|^2 - \|U_{N+1}x\|^2 - \epsilon \sum_{n \leq N} \|U_n\|^2 \leq \|S_N^J x\|^2 . \tag{2.45}$$

From supra, we have the first conclusion. If $\epsilon = 0$, we obtain the claim, by density of functions with compact support combined with Prop. 2.4. □

We now focus on stability to deformations, which means here that the Scattering Transform is Lipschitz with respect deformations.

Theorem 2.6 (Stability to deformation)**.** *Assume that ψ, ϕ are smooth with ψ, ϕ and their derivatives having a fast decay. And assume also that $\|W_J\| \leq 1$. If $\int \psi = 0$, there exists $C > 0$, such that for any $J \in \mathbf{N}$, we get:*

$$\|S_n^J x - S_n^J \mathcal{L}_\tau x\| \leq n^{3/2} C(\|\nabla \tau\|_\infty + \|\Delta \tau\|_\infty + 2^{-J}\|\tau\|_\infty)\|x\| \tag{2.46}$$

Proof. For a fixed J, we will consider each order n independantly:

$$\begin{aligned}
&\|A_J|V_J|...|V_J|\mathcal{L}_\tau - A_J|V_J|...|V_J|\| \leq \|A_J(\mathcal{L}_\tau - \mathbf{I})|V_J|...|V_J|\|+ \\
&\sum_{m=0}^{n-1} \|A_J \overbrace{|V_J|...|V_J|}^{m \text{ times}} \mathcal{L}_\tau \overbrace{|V_J|...|V_J|}^{n-m \text{ times}} - A_J \overbrace{|V_J|...|V_J|}^{m+1 \text{ times}} \mathcal{L}_\tau \overbrace{|V_J|...|V_J|}^{n-m-1\text{times}} \| \\
&\leq n\|[V_J, \mathcal{L}_\tau]\| + \|A_J - A_J\mathcal{L}_\tau\|
\end{aligned}$$

where the last inequality commes from the non-expansivity of the operators. Next, we note that:

$$\|S_m^J x - S_m^J \mathcal{L}_\tau x\|^2 = \sum_{n=1}^{m} \|A_J|V_J|...|V_J|\mathcal{L}_\tau - A_J|V_J|...|V_J|\|^2 \tag{2.47}$$

$$\leq \sum_{n=1}^{m} (n\|[V_J, \mathcal{L}_\tau]\| + \|A_J - A_J\mathcal{L}_\tau\|)^2 \tag{2.48}$$

The next Lemma allow to upper bound this quantity. □

Remark 2.9. *This inequality is clearly suboptimal, in particular because it involves the order n of the Scattering Transform. In fact, [26] shows the bound can be independant of n but this is more technical.*

In the following, we leverage the assumption that $\|W_J\| \leq 1$, which can be obtained via a renormalization of the mother wavelet.

Lemma 2.8 (Stability of a low-pass filter to deformations)**.** *Assume $\nabla\phi$ and ϕ are integrable. There is $C > 0$, such that for any $J \in \mathbb{N}$, we get:*

$$\|A_J - A_J\mathcal{L}_\tau\| \leq C(2^{-J}\|\tau\|_\infty + \|\nabla\tau\|) \tag{2.49}$$

Proof. For proving this result, we exhibit the kernel of the integral operator $A_J - A_J\mathcal{L}_\tau$, which is given by:

$$A_J x - A_J\mathcal{L}_\tau x(u) = \int_{\mathbb{R}^d} x(v)\phi_J(u-v) - x(v-\tau(v))\phi(u-v)\, dv \tag{2.50}$$

$$= \int_{\mathbb{R}^d} x(v-\tau(v))\phi_J(u-v-\tau(v))\det(\mathbf{I} - \nabla\tau(v)) - x(v-\tau(v))\phi(u-v)\, dv\,. \tag{2.51}$$

$$= K_1 L_\phi x + K_2 L_\phi x \tag{2.52}$$

where we used $v' - \tau(v') = v$ for the left term of the second line and:

$$K_1 x = \int_{\mathbb{R}^d} \det(\mathbf{I} - \nabla\tau(v))(\phi_J(u - v - \tau(v)) - \phi_J(u - v)x(v)\, dv$$

and

$$K_2 x = \int_{\mathbb{R}^d} (\det(\mathbf{I} - \nabla\tau(v)) - 1)\phi_J(u - v)x(v)\, dv\,.$$

Now, we will use a lot the Schur's Lemma, with Lemma 2.6 and $u' = u - v - t\tau(u)$
:

$$\begin{aligned}
\int_{\mathbb{R}^d} |\det(\mathbf{I} - \nabla\tau(v))\phi_J(u - v) - \phi(u - \tau(u) - v)|\, du &\leq 2^d \int_{\mathbb{R}^d} |\int_0^1 \langle \nabla\phi_J(u - v - t\tau(u)), \tau(u)\rangle\, dt|\, du \\
&\leq 2^d \int_{\mathbb{R}^d} \int_0^1 \|\tau\|_\infty \|\nabla\phi_J(u)\| |\det{}^{-1}(\mathbf{I} - t\nabla\tau)((\mathbf{I} - t\tau)^{-1}(u + v)))|\, dudt\,, \\
&\leq \|\tau\|_\infty 2^{2d} \int_{\mathbb{R}^d} \|\nabla\phi_J(u)\|\, du
\end{aligned}$$

Next, we again get:

$$\begin{aligned}
\int_{\mathbb{R}^d} |\det(\mathbf{I} - \nabla\tau(v))||\phi_J(u - v) - \phi(u - \tau(u) - v)|\, dv &\leq 2^d \int_{\mathbb{R}^d} |\int_0^1 \langle \nabla\phi_J(u - v + t\tau(u))^T, \tau(v)\rangle\, dt|\, dv \\
&\leq 2^d \int_{\mathbb{R}^d} \|\nabla\phi_J(v)\| \|\tau\|_\infty\, dv\,.
\end{aligned}$$

Thus, by Schur Lemma, $\|K_1\| \leq \|\tau\|_\infty 2^{3/2d} \|\nabla\phi_J\|_1$. For the second kernel, $\int_{\mathbb{R}^d} |k_2(u, v)|\, du \leq d\|\nabla\tau\|_\infty \|\phi_J\|_1$, and by Symmetry and Schur Lemma, $\|K_2\| \leq d\|\nabla\tau\|_\infty$ Now, knowing that $\|\nabla\phi_J\|_1 = \frac{1}{2^J}\|\nabla\phi\|_1$ and $\|\phi_J\|_1 = \|\phi\|_1$, we can now conclude. □

Lemma 2.9 (Commutation of a wavelet transform with diffeomorphism)**.** *Let $\psi \in L^2(\mathbb{R}^d)$, and let $\psi_j(u) = \frac{1}{2^{jd}}\psi(\frac{u}{2^j})$ and assume that $\int_{\mathbb{R}^d} \psi(u)\, du = 0$ and that ψ is $\mathcal{C}^1$, with ψ and its derivatives with a fast decay. Write $V_J x = \{x \star \psi_j\}_{J \geq j \geq 0}$. Then, there exists a constant C such that for any J:*

$$\|[V_J, L_\tau]\| \leq C\big(\|\nabla\tau\|_\infty + \|\Delta\tau\|_\infty\big)\,. \tag{2.53}$$

Proof. First, note that:

$$\|[V_J, L_\tau]\|^2 = \|\sum_{j=0}^{J} [K_j, L_\tau]^*[K_j, L_\tau]\|\,,$$

where we write $K_j x = x \star \psi_j$ for $0 \leq j \leq J$. We begin with the following technical Lemma:

Lemma 2.10 (Technical)**.** *Let $\tilde{K}_j$ with kernels $\tilde{k}_j(u,v) = a(v)\psi_j(u-v)$ then there is $C > 0$ which depends on ψ such that:*

$$\| \sum_j \tilde{K}_j^* \tilde{K}_j \| \leq \|a\|_\infty^2 C \,.$$

Proof. First, we note that:

$$\tilde{K}_j x(u) = \int_v a(v)x(v)\psi_j(u-v)\,dv = K_j(ax)(u)$$

Next, because ψ is $\mathcal{C}^1$ with fast decay, we can find $c > 0$ such that $|\hat{\psi}(\omega)|^2 \leq \frac{c}{\|\omega\|^2}$. Indeed, $\widehat{\nabla\psi}$ is a bounded function, as $\nabla\psi$ is L^1. At the same time, because $u \to |\psi(u)|\|u\|$ is integrable, $\hat{\psi}$ is $\mathcal{C}^1$ and because $\hat{\psi}(0) = 0$, we know there is $c' > 0$, $|\hat{\psi}(\omega)|^2 \leq c'\|\omega\|^2$ for $\|\omega\| \leq 1$. Thus, there is $\tilde{c}$, such that $|\hat{\psi}|^2(\omega) \leq \tilde{c}\min(\|\omega\|^2, \frac{1}{\|\omega\|^2})$. Now, write $r = \|\omega\|^2$. Either $r \geq 1$, in which case:

$$\sum_{j\geq 0} |\hat{\psi}(2^j\omega)|^2 \leq \tilde{c}\sum_{j\geq 0} 2^{-2j} \leq 2\tilde{c}\,.$$

Either $r < 1$, in which case there is $j_0 \in \mathbb{N}$, $2^{-2j_0-2} < r < 2^{-2j_0}$. But then:

$$\sum_{j=0}^{\infty} \min(2^{2j}r, \frac{1}{r2^{2j}}) \leq \sum_{j=0}^{j_0} 2^{2j}r + \sum_{j=j_0+1}^{\infty} \frac{1}{r2^{2j}} \leq 4$$

In both cases:

$$\forall\omega, \sum_{j\geq 0} |\hat{\psi}(2^j\omega)|^2 \leq C$$

Now, we note that:

$$\sum_j \|\tilde{K}_j x\|^2 = \sum_j \|K_j(ax)\|^2 \leq \|a\|_\infty^2 \sum_j \|K_j x\|^2 = C\|a\|_\infty^2\|x\|^2 \tag{2.54}$$

In particular, $\| \sum_j \tilde{K}_j\tilde{K}_j \| \leq C\|a\|_\infty^2$.

□

We note that:

$$(L_\tau K_j - K_j L_\tau)x(u) = \int_v \psi_j(u-v-\tau(u))x(v) - x(v-\tau(v))\psi_j(u-v)\,dv \tag{2.55}$$

$$= \int_v \psi_j(u-v'+\tau(v')-\tau(u))x(v'-\tau(v'))\det(\mathbf{I}-\nabla\tau)(v')dv' \tag{2.56}$$

$$- \int_v x(v-\tau(v))\psi_j(u-v)\,dv \tag{2.57}$$

$$\tag{2.58}$$

with $v = (\mathbf{I} - \tau)(v')$. We note that thus:

$$[L_\tau, V_j] = \tilde{K}_j L_\tau$$

and we write:

$$K_j^1 x(u) = \int_v x(v)[\psi_j(u - v + \tau(v) - \tau(u)) - \psi_j(u - v)]\det(\mathbf{I} - \nabla\tau)(v)\, dv$$

and

$$K_j^2 x(u) = \int_v x(v)\psi_j(u - v)(\det(\mathbf{I} - \nabla\tau)(v) - 1)\, dv$$

such that $\tilde{K}_j = K_j^1 + K_j^2$, with kernels k_j^1, k_j^2. In this case,

$$\|[V_J, L_\tau]\|^2 = \|\sum_j L_\tau^* \tilde{K}_j^* \tilde{K}_j L_\tau\| \tag{2.59}$$

$$\leq \|L_\tau\|^2 \|\sum_j \tilde{K}_j^* \tilde{K}_j\| \tag{2.60}$$

$$= \|L_\tau\|^2 \|\sum_j (\tilde{K}_j^{1,*} + \tilde{K}_j^{2,*})(\tilde{K}_j^1 + \tilde{K}_j^2)\| \tag{2.61}$$

$$\leq \|L_\tau\|^2 (\|\sum_j \tilde{K}_j^{2,*} K_j^2\| + 2\sum_j \|\tilde{K}_j^1\|\|\tilde{K}_j^2\| + \sum_j \|\tilde{K}_j^1\|^2) \tag{2.62}$$

At this stage, we write:

$$|k_j^1(u, v)| = |\int_0^1 \langle \nabla\psi_j(u - v + t(\tau(v) - \tau(u))), \tau(v) - \tau(u)\rangle\, dt \det(\mathbf{I} - \nabla\tau)(v)| \tag{2.63}$$

$$\leq 2^d \|\Delta\tau\|_\infty \int_0^1 \|\nabla\psi_j(u - v + t(\tau(v) - \tau(u)))\| \tag{2.64}$$

where we used that $|\det(\mathbf{I} - \nabla\tau)(v)| \leq 2^d$ via Lemma 2.6. Next, if $u' = u - t\tau(u) - v + t\tau(v)$, then:

$$\int_u |k_j^1(u, v)|\, du \leq 2^d \|\Delta\tau\|_\infty \int_{u'} \|\nabla\psi_j(u')\| \frac{1}{\det(\mathbf{I} - t\nabla\tau)((\mathbf{I} - \tau)^{-1}(u' + v - t\tau(v)))} \tag{2.65}$$

$$\leq 2^{2d-J} \|\Delta\tau\|_\infty \|\nabla\psi\|_1 \tag{2.66}$$

where we used Lemma 2.6. By symmetry and Schur Lemma, $\|K_j^1\| \leq 2^{2d-J}\|\Delta\tau\|_\infty\|\nabla\psi\|_1$. Next k_j^2 writes:

$$k_j^2(u, v) = \psi_j(u - v)a(v)$$

with $a(v) = \det(\mathbf{I} - \nabla\tau)(u) - 1$. However, $\|a\|_\infty \leq \max(1 - (1 - \|\nabla\tau\|_\infty)^d, (1 + \|\nabla\tau\|)^d - 1)) \leq d\|\nabla\tau\|_\infty$. Using the technical Lemma, this implies that:

$$\| \sum_j \tilde{K}_j^{2,*} \tilde{K}_j^2 \| \leq d^2 C \|\nabla \tau\|_\infty^2$$

Next, it is clear that:

$$\|\tilde{K}_j^2\|^2 \leq \| \sum_j \tilde{K}_j^{2,*} K_j^{2,*} \|$$

Furthermore, we have using Corollary 2.1:

$$\|L_\tau\| \leq 2^d$$

All combined, and summing over j, we get a constant $C > 0$ which depends on d, ψ such that:

$$\|[V_J, L_\tau]\| \leq C\big(\|\nabla \tau\|_\infty + \|\Delta \tau\|_\infty\big)\,.$$

□

2.2.3 Roto-translation scattering

We now discuss how to obtain some invariance on a specific group, which is non-commutative and not compact: the roto-translation group. First, observe that thanks to the section above, we can define the convolution along $SL(E)$. Here, the first layer is given by a coordinate mapping of the Euclidean Scattering Transform define in Sec. 2.2.2 on $SL(E)$:

$$U_1 x[j,\theta](u) \triangleq \tilde{U}_1 x[j](u,\theta) = |x \star \psi(u,\theta)| \tag{2.67}$$

Note that applying an operator along angles without non-linearity would simply lead to an isotropic filter. This is the reason why $x \star \phi_J(u)$ is invariant to rotations, if ϕ is for instance an isotropic Gaussian filter. Instead, we consider a mother filter $\Psi \in L^2(SL(E))$ and we peform a convolution along $SL(E)$ as defined in Def. 2.1 leads to:

$$(\tilde{x} \circledast \Psi)(u,\theta) = \int_{\mathbb{R}^2} \int_{[0,2\pi]} \Psi(r_{-\theta'}(u-u'), \theta - \theta') \tilde{x}(u', \theta') \tag{2.68}$$

We thus have a very natural formulations of our filters, via (assuming for now the parameters haven't been discretized):

$$\Psi_{j_2,\theta_2,k_2}(u,\theta) = \psi_{j_2,\theta_2}(u) \breve{\psi}_{k_2}(\theta)\,. \tag{2.69}$$

In this case, we note that the convolution can be naturally casted as follow:

$$\tilde{U}_2 x[\lambda_2, \lambda_1](u,\theta) = |\int_{\mathbb{R}^2} \int_{[0,2\pi]} \psi_{j_2,\theta_2+\theta'}(u-u') \breve{\psi}_{k_2}(\theta - \theta') \tilde{U}_1 x[\lambda_1](u',\theta')|\,. \tag{2.70}$$

In order to get invariant coefficients over the roto-translation groups, we can average them along the orbit of the group, using a simple separable averaging, where ϕ_J, ϕ_k are two averaging along spatial and angular variabilities:

$$\Phi_{J,K}(u,\theta) = \phi_J(u)\phi_K(\theta)\,, \tag{2.71}$$

and we define:

$$\begin{aligned}
&S_0^{J,K}x = x \star \phi_J \\
&S_1^{J,K}x = \tilde{U}_1 x \circledast \Phi_{J,K} \\
&S_2^{J,K}x = \tilde{U}_2 x[\lambda_2, \lambda_1] \circledast \Phi_{J,K}\,.
\end{aligned}$$

This is the invariant Scattering on $SL(E)$ as introduced by [33].

Chapter 3

Graphs Neural Networks and Manifold data

In this chapter, we propose to analyze Graph Neural Networks with Signal Processing tools, and in the particular setting for which the graph is sampled from an underlying manifold structure. We will study particularly the example of the Laplacian on the sphere $\mathcal{S}^{d-1}$.

3.1 The Laplacian and Graph Signal Processing

3.1.1 Basics on graphs

Definition 3.1. *The adjacency matrix $\mathcal{A}$ of an undirected graph $\mathcal{G} \subset \{1, ..., n\}$ and $\mathcal{E} \subset \mathcal{G} \times \mathcal{G}$ is any symmetric matrix which satisfies:*

$$\mathcal{A}_{i,j} \geq 0\,, \tag{3.1}$$

with $(i, j) \in \mathcal{E}$ if and only if $\mathcal{A}_{i,j} > 0$. Furthermore, we assume that:

$$\sum_i \mathcal{A}_{i,j} = 1\,. \tag{3.2}$$

The Laplacian is given by:

$$\Delta_{\mathcal{G}} = \mathbf{I} - \mathcal{A}\,. \tag{3.3}$$

Furthermore, we write $L^2(\mathcal{G})$ the set of integrable signals $\mathcal{G} \to \mathbb{R}$.

