The Scattering Transform

February 27, 2020

Abstract

In this Chapter we describe Scattering Representations, a signal representation built using wavelet multiscale decompositions with a deep convolutional architecture. Its construction highlights the fundamental role of geometric stability in deep learning representations, and provides a mathematical basis to study CNNs. We describe its main mathematical properties, its applications to computer vision, speech recognition and physical sciences, as well as its extensions to Lie Groups and non-Euclidean domains. Finally, we discuss recent applications to modeling high-dimensional probability densities.

Contents

1 Introduction

Understanding the success of deep learning in challenging data domains such as computer vision, speech recognition or natural language processing remains a major unanswered question, that requires a tight integration of different theoretical aspects of the learning algorithm: approximation, estimation and optimization. Amongst the many pieces responsible for such success, an important element comes from the extra structure built into the neural architecture as a result of the input signal structure. Images, sounds and text are signals defined over low-dimensional domains, such as grids or their continuous Euclidean counterparts. In these domains one can articulate specific priors of data distributions and tasks, which are leveraged in neural networks through convolutional layers.

This requires developing a signal processing theory of deep learning. In order to gain a mathematical understanding of the interplay between geometric properties of the input domain and convolutional architectures, in this chapter we set aside the optimization and data-adaptivity pieces of the puzzle, and take an axiomatic approach to build high-dimensional signal representations with prescribed properties that make them amenable to complex recognition and classification tasks.

The first step is to develop the notion of geometric stability (Section 2). In essence, a signal representation defined on a metric domain is geometrically stable if small perturbations in the metric structure result in small changes in the output features. In Euclidean domains, geometric stability can be expressed in terms of diffeomorphisms, which model many naturally occurring transformations

in computer vision and speech recognition, such as changes in viewpoint, local translations, or pitch transpositions.

Stability to the action of diffeomorphisms is achieved by separating scales, leading to multiscale signal decompositions. Section 3 describes Scattering Representations on the Euclidean Translation Group. First introduced in [Mal12], they combine wavelet multiscale decompositions with point-wise modulus activation functions. We describe its main mathematical properties and applications to computer vision. Scattering transforms are natural generalisations of multiscale representations of stochastic processes, in which classical high-order polynomial moments are replaced by stable non-linear transforms. Section 4 reviews Stochastic Scattering representations and their main applications to multifractal analysis.

Euclidean Scattering representations serve as a mathematical basis to study CNNs on image and audio domains. In many areas of physical and social sciences, however, data is rarely defined over regular Euclidean domains. As it turns out, one can extend the formalism of geometric stability and wavelet scattering representations on two important directions: first, to more general Lie Groups of transformations (Section 5), and then to graphs and manifolds (Section 5.3).

We conclude this chapter by focusing on two important applications of scattering representations. Thanks to their ability to capture key geometrical properties of high-dimensional signals with stability guarantees, they may be used in unsupervised learning to perform high-dimensional density estimation and implicit modeling, as described in Section 6.

2 Geometric Stability

This Section describes the notion of geometric stability in signal representations. We begin with the Euclidean setting (subsection 2.1), where this stability is expressed in terms of diffeomorphisms of the signal domain. We then discuss how to extend this notion to general metric domains in subsection 2.3, and then highlight the limitations of several standard high-dimensional signal representations in regards to geometric stability (subsection 2.4).

2.1 Euclidean Geometric Stability

Consider a compact d-dimensional Euclidean domain $\Omega = [0,1]^d \subset \mathbb{R}^d$ on which square-integrable functions $\mathbf{x} \in L^2(\Omega)$ are defined (for example, in image analysis applications, images can be thought of as functions on the unit square $\Omega = [0,1]^2$). We consider a generic supervised learning setting, in which an unknown function $f : L^2(\Omega) \to \mathcal{Y}$ is observed on a training set $\{\mathbf{x}_i \in L^2(\Omega), f_i = f(\mathbf{x}_i)\}_{i \in \mathcal{I}}$. In the vast majority of computer vision and speech analysis tasks, the unknown function f satisfies crucial regularity properties expressed in terms of the signal domain Ω.

Global Translation Invariance: Let $\mathcal{T}_v \mathbf{x}(u) = \mathbf{x}(u - v), \quad u, v \in \Omega$, be a *translation operator*[1] acting on functions $\mathbf{x} \in L^2(\Omega)$. Our first assumption is that the function f is either *invariant*, ie $f(\mathcal{T}_v \mathbf{x}) = f(\mathbf{x})$ for any $\mathbf{x} \in L^2(\Omega)$ and $v \in \Omega$, or *equivariant*, ie $f(\mathcal{T}_v \mathbf{x}) = \mathcal{T}_v f(\mathbf{x})$, with respect to translations, depending on the task. Translation invariance is typical in object classification tasks, whereas equivariance arises when the output of the model is a space in which translations can act upon (for example, in problems of object localization, semantic segmentation, or motion estimation).

The notion of global invariance/equivariance can be easily extended to other transformation groups beyond translations. Section 5 discusses one such extension, to the group of rigid motions generated by translations and rotations in Ω.

However, global invariance is not a strong prior in the face of high-dimensional estimation. Ineed, global transformation groups are typically low-dimensional; in particular, in signal processing, they often correspond to subgroups of the affine group $\