Remark 3.1. *In general the adjacency matrix is not normalized, and there are many variants of possible normalization, which all have a theoretical or practical fundation. For instance, if $\mathcal{D}_{i,i} = \sum_j \mathcal{A}_{i,j}$ is the diagonal weight matrix, [22] writes it $\mathbf{I} - \mathcal{D}^{-1/2}\mathcal{A}\mathcal{D}^{-1/2}$ and sometimes the Laplacian is simply $\mathcal{D}^{-1}(\mathbf{I} - \mathcal{A})$ [].*

Proposition 3.1. *If $\mathcal{G}$ is a cycle, then $\Delta_{\mathcal{G}}$ is a first order approximation of $-\Delta_{\mathbb{R}}$, the Laplacian on $\mathcal{S}^1$ and its eigen-basis is given by the DCT.*

Proof. Indeed, we note that the Laplacian is given by (with a circular symmetry), for $0 \leq n \leq N-1$:

$$\Delta_{\mathcal{G}} x[n] = x[n] - \frac{x[n+1] + x[n-1]}{2} \tag{3.4}$$

Now, introducing the Toeplitz matrix $Je_n = e_{n+1}$, we know the eigenvectors are given by $[e^{i\omega_N m}]_m$, with $\omega_n = \frac{2\pi}{N}$ and summing the two conjugate eigenvectors lead to $[\cos(\omega_N m)]_m$. □

Proposition 3.2 (Perron-Frobenius)**.** *Let $\lambda_1 \leq ... \leq \lambda_n$ the spectrum of $\mathcal{A}$, then $\lambda_n = 1$ and $|\lambda_k| \leq 1$. Also, $\lambda_{n-1} < 1$ if and only if the graph has a single connected component.*

Proof. Due to the normalization, if (λ, e) is an eigen-couple, $|\lambda e| = |\mathcal{A}e| \leq \mathcal{A}|e|$ thus taking the ℓ^∞-norm, $|\lambda| \leq 1$ and it's clear that $\lambda_n = 1$ (due to the normalization). If the graph has two connected components (or more), then $\mathcal{A}$ is equivalent to a block-diagonal matrix and has thus 2 eigen-vectors related to 1. Reciprocally, assume one instant that $\mathcal{A}_{i,j} > 0, \forall i, j$. If the graph has a single connected component let e the eigenvector of λ_{n-1} and assume one instant that $\lambda_{n-1} = 1$ and let i_0 s.t. $|e_{i_0}| = \max_i |e_i|$. Then, $|e_{i_0}| = |\sum_j \mathcal{A}_{i_0,j} e_j| \leq \sum_j \mathcal{A}_{i_0,j}|e_j| \leq \sum_j \mathcal{A}_{i_0,j}|e_{i_0}| = |e_{i_0}|$, which implies that $|e_j| = |e_{i_0}|$ for any j, and have the same sign: they are equal. Note that it is possible to assume that $\mathcal{A}_{i,j} > 0$, because $\mathcal{A}^k$ will be so for some k, as its entries are non 0 iff there exists a path of length k and we assumed here the graph has a single component. □

Here, λ_{n-1} is often called the **spectral gap** of the graph [] and is a geometrical quantity that describes the connectivity of a given graph (in particular, how far the behavior of the graph is from a two connected components graph). This can be noticed as a diffusion is simply controlled by:

$$\|\mathcal{A}^k x - (1, ..., 1)^T\| \leq \lambda_{n-1}^k \|x - (1, ..., 1)^T\| . \tag{3.5}$$

and the higher is the connectivity (and thus smaller is λ_{n-1}), the faster is the convergence.

3.1.2 The Laplacian on a manifold

In this section, we write $\Delta_{\mathbb{R}^d}$ or only Δ if no confusion is possible with other variants of Laplacians.

Theorem 3.1 (Divergence Theorem)**.** *Let $u, v \in C^\infty(\mathbb{R}^d)$, then for any open set $\Omega \subset \mathbb{R}^d$ s.t. the oriented boundary $\partial\Omega$ is regular or empty, we get the first Green identity:*

$$\int_\Omega u\Delta v + \nabla u.\nabla v d\lambda = \int_{\partial\Omega} u(\nabla v.\boldsymbol{n})\, d\sigma \tag{3.6}$$

where $\boldsymbol{n}$ is a normal vector pointed outward, σ is the desintegration of λ on $\partial\Omega$.

This also leads, under the same assumptions to the second Green identity:

$$\int_\Omega u\Delta v - v\Delta u \, d\lambda = \int_{\partial\Omega} u(\nabla v.\mathbf{n}) - v(\nabla u.\mathbf{n}) \, d\sigma \,. \tag{3.7}$$

Let $(\mathcal{M}, \mu)$ be a compact sub-manifold of $\mathbb{R}^d$ with measure μ, with no boundary (ie $\partial\mathcal{M} = \emptyset$, meaning locally each neighborhood is diffeomorph to an Euclidean neighrborhood). It implies that there exists for $\forall x \in \mathcal{M}$ a scalar product $\langle ., . \rangle_{T\mathcal{M}_x}$ linked to the tangent space $T\mathcal{M}_x$ at x.

Definition 3.2 (Gradient)**.** *For any smooth $f : \mathcal{M} \to \mathbb{R}$, we define $\nabla_\mathcal{M} f$ as the unique 1-form such that $df(x)y = \langle \nabla_\mathcal{M} f(x), y \rangle_{T\mathcal{M}x}, \forall x \in \mathcal{M}, y \in T\mathcal{M}_x$.*

The Laplacian $\Delta_\mathcal{M} : L^2(\mathcal{M}) \to L^2(\mathcal{M})$ on $\mathcal{M}$ is defined formally as the unique operator satisfying, for f, g smooth:

$$\int_\mathcal{M} f(x)\Delta_\mathcal{M} g(x) d\mu(x) = -\int_\mathcal{M} \langle \nabla_\mathcal{M} f(x), \nabla_\mathcal{M} g(x) \rangle_{T\mathcal{M}_x} d\mu(x) \,. \tag{3.8}$$

(to do so, one can extend the notion of Laplacian by density to the C^∞ functions on $\mathcal{M}$, yet this is difficult) Its existence or unicity is beyond the scope of this class. Sometimes, the Laplacian can be computed in a simpler manneer. Let us consider $f \circ i$ where $i : \mathbb{R}^d \to \mathcal{M}$ is a smooth and surjective application, such that a subset of $\mathbb{R}^d$ is isometrical to $\mathcal{M}$ (meaning that $x \to di(x)$ is a unitary transform along this subset). Since $d(f \circ i) = di^T \nabla_\mathcal{M} f$ and from the Eq. (3.8), we get:

$$\Delta_\mathcal{M} f(i(x)) = \Delta_{\mathbb{R}^d}(f \circ i)(x) \,. \tag{3.9}$$

Furthermore, as an operator of $L^2(\mathcal{M})$, we obtain:

Proposition 3.3. *$-\Delta_\mathcal{M}$ is a non-positive symmetric operator.*

Proof. It is clear from the Eq. (3.8), with $f = g$. □

If $\mathcal{M}$ is compact, then $\Delta_\mathcal{M}$ is a compact operator (this can be proved via Rellich's Lemma [37] and this is beyond the scope of the class) and consequently, we can find an orthonormal basis s.t.:

$$\Delta_\mathcal{M} e_i = -\lambda_i e_i \,, \tag{3.10}$$

$L^2(\mathcal{M}) = \bigoplus_{i>0} \mathrm{Vect}(e_i)$ and $\forall i, \lambda_i \geq 0$.

Remark 3.2. *It is possible to obtain a close form of the Laplacian via Christoffel symbols yet this is clearly beyond the scope of this lecture.*

If we let $K_\sigma(x, y) = e^{-\frac{\|x-y\|^2}{\sigma^2}}$ a Gaussian kernel with bandwidth σ, and if we consider the graph Laplacian s.t. $\mathcal{A}_{i,j} = K_\sigma(x_i, x_j)$ then with the appropriate normalization, we get for a smooth f:

$$\Delta_\mathcal{G} f(x) = \frac{\sum_{i=1}^N K_\sigma(x_i, x) f(x_j)}{\sum_{i=1}^N K_\sigma(x_i, x)} - f(x) \,. \tag{3.11}$$

In this case, [34, 7] show that as $\sigma \to 0, N \to \infty$, then:

$$\Delta_{\mathcal{G}} f(x) \to \Delta_{\mathcal{M}} f(x)\,. \tag{3.12}$$

This last observation justifies how we can handle non-Euclidean data with a Manifold structure via graph method. In fact, it is possible to define a notion of convolution, via:

Definition 3.3 (Convolution on $\mathcal{M}$)**.** *The convolution of two signals $f \in L^2(\mathcal{M})$ and $g \in L^2(\mathbb{N})$ is given by:*

$$f \star g = \sum_i \langle f, e_i \rangle g[i] e_i\,. \tag{3.13}$$

By the Bessel inequality, the previous quantity is well defined. Those tools are the basis of many works that perform signal processing on manifolds or graphs, and we refer the reader to [12] for a more complete review.

3.1.3 Wavelet transforms on Graphs

Following the construction of [14, 19], we will construct filters $\psi_j(m)$ s.t. if $Wx = \{x \star \psi_j\}$, then W should be a frame:

Definition 3.4. *We say that $\{\psi_j\}_j \subset L^2(\mathcal{G})$ is a frame if:*

$$A\|x\| \leq \|Wx\| \leq B\|x\|\,. \tag{3.14}$$

The frame is unitary if $A = B = 1$.

Diffusion wavelets Similarly to a setting in the Euclidean grid, we can introduce wavelets which are a difference of two dilated low-pass filter (*Difference of Gaussians* [24])

$$\psi_j = \mathcal{A}^{j+1} - \mathcal{A}^j \tag{3.15}$$

and

$$\phi = \mathbf{I} + \mathcal{A}\,. \tag{3.16}$$

Then, note:

$$\|Wx\|^2 = \|\phi x\|^2 + \sum_{j \geq 0} \|\psi_j x\|^2 = x^T\big((\mathbf{I} + \mathcal{A})^2 + (\mathbf{I} - \mathcal{A})^2 \sum_{j \geq 0} \mathcal{A}^{2j}\big) x\,, \tag{3.17}$$

Now, having in mind that $0 \preccurlyeq (\mathbf{I} - \mathcal{A})^2 \preccurlyeq \mathbf{I} - \mathcal{A}^2$, we can chose an appropriate B, and for the left term A can be chosen non-negative.

Extrapolating real wavelets on graphs Following [19], a standard approach is to consider a real valued wavelet transform $\{\tilde{\psi}_j\}_j \subset L^2(\mathbb{R})$, and we introduce $\psi_j = \sum_{i\geq 0} \hat{\tilde{\psi}}_j(\lambda_i)e_i$. If the wavelet transform is admissible, then:

$$\|Wx\|^2 = \sum_i \sum_j |\hat{\tilde{\psi}}_j(\lambda_i)|^2 \langle x, e_i\rangle^2 \tag{3.18}$$

and given that there is an $\epsilon > 0$ s.t.:

$$1 - \epsilon \leq \sum_j |\hat{\tilde{\psi}}_j(\lambda_i)|^2 \leq 1\,, \tag{3.19}$$

we obtain the fact that this is also a frame.

Remark 3.3. *Via those constructions, it is possible to define a notion of Scattering Transform, see [16, 17].*

3.1.4 Graph Convolutional Networks

We shortly discuss Spectral GCNs as introduced by [22].

Definition 3.5 (Convolution on a graph)**.** *A GCN is given by:*

$$X_{j+1} = \rho(\mathbf{I} + \mathcal{A})X_j W_j\,, \tag{3.20}$$

where X_j is of size $n \times P_j$ where n is the number of nodes of $\mathcal{A}$ and P_j is the number of features, and W_j is learned through supervision and ρ is a pointwise non-linearity.

Thanks to the formalism we developed above, we can understand now $\mathbf{I}+\mathcal{A}$ as a smoothing operator, which might not be desirable in many applications, because it oversmoothes a signal.

3.2 A cryptic example: $\mathcal{S}^{d-1}$

3.2.1 Laplacian on $\mathcal{S}^{d-1}$

We now go through an important example, both from the perspective of applications and theoretical insights: the $d-1$ dimensional sphere. For $f \in L^2(\mathcal{S}^{d-1})$, we write $\Delta_{\mathcal{S}^{d-1}} f$ the Laplacian on the $d-1$ dimensional unit sphere. We always assume $d \geq 3$.

Proposition 3.4. *Let $f : \mathbb{R}^d \to \mathbb{R}$. Then:*

$$\Delta_{\mathbb{R}^d} f(\frac{x}{\|x\|}) = \Delta_{\mathcal{S}^{d-1}} f(\frac{x}{\|x\|}). \tag{3.21}$$

Proof. From the remark above, it's enough to show that a restriction of $i(x) = \frac{x}{\|x\|}$, $i : \mathbb{R}^d \to \mathcal{S}^{d-1}$ is an isometry on $\mathcal{S}^d$. Straightforward computations show that if $\|x\| = 1$, then $\frac{x+\epsilon}{\|x+\epsilon\|} = x + \epsilon - x\langle x, \epsilon\rangle + o(\epsilon)$: the differential is an orthogonal projection on the orthogonal of x and consequently, i restricted to $x^\perp$ is locally a diffeomorphism on $\mathcal{S}^{d-1}$ as $di(x).\epsilon = \epsilon, \forall \epsilon \perp x$. We have thus the conclusion using Eq. (3.9). □

Remark 3.4. *There is a more general "trick" linked to the notion of normal tangent, that can be embedded isometrically, see [].*

We note that the related (complex) scalar product $(\mathcal{S}^{d-1}, \langle ., . \rangle)$ is given by:

$$\langle f, g\rangle_{\mathcal{S}^{d-1}} = \int_{\mathcal{S}^{d-1}} \bar{f}(x)g(x)\, d\sigma(x) \tag{3.22}$$

$$= \frac{1}{\Lambda_d} \int_0^\pi ... \int_0^\pi \int_0^{2\pi} \bar{f}(\varphi)g(\varphi) \sin^{d-2}(\varphi_1)... \sin^1(\varphi_{d-1}) d\varphi_1 ... d\varphi_{d-1}\,, \tag{3.23}$$

where $\Lambda_d = \frac{2\pi^{\frac{d+1}{2}}}{\Gamma(\frac{d+1}{2})}$ and $\varphi = (\varphi_1, ..., \varphi_{d-1})$. It is adjusted such that $\|1\|_{\mathcal{S}^{d-1}} = 1$.

Definition 3.6 (Harmonic homogeneous polynomials)**.** *A homogeneous polynomial of degree n on $\mathbb{R}^d$ is a polynomial Y that satifies:*

$$\forall x \in \mathbb{R}^d, \forall \lambda \in \mathbb{R}, Y(\lambda x) = \lambda^n Y(x)\,. \tag{3.24}$$

We say that a homogeneous polynomial is harmonic, if in addition:

$$\Delta Y = 0\,. \tag{3.25}$$

and we write $\mathcal{H}_n^d \subset \mathbb{R}[X_1, ..., X_d]$ the sub-vector space of harmonic and homogenous polynomials of degree n.

Lemma 3.1 (Dimension of harmonic homogeneous polynomials)**.** *We have dim $\mathcal{H}_n^d = (2n + d - 2)\frac{(n+d-3)!}{n!(d-2)!}$.*

Proof. See [2, Prop 5.8, p.78]. □

Proposition 3.5 (Harmonic functions on the unit ball)**.** *If P is harmonic on $\mathcal{B}(0, 1)$, then for any $0 \le r < 1$ and $x \in \mathcal{S}^{d-1}$:*

$$P(rx) = \int_{\mathcal{S}^{d-1}} P(y) \frac{1 - r^2}{\|rx - y\|^d} d\sigma(y) \tag{3.26}$$

Proof. This is here tricky. First, we note that if $v(x) = \frac{1}{\|x\|^d}$, then $\nabla v = -\frac{dx}{\|x\|^{d+2}}$ and $\Delta v = \frac{2d}{\|x\|^{d+2}}$. We consider:

$$u(x, y) = \frac{1}{\|rx - y\|^{d-2}} - \frac{1}{\|ry - x\|^{d-2}}\,. \tag{3.27}$$

We note that thus: $\Delta_x u(x, y) = -\Delta_y u(x, y)$ by symmetry.

Lemma 3.2 (Symmetry Lemma)**.** *If $x, y \in \mathcal{S}^{d-1}$, then $u(x,y) = 0$.*

Now, we apply Theorem 3.1 on $\mathcal{B}_\epsilon \triangleq \mathcal{B}(0,1)\backslash\mathcal{B}(rx,\epsilon)$ (and writing $\mathcal{S}^{d-1}(x,\rho)$ the sphere centered in x of radius ρ) for ϵ small enough, and despite u having a singularity in rx, we get, taking the normal n w.r.t. the y variable:

$$\int_{\mathcal{B}_\epsilon} \Delta_y u(x,y)P(y)\,d\lambda(y) - \int_{\mathcal{B}_\epsilon} \Delta P(y)u(x,y), d\lambda(y) = \int_{\mathcal{S}^{d-1}} P(y)\frac{\partial u}{\partial n}(x,y)\,d\sigma(y) - \int_{\mathcal{S}^{d-1}} u(x,y)\frac{\partial P}{\partial n}(y)\,d\sigma(y)$$
$$- \int_{\mathcal{S}^{d-1}(rx,\epsilon)} P(y)\frac{\partial u}{\partial n}(x,y)\,d\sigma_\epsilon(y) + \int_{\mathcal{S}^{d-1}(rx,\epsilon)} u(x,y)\frac{\partial P}{\partial n}(y)\,d\sigma_\epsilon(y)$$

Now, note that since all the functions are smooth and P is harmonic:

$$\begin{aligned}
\int_{\mathcal{B}_\epsilon} \Delta_y u(x,y)P(y)\,d\lambda(y) &= -\int_{\mathcal{B}_\epsilon} \Delta_x u(x,y)P(y)\,d\lambda(y) \\
&= -\Delta_x \int_{\mathcal{B}_\epsilon} u(x,y)P(y)\,d\lambda(y) \\
&= -\Delta_x \int_{\mathcal{B}_\epsilon} \big(\frac{1}{\|rx-y\|^{d-2}} - \frac{1}{\|x-ry\|^{d-2}}\big)P(y)\,d\lambda(y) \\
&= -\Delta_x \int_{\mathcal{L}_{rx}\mathcal{B}_\epsilon} \frac{1}{\|y\|^{d-2}}P(rx-y)\,d\lambda(y) + \Delta_x \int_{\mathcal{L}_{\frac{x}{r}}\mathcal{B}_\epsilon} \frac{1}{r^{d-2}\|y\|^{d-2}}P(\frac{x}{r}-y)\,d\lambda(y) \\
&= \int_{\mathcal{L}_{rx}\mathcal{B}_\epsilon} \frac{1}{\|y\|^{d-2}}\Delta_x P(rx-y) - \int_{\mathcal{L}_{\frac{x}{r}}\mathcal{B}_\epsilon} \frac{1}{r^d\|y\|^{d-2}}\Delta_x P(\frac{x}{r}-y)\,d\lambda(y) \\
&= 0\,.
\end{aligned}$$

Next:

$$\begin{aligned}
\int_{\mathcal{S}^{d-1}} P(y)\frac{\partial u}{\partial n}(x,y)\,d\sigma(y) &= \int_{\mathcal{S}^{d-1}} P(y)d\big(\frac{(r(ry-x))}{\|rx-y\|^d} - \frac{((y-rx))}{\|rx-y\|^d}\big).\,yd\sigma(y) \\
&= \int_{\mathcal{S}^{d-1}} P(y)d\frac{r^2-1}{\|rx-y\|^d}\,d\sigma(y)\,.
\end{aligned}$$

Then, because the normal vector is $\frac{y-rx}{\epsilon}$ on $\mathcal{S}^{d-1}(rx,\epsilon)$ and the right term is continuous:

$$\begin{aligned}
\int_{\mathcal{S}^{d-1}(rx,\epsilon)} P(y)\frac{\partial u}{\partial n}(x,y)\,d\sigma_\epsilon(y) &= \int_{\|rx-y\|=\epsilon} P(y)\frac{1}{\epsilon^d}(d\epsilon)\,d\sigma_\epsilon(y) - \int_{\|rx-y\|=\epsilon} P(y)\frac{\partial}{\partial n}.\frac{1}{\|ry-x\|^{d-2}}\,d\sigma_\epsilon(y) \\
&= d\int_{\mathcal{S}^{d-1}} P(rx+\epsilon y)d\sigma(y) \to dP(rx) \text{ as } \epsilon \to 0
\end{aligned}$$

Last and least:

$$\int_{\mathcal{S}^{d-1}(rx,\epsilon)} u(x,y)\frac{\partial P}{\partial n}(y)\,d\sigma_\epsilon(y) = \int_{\mathcal{S}^{d-1}(rx,\epsilon)} \big(\frac{1}{\|rx-y\|^{d-2}} - \frac{1}{\|ry-x\|^{d-2}}\big)\frac{\partial P}{\partial n}(y)\,d\sigma_\epsilon(y)$$

The right term of the integrand is continuous, thus we only need to take care of:

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

Taking $\epsilon \to 0$ allows to conclude. □

Proposition 3.6 (Density of harmonic polynomials in $L^2(\mathcal{S}^{d-1})$)**.** *Harmonic polynomials are dense in $L^2(\mathcal{S}^{d-1})$.*

Proof. Fix the space of polynomials of degree less than p as: $\mathbb{R}_p[X_1, ..., X_d]$. We note $\Delta_p : \mathbb{R}_p[X_1, ..., X_d] \to \mathbb{R}_{p-2}[X_1, ..., X_d], X \to \Delta X$ is linear and $\mathrm{Ker}\Delta_p \cap (\|X\|^2 - 1)\mathbb{R}_{p-2}[X_1, ..., X_d] = \{0\}$ from supra (because it values 0 on the sphere and we can use Prop. 3.5). $\mathbb{R}_p[X_1, ..., X_d] = \mathrm{Ker}\Delta_p + (\|X\|^2 - 1)\mathbb{R}_{p-2}[X_1, ..., X_d]$. Now let $f \in L^2(\mathcal{S}^{d-1})$, and write $\tilde{f}(x) = \|x\| f(\frac{x}{\|x\|})$, which is a L^2 extension of f to $\mathcal{K} = \overline{\mathcal{B}(0, \frac{3}{2}) \backslash \mathcal{B}(0, \frac{1}{2})}$. For $\epsilon > 0$, by Bolzano-Weierstrass, there exists P s.t. $\sup_{x \in \mathcal{K}} \|P(x) - \tilde{f}(x)\| < \epsilon$ (you can regularize $\tilde{f}$ first with a unit approximation up to a precision $\epsilon > 0$). Then, from supra, we note the restriction of P to $\mathcal{S}^{d-1}$ coincides with a harmonic polynomial, as $P = Q + (\|X\|^2 - 1)R$ with $\Delta Q = 0$. □

Proposition 3.7. *If Y is harmonic and homogenous of degree k, then:*

$$\Delta_{\mathcal{S}^{d-1}} Y = -k(k + d - 2)Y \tag{3.28}$$

Proof. Here, $Y(\|x\| \frac{x}{\|x\|}) = \|x\|^k Y(\frac{x}{\|x\|})$ and using $\Delta(uv) = v\Delta u + 2\nabla u.\nabla v + u\Delta v$, and $x.\nabla P(x) = kP(x)$ for P homogeneous, combined with Prop. 3.4 we get: $\Delta_{\mathcal{S}^{d-1}} Y = -k(k + d - 2)Y$. □

This implies that if P polynomial is homogeneous of degree k and harmonic, we get:

$$r^k P(x) = \int_{\mathcal{S}^{d-1}} P(y) \frac{1 - r^2}{\|rx - y\|^d} d\sigma(y)\,, \tag{3.29}$$

and thus we can identify the term of degree k of the serie.

Proposition 3.8 (Poisson Kernel)**.** *For $0 \le r < 1$, we write:*

$$P_r(x, y) = \frac{1 - r^2}{(1 - 2r\langle x, y\rangle + r^2)^{d/2}}\,, \tag{3.30}$$

such that if $P_r(x, y) = \sum_{n=0}^{\infty} P_n(\langle x, y\rangle) r^n$, then P_n is a (real) polynomial of degree n and we introduce for $f \in L^2(\mathcal{S}^{d-1})$:

$$K_n(f)(x) = \int_{\mathcal{S}^{d-1}} P_n(\langle x, y\rangle) f(y) d\sigma(y)\,. \tag{3.31}$$

Then:

$$P_r(x, y) \ge 0\,, \tag{3.32}$$

and:

$$\int_{\mathcal{S}^{d-1}} P_r(x,y)d\sigma(x) = 1\,, \tag{3.33}$$

and:

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

and for $m \neq n$, $K_m(P) = 0$ *if* $P \in \mathcal{H}_m^d$, *and finally:*

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

Proof. First, each P_n is a polynomial by composition of the Taylor expansions of $r \to 1-2r\langle x,y\rangle + r^2$ and $x \to \frac{1}{(1+x)^{d/2}}$. Note that K_n is well-defined because $f \in L^1(\mathcal{S}^{d-1})$ by Jensen inequality. We also note that for any $f \in L^2(\mathcal{S}^{d-1}), K_n(f)$ is a polynomial (by linearity of the ingral). The first claim is clear. The second claim is obtained by inserting in Eq. (3.29) $P = 1$ which is clearly harmonic. The third claim is also obtained by using Eq. (3.29) and by identifying the n-term of the development of P_r and checking its degree. We also observe that for any $\Delta P = 0$:

$$P(rx) = \sum_{n=0}^{\infty} r^n K_n(P)(x)$$

We deduce by identification, since $K_n(P)$ is a polynomial in x, and also that $\Delta K_n(P) = 0$ by differentiation along x. On the left, one has a finite degree harmonic polynomial, and on the right, a sum of orthonormal element of harmonic and homogeneous polynomials: the right term must be finite. Thus, taking $r \to 1$ is legit, and we have:

$$P = \sum_{n=0}^{\infty} K_n(P)$$

Thus, $\{K_n\}_n$ spans harmonic polynomials. Note that then, on the set of harmonic polynomials: $K_n^* = K_n$, $K_n K_m = K_n \delta_{m=n}$ and $\{K_n\}$ is thus an orthogonal family of projectors of harmonic polynomials: it implies that if $f \in L^2(\mathcal{S}^{d-1})$ and $m \leq N$:

$$\langle f - \sum_{n=0}^{N} K_n(f), K_m(f)\rangle = \langle f, K_m(f)\rangle - \|K_m(f)\|^2 = 0\,,$$

because $\|K_m(f)\|^2 = \langle f, K_m^*(K_m(f))\rangle = \langle f, K_m(f)\rangle$. Thus, $f - \sum_{n=0}^{N} K_n(f)$ is orthogonal with $\sum_{n=0}^{N} K_n(f)$ and:

$$\|f\|^2 = \|f - \sum_{n=0}^{N} K_n(f)\|^2 + \|\sum_{n=0}^{N} K_n(f)\|^2 \geq \sum_{n=0}^{N} \|K_n(f)\|^2$$

From Prop. 3.6, harmonic polynomials are dense in $L^2(\mathcal{S}^{d-1})$ and thus, for $\epsilon > 0$ there is $P, \Delta P = 0$:

$$\|f - P\| < \epsilon\,.$$

All above combined together implies that:

$$\|f - \sum_{n=0}^{\infty} K_n(f)\|^2 = \|f - P + \sum_{n=0}^{\infty} K_n(f-P)\|^2 \leq 2\|f-P\|^2 + 2\|f-P\|^2 = 4\epsilon^2\,.$$

□

From the previous proposition and Bessel identity, we have a Perseval identity for $f \in L^2(\mathcal{S}^{d-1})$:

$$\|f\|^2 = \sum_{m=0}^{\infty} \|K_m(f)\|^2\,. \tag{3.34}$$

Note also that if $\sum_n \|P^n\|^2 < \infty$, then let $f = \sum_n P^n \in L^2(\mathcal{S}^{d-1})$ and $K_n(f) = P^n$: we thus have built an isomorphism. We introduce $\Lambda_d = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}$ the area of $\mathcal{S}^{d-1}$ for the Lebesgue measure. We will use the following Lemma:

Lemma 3.3. *We obtain for any $x, y \in \mathcal{S}^{d-1}$:*

$$K_n(P_r(\langle x, .\rangle)(y) = r^n P_n(\langle x, y\rangle)$$

and

$$K_n(P_m(\langle x, .\rangle))(y) = \delta_{m=n} P_m(\langle y, x\rangle)\,. \tag{3.35}$$

Proof. Here, for $x, y \in \mathcal{S}^{d-1}$:

$$K_n(P_r(\langle x, .\rangle)(y) = \int_{\mathcal{S}^{d-1}} \sum_m r^m P_m(\langle x, y\rangle) P_n(\langle y, z\rangle)\, d\sigma(z) \tag{3.36}$$

$$= \sum_m r^m \int_{\mathcal{S}^{d-1}} P_m(\langle x, y\rangle) P_n(\langle y, z\rangle)\, d\sigma(z) \tag{3.37}$$

Now, we note that:

$$K_n(K_m(f))(y) = \int_{\mathcal{S}^{d-1}} P_n(\langle x, y\rangle) \int_{\mathcal{S}^{d-1}} P_m(\langle x, z\rangle) f(z) d\sigma(z)\, d\sigma(x) \tag{3.38}$$

$$= \int_{\mathcal{S}^{d-1}} \Big(\int_{\mathcal{S}^{d-1}} P_n(\langle x, y\rangle) P_m(\langle x, z\rangle) \Big) d\sigma(x)\, f(z) d\sigma(z) \tag{3.39}$$

$$= \delta_{m=n} K_n(f)(y) \tag{3.40}$$

This being true for any f, we get that:

$$\int_{\mathcal{S}^{d-1}} P_n(\langle x, y\rangle) P_m(\langle x, z\rangle) d\sigma(x) = \delta_{m=n} P_m(\langle y, z\rangle)\,. \tag{3.41}$$

Combining the two identity, we get:

$$K_n(P_r(\langle x, .\rangle))(y) = r^n P_n(\langle x, y\rangle)$$

□

The previous computations imply that:

Proposition 3.9 (Zonal harmonics)**.** *With the notations of the previous proposition, we have* $\deg P_n = n$*,* $P_n(1) = dim(\mathcal{H}_m^d)$ *and*

$$\int_{-1}^{1} P_n(t)P_m(t)(1-t^2)^{\frac{d-3}{2}}\, dt = \delta_{m=n} dim\ \mathcal{H}_m^d \frac{\Lambda_d}{\Lambda_{d-1}}. \tag{3.42}$$

This family of polynomial is thus uniquely defined.

Proof. Using the lemma above, for $x = y \in \mathcal{S}^{d-1}$, this leads to:

$$\int_{\mathcal{S}^{d-1}} P_n(\langle x, z\rangle)P_m(\langle x, z\rangle)d\sigma(z) = \delta_{m,n}P_m(1)\,. \tag{3.43}$$

This writes, via a change of variable and given this quantity is independant of the initial direction $x \in \mathcal{S}^{d-1}$:

$$\begin{aligned}
&\int_{\mathcal{S}^{d-1}} P_n(\langle x, y\rangle)P_m(\langle x, y\rangle)d\sigma(y)\\
&= \frac{1}{\Lambda_d}\int_0^{\pi} ... \int_0^{\pi}\int_0^{2\pi} P_n(\cos(\varphi_1))P_m(\cos(\varphi_1))\sin^{d-2}(\varphi_1)...\sin^1(\varphi_{d-1})d\varphi_1...d\varphi_{d-1}\\
&= \frac{\Lambda_{d-1}}{\Lambda_d}\int_{-1}^{1} P_n(t)P_m(t)(1-t^2)^{\frac{d-3}{2}}dt\,,
\end{aligned}$$

with the change of variable $t = \cos(\varphi_1)$. Now, because K_m is a projector on $\mathcal{H}_m^d$, if $\{e_i\}$ is an orthonormal basis of $\mathcal{H}_m^d$, then:

$$K_m(P) = \sum_i \langle e_i, P\rangle e_i$$

and having in mind that: $K_m(e_i)(x) = \langle P_m(\langle x, .\rangle), e_i\rangle = e_i(x)$ thus:

$$K_m(P_m(\langle x, .\rangle))(x) = \sum_i \langle e_i, P_m(\langle x, .\rangle)\rangle e_i(x) = \sum_i |e_i(x)|^2$$

Now, we conclude because:

$$\begin{aligned}
\dim(\mathcal{H}_m^d) &= \mathrm{Tr}(K_m) && (3.44)\\
&= \sum_i \|e_i\|^2 && (3.45)\\
&= \int_{\mathcal{S}^{d-1}} K_m(P_m(\langle x, .\rangle))(x)\, d\sigma(x) && (3.46)\\
&= \int_{\mathcal{S}^{d-1}} P_m(\langle x, y\rangle)P_m(\langle x, y\rangle)\, d\sigma(y)\, d\sigma(x) && (3.47)\\
&= P_m(1)\,. && (3.48)
\end{aligned}$$

□

Note that the above formula characterizes exactly the family $\{P_n\}_n$. It is possible to observe that this family satisfies:

Proposition 3.10 (Rodrigues representation formula)**.** *For a given $d \in \mathbb{N}$, the only family satisfying the constraints above is given by:*

$$P_n(t) = (-1)^n \frac{\Gamma(\frac{d-1}{2})}{2^n \Gamma(n + \frac{d-1}{2})} (1-t^2)^{\frac{3-d}{2}} \frac{d}{dt^n} \left((1-t)^{n+\frac{d-3}{2}}\right). \tag{3.49}$$

Proof. See the exercise sheet. □

Remark 3.5. *Note that this could have been inferred from another variant of Rodrigues' formula that links a family orthogonal for a given scalar product to an ODE satisfied by the system of polynomials, see* `https://fr.wikipedia.org/wiki/Th%C3%A9orie_de_Sturm-Liouville`.

Proposition 3.11. *Consider the action $\rho : SO_d(\mathbb{R}) \to L^2(\mathcal{S}^{d-1}), r_\theta \to (u \to x(r_{-\theta}u))$. Then its invariant irreducible subspaces are $\mathcal{H}_n^d$.*

Proof. Here, note that if $P \in \mathcal{H}_n^d$, then:

$$\Delta(\rho(r_\theta)(P)) = 0$$

and

$$\rho(r_\theta)(P)(\lambda x) = \lambda^n \rho(r_\theta)(P)(x)$$

Thus, $\rho(r_\theta)\mathcal{H}_n^d \subset \mathcal{H}_n^d$. See [15, Theorem 1.17] to obtain the irreducibility. □

Remark 3.6. *We recover that $SO_d(\mathbb{R})$ is not commutative, as the irreducible subspaces are not of dimension 1, except for $d = 2$.*

3.2.2 Fourier analysis on $\mathcal{S}^{d-1}$

Now, for $f \in L^2(\mathcal{S}^{d-1})$ and $g \in L^1\left([-1,1], (1-t^2)^{\frac{d-1}{2}} d\lambda(t)\right)$, we can introduce a convolutional operator:

$$f \circledast g(x) \triangleq \int_{y \in \mathcal{S}^{d-1}} f(y) g(\langle x, y \rangle) d\sigma(y), \tag{3.50}$$

which is clearly covariant with $SO_d(\mathbb{R})$.This is formalized by the following proposition and formula:

Proposition 3.12 (Funk-Hecke formula)**.** *For $f \in L^2(\mathcal{S}^{d-1})$ and $g \in L^2\left([-1,1], (1-t^2)^{\frac{d-1}{2}} d\lambda(t)\right)$, then $f \circledast g \in L^2(\mathcal{S}^{d-1})$ and:*

$$f \circledast g = \sum_{n=0}^{\infty} \lambda_n(g) K_n(f), \tag{3.51}$$

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

Proof. First, we note that by Cauchy-Schwartz, if $f \in L^2(\mathcal{S}^{d-1})$, then $|(f \circledast g)(x)| \leq \int_{y\in\mathcal{S}^{d-1}} |f|^2(y)\, d\sigma(y) \int_{y\in\mathcal{S}^{d-1}} |g(\langle x,y\rangle)|^2\, d\sigma(y) \leq \|f\|^2\|g\|^2$. Thus the function is bounded, and thus $f \circledast g \in L^2(\mathcal{S}^{d-1})$. From the proposition above and by density, we have by decomposition g:

$$g(\langle x,y\rangle) = \sum_n \lambda_n(g) P_n(\langle x,y\rangle)\,, \tag{3.52}$$

where $\lambda_n(g) = \frac{\Lambda_{d-1}}{\Lambda_d \dim \mathcal{H}_n^d} \int_{-1}^1 P_n(t) g(t)(1-t^2)^{\frac{d-1}{2}} dt$. We're thus allowed to compute, using $K_m K_n = \delta_{m=n} K_n$:

$$\begin{aligned} K_n(f \circledast g) &= \int_{y\in\mathcal{S}^{d-1}} P_n(\langle x,y\rangle) \int_{z\in\mathcal{S}^{d-1}} \sum_{m\geq 0} \lambda_m(g) P_m(\langle z,y\rangle) f(z) d\sigma(z) d\sigma(x) \\ &= K_n(f)\lambda_n(g)\,. \end{aligned}$$

□

We remind the following Lemma, that shows that P_r acts a unit approximation for the L^∞-norm:

Lemma 3.4. *If f is continuous in x, then one has a point-wise convergence:*

$$f(x) = \lim_{r\to 1} \int_{\mathcal{S}^{d-1}} P_r(x,y) f(y) d\sigma(y)\,. \tag{3.53}$$

Furthermore, if f is continuous on $\mathcal{S}^{d-1}$, the convergence above is uniform.

Proof. Assume that f is continuous in x, we note that for $\|x-y\| > \delta$, so that $-2\langle x,y\rangle > \delta^2 - 2$ and $1 - 2r\langle x,y\rangle + r^2 > 1 + r^2 + r(\delta^2-2) > 0, \forall r \in [0,1]$ and then:

$$\begin{aligned} |\int_{\mathcal{S}^{d-1}} P_r(x,y)(f(x)-f(y)) d\sigma(y)| &\leq \int_{\mathcal{S}^{d-1}} P_r(x,y)|f(x)-f(y)| d\sigma(y) \\ &\leq \int_{\|x-y\|\leq\delta} P_r(x,y)|f(x)-f(y)| d\sigma(y) \\ &+ \int_{\|x-y\|>\delta} P_r(x,y)|f(x)-f(y)| d\sigma(y) \\ &\leq \int_{\|x-y\|\leq\delta} P_r(x,y)\epsilon d\sigma(y) + \mathcal{O}\big((1-r)^2\|f\|_\infty\big) \\ &\leq \epsilon + \mathcal{O}\big((1-r)^2\|f\|_\infty\big) \end{aligned}$$

Thus, for r close enough to 1, one gets:

$$f(x) = \lim_{r\to 1} \int_{\mathcal{S}^{d-1}} P_r(x,y) f(y) d\sigma(y) \tag{3.54}$$

By compacity, the second part of the proposition is straightforward. □

Similarly to Fourier, we can characterize the decay of differentiable functions:

Proposition 3.13. *Assume that f is $\mathcal{C}^{2p}$ (meaning its $2p$ differentiable with continuous $2p$ differential, on the sphere), then:*

$$\|K_n(f)\| = o(\frac{1}{n^{2p}})\,. \tag{3.55}$$

Proof. First, we observe that:

$$\forall P \in \mathcal{H}_n^d, -n(n+d-2)P = \Delta_{\mathcal{S}^{d-1}} P = K_n(\Delta_{\mathcal{S}^{d-1}} P) \tag{3.56}$$

and cancels if $P \in \mathcal{H}_m^d, m \neq n$. Thus, $K_n(\Delta_{\mathcal{S}^{d-1}} f) = \Delta_{\mathcal{S}^{d-1}} K_n(f)$, by density. Now, we have the convergence in L^2 of:

$$\Delta^p_{\mathcal{S}^{d-1}} f = \sum_n K_n(\Delta_{\mathcal{S}^{d-1}} f) = \sum_n \Delta_{\mathcal{S}^{d-1}} K_n(f) = \sum_n (-n(n+d-2))^p K_n(f)\,. \tag{3.57}$$

Thus $\|\Delta_{\mathcal{S}^{d-1}} K_n(f)\| \to 0$(by Cauchy criterium) and $\|K_n(f)\| = o(\frac{1}{n^{2p}})$. □

Reciprocally, we have:

Proposition 3.14. *Assume that $\|K_n(f)\|_\infty = o(\frac{1}{n^{2(p+1)}})$, then f is $\mathcal{C}^{2p}$.*

Proof. We thus have that $\sum_n \|\Delta^p K_n(f)\|_\infty$ converges normally and given $\sum_n K_n(f)$ also converges (and all its derivatives) and that each term is smooth, we have that $f = \sum_n K_n(f)$ is C^{2p}. □

Note that we have also been able to exploit the Lie group structure, as discussed in Chapter 2.

Chapter 4

Approximation properties of (shallow) Neural Networks

In this chaper, we will derive several bounds concerning the approximation and generalization power shallow neural networks. Notably, those bounds are based on [3]. We will mainly use the work done in the Chaper 3 of this document, that will allow us to derive some rates for spherical functions. We will then note that the weights of neural networks inherit a spherical symmetry, and thus are more amenable to be expressed as observed by [3, 10].

4.1 Convex Infinite width shallow neural networks

4.1.1 From $\mathbb{R}^d$ to $\mathcal{S}^{d-1}$

We now focus on the specific case of generic 1-hidden layer neural networks. We assume the neural network has no bias, as the bias can be removed via $w^T x + b = \begin{bmatrix} w \\ b \end{bmatrix}^T \begin{bmatrix} x \\ 1 \end{bmatrix}$. If Φ is a 1-hidden layer Neural Network with ReLU non-linearity, we introduce $w_k = \theta_k \|w_k\| \in \mathbb{R}^d$, s.t. $\theta_k \in \mathcal{S}^{d-1}$ and we note that

$\forall x \in \mathbb{R}^d$ and $\Gamma = [w_1, ..., w_K, \alpha_1, ..., \alpha_K]$:

$$\Phi(x;\Gamma) = \sum_{k \leq K} \alpha_k \rho(\langle x, w_k \rangle) \tag{4.1}$$

$$= \sum_{k \leq K} \frac{\alpha_k}{\|w_k\|} \rho(\langle x, \theta_k \rangle) \tag{4.2}$$

$$= \int_{\mathcal{S}^{d-1}} \rho(\langle x, \theta \rangle)\, d\mu_K(\theta)\,, \tag{4.3}$$

where $\mu_K = \sum_{k \leq K} \frac{\alpha_k}{\|w_k\|} \delta_{\theta_k}$ is a discrete measure. In order to apply the results of Chapter 2, let us assume that $\Phi(x;w)$ is only defined over half the unit ball of $\mathbb{R}^d$. It implies, by homogeneity of ρ, that if $\|x\| \leq \frac{1}{2}$, for $|t| = \sqrt{1 - \|x\|^2}$, we can define:

$$\tilde{\Phi}((x,t);\mu_K) \triangleq |t| \Phi(\frac{x}{t};\Gamma) = \Phi(x;\Gamma) \text{ if } t > 0\,. \tag{4.4}$$

By using $\tilde{\Phi}$, which is defined over the sphere and is parametrized by the sphere, we have thus used fully the symmetry of the problem. In fact, we will consider signals from $L^2(\mathcal{S}^{d-1})$. Indeed, fix as a reference measure the uniform measure σ on $\mathcal{S}^{d-1}$ as in Chapter 3, and for some given target function $f : \mathcal{S}^{d-1} \to \mathbb{R}$ (we will specify later the target set), we relax Eq. 4.3 in order to approximate f via:

$$f(x) \approx \int_{\mathcal{S}^{d-1}} \rho(\langle x, \theta \rangle) p(\theta) d\sigma(\theta) = \rho \circledast p(x)\,, \tag{4.5}$$

for $p \in L^2(\mathcal{S}^{d-1})$. This is a relaxation because $p(\theta)d\sigma(\theta)$ defines a finite measure (by Jensen inequality) for every $p \in L^2(\mathcal{S}^{d-1})$. We thus introduce:

$$\mathcal{F}_2 = \{\rho \circledast p, p \in L^2(\mathcal{S}^{d-1})\}, \tag{4.6}$$

that will study below. Interestingly, this is structured as a vector space.

4.1.2 $\mathcal{F}_2$ as a RKHS

In this section, we will see that we can derive many properties in a similar fashion to what is done in a Fourier-like framework. For $p \in L^2(\mathcal{S}^{d-1})$, using Prop. 3.12, we have:

$$K_n(\rho \circledast p)(x) = \lambda_n(\rho) K_n(p)(x)\,. \tag{4.7}$$

This implies the following Lemma:

Lemma 4.1. *Let $f \in L^2(\mathcal{S}^{d-1})$, then $f \in \mathcal{F}_2$ if and only if $\sum_{\lambda_n(\rho) \neq 0} \frac{\|K_n(f)\|^2}{\lambda_n(\rho)^2} < \infty$.*

Proof. Indeed, if $f \in \mathcal{F}_2$, then there is $p \in L^2(\mathcal{S}^{d-1})$ such that $f = \rho \circledast p$, and thus $\sum_{\lambda_n \neq 0} \|K_n(p)\|^2 = \sum_{\lambda_n \neq 0} \frac{\|K_n(f)\|^2}{\lambda_n(\rho)^2} < \infty$. Reciprocally, we can consider $\tilde{p} = \sum_{\lambda_n \neq 0} \frac{K_n(f)}{\lambda_n} \in L^2(\mathcal{S}^{d-1})$, by assumption. Yet, $f = \rho \circledast p$. □

Proposition 4.1. *$\mathcal{F}_2$ is a RKHS with norm given by*

$$\|f\|_{\mathcal{F}_2} = \inf_{f=\rho\circledast p, p\in L^2(\mathcal{S}^{d-1})} \|p\|. \tag{4.8}$$

and the corresponding kernel k is given by:

$$k(x,y) = \int_{\mathcal{S}^{d-1}} \rho(\langle x,z\rangle)\rho(\langle y,z\rangle)\, d\sigma(z)\,. \tag{4.9}$$

Proof. Indeed, let $f \in \mathcal{F}_2$, for a given decomposition $f = \rho \circledast p(x)$, given that $|\rho(\langle x,\theta\rangle)| \leq 1$, we have:

$$|\rho \circledast p(x)| \leq \|p\|_2\,.$$

In particular,

$$|\rho \circledast p(x)| \leq \inf_{f=\rho\circledast p, p\in L^2(\mathcal{S}^{d-1})} \|p\|_2\,.$$

Let $Tp = \rho \circledast p = \sum_n \lambda_n(\rho)K_n(f)$, s.t. $T^* = T$, and introduce next $Uf = \sum_{\lambda_n(\rho)>0} \lambda_n^{-1}(\rho)K_n(f)$. It's clear that:

$$TUT = T \text{ and } UTU = U$$

and $UT = TU$ thus $(UT)^* = UT$ and $(TU)^* = TU$. Thus, U is the pseudo-inverse of T. In fact, one has: $\inf_{f=\rho\circledast p, p\in L^2(\mathcal{S}^{d-1})} \|p\| = \|Uf\|$ which defines a norm on $\mathcal{F}_2$ (as $Uf = 0$ implies $f = 0$ for $f \in \mathcal{F}_2$). It's thus natural to introduce: $\langle f,g\rangle_{\mathcal{F}_2} = \langle Uf, Ug\rangle$. Now, we note that:

$$\begin{aligned}
f(x) &= Tp(x) = TUTp(x)\\
&= TUf\\
&= \int_{\mathcal{S}^{d-1}} \rho(\langle x,y\rangle)Uf(x)\, d\sigma(y)\\
&= \langle \rho(\langle x,.\rangle), Uf\rangle\\
&= \langle \rho(\langle x,.\rangle), UTUf\rangle\\
&= \langle (UT)^*\rho(\langle x,.\rangle), Uf\rangle\\
&= \langle U(T\rho(\langle x,.\rangle)), Uf\rangle\\
&= \langle T\rho(\langle x,.\rangle), f\rangle_{\mathcal{F}}
\end{aligned}$$

And thus, the kernel is given by:

$$k(x,y) = Tf\rho(\langle x,.)(y) = \int_{\mathcal{S}^{d-1}} \rho(\langle x,z\rangle)\rho(\langle y,z\rangle)d\sigma(z)$$

□

The first necessary tool to understand better $\mathcal{F}_2(\sigma)$ is to compute explicitely $\{\lambda_n(\rho)\}_n$. This is done via:

Proposition 4.2. *Assuming that $\rho(x) = \max(0, x)$, we get:*

$$\lambda_{2n+2}(\rho) = \frac{\Gamma(\frac{d-1}{2})}{(-2)^n \Gamma(n + \frac{d-1}{2})} (-1)^{\frac{n}{2}+1} \binom{n + \frac{d-3}{2}}{\frac{n}{2}} n! \sim C(d)(-1)^{n+1}(2n)^{-\frac{d+3}{2}}, \tag{4.10}$$

where $C(d) = \Gamma(\frac{d-1}{2})\sqrt{\frac{2}{n\pi}} 2^{\frac{d-3}{2}} e^{\frac{d-3}{2}}$ and $\lambda_{2n+3} = 0$.

Proof. Using Prop. 3.10, we need to compute:

$$\lambda_n(\rho) = \int_{-1}^{1} \max(0, t) P_n(t)(1 - t^2)^{\frac{d-3}{2}} dt = \frac{\Gamma(\frac{d-1}{2})}{(-2)^n \Gamma(n + \frac{d-1}{2})} \int_0^1 t \frac{d}{dt^n} \big((1 - t^2)^{n+\frac{d-3}{2}}\big)\, dt\,.$$

It follows:

$$\begin{aligned}
\int_0^1 t \frac{d}{dt^n} \big((1 - t^2)^{n+\frac{d-3}{2}}\big)\, dt &= \int_0^1 t \frac{1}{2^n} \frac{d}{dt^n} (1 - t^2)^{n+\frac{d-3}{2}}\, dt \\
&= \big([t \frac{d}{dt^{n-1}} (1 - t^2)^{n+\frac{d-3}{2}}]_0^1 - \int_0^1 \frac{d}{dt^{n-1}} (1 - t^2)^{n+\frac{d-3}{2}}\, dt\big) \\
&= -[\frac{d}{dt^{n-2}} (1 - t^2)^{n+\frac{d-3}{2}}]_0^1 \\
&= -\sum_{k=0}^{\infty} \binom{n + \frac{d-3}{2}}{k} (-1)^k [\frac{d}{dt^{n-2}} t^{2k} dt]_0^1
\end{aligned}$$

Now, for $n > 1$, we have:

$$\int_0^1 t \frac{d}{dt^n} \big((1 - t)^{n+\frac{d-3}{2}}\big)\, dt = \begin{cases} (-1)^{\frac{n}{2}+1} \binom{n+\frac{d-3}{2}}{\frac{n}{2}-1} n!, & \text{if } n - 2 = 2k \text{ or } n = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{4.11}$$

Note that $\binom{n}{k} = \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)}$, even for non-integer values. Using Lemma 3.1, we get an equivalent given by (having in mind that $\Gamma(z+1) \sim (2\pi z)^{1/2}(\frac{z}{e})^z$)(for

odd values of n):

$$\begin{aligned}
\lambda_n(\rho) &\sim \frac{\Gamma(\frac{d-1}{2})}{(-2)^n\Gamma(n+\frac{d-1}{2})}(-1)^{\frac{n}{2}+1}\frac{\Gamma(n+\frac{d-1}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{n}{2}+\frac{d+1}{2})}\Gamma(n+1)\\
&\sim \frac{\Gamma(\frac{d-1}{2})}{2^n}(-1)^{\frac{n}{2}+1}\frac{\Gamma(n+1)}{\Gamma(\frac{n}{2})\Gamma(\frac{n}{2}+\frac{d+1}{2})}\\
&\sim \frac{\Gamma(\frac{d-1}{2})}{2^n}(-1)^{\frac{n}{2}+1}\frac{\sqrt{2\pi n}\frac{n^n}{e^n}}{\sqrt{2\pi(\frac{n}{2}-1)}\frac{(\frac{n}{2}-1)^{\frac{n}{2}-1}}{e^{\frac{n}{2}-1}}\sqrt{2\pi\frac{(n+d-1)}{2}}\frac{\frac{n+d-1}{2}^{\frac{n+d-1}{2}}}{e^{\frac{n+d-1}{2}}}}\\
&\sim \Gamma(\frac{d-1}{2})\sqrt{\frac{2}{n\pi}}2^{\frac{d-3}{2}}e^{\frac{d-3}{2}}(-1)^{\frac{n}{2}+1}n^{\frac{3-d}{2}}\,.
\end{aligned}$$

We need also to compute:

$$\lambda_0 = \int_0^1 t(1-t^2)^{\frac{d-3}{2}}\,dt = [\frac{(1-t^2)^{\frac{d-1}{2}}}{\frac{d-1}{2}}]_0^1 = \frac{2}{d-1} \tag{4.12}$$

and

$$\begin{aligned}
\lambda_1 &= \frac{1}{1-d}\int_0^1 t\frac{d}{dt}\big((1-t^2)^{\frac{d-1}{2}}\big)\,dt\\
&= \int_0^1 t^2(1-t^2)^{\frac{d-3}{2}}\,dt > 0
\end{aligned}$$

where we just needed to verify it is not 0.

□

Now, we can then exhibit several elements which belong to $\mathcal{F}_2$.

Proposition 4.3 (Elements in $\mathcal{F}_2(\sigma)$)**.** *Let $f : \mathcal{S}^{d-1} \to \mathbb{R}$, of class $\mathcal{C}^{2k}$ with $2k \geq \frac{d+5}{2}$ and f odd, then $f \in \mathcal{F}_2(\sigma)$.*

Proof. From Prop. 3.13, $f \in L^2(\mathcal{S}^{d-1})$ and $\|K_n(f)\| = o(\frac{1}{n^{2k}})$. This implies from supra that $\frac{\|K_{2n}(f)\|^2}{\lambda_{2n}^2} \sim (2n)^{d+3-4k}$, and thus $f \in \mathcal{F}_2(\sigma)$ because of our assumption on k. □

4.2 Approximation properties of $\mathcal{F}_2$

4.2.1 Lipschitz function approximations

Now, we show that Lipschitz function can be well approximated in $\mathcal{F}_2$ and we derive the rate of convergence in the following proposition:

Proposition 4.4. *There exists $\delta > 0$ (that depends only on d) such that for any f η-Lipschitz, with $f(0) = 0$ and which is odd, there exists $p \in L^2(\mathcal{S}^{d-1})$ with $\|p \circledast \rho\|_{\mathcal{F}} \leq \delta$, and:*

$$\|\rho \circledast p - f\|_\infty \leq \mathcal{O}\big(\eta\log(\frac{\delta}{\eta})(\frac{\delta}{\eta})^{-\frac{2}{d-3}}\big)\,. \tag{4.13}$$

Proof. f is Lipschitz, thus it is an element of $L^2(\mathcal{S}^{d-1})$ as it is bounded because continuous on a compact. Furthermore, from the assumptions, $\|f\| \leq 2\eta$. We consider:

$$\begin{aligned}
|f \circledast P_r(x) - f(x)| &= |\int_{\mathcal{S}^{d-1}} P_r(\langle x, y\rangle)\big(f(y) - f(x)\big)d\sigma(x)| \\
&\leq \eta \int_{\mathcal{S}^{d-1}} P_r(\langle x, y\rangle)\|x - y\|d\sigma(x) \\
&= \sqrt{2}\eta \int_{\mathcal{S}^{d-1}} P_r(\langle x, y\rangle)\sqrt{1 - \langle x, y\rangle}d\sigma(x) \\
&\leq \mathcal{O}(d)\sqrt{2}\eta(1 - r^2) \int_{-1}^{1} \frac{(1 - t^2)^{\frac{d-1}{2}}}{(1 + r^2 - 2tr)^{\frac{d}{2}}}\sqrt{1 - t}\, dt
\end{aligned}$$

Lemma 4.2.

$$\int_{-1}^{1} \frac{(1 - t^2)^{\frac{d-1}{2}}}{(1 + r^2 - 2tr)^{\frac{d}{2}}}\sqrt{1 - t}\, dt = \mathcal{O}(\ln(1 - r))\,. \tag{4.14}$$

Proof. Here, we do $u = \sqrt{1 - t}$ leading to:

$$\begin{aligned}
\int_{-1}^{1} \frac{(1 - t^2)^{\frac{d-1}{2}}}{(1 + r^2 - 2tr)^{\frac{d}{2}}}\sqrt{1 - t}\, dt &= \int_0^{\sqrt{2}} \frac{\big(1 - (1 - u^2)^2\big)^{\frac{d-1}{2}}}{\big(1 + r^2 - 2r(1 - u^2)\big)^{\frac{d}{2}}}u^2 du \\
&\leq \mathcal{O}(d) \int_0^{\sqrt{2}} \frac{u^{d-1}}{\big((1 - r)^2 + 2ru^2)\big)^{\frac{d}{2}}}du \\
&\leq \mathcal{O}(d) \int_0^{\sqrt{2}} \frac{u^{2d-2}}{(1 - r)^d + r^{\frac{d}{2}}(\sqrt{2}u)^d}du \\
&\leq \mathcal{O}(d) \int_0^{2} \frac{1}{(1 - r)^d + r^{\frac{d}{2}}v}dv \\
&\leq \mathcal{O}(d)\frac{1}{r^{\frac{d}{2}}}\ln(2 + (\frac{1 - r}{2\sqrt{r}})^d) = \mathcal{O}(\ln(1 - r))
\end{aligned}$$

□

Thus, $\|f \circledast P_r - f\|_\infty \leq \mathcal{O}\big(\eta(1-r)\ln(1-r)\big)$. Let us now compute $\|P_r \circledast f\|$. Prop. 3.12 leads to:

$$K_n(f \circledast P_r) = r^n K_n(f)\,. \tag{4.15}$$

and thus, the norm of the candidate p is no more than:

$$\sum_{n\geq 0} \lambda_{2n}^{-2} r^{2n}\|K_n(f)\|^2 \leq \sup_n \lambda_{2n}^{-2} r^{2n}\eta^2 \tag{4.16}$$

$$= \mathcal{O}(\sup_n (2n)^{d+3} r^{2n})\eta^2 \tag{4.17}$$

$$= \mathcal{O}\big((1 - r)^{-d-3}\eta^2\big)\,, \tag{4.18}$$

where we have used the following lemma:

Lemma 4.3. *If $0 < r < 1, \alpha > 0$, then* $\sup_{x>0} x^\alpha r^x = \mathcal{O}((1-r)^{-\alpha})$.

Proof. Here, let $f(x) = x^\alpha r^x$, then $f'(x) = \alpha x^{\alpha-1} r^x + \ln r r^x x^\alpha = x^{\alpha-1} r^x (x \ln r + \alpha)$. Thus, there is a maximum at $x = -\frac{\alpha}{\ln r} > 0$ (because $\ln r \leq 0$), which is s.t.:

$$f(\frac{\alpha}{\ln r}) = \big(-\frac{\alpha}{\ln r}\big)^\alpha r^{-\frac{\alpha}{\ln r}} = \mathcal{O}((\ln \frac{1}{r})^{-\alpha}) = \mathcal{O}((1-r)^{-\alpha}) \tag{4.19}$$

□

Thus, if $(1-r)^{-\frac{d-3}{2}}\eta = \delta$, then $\big(\frac{\eta}{\delta}\big)^{\frac{2}{d-3}} = (1-r)$ and $\|f \circledast P_r - f\|_\infty = \mathcal{O}\big(\log(\frac{\delta}{\eta})(\frac{\delta}{\eta})^{-\frac{2}{d-3}}\big)$. □

Remark 4.1. *The bound could have been tighter by using that there exists $h \in L^2(\mathcal{S}^{d-1})$ s.t. $f = \Delta^{\frac{1}{2}} h$ yet I chosed to simplify the exposition.*

4.2.2 Finite neurons approximation

Now that we have derived several approximation properties in $\mathcal{F}_2$, we derive some results for finite width Neural Networks. This can be technically challenging, as typically, $x \to \langle x, \theta \rangle \notin \mathcal{F}_2$.

Proposition 4.5 (Random sampling, L^2-norm)**.** *Let $f \in L^2(\mathcal{S}^{d-1})$, then for any n, there exists $v_1, ..., v_n$ s.t.:*

$$\|f \circledast \rho - \sum_{i=1}^n \rho(v_i^T .) f(v_i)\| \leq \sqrt{\frac{8\pi d}{n}} \|f\| . \tag{4.20}$$

Proof. We follow a very standard scheme of proof. We introduce the r.v.s $v_i \sim \sigma$. We let $\delta_n(x) = \frac{1}{n}\sum_{i=1}^n \rho(v_i^T x) f(v_i) - f \circledast \rho(x)$ then $\mathbb{E}[\delta_n(x)] = 0, \forall x \in \mathcal{S}^{d-1}$. Furthermore:

$$\mathbb{E}[\|\delta_n\|^2] = \frac{1}{n}\mathbb{E}_v[\|\rho(.^T v) f(v)\|^2] - \|\rho \circledast f\|^2 , \tag{4.21}$$

yet here:

$$\mathbb{E}_v[\|\rho(.^T v) f(v)\|^2] = \int_{\mathcal{S}^{d-1}} \int_{\mathcal{S}^{d-1}} \rho^2(v^T x) f^2(v) d\sigma(v) d\sigma(x) = \|\rho\|^2 \|f\|^2 . \tag{4.22}$$

Consequently, we obtain:

$$\mathbb{E}[\|\delta_n\|^2] \leq \frac{1}{n}\big[\|\rho\|^2\|f\|^2 - \|\rho \circledast f\|^2\big] . \tag{4.23}$$

It implies that there exists $v_1, ..., v_n$ s.t.:

$$\|\frac{1}{n}\sum_{i=1}^n \rho(v_i^T x) f(v_i) - f \circledast \rho(x)\|^2 \leq \frac{1}{n}\big[\|\rho(\langle e, \rangle)\|^2 \|f\|^2 - \|\rho \circledast f\|^2\big] , \tag{4.24}$$

where e is any fixed vector. Now, it's enough to note that:

$$\|\rho\|^2 = \int_{\mathcal{S}^{d-1}} \rho^2(x^T y) d\sigma(y) = \frac{\Lambda_{d-1}}{\Lambda_d} \int_0^{\frac{\pi}{2}} \sin^{d-2}(\theta) - \sin^d(\theta) d\theta = \frac{\Lambda_{d-1}}{\Lambda_d} \frac{1}{d-1} W_d \tag{4.25}$$

where $W_d = \int_0^{\frac{\pi}{2}} \sin^d(\theta) d\theta = \frac{\Gamma(\frac{d+1}{2})\sqrt{\pi}}{2\Gamma(\frac{d}{2}+1)}$. (we have recognized a Wallis integral...) Thus,

$$\|\rho\|^2 = \frac{\Gamma(\frac{d}{2}) 2\pi^{d/2}}{\Gamma(\frac{d-1}{2}) 2\pi^{(d-1)/2}} \frac{\Gamma(\frac{d+1}{2})\sqrt{\pi}}{2\Gamma(\frac{d}{2}+1)} \frac{1}{d-1} = \frac{\pi d}{8}. \tag{4.26}$$

□

Remark 4.2. *It is crucial to observe that our bounds are obtained with the L^2-norm rather than the L^∞-norm. In fact, by considering functions parametrized by a finite measure μ (which means that $|\mu|(\mathcal{S}^{d-1}) < \infty$), we consider could instead for such a measure μ:*

$$f(x) = \int_{\mathcal{S}^{d-1}} \rho(\langle x, y \rangle) \, d\mu(y) \, . \tag{4.27}$$

Then, [3] shows tighter bounds and adaptativity of the approximation bounds to the target function regularity.

Chapter 5

Lazy regime to train Neural Networks

In this chapter, we propose to analyze the training of Neural Networks in several particular cases where they behave like their linearized counter-parts in a neighborhood of their initialization. First, we discuss some basic properties of the gradient of those highly non-linear models, then we will discuss several standard properties of the Neural Tangent Kernel [21] which is a particular case of the Lazy Training regime [13].

5.1 Training a Neural Network

5.1.1 A note on the back-propagation mechanism

First, we review differentiable properties of deep models, viewed as a real-valued function, and in particular how to compute the gradient w.r.t. their parameters. We will assume by now that $\Phi : \mathbb{R}^p \times \Omega \to \mathbb{R}^k$ is a.s. differentiable, and that the loss function $\ell : \mathbb{R}^k \to \mathbb{R}_+$ is L-Lipschitz.

We remind that the gradient of $\ell \circ \Phi$ taken at $\mathbf{p} = (W, x)$, where $W \in \mathbb{R}^d, x \in \Omega$ along $\mathbf{x}$ for a cost ℓ is obtained from the differential via:

$$\langle \nabla_x (\ell \circ \Phi)(\mathbf{p}), \mathbf{x} \rangle \triangleq \partial_x (\ell \circ \Phi)_{\mathbf{p}}(\mathbf{x}) \,,$$

where $\langle ., . \rangle$ is here the standard Euclidean scalar product. On the other hand, by the chain rule, we get:

$$\partial_x (\ell \circ \Phi)_{\mathbf{p}}(\mathbf{x}) = \partial \ell_{\Phi(\mathbf{p})} \circ \partial_x \Phi_{\mathbf{p}}(\mathbf{x}) = \langle \nabla \ell(\Phi(\mathbf{p})), \partial_x \Phi_{\mathbf{p}}(\mathbf{x}) \rangle = \langle \partial_x \Phi_{\mathbf{p}}^T \nabla \ell(\Phi(\mathbf{p})), \mathbf{x} \rangle \,,$$

which implies that:

$$\nabla_x (\ell \circ \Phi)(\mathbf{p}) = \partial_x \Phi_{\mathbf{p}}^T \nabla \ell(\Phi(\mathbf{p})) \,.$$

This has several implications on $\nabla_x (\ell \circ \Phi)(\mathbf{p})$: first, one has to compute $\Phi(\mathbf{p})$ (*forward* pass) from the first to the last layer, then, one evaluate $\partial_x \Phi_{\mathbf{p}}^T$ (*backward*

pass) from the last to the first layer. Let $f_1, ..., f_J$ be some functions with parameters $(\theta_1, ..., \theta_J)$ such that $f_{j+1} \circ f_j(x_j) \triangleq f_{j+1}(f_j(x_j; \theta_j); \theta_{j+1})$ makes sense (note we explicitly remove the dependency in θ_j, θ_{j+1} here). This implies that if $\Phi x = f_J \circ ... \circ f_1(x)$, then:

$$\partial_x \Phi^T \mathbf{x} = (\partial_x f_1)_x^T ... (\partial_{x_{J-1}} f_J)_{f_{J-1}(x)}^T \mathbf{x} \,. \tag{5.1}$$

This leads to the celebrated *back-propagation* algorithm, and now well-known automatic differentiation tools: note that for each f_j, we need to know $f_j, \partial f_j$ to further obtain:

Proposition 5.1. *If $\Theta = (\theta_1, ..., \theta_J)$, we have:*

$$\nabla_{\theta_j}(\ell \circ \Phi)(\Theta; x) = (\partial_{\theta_j} f_j)_{f_{j-1}(x)}^T (\partial_{x_j} f_{j+1})_{f_j(x)}^T ... (\partial_{x_{J-1}} f_J)_{f_{J-1}(x)}^T \nabla \ell(\Phi x) \tag{5.2}$$

Furthermore, for instance if $f_j(x_j, W_j) = \rho W_j x_j$, we have:

$$\nabla_{W_j} \ell(\Phi x) = [\partial \rho]_{W_j f_j(x)} \nabla_{x_j} \ell(f_J \circ ... \circ f_{j+1}) f_j(x)^T \tag{5.3}$$

Proof. Almost everything is direct, and the last claim follows via: $\langle \nabla_{W_j} \ell(\Phi x), W \rangle = \mathrm{Tr}(x_j \nabla_{x_j} \ell(f_J \circ ... \circ f_{j+1})^T [\partial \rho]_{W_j x_j} W)$ □

Note that the scalar product linked to A, B matrix is here $\mathrm{Tr}(A^T B)$.

5.1.2 The "best" non-convex convergence rate with SGD

We now discuss a standard convergence rate in non-convex optimization [8], which requires to define and study the risk $\mathcal{R}$:

$$\mathcal{R}(\Phi(W)) \triangleq \mathbb{E}_X[\ell \circ \Phi(X; W)] \,, \tag{5.4}$$

Proposition 5.2. *If ℓ is L-Lipschitz, then $\mathcal{R}$ is L-lipschitz and non-negative.*

Our goal, in the following, will be to minimize $\mathcal{R}$:

$$\inf_W \mathcal{R}(\Phi(W)). \tag{5.5}$$

Here, we show under minimal assumptions that a gradient descent will converge to a local minimum. For a batch $\mathcal{B} = (x_1, ..., x_{|\mathcal{B}|})$ of data, consider the Stochastic Gradient Descent step given by:

$$W_{t+1} = W_t - \eta_t \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \nabla_W (\ell \circ \Phi)(W_t; X_i) \,. \tag{5.6}$$

Following [9], we would like to design a Lyapunov function that will constrain the optimization path to a local minimum. We thus introduce $V : \mathbb{R}^d \to \mathbb{R}_+$:

$$V(W) = (\mathcal{R} \circ \Phi)(W) - \inf_{\tilde{W}} (\mathcal{R} \circ \Phi)(\tilde{W}) \,, \tag{5.7}$$

and we note that if:

$$g(W) \triangleq \frac{1}{\mathcal{B}} \sum_{i=1}^{|\mathcal{B}|} \nabla_W (\mathcal{R}_{\mathcal{B}} \circ \Phi)(W; X_i)\,, \tag{5.8}$$

then,

$$\mathbb{E}[g(W)] = V'(W)\,. \tag{5.9}$$

Via Prop. 5.2, we know that V is differentiable, with gradients L-Lipschitz, in other words that V is L-smooth, implying that:

$$V(W) \leq V(\tilde{W}) + V'(\tilde{W})^T (W - \tilde{W}) + \frac{L}{2} \|W - \tilde{W}\|^2 \tag{5.10}$$

Then the computations are straighforward, as $W_{t+1} = W_t - \eta_t g(W_t)$, and:

$$\mathbb{E}[V(W_{t+1})|W_t] \leq V(W_t) - \eta_t \mathbb{E}[g(W_t)]^T V'(W_t) + \eta_t^2 \frac{L}{2} \|g(W_t)\|^2\,, \tag{5.11}$$

Proposition 5.3. *If $\sum_t \eta_t^2 < \infty$, $\sum_t \eta_t = \infty$ and assume that g has a bounded variance, i.e., $\exists M > 0 : \mathbb{E}[\|g(W)\|^2] \leq M + \|V(W)\|^2$, then:*

$$\liminf_t \mathbb{E}[\|V'(W_t)\|^2] = 0\,. \tag{5.12}$$

Proof. From supra, we get:

$$\mathbb{E}[V(W_{t+1})] \leq \mathbb{E}[V(W_t)] - (\eta_t - \frac{\eta_t^2 L}{2}) \mathbb{E}[\|V'(W_t)\|^2] + \frac{\eta_t^2 L}{2} M\,, \tag{5.13}$$

and summing up to T leads to:

$$\sum_{t=0}^{T-1} (\eta_t - \frac{\eta_t^2 L}{2}) \inf_{0 \leq t \leq T} \mathbb{E}[\|V'(W_t)\|^2] \leq \sum_{t=0}^{T-1} (\eta_t - \frac{\eta_t^2 L}{2}) \mathbb{E}[\|V'(W_t)\|^2] \tag{5.14}$$

$$\leq \mathbb{E}[V(W_0)] + \sum_{t=0}^{T-1} \frac{\eta_t^2 L}{2} M - \mathbb{E}[V(W_T)] < \infty \tag{5.15}$$

which implies, as $T \to \infty$, that $\inf_{t \leq T} \mathbb{E}[\|V'(\omega_t)\|^2] \to 0$. □

We note those bounds are mainly vacuous because they do can not guarantee anything about the quality of the (only local) optimum reached.

5.1.3 Compacity of the training path on a finite horizon

By now, use a more general formalism. Let $(\mathcal{F}, \langle ., . \rangle_{\mathcal{F}})$ be some Hilbert space of functions defined over an open set $\Omega \subset \mathbb{R}^d$. It could be, for instance, the Sobolev space $\mathcal{H}^1(\Omega) = \{f \in L^2(\Omega), \partial_{u_i} f \in L^2(\Omega), 1 \leq i \leq d\}$. We also assume that Ω is bounded (thus our data are assumed to belong to a compact, which is the case for practical applications). In this particular case, $\mathbf{I} \in \mathcal{F}$ (which will be an useful identification). Note that $\mathcal{F}$ could also potentially have a RKHS structure, meaning that $\delta_x : x \to f(x)$ is continuous for $\|.\|_{\mathcal{F}}$.

Remark 5.1. *In this context, it is natural to consider the squared loss* $\ell(\Phi) = \|\Phi - \Phi^*\|_{\mathcal{F}}^2$ *for some target* $\Phi^* \in \mathcal{F}$.

We now assume that $W \to \Phi(W) \in \mathcal{F}$ is a.s. differentiable with differentiable locally-Lipschitz, and that $\mathcal{R} : \mathcal{F} \to \mathbb{R}_+$ is differentiable and L-smooth. We consider the training ODE (which is the continuous version of the section above) given by:

$$\frac{d}{dt}W(t) = -\lambda \nabla_W (\mathcal{R} \circ \Phi)(W(t))\,, \tag{5.16}$$

where $\lambda > 0$ is a step length. In all generality, we have the following proposition:

Proposition 5.4. *For a finite horizon* T*, the risk is bounded along the path of optimization for* $t \in [0, T]$*:*

$$\mathcal{R}(\Phi(W(t)) \leq \mathcal{R}(\Phi(W(0)))\,, \tag{5.17}$$

the trajectory of W *is bounded along the path of optimization:*

$$\|W(t) - W(0)\| \leq \sqrt{T\lambda \mathcal{R}(\Phi(W(0))}\,, \tag{5.18}$$

Proof. For the first claim, it's standard to notice that $\frac{d}{dt}(\mathcal{R} \circ \Phi)(W(t)) = -\frac{1}{\lambda}\|\frac{d}{dt}W(t)\|^2 = -\lambda\|\nabla_W(\mathcal{R} \circ \Phi)(W(t))\|^2$, thus $\mathcal{R}(\Phi(W(t)) \leq \mathcal{R}(\Phi(W(0))$. We see that, by Cauchy-Schwartz:

$$\|W(t) - W(0)\|^2 \leq \big(\int_0^t \|\frac{d}{dt}W(t)\|\big)^2 \leq -\lambda T \int_0^t \frac{d}{dt}(\mathcal{R} \circ \Phi)(W(t)) \leq T\lambda \mathcal{R}(\Phi(W(0)))\,. \tag{5.19}$$

□

We also remind:

Lemma 5.1 (Grönwall's lemma)**.** *If for any* $t \geq 0$*,* $y'(t) \leq ay(t) + b(t), a > 0$*, then, for* $t \geq 0$*,* $y(t) \leq y(0)e^{at} + e^{at}\int_0^t b(u)\,du$*.*

Proof. We note that this is equivalent to: $e^{-at}(y'(t) - ay(t)) \leq b(t)e^{-at} \Rightarrow \frac{d}{dt}(e^{-at}y(t)) \leq b(t)e^{-at}$, thus $y(t) \leq y(0)e^{at} + e^{at}\int_0^t b(u)e^{-au}\,du \leq y(0)e^{at} + e^{at}\int_0^t b(u)\,du$. □

5.2 Wide linear networks

Lazy regimes correspond to training regimes of neural networks in which the neural network behaves as a linearization around its (random) initialization, like a linear model. This is in particular the case when the width of the neural network becomes arbitrary large, as we will see below.

5.2.1 Neural Tangent Kernels (NTKs)

We will train a real-valued neural network *without bias*, using the NTK rescaling, given for $W = [W_1, ..., W_J]$ by:

$$\Phi(W; x) = W_J \frac{1}{\sqrt{w_J}} \rho W_{J-1} \frac{1}{\sqrt{w_{J-1}}} \rho ... \rho W_1 \frac{1}{\sqrt{w_1}} x \,, \tag{5.20}$$

where $\sqrt{w_1}, ... \sqrt{w_J}$ correspond to the input width of each respective layer $W_1, ..., W_J$. For instance w_1 is equal to the dimension of x. We will also write $\Phi_j(W_1 ..., W_j; x) = \frac{1}{\sqrt{w_j}} W_j \rho ... \rho \frac{1}{\sqrt{w_1}} W_1 x$ s.t. $\Phi = \Phi_J$. We assume that our objective is to minimize a risk $\mathcal{R}(\Phi)$ as defined above, we thus consider:

$$\frac{d}{dt} W(t) = -\lambda \nabla_W \big(\mathcal{R} \circ \Phi)(W(t)\big) \tag{5.21}$$

where $\lambda > 0$ is some step size. The dynamic also writes:

$$\frac{d}{dt} W(t) = -\lambda \partial_W \Phi(W(t))^T \nabla \mathcal{R}(\Phi(W(t)))\,, \tag{5.22}$$

and we can introduce the Neural Tangent Kernel:

Definition 5.1 (Neural Tangent Kernel (NTK))**.** *The NTK at W is an operator of $\mathcal{F}$ defined by:*

$$K_W \triangleq \partial_W \Phi(W) \partial_W \Phi(W)^T \,. \tag{5.23}$$

In this case, we have:

$$\frac{d}{dt} \Phi(W(t)) = -\lambda K_{W(t)} \nabla \mathcal{R}(\Phi(W(t))\,. \tag{5.24}$$

For the real valued case, assuming an underlying RKHS structure, we assume that $\delta_x \in \mathcal{F}$, which implies that for any vector $e \in \mathbb{R}^p$:

$$\langle \delta_x, \partial_W \Phi(W).e \rangle_{\mathcal{F}} = \nabla_W \Phi(W; x)^T e \,. \tag{5.25}$$

Proposition 5.5. *Assume that $\mathcal{F}$ is a RKHS. Then, K_W is completely determined by $x \to \nabla_W \Phi(W; x)$ and:*

$$\langle \delta_x, K_W \delta_{x'} \rangle_{\mathcal{F}} = \nabla_W \Phi(W; x)^T \nabla_W \Phi(W; x')\,. \tag{5.26}$$

Proof. Using supra, we have:

$$\begin{aligned} \langle \delta_x, K_W \delta_{x'} \rangle_{\mathcal{F}} &= \langle \partial_W \Phi^T \delta_x, \partial_W \Phi^T \delta_{x'} \rangle \\ &= \nabla_W \Phi(W; x)^T \nabla_W \Phi(W; x')\,, \end{aligned}$$

and since this is true for any x, x', we get the result. □

We can thus canonically abuse of the notation $K_W(x,x') = \nabla_W \Phi(W;x)^T \nabla_W \Phi(W;x')$, as a real-valued kernel. We thus decide to denote:

$$\Sigma_1(x,x') = \frac{1}{w_1} x^T x', \tag{5.27}$$

and also:

$$\Sigma_{j+1}(x,x') = \mathbb{E}_{(u,v)\sim\mathcal{N}(0,\begin{bmatrix}\Sigma_j(x,x) & \Sigma_j(x',x)\\ \Sigma_j(x,x') & \Sigma_j(x',x')\end{bmatrix})}[\rho(u)\rho(v)], \tag{5.28}$$

and:

$$\dot{\Sigma}_{j+1}(x,x') = \mathbb{E}_{(u,v)\sim\mathcal{N}(0,\begin{bmatrix}\Sigma_j(x,x) & \Sigma_j(x',x)\\ \Sigma_j(x,x') & \Sigma_j(x',x')\end{bmatrix})}[\dot{\rho}(u)\dot{\rho}(v)]]\,. \tag{5.29}$$

We finally assume that for any depth j, $(W_j)_{ab} \sim \mathcal{N}(0,1)$. We will need the concept of Gaussian Process:

Definition 5.2 (Gaussian Process)**.** *A random process $\{f(x)\}_{x\in\mathcal{X}}$ is a Gaussian Process if and only if for any $x_1,...,x_n \in \mathcal{X}$, $(f(x_1),...,f(x_n))$ is a multi-variate Gaussian. In this case, f is completely determined by its kernel K, s.t. for $x,x' \in \mathcal{X}$ $K(x,x') = \mathbb{E}[f(x)f(x')]$.*

We now provide an explicit equation of the NTK kernel at the initialization when the widths of the layers grow to infinity, as done in [21]:

Proposition 5.6. *If $W(0)$ corresponds to the random parameters of a neural network such that each entry is initialized as $\mathcal{N}(0,1)$, in the infinite width limit, we get*

$$\exists \lim_{w_J\to\infty} ...\exists \lim_{w_2\to\infty} K_{W(0)}(x,x') \overset{law}{\to} K_{NTK}(x,x')\,, \tag{5.30}$$

(the convergence in law and the order of the limits are important), where K_{NTK} is deterministic. Furthermore, in this limit, $\forall j, \Phi_j(x,W(0))$ is a Gaussian process with covariance $\Sigma_j(x,x')$. Finally, we have:

$$K_{NTK} = \sum_{j=1}^{J} \Sigma_j \dot{\Sigma}_{j+1} ... \dot{\Sigma}_J\,. \tag{5.31}$$

Proof. We show by recursion on the depth j that the j-th output and limiting kernels are deterministic. For $j=1$, the result is obvious as W_1 is a projection and:

$$= \frac{1}{w_1} x^T \mathbb{E}[W_1^T W_1] x' = \mathbf{I}_{w_2} \frac{1}{w_1} x^T x' = \mathbf{I}_{w_2} \Sigma_1(x,x')\,. \tag{5.32}$$

and $K_{NTK}(x,x') = \Sigma_1(x,x')\mathbf{I}_{w_2}$ is deterministic. It is clear that Φ_1 is a Gaussian process since it is constant in x. Assuming it's true at rank J, we then get at rank $J+1$, by differentiation:

$$K_{W(0)}(x,x') = \frac{1}{w_{J+1}} \rho\Phi_J(x)^T \rho\Phi_J(x') \mathbf{I}_{w_{J+2}} \tag{5.33}$$

$$+ \frac{1}{w_{J+1}} W_{J+1} [\partial\rho]_{\Phi_J} \partial_W \Phi_J(x) \partial_W \Phi_J(x')^T [\partial\rho]_{\Phi_J}^T W_{J+1}^T \tag{5.34}$$

Note that $[\partial\rho]_{\Phi_J}$ is a diagonal matrix of size w_{J+1}. First, we note that: $\Phi_{J+1} = \frac{1}{\sqrt{w_{J+1}}} W_{J+1}\rho\Phi_J$, and $\mathbb{E}[\rho(\Phi_J x)^T \rho(\Phi_J x')] = \Sigma_{J+1}(x, x')$, by induction. Furthermore, each $(\Phi_J x)_k$ is sampled from iid centered Gaussians: and thus from central limit theorem (as the limit is independent from the activations), $\frac{1}{\sqrt{w_{J+1}}} W_{J+1}\rho\Phi_J$ tends in law to a Gaussian with expectation:

$$\mathbb{E}[W_{J,k}\rho(\Phi_j x)_k] = 0\,.$$

By hypothesis and the law of large numbers, $\frac{1}{w_{J+1}}(\rho\Phi_J)^T(\rho\Phi_J)\mathbf{I}_{w_{J+2}} \to \Sigma_{J+1}\mathbf{I}_{w_{J+2}}$, $\partial_W\Phi_J(x)\partial_W\Phi_J(x')^T \to \sum_{j=1}^J \Sigma_j\dot{\Sigma}_{j+1}...\dot{\Sigma}_J\mathbf{I}_{w_{J+1}}$. Furthermore, by induction hypothesis $\frac{1}{w_{J+1}} W_{J+1}[\partial\rho]_{\Phi_J x}[\partial\rho]^T_{\Phi_J x'} W^T_{J+1} \to \dot{\Sigma}_{J+1}(x, x')\mathbf{I}_{w_{J+2}}$, and the law of large numbers now for $w_{J+1} \to \infty$ combined to the inducion hypothesis leads to:

$$\frac{1}{w_{J+1}} W_{J+1}[\partial\rho]_{\Phi_J}\partial_W\Phi_J(x)\partial_W\Phi_J(x')^T[\partial\rho]^T_{\Phi_J} W^T_{J+1} \to \dot{\Sigma}_{J+1}\big(\sum_{j=1}^J \Sigma_j\dot{\Sigma}_{j+1}...\dot{\Sigma}_J\big)\mathbf{I}_{w_{J+2}}\,. \tag{5.35}$$

Combining all the two equations allow to conclude. □

Proposition 5.7. *For ρ a ReLU non-linearity, K_{NTK} restricted to $\mathcal{S}^{d-1}$ is positive definite.*

Proof. See the exercise sheet. □

5.2.2 Infinite width Neural Networks

Now, we will show that under a NTK rescaling, wide neural networks weights do not move much from their initialization, and that the corresponding neural network behaves essentially like its the linearization of its initialization. We'll need and use the following lemma:

Lemma 5.2. *Fix w_1, W_{J+1}. We have $C > 0$ such that for any $w_2, ..., w_J$, if $\rho(0) = 0$ and is Lipschitz then:*

$$\mathbb{E}[\|\Phi(x; W(0))\|] \leq C\,. \tag{5.36}$$

Proof. Indeed, as for any vector $x \in \mathbb{R}^{n_j}$ and W_j with iid standardized Gaussians, we get $\mathbb{E}xW_j^T W_j x = w_{j+1}\|x\|^2$ and assumign ρ is L-Lipschitz:

$$\mathbb{E}[\|\Phi(x; W)\|]^2 \leq \mathbb{E}[\|\Phi(x; W)\|^2] \tag{5.37}$$

$$= \mathbb{E}[\|\Phi_J(x; W)\|^2] \tag{5.38}$$

$$= \mathbb{E}[\|W_J\rho\Phi_{J-1}x\|^2] \leq w_{J+1}\mathbb{E}[\|\rho\frac{1}{\sqrt{w_J}}\Phi_{J-1}x\|^2] \leq \frac{w_{J+1}}{w_J} wL\mathbb{E}[\|\Phi_{J-1}x\|^2] \tag{5.39}$$

and we get the result by induction, as w_{J+1} is constant. □

In this subsection, we assume all the layers have the same width $w = w_J = ... = w_2$ and that $w_1 = 1$ (it is just a constant rescaling factor anyway). We also assume that that the learning rate is constant.

Proposition 5.8. *We consider the setting of Sec. 5.1.3. Assume that* $\lambda \int_0^T \|\nabla \mathcal{R}(\Phi(W(t)))\| \, dt = \mathcal{O}(1)$*, as the layers grow. Assume that* ρ *is* 1*-Lipschitz. Then:*

$$\sup_{t\in[0,T]} \|K_{W(t)} - K_{W(0)}\| = \mathcal{O}(\frac{1}{w^{1/2}}) \tag{5.40}$$

Proof. For $t \in [0, T]$ and $1 \leq j \leq J$, we have:

$$\frac{d}{dt} W_j(t) = -\lambda \frac{1}{\sqrt{w}} \big(\partial \rho^T_{W_j \Phi_{j-1}} ... \frac{1}{\sqrt{w}} W^T_{J-1} \partial \rho^T_{\Phi_{J-1}} \frac{1}{\sqrt{w}} W^T_J \nabla \mathcal{R}(\Phi(W(t))\big) \Phi^T_{j-1} \, . \tag{5.41}$$

Here, an identification $\langle a, Wb\rangle_{\mathcal{F}} = a^T bW$ has been done. We introduce: $u(t) = (\|W_1(t) - W_1(0)\| + \|W_1(0)\|, ..., \|W_{J-1}(t) - W_{J-1}(0)\| + \|W_{J-1}(0)\|, \|W_J(t) - W_J(0)\| + \|W_J(0)\|)$. From the assumptions, as ρ is 1-Lipschitz, we get for $J > j \geq 1$:

$$\begin{aligned}
\|\frac{d}{dt} W_j(t)\| &\leq \frac{\lambda_0}{w^{1/2}} \prod_{k\neq j, k<J} \frac{\|W_k(t)\|}{\sqrt{w}} \|\|W_J(t)\|\|\nabla \mathcal{R}(\Phi(W(t))\| \\
&\leq \frac{1}{w^{(J-1)/2}} \big(\sqrt{\sum_{k\neq j} \frac{1}{(J-1)} \|W_k(t)\|^2 + \|W_J(t)\|^2}\big)^{J-1} \|\nabla \mathcal{R}(\Phi(W(t))\| \\
&\leq \frac{C'}{w^{(J-1)/2}} \|u(t)\|^{J-1} \|\nabla \mathcal{R}(\Phi(W(t))\|
\end{aligned}$$

and for $j = J$:

$$\begin{aligned}
\|\frac{d}{dt} W_J(t)\| &\leq \frac{1}{w^{(J-1)/2}} \big(\sqrt{\sum_{k<J} \frac{1}{(J-1)} \|W_k(t)\|^2}\big)^{J-1} \|\nabla \mathcal{R}(\Phi(W(t))\| \\
&\leq \frac{C'}{w^{(J-1)/2}} \|u(t)\|^{J-1} \|\nabla \mathcal{R}(\Phi(W(t))\|
\end{aligned}$$

where $C' > 0$ is some constant which depends on J. In this case:

$$|\frac{d}{dt} \frac{1}{\|u(t)\|^{J-2}}| \leq J \frac{|u(t).\frac{d}{dt} u(t)|}{\|u(t)\|^J} \leq J \frac{\|\frac{d}{dt} u(t)\|}{\|u(t)\|^{J-1}} \leq \frac{JC'}{w^{(J-1)/2}} \|\nabla \mathcal{R}(\Phi(W(t))\| \tag{5.42}$$

Now, we know that $|\frac{1}{\|u(0)\|^{J-2}} - \frac{1}{\|u(t)\|^{J-2}}| = \mathcal{O}(\frac{1}{w^{(J-1)/2}})$, thus, as $0 < \|u(0)\|$ is bounded, we get that $\|u(t)\|$ is bounded. It implies in particular that for $j \leq J$, we can find $\tilde{C} > 0$ such that:

$$\frac{d\|W_j(t) - W_j(0)\|}{dt} \leq \|\frac{d}{dt} W_j(t)\| \leq \frac{\tilde{C}}{w^{(J-1)/2}} \|\nabla \mathcal{R}(\Phi(W(t)))\|$$

Now, we see we can upper bound each $\|W_j(t) - W_j(0)\|$ by $\frac{\tilde{C}}{w^{(J-1)/2}}$. To conclude simply note $\partial_{W_J}\Phi$ has bounded variations w.r.t. to parameters.

Now, since $\|uu^T - vv^T\| = \|(u+v)(v-u)^T\| \leq (\|u\| + \|v\|)\|u - v\|$, we have:

$$\sup_{t\in[0,T]} \|K_{W(t)} - K_{W(0)}\| = \mathcal{O}(\frac{1}{w^{1/2}})\,. \tag{5.43}$$

□

Remark 5.2. *This proposition implies in particular that* $\|K(W(t)) - K(W(0))\| \to 0$. *Observe that this model has an implicit rescaling* $\Phi' = \frac{1}{w^{\frac{J-1}{2}}}\Phi$.

This seems to imply that the dynamic of an infinite width neural networks is linear. Let's go further and consider:

$$\frac{d}{dt}\Phi^t = -\lambda\partial_W(\Phi(W(0))\partial_W(\Phi(W(0))^T\nabla\mathcal{R}(\Phi^t) \tag{5.44}$$

which is well-defined because $K_{W(0)}$ is continuous from Prop. 5.6. Then:

Proposition 5.9. *For a given* $t \in [0,T]$, *we have that:*

$$\sup_t |\Phi^t - \Phi(W(t))| = \mathcal{O}(\frac{1}{w^{1/2}}) \tag{5.45}$$

Proof. Write $\Delta(t) = \|\Phi^t - \Phi(W(t))\|$. We note that:

$$\begin{aligned}\frac{d}{dt}\Delta(t) &\leq \lambda\|K_{W(0)}\nabla\mathcal{R}(\Phi^t) - K_{W(t)}\nabla\mathcal{R}(\Phi(W(t)))\| \\ &\leq \lambda\|(K_{W(t)} - K_{W(0)})\nabla\mathcal{R}(\Phi(W(t)))\| + \lambda\|K_{W(0)}\|L\|\Phi^t - \Phi(W(t))\| \\ &\leq \lambda\|\nabla\mathcal{R}(\Phi(W(t))\| + C'\Delta(t)\end{aligned}$$

Thus, from Lemma 5.1:

$$\Delta(t) = \mathcal{O}(\frac{1}{w^{1/2}}) \tag{5.46}$$

□

In other words, the asymptotic dynamic of this randomly initialized Neural Network is linear. Indeed, take for instance $\mathcal{R}(\Phi) = \mathbb{E}_{\mathcal{X}}\|\Phi - \Phi^*\|^2$ s.t. $\nabla\mathcal{R}(\Phi) = \Phi - \Phi^*$, which satisfies the assumption of Prop. 5.8. Then it's clear that 5.44 is of type $\frac{d}{dt}\Phi^t(x) = A(\Phi^t - \Phi^*)$ for A some linear operator, whose solution is given by $\Phi^t(x) = \Phi^* + e^{tA}(\Phi^0 - \Phi^*)$, and the convergence is clear at the moment that A has a positive spectrum. This is in particular the case if A is given by Prop. 5.7.

5.3 Lazy training

We have found out that wide neural networks, under the NTK renormalization, behave like a kernel defined by by their initialization. It is possible to deduce a more general property. Indeed, we note that the previous formulation writes:

$$\frac{d}{dt}W(t) = -\lambda \nabla_W (\mathcal{R} \circ (\frac{1}{\sqrt{w_J}}\Phi))(W(t)) \tag{5.47}$$

with the initialization $W_J \sim \mathcal{N}(0, \mathbf{I}_{w_J})$. We note that during training, each step size is rescaled by a constant factor $\frac{1}{\sqrt{w_J}}$. We have observed that letting w_J growing leads to an optimization path that is dominated by the initialization, via a renormalization phenomenon. We will show this phenomenon is not specific to NTK, and to do so, we introduce the linearization of Φ (of its parameters) around the initialization:

$$\bar{\Phi}(W) = \Phi(W(0)) + \partial_W \Phi(W(0))^T (W - W(0)) \,. \tag{5.48}$$

We will show that the variation of the loss can be huge compared to the variations of the weight of the neural network. This can be intuitively quantified, via the first iteration step $W_1 = W_0 - \lambda \nabla(\mathcal{R} \circ \Phi)(W_0)$ along a discrete optimization path. Note that the linearization around W_1 and W_0 remains close if:

$$\Phi(W_1) + \partial_W \Phi(W_1)^T (W - W_1) \approx \Phi(W_0) + \partial_W \Phi(W_0)^T (W - W_0) \,. \tag{5.49}$$

This implies in particular that the order 1 coefficient, $\partial_W \Phi(W_1) - \partial_W \Phi(W_0)$, must remain small, which is quantified via a 2nd order approximation:

$$\frac{\partial_W \Phi(W_1) - \partial_W \Phi(W_0)}{\partial_W \Phi(W(0))} \sim \lambda \|\nabla_W (\mathcal{R} \circ \Phi)(W_0)\| \frac{\|\mathcal{H}\Phi(W_0)\|}{\|\partial_W \Phi(W_0)\|} \,. \tag{5.50}$$

It has to be small, compared to the variations of the loss:

$$\frac{\mathcal{R}(\Phi(W_1) - \mathcal{R}(\Phi(W_0))}{\mathcal{R}(\Phi(W_0))} \sim \frac{\lambda \|\nabla(\mathcal{R} \circ \Phi)(W_0)\|^2}{\mathcal{R}(\Phi(W_0))} \,. \tag{5.51}$$

Note that we assumed here that W_0 is not a local minimum. For instance, in the case of the MSE-loss $\mathcal{R}(\Phi) = \frac{1}{2}\|\Phi - \Phi^*\|^2$, this leads to the informal criterium:

$$\kappa_\Phi(W_0) = \|\Phi(W_0) - \Phi^*\| \frac{\|\mathcal{H}\Phi(W_0)\|}{\|\partial_W \Phi(W_0)\|^2} \ll 1 \,. \tag{5.52}$$

Note that if $\Phi(W_0) = 0$, it is possible to set a model in its lazy regime simply via a rescaling, as shown by the following quantity which can be arbitrary small:

$$\kappa_{\alpha\Phi}(W_0) = \frac{1}{\alpha}\|\Phi^*\| \frac{\|\mathcal{H}\Phi(W_0)\|}{\|\partial_W \Phi(W_0)\|^2} \,. \tag{5.53}$$

In the case of a 1-hidden layer neural network, this can be explicitely measured:

Proposition 5.10. *Assume that* $\forall f \in \mathcal{F}, \|f\|_\infty \leq \|f\|$, $W = (w_1, ..., w_n)$, *and:*

$$\Phi(W; x) = \alpha(n) \sum_{i=1}^{n} \varphi_i(x; w_i) \tag{5.54}$$

Assume that each φ_i *is i.i.d. with finite variance and 0-mean, s.t.* $|\frac{\partial^2}{\partial^2 w}\varphi_i(x; w))| \leq L$. *Assume that* $0 \neq |\frac{\partial}{\partial w}\varphi_i(x; w_0))|$ *on some neighborhood of some* w_0. *Then there exists* c *s.t.:*

$$\mathbb{E}[\kappa_\Phi(W) 1_{\|\nabla\Phi(W)\|>0}] \leq \frac{c}{\alpha(n)n} \tag{5.55}$$

Proof.

$$\mathbb{E}[\|\Phi\|^2] \leq \alpha^2(n) \sum_{i=1}^{n} \mathbb{E}[\|\varphi_i\|^2] = n\alpha^2(n)\mathbb{E}[\|\varphi_0\|^2] \tag{5.56}$$

Then, under the spectral norm(which as seen in Sec. 2 is tighter than others), we get:

$$\|\mathcal{H}\Phi(W)\mathbf{u}\|^2 = \mathbf{u}^T \mathcal{H}\Phi(W)^T \mathcal{H}\Phi(W)\mathbf{u} = \alpha^2(n) \sum_i u_i^2 \|\frac{\partial^2}{\partial_w^2}\varphi(.; w_i)\|^2 \leq \alpha^2(n) L^2 \|\mathbf{u}\|^2 \tag{5.57}$$

furthermore, as φ_i is at least piece-wise continuous along w, we get $\|\frac{\partial}{\partial w}\varphi(., w)\| \sup_x |\frac{\partial}{\partial w}\varphi(x, w)| \geq c$ for $w \in \Omega$ with $\mathbb{P}(\Omega) > 0$ (we write $W \in \Omega^n$ the event $\forall i, W_i \in \Omega$). Consequently, we obtain: $\|\nabla\Phi(W_0)\|^2 \geq n\alpha(n)^2 c$ Combining this and with the triangular inequality and Cauchy-Schwartz, we get:

$$\mathbb{E}[\kappa_\Phi(W) 1_{\|\nabla\Phi(W)\|>0}] \leq \frac{n\alpha^2(n) L \mathbb{E}[\|\varphi_0\|^2]}{\alpha^2(n) c^2 n^2} + \frac{\|\Phi^*\| \alpha(n) L}{\alpha^2(n) c^2 n} \tag{5.58}$$

and this leads us to the result. □

The main focus of this proposition corresponds to a neuron $\varphi(x, w) = w^1 \max(0, w^{2T}x + w^3)$, which is 2-homogeneous in w and satisfies the condition above. Then, rescaling the variance of the initialization W_0 by λ leads to a rescaling factor of λ^2. In the NTK regime, we actually had $\alpha(n) = \frac{1}{\sqrt{n}}$, thus we were indeed in a lazy regime. We shed more light on the lazy regime by considering the rescaled loss:

$$\mathcal{R}^\alpha(\Phi) = \frac{1}{\alpha^2}\mathcal{R}(\alpha\Phi)\,, \tag{5.59}$$

where the rescaling factor α is a normalization factor that allows to set a given model in its lazy regime, asymptotically as $\alpha \to \infty$. We also consider the rescaled dynamic (λ is constant!):

$$\frac{d}{dt}W^\alpha(t) = -\lambda \nabla_W (\mathcal{R}^\alpha \circ \Phi)(W^\alpha(t))\,, \tag{5.60}$$

which also writes, for the rescaled function: $\Phi^\alpha(t) = \Phi(W^\alpha(t))$:

$$\begin{aligned}\frac{d}{dt}\Phi^\alpha(t) &= -\frac{\lambda}{\alpha}\partial_W\Phi(W^\alpha(t))\partial_W\Phi(W^\alpha(t))^T\nabla\mathcal{R}(\alpha\Phi^\alpha)\\ &= -\frac{\lambda}{\alpha}K_{W^\alpha(t)}\nabla\mathcal{R}(\alpha\Phi^\alpha)\,,\end{aligned}$$

as well as the rescaled linearized dynamic:

$$\frac{d}{dt}\bar{W}^\alpha(t) = -\lambda\nabla_W(\mathcal{R}^\alpha\circ\bar{\Phi})(\bar{W}^\alpha(t))) \tag{5.61}$$

Here, we will only need that Φ is a.s. differentiable with differentiable locally-lipschitz, $\mathcal{R}$ is also differentiable and L-smooth.

Proposition 5.11. *Assume that $\Phi(W(0)) = 0$, then for any fixed time horizon $T > 0$, $\sup_{t\in[0,T]}\|W^\alpha(t) - W(0)\| = \mathcal{O}(\frac{1}{\alpha})$, $\sup_{t\in[0,T]}\|W^\alpha - \bar{W}^\alpha(t)\| = \mathcal{O}(\frac{1}{\alpha^2})$ and $\sup_{t\in[0,T]}\|\Phi(W^\alpha(t)) - \bar{\Phi}(\bar{W}^\alpha(t))\| = \mathcal{O}(\frac{1}{\alpha^2})$.*

Proof. The first part of the equation is obtained by using Prop. 5.4 and noting that $\mathcal{R}^\alpha(\Phi(W_0)) = \frac{1}{\alpha^2}\mathcal{R}(0)$. Next, let $\Delta(t) = \|\Phi(W^\alpha(t)) - \bar{\Phi}(\bar{W}^\alpha(t))\|$, then:

$$\begin{aligned}\frac{d}{dt}\Delta(t) &\leq \frac{\lambda}{\alpha}\big\|K_{W^\alpha(t)}\mathcal{R}'(\alpha\Phi^\alpha) - K_{W(0)}\mathcal{R}'(\alpha\bar{\Phi}^\alpha)\big\|\\ &\leq \frac{\lambda}{\alpha}\big\|K_{W^\alpha(t)}\nabla\mathcal{R}(\alpha\Phi^\alpha) - K_{W(0)}\mathcal{R}'(\alpha\Phi^\alpha)\big\|\\ &+ \frac{\lambda}{\alpha}\big\|K_{W(0)}\nabla\mathcal{R}(\alpha\Phi^\alpha) - K_{W(0)}\nabla\mathcal{R}(\alpha\bar{\Phi}^\alpha)\big\|\end{aligned}$$

Then, we note that $\nabla\mathcal{R}(\alpha\Phi^\alpha) \leq L\|\Phi(W^\alpha(t)) - \Phi(W(0))\| + |\nabla\mathcal{R}(0)| \leq \frac{c}{\alpha} + c'$ because Φ is Lipschitz on a compact surrounding the trajectory of W^α. Similarly, $W \to K_W$ is Lipschitz on this set for the same reason. Then, we conclude:

$$\frac{d}{dt}\Delta(t) \leq \frac{c'}{\alpha^2} + \Delta(t) \tag{5.62}$$

As $\Delta(0) = 0$, we conclude that $\Delta(t) = \mathcal{O}(\frac{1}{\alpha}^2)$ thanks to Lemma 5.1. Finally, bounding $\frac{d}{dt}\|W^\alpha - \bar{W}^\alpha(t)\|$, where the analysis before have shown all the terms in the norm are bounded, thus integrating leads to the conclusion. □

Chapter 6

Generalization properties of (deep) Neural Networks

In this chapter, we discuss the generalization properties of neural networks via complexity bounds. We follow mainly the works of Bartlett [5, 6], and we simply simplify the exposition of several proofs, in the case of ReLU Neural Networks without bias. We will be mainly interested in two complexity measures: the VC dimension, which is function dependent, and the Rademacher complexity, which is data dependent.

6.1 Statistical learning reminders

6.1.1 Bias-variance decomposition

We now discuss the generalization property of Neural Networks, and in particular how to relate the empirical error to the estimated error. We introduce the expected risk:

$$\mathcal{R}(\Phi) = \mathbb{E}_{X,Y}\left[\ell(\Phi X, Y)\right], \tag{6.1}$$

as well as its empirical risk, for iid samples (X_i, Y_i),

$$\mathcal{R}_n(\Phi) = \frac{1}{n}\sum_{i\leq n} \ell(\Phi X_i, Y_i). \tag{6.2}$$

Clearly, the empirical risk is an unbiased estimator of the expected risk. We are always interested in finding models Φ which minimize the expected risk, and typically we estimate a model via the empirical risk. We consider $\tilde{\Phi} \in \mathcal{F}$ and we assume that $\mathcal{R}_n(\Phi_n) = \inf_{\hat{\Phi}} \mathcal{R}_n(\hat{\Phi})$ and $\mathcal{R}(\Phi^*) = \inf_{\Phi} \mathcal{R}(\Phi)$, we then bound, similarly to [8], the *expected risk of our model* from the *minimal expected risk*, which measures the generalization properties of $\tilde{\Phi}$:

$$\mathcal{R}(\tilde{\Phi}) - \mathcal{R}(\Phi^*) = \underbrace{\mathcal{R}(\tilde{\Phi}) - \mathcal{R}(\Phi_n)}_{\text{Optimization error}} + \underbrace{\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*)}_{\geq 0,\text{Estimation error}}. \tag{6.3}$$

The *estimation error* measures the error due to minimizing the empirical risk rather than the expected risk. We assume that the optimization error is small enough to be neglectible. Then, we upper bound the estimation error:

$$\begin{aligned}\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*) &= \mathcal{R}(\Phi_n) - \mathcal{R}_n(\Phi_n) + \mathcal{R}_n(\Phi_n) - \mathcal{R}_n(\Phi^*) + \mathcal{R}_n(\Phi^*) - \mathcal{R}(\Phi^*) \\ &\leq 2 \sup_{\Phi} |\mathcal{R}(\Phi) - \mathcal{R}_n(\Phi)| \,,\end{aligned}$$

because $\mathcal{R}_n(\Phi_n) \leq \mathcal{R}_n(\Phi^*)$. Next, we get the generalization error via:

$$\mathbb{E}[\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*)] \leq 2\mathbb{E}[\sup_{\Phi} |\mathcal{R}(\Phi) - \mathcal{R}_n(\Phi)|] \,. \tag{6.4}$$

Now we can take the expectation on the Eq. (6.3). Our goal is to refine this generalization bound in the case of neural networks.

6.1.2 Estimation Error

We now discuss how to deal with the estimation term error. We first note that:

$$\mathbb{E}[\sup_{\Phi \in \mathcal{F}} |\mathcal{R}_n(\Phi) - \mathcal{R}(\Phi)|] = \mathbb{E}_{X_i}[\sup_{\Phi \in \mathcal{F}} \sum_n \frac{1}{n} | \sum_{i=1}^n \mathbb{E}_{\tilde{X}_i}[\ell(\Phi, X_i) - \ell(\Phi, \tilde{X}_i)]|] \,. \tag{6.5}$$

where $\{X_i, \tilde{X}_i\}_i$ are independent. Now, let $\varepsilon_1, ..., \varepsilon_n$ be some iid Rademacher variables, such that $\varepsilon_i(\ell(\Phi, X_i) - \ell(\Phi, \tilde{X}_i))$ has same law as $\ell(\Phi, X_i) - \ell(\Phi, \tilde{X}_i)$. We get:

$$\mathbb{E}[\sup_{\Phi \in \mathcal{F}} \sum_n \frac{1}{n} | \sum_{i=1}^n \mathbb{E}_X[\ell(\Phi, X_i) - \ell(\Phi, X)]|] \leq \frac{2}{n} \mathbb{E}[\sup_{\Phi \in \mathcal{F}} | \sum_{i=1}^n \varepsilon_i \ell(\Phi, X_i)|] \,. \tag{6.6}$$

Then fix $X_1^n = X_1, ..., X_n$, we get a standard upper bound, assuming that ℓ is L-Lipschitz and $\Phi \in \mathcal{F}$ iff $-\Phi \in \mathcal{F}$, we will show that it is possible to employ a simpler complexity measure that does not involve ℓ:

$$\begin{aligned}\mathbb{E}_{\varepsilon_i}[\sup_{\Phi \in \mathcal{F}} | \sum_{i=1}^n \varepsilon_i \ell(\Phi, X_i)|] &= \mathbb{E}[\sup_{\Phi \in \mathcal{F}} \sum_{i=1}^n \varepsilon_i \ell(\Phi, X_i)] \\ &= \frac{1}{2}(\mathbb{E}[\sup_{\Phi \in \mathcal{F}} \sum_{i=1}^n \varepsilon_i \ell(\Phi, X_i) | \varepsilon_n = 1] + \mathbb{E}[\sup_{\Phi \in \mathcal{F}} \sum_{i=1}^n \varepsilon_i \ell(\Phi, X_i) | \varepsilon_n = -1]) \\ &= \mathbb{E}[\sup_{\Phi, \tilde{\Phi} \in \mathcal{F}} \frac{\ell(\Phi, X_n) - \ell(\tilde{\Phi}, X_n)}{2} + \sum_{i=1}^{n-1} \varepsilon_i (\ell(\Phi, X_i) + \ell(\tilde{\Phi}, X_i))] \\ &\leq \mathbb{E}[\sup_{\Phi, \tilde{\Phi} \in \mathcal{F}} \frac{L|\Phi X_n - \tilde{\Phi} X_n|}{2} + \sum_{i=1}^{n-1} \varepsilon_i (\ell(\Phi, X_i) + \ell(\tilde{\Phi}, X_i))] \\ &= \mathbb{E}[\sup_{\Phi, \tilde{\Phi} \in \mathcal{F}} \frac{L\Phi X_n - L\tilde{\Phi} X_n}{2} + \sum_{i=1}^{n-1} \varepsilon_i (\ell(\Phi, X_i) + \ell(\tilde{\Phi}, X_i))] \\ &\leq ... \\ &\leq L\mathcal{R}ad(\mathcal{F}_{|X_1^n})\end{aligned}$$

where we have used the Rademacher complexity defined by:

Definition 6.1. *For a n-tuple of variables and a functional class $\mathcal{F}$, the Rademacher complexity of $X_1^n \triangleq (X_1, ..., X_n)$ is given by:*

$$\mathcal{R}ad(\mathcal{F}_{|X_1^n}) = \mathbb{E}_{\epsilon_i}[|\sup_{\Phi\in\mathcal{F}} \sum_{i=1}^{n} \varepsilon_i \Phi X_i|]\,. \tag{6.7}$$

The Rademacher complexity measures the richness of $\mathcal{F}$, in particular it is worth introducing:

Definition 6.2. *For a n-tuple of variables and a functional class $\mathcal{F}$, the Rademacher complexity of $\mathcal{F}$ is given by:*

$$\mathcal{R}ad_n(\mathcal{F}) = \mathbb{E}_{(X_i,\epsilon_i)}[\sup_{\Phi\in\mathcal{F}} |\sum_{i=1}^{n} \varepsilon_i \Phi X_i|]\,. \tag{6.8}$$

6.2 Measures of complexity

For the sake of simplicity, we will consider neural networks without biases, which simplifies the formalisms of the proofs, without losing much in generality. Consequently, our Neural Networks of depth J can be written: $\Phi x = W_J \rho W_{J-1} \rho ... \rho W_1 x$, where ρ is a ReLU and $W_1, ..., W_J$ some linear layers. We employ the same formalism as in Chapter 5.

6.2.1 Rademacher complexity

In this section, we remind some basic properties of the Rademacher complexity.

Definition 6.3 (SubGaussian processes)**.** *Fix $(T, \|.\|)$, then $(X_t)_t$ is a Subgaussian process, if for any $t \in T$, $\mathbb{E}X_t = 0$ and $X_{t_1} - X_{t_2}$ is $\|t_1 - t_2\|^2$-SubGaussian for any $t_1, t_2 \in T$.*

We remind that $\mathcal{N}(T, \epsilon, \|.\|)$ is the smallest number of ϵ-ball for the norm $\|.\|$ needed to cover T (also said *ε-net*). We have the following standard result:

Lemma 6.1 (Dudley's entropy)**.** *Let $(X_t)_{t\in T}$ be a sub-Gaussian process for $(T, \|.\|)$, assume that $D = \sup_{t_1,t_2\in T} \|t_1 - t_2\| < \infty$, then there exists a universal constant $C > 0$:*

$$\mathbb{E}[\sup_{t_1,t_2\in T} (X_{t_1} - X_{t_2})] \leq C \int_0^D \sqrt{\log \mathcal{N}(T, \epsilon, \|.\|)} d\epsilon\,. \tag{6.9}$$

Proof. See the proof in [38]. □

We further note that $X_f = \frac{1}{\sqrt{n}} \sum_{i\leq n} \varepsilon_i f(x_i)$ is a centered sub-Gaussian process for $\|f\| = \sup_{x\in\mathcal{X}} |f(x)|$, as by independence:

$$\mathbb{E}_{\varepsilon_i} e^{\lambda(X_f - X_g)} = \prod_i \mathbb{E}_{\varepsilon_i}[e^{\frac{\lambda\varepsilon_i}{\sqrt{n}}\left(f(x_i)-g(x_i)\right)}] \leq e^{\frac{\lambda^2}{2n}\sum_{i=1}^{n}|f(x_i)-g(x_i)|^2} \leq e^{\lambda^2 \frac{\|f-g\|^2}{2}}\,, \tag{6.10}$$

where we used $\cosh x \leq e^{\frac{x^2}{2}}$. This implies in particular that if $D = 2\sup \|X_i\|$, then by application of the Lemma 6.1 we have:

$$\mathbb{E}[\mathcal{R}ad_n(\mathcal{F}_{|X_1^n})] = \mathbb{E}[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i\leq n} \epsilon_i f(X_i)] \leq \frac{C}{\sqrt{n}} \int_0^D \sqrt{\log \mathcal{N}(T, \epsilon, \|.\|)} d\epsilon \,. \quad (6.11)$$

We have the following generalization bound, that we incorporate here for the sake of illustration:

Proposition 6.1. *With probability $1-\delta$, assuming that $\mathcal{F} \subset [0,1]^{\mathcal{X}}$, we can upper bound the generalization error via:*

$$\sup_{\Phi\in\mathcal{F}} |\mathcal{R}_n(\Phi) - \mathcal{R}(\Phi)| \leq 2L\mathcal{R}ad_n(\mathcal{F}) + L\sqrt{2\frac{\ln\frac{1}{\delta}}{n}} \quad (6.12)$$

Proof. Indeed, from Sec. 6.1.2, we know that

$$\sup_{\Phi\in\mathcal{F}} |\mathcal{R}_n(\Phi) - \mathcal{R}(\Phi)| \leq 2L\mathcal{R}ad(\mathcal{F}_{|X_1^n}) \quad (6.13)$$

and we further note that, if $\tilde{X}_1^n = (X_1, ..., \tilde{X}_i, ..., X_n)$, then because the functions are bounded by 1:

$$\mathcal{R}ad(\mathcal{F}_{|X_1^n}) - \mathcal{R}ad(\mathcal{F}_{|\tilde{X}_1^n}) \leq \frac{1}{n}\,, \quad (6.14)$$

Thus, we can apply McDiarmid concentration inequality [38] and we get:

$$\mathcal{R}ad(\mathcal{F}_{|X_1^n}) \leq \mathbb{E}_{X_i}\mathcal{R}ad(\mathcal{F}_{|X_1^n}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\,. \quad (6.15)$$

Combining those bounds allows to conclude. □

6.2.2 Vapnik-Chervonenkis (VC) dimension

The VC dimension is an explicit measure of the ability of a functional set to shatter n points:

Definition 6.4. *The VC dimension of a functional set $\mathcal{F} \subset \{-1,1\}^{\mathcal{X}}$ is the largest n s.t.*

$$\sup_{(a_1,...,a_n)} \#\{\big(f(a_1), ..., f(a_n)\big), f \in \mathcal{F}\} = 2^n\,. \quad (6.16)$$

We denote the VC dimension as dim VC($\mathcal{F}$).

We first remind a relation between the VC dimension and the Rademacher complexity, for $\{-1,1\}$ valued functions.

Proposition 6.2. *We have $\mathcal{R}ad_n(\mathcal{F}) = \sqrt{\frac{2dim\ VC(\mathcal{F})\log\frac{en}{dim\ VC(\mathcal{F})}}{n}}$.*

Proof. We apply successively the two lemma below, since the diameter is at most $\sqrt{n}$

$$\mathbb{E}_\varepsilon[\frac{1}{n}\sup_{\Phi\in\mathcal{F}}\sum_{i=1}^{n}\varepsilon_i\Phi(x_i)] \leq \mathbb{E}_\varepsilon[\frac{1}{n}\sup_{(a_i)=(\Phi(x_1),...,\Phi(x_n))\in\{-1,1\}^n}\sum_{i=1}^{n}\varepsilon_i a_i]$$
$$\leq \sqrt{\frac{2\text{dim VC}(\mathcal{F})\log\frac{en}{\text{dim VC}(\mathcal{F})}}{n}}$$

□

Lemma 6.2 (Sauer's Lemma)**.**

$$\sup_{(x_1,...,x_n)}\#\{\big(f(x_1),...,f(x_n)\big), f\in\mathcal{F}\} \leq (\frac{en}{dim\ VC(\mathcal{F})})^{dim\ VC(\mathcal{F})}. \tag{6.17}$$

Proof. See []. □

Lemma 6.3 (Massart Lemma)**.** *Let $\mathcal{A}$ be a finite subset and assume $r = \sup_{a\in\mathcal{A}}\|a\|$, then:*

$$\mathbb{E}_\varepsilon[\frac{1}{n}\sup_{a\in\mathcal{A}}\sum_{i=1}^{n}\epsilon_i a_i] \leq \frac{r\sqrt{2\log|\mathcal{A}|}}{n}. \tag{6.18}$$

Proof. See []. □

Remark 6.1. *It is also possible to obtain generalization bounds without using the Rademacher complexity yet the VC dimension, see [].*

Now, we are ready to address the problem of computing the VC dimension of a standard Multi-Layer Perceptron with ReLU non-linearity. The following lemma will be the core of our proof:

Lemma 6.4. *Let $P_1[X_1,...,X_n],...,P_m[X_1,...,X_n]$ multinomials of degree d with $n\leq m$. Then, for all $a\in\mathbb{R}^n$:*

$$\#\{\big(sign\ (P_1(a)),...,sign\ (P_n(a))\big), a\in\mathbb{R}^n\} \leq 2(\frac{2emd}{n})^n. \tag{6.19}$$

Proof. We admit this Lemma but a proof can be found in []. □

The main idea is to find a sign-invariant partition of the parameters of the neural network. This partition not being shatterable, its size will give a (pessimistic) upper bound on the VC-dimension of the feed-forward neural network.

Theorem 6.1 (Adapted from Bartlet, see [6].)**.** *Consider the feedforward neural networks with J (hidden) layers of width at most K, and ReLU non-linearity, then, writing $\mathcal{F} = \{x\to W_J\rho W_{J-1}\rho...\rho W_1 x\}$ their set:*

$$dim\ VC(\mathcal{F}) \leq \mathcal{O}(K^2J^2\log(KJ^2)) \tag{6.20}$$

Proof. Fix $(x_1, ..., x_m)$. We will consider Θ^j, the subset of parameters to parametrize $\Phi_j(., \theta), \theta \in \Theta^j$. Let's assume we have obtained a partition

$\Theta_1^j ... \Theta_{n_j}^j$ such that $\forall q, \forall k, \exists P_{q,k} : \forall \theta \in \Theta_i^j, P_{q,k}(\theta) = [\Phi_j(x_k, \theta)]_q$, and

$$\forall \theta \in \Theta_i^j, \big(\text{sign}([\Phi_j(x_1;\theta)]_1, ..., \text{sign}([\Phi_j(x_k;\theta)]_1),$$
$$..., \text{sign}([\Phi_j(x_1;\theta)]_K, ..., \text{sign}([\Phi_j(x_k;\theta)]_K)\big) = \text{constant}$$

Then, at rank $j+1 < J$, we consider Θ_i^j and we partition $\{(\theta, \eta) \subset \Theta^{j+1} : \theta \in \Theta_i^j\}$ in $\{\eta_p\}_p$ such that $\forall q, k, \exists P_{q,k} : P_{q,k}(\theta, \eta) = [\Phi_{j+1}(x_k, (\theta, \eta))]_q, \forall \eta \in \eta_j$. We note that for θ fixed, there is a bijection between $\{\eta_p\}$ and

$$\{(\text{sign}([W_j(\eta)\Phi_j(x_1;\theta)]_1, ..., \text{sign}([W_j(\eta)\Phi_j(x_k;\theta)]_1), ..., \text{sign}([W_j(\eta)\Phi_j(x_1;\theta)]_K, ..., \text{sign}([W_j(\eta)\Phi_j(x_k;\theta)]_K)\big)\} \subset \{-1, 1\}^{Km}$$

because for $q \leq K$ $[W_j(.)\Phi_j(x_k;.)]_q$ is a polynomial (of degree $j+1$, in $K^2(j+1)$ variables) and the ReLU ρ set to 0 negative values. It implies by the Lemma 6.4, that:

$$n_{j+1} \leq n_j 2(\frac{2eKm(j+1)}{K^2(j+1)})^{K^2(j+1)}. \tag{6.21}$$

Now, the only difference between the intermediary layers and the final layer is the output dimension (which is 1), thus the sign partition has cardinality n given, at most, by:

$$n \leq n_J 2(\frac{2em(J+1)}{K^2J+K})^{K^2J+K} \tag{6.22}$$

As $n_0 = 1$, this leads to:

$$n_J \leq 2^J(\frac{2em}{K})^{\frac{1}{2}K^2J(J+1)} \tag{6.23}$$

Thus with a logarithm in base 2,

$$\log n \leq (J+1) + (\frac{3}{2}K^2J + K + \frac{K^2J^2}{2})\log(\frac{2em}{K}) + (K^2J+K)\log(\frac{J+1}{K^2J+K}). \tag{6.24}$$

Now, we find an upper bound of the VC dimension. Finding the smallest n such that our model can not shatter the m points $(x_1, ..., x_m)$ leads to finding the smallest n s.t. $n = 2^m$ is smaller than the right term of Eq. (6.24) and the VC dimension can not be larger than $\log n$. We will use the following lemma:

Lemma 6.5. $x \leq a + b\log x \Rightarrow x \leq 2(a + b\log b)$.

By direct application of this lemma, m can not be larger than:

$$m \leq J + 1 + (\frac{3}{2}K^2J + K + \frac{1}{2}K^2J^2)\log(eKJ^2) + (K^2J+K)\log(\frac{J+1}{K^2J+K})$$
$$= \mathcal{O}(J + K^2J^2\log(KJ^2)).$$

□

Unfortunately, combining with 6.2, the quantity obtained in this bound remains still large if we apply it directly to state-of-the-art neural networks. In the next subsection we propose to refine this bound.

6.2.3 Spectral Norm-based bounds

We now propose a different proof directly inspired from [5]. We however simplify our setting because we restrict ourself to ℓ^2-norm bounds. For some $\mathcal{X}$, we introduce $\|f\|_{\mathcal{X}} = \sup_{x\in\mathcal{X}} \|f(x)\|$ and $\kappa = \text{diam}\,(\mathcal{X})$. We will now propose a bound on the sample complexity of neural networks: the idea will be to compute an ϵ-covering of the parametrization of our neural network, and to combine it with Eq. (6.11). Our bound will then highlight a quantity that depends on the spectral norm of the layers $W_1, ..., W_J$: we will restrict our analysis to models such that $\|W_j\| \leq \alpha_j$ for some predefined α_j.

Proposition 6.3. *Let $\epsilon > 0$ and consider $\mathcal{F}_J = \{\Phi_J = W_J\rho W...\rho W_1, \|W_j\| \leq \alpha_j, j \leq J\}$ then:*

$$\mathcal{N}(\mathcal{F}_J, \epsilon, \|.\|_{\mathcal{X}}) \leq \prod_{j=0}^{J-1} \sup_{\Phi_j\in\mathcal{F}_j} \mathcal{N}(\{W_{j+1}\rho\Phi_j, \|W_{j+1}\| \leq \alpha_{j+1}\}, \epsilon_j, \|.\|_{\mathcal{X}}), \quad (6.25)$$

where $\epsilon_j = \frac{2^{j-J}\epsilon}{\prod_{J\geq i>j+1}\alpha_i}$ for $0 \leq j \leq J-1$.

Proof. We prove this result by induction on the depth J. For $J = 1$, as $\Phi_0 = \mathbf{I}$, the result is true with $\epsilon_0 = \frac{1}{2}\epsilon$. Assuming the formula is true at rank J, we consider a covering of $\{W_{J+1}\Phi_J., \|W_{J+1}\| \leq \alpha_{J+1}\} = \Gamma_{\Phi_J X}$ (for $\epsilon_J = \frac{1}{2}\epsilon$) and of $\mathcal{F}_J$(for $\tilde{\epsilon}_{J+1} = \frac{\epsilon}{2\alpha_{J+1}}$), thus we can find $\tilde{W}_{J+1} \in \Gamma_{\Phi_J X}$ and $\tilde{\Phi}_J \in \mathcal{F}_J$, such that for all $i \leq n$:

$$\begin{aligned}\|\Phi_{J+1}X_i - \tilde{W}_{J+1}\rho\tilde{\Phi}_J X_i\| &\leq \|\tilde{W}_{J+1}\rho\Phi_J X_i - W_{J+1}\rho\Phi_J X_i\| + \|\tilde{W}_{J+1}\|\|\Phi_J X_i - \tilde{\Phi}_J X_i\| \\ &\leq \epsilon_J + \tilde{\epsilon}_J\alpha_{J+1} = \epsilon\end{aligned}$$

By induction, this is obtained for the right term for $\epsilon_j = (\frac{2^{j-J}}{\prod_{J\geq i>j+1}\alpha_i})(\frac{\epsilon}{2\alpha_{J+1}}) = \frac{2^{j-J-1}\epsilon}{\prod_{J+1\geq i>j+1}\alpha_i}, 0 \leq j \leq J-1$. We note that $\{\tilde{W}_{J+1}, \tilde{\Phi}_J\}$ provides a covering with desired cardinality of $\mathcal{F}_{J+1}$, leading to the result. □

Now, from this proposition, we see it is enough to bound:

$$\mathcal{N}(\{W., \|W\| \leq \alpha\}, \epsilon, \|.\|_{\mathcal{X}}). \quad (6.26)$$

This is done by the following proposition:

Proposition 6.4. *Assuming $\|X_i\| \leq C, \forall i$, we have for $\alpha > 0$:*

$$\mathcal{N}(\{W., \|W\| \leq \alpha\}, \epsilon, \|.\|_{\mathcal{X}}) \leq \frac{(3C\alpha)^{K^2}}{\epsilon^{K^2}}. \quad (6.27)$$

Proof. We note that:

$$\|WX_i\| \leq \|W\|C \tag{6.28}$$

Thus, we can consider a covering with radii $\frac{\epsilon}{C}$ of the ball $\mathcal{B}_\alpha$ for $\|.\|$, given the ambient dimension is K^2, this leads to the bound (see [38]):

$$\mathcal{N}(\{W., \|W\| \leq \alpha\}, \epsilon, \|.\|_{\mathcal{X}}) \leq \frac{(3C\alpha)^{K^2}}{\epsilon^{K^2}}. \tag{6.29}$$

□

Remark 6.2. *This bound is extremely naive, and [5] shows it is possible to obtain tighter inequalities via more carefully chosen ϵ-net, notably with a $\ell^1 - \ell^2$ type bound.*

Proposition 6.5. *Writing $\mathcal{S} = \prod_{1 \leq j \leq J} \alpha_j$ and again $\mathcal{F}_J = \{\Phi_J = W_J \rho W ... \rho W_1, \|W_j\| \leq \alpha_j, j \leq J\}$, we have:*

$$\mathcal{R}ad_n(\mathcal{F}_J) = \mathcal{O}(\frac{\mathcal{S}K\sqrt{J}(\sqrt{J}\log\mathcal{S} + 1)}{\sqrt{n}}) \tag{6.30}$$

Proof. Combining all those bounds, we get:

$$\begin{aligned}
\mathcal{R}ad_n(\mathcal{F}_J) &\leq \frac{C}{\sqrt{n}} \int_0^{\kappa\mathcal{S}} \sqrt{\log(\mathcal{N}(\mathcal{F}_J, \epsilon, \|.\|_{\mathcal{X}})} d\epsilon \\
&\leq \frac{C}{\sqrt{n}} \int_0^{\kappa\mathcal{S}} \sqrt{\sum_{j=0}^{J-1} \log(\mathcal{N}(\{W_{j+1}\Phi_j, \|W_{j+1}\| \leq \alpha_{j+1}\}, \epsilon_j, \|.\|_{\mathcal{X}})} d\epsilon \\
&\leq \frac{C}{\sqrt{n}} \int_0^{\kappa\mathcal{S}} \sqrt{\sum_{j=0}^{J-1} K^2 \log(3\kappa\alpha_1 ... \alpha_j \alpha_{j+1}/\epsilon_j)} d\epsilon \\
&\leq \frac{C}{\sqrt{n}} \int_0^{\kappa\mathcal{S}} \sqrt{\sum_{j=0}^{J-1} K^2 \log(\frac{3\kappa\mathcal{S}}{2^{j-J}\epsilon})} d\epsilon \\
&= \frac{CK}{\sqrt{n}} \int_0^{\kappa\mathcal{S}} \sqrt{J\log(3\kappa\mathcal{S}) - J(J-1)/2\log 2 - J\log(\epsilon)} d\epsilon \\
&= \mathcal{O}(\frac{\kappa\mathcal{S}K(\sqrt{J} + \sqrt{J}\log\mathcal{S} + J + \sqrt{J}\log(\kappa\mathcal{S}) + 1)}{\sqrt{n}})
\end{aligned}$$

□

Compared naively to the previous bound obtained in Prop. 6.4 and Prop. 6.1, we see that if $\mathcal{S} \ll 1$, then this bound is better. Yet it is unclear if this is systematically true.

Acknowledgements

I would like to thank much Mathieu Andreux, Lénaïc Chizat, Michael Eickenberg, Georgios Exarchakis, Louis Fournier, Louis Leconte, Thomas Pumir, Aladin Virmaux, Irène Waldspurger for helpful comments, and, for the most dedicated (e.g., Mathieu, Michael), proofreading some parts of this manuscript.

Bibliography

[1] S. Aguilar. A survey on representation theory. 2015.

[2] Sheldon Axler, Paul Bourdon, and Ramey Wade. *Harmonic function theory*, volume 137. Springer Science & Business Media, 2013.

[3] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[4] Christoph Bandt. Metric invariance of haar measure. *Proceedings of the American Mathematical Society*, pages 65–69, 1983.

[5] Peter Bartlett, Dylan J Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1706.08498*, 2017.

[6] Peter L Bartlett, Vitaly Maiorov, and Ron Meir. Almost linear vc-dimension bounds for piecewise polynomial networks. *Neural computation*, 10(8):2159–2173, 1998.

[7] Mikhail Belkin. Problems of learning on manifolds. 2004.

[8] Léon Bottou and Olivier Bousquet. 13 the tradeoffs of large-scale learning. *Optimization for machine learning*, page 351, 2011.

[9] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[10] Jean Bourgain and Joram Lindenstrauss. Projection bodies. In *Geometric Aspects of Functional Analysis*, pages 250–270. Springer, 1988.

[11] B Brainerd and RE Edwards. Linear operators which commute with translations. part i: representation theorems. *Journal of the Australian Mathematical Society*, 6(3):289–327, 1966.

[12] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[13] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

[14] Ronald R Coifman and Mauro Maggioni. Diffusion wavelets. *Applied and Computational Harmonic Analysis*, 21(1):53–94, 2006.

[15] Jean Gallier. Notes on spherical harmonics and linear representations of lie groups. *preprint*, 2009.

[16] Fernando Gama, Joan Bruna, and Alejandro Ribeiro. Stability of graph scattering transforms. *arXiv preprint arXiv:1906.04784*, 2019.

[17] Feng Gao, Guy Wolf, and Matthew Hirn. Geometric scattering for graph data analysis. In *International Conference on Machine Learning*, pages 2122–2131. PMLR, 2019.

[18] Jonathan Gleason. Existence and uniqueness of haar measure. *University of Chicago*, 30, 2010.

[19] David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.

[20] Lars Hörmander. Estimates for translation invariant operators in l p spaces. *Acta Mathematica*, 104(1):93–140, 1960.

[21] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

[22] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[23] Yann LeCunn, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[24] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[25] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

[26] Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

[27] Stéphane Mallat. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150203, 2016.

[28] Keihachiro Moriyasu. *An elementary primer for gauge theory*. World Scientific, 1983.

[29] Edouard Oyallon, Eugene Belilovsky, Sergey Zagoruyko, and Michal Valko. Compressing the input for cnns with the first-order scattering transform. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 301–316, 2018.

[30] Richard S Palais. Natural operations on differential forms. *Transactions of the American Mathematical Society*, 92(1):125–141, 1959.

[31] Frédéric Paulin. Compléments de théorie spectrale et d'analyse harmonique.

[32] Laure Saint-Raymond. Analyse fonctionnelle.

[33] Laurent Sifre and Stéphane Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1233–1240, 2013.

[34] Amit Singer. From graph to manifold laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.

[35] Mitsuo Sugiura. *Unitary representations and harmonic analysis: an introduction*. Elsevier, 1990.

[36] Terence Tao. Lecture notes 10 for 247b.

[37] Michael Taylor. *Partial differential equations II: Qualitative studies of linear equations*, volume 116. Springer Science & Business Media, 2013.

[38] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[39] Irene Waldspurger. Exponential decay of scattering coefficients. In *2017 international conference on sampling theory and applications (SampTA)*, pages 143–146. IEEE, 2017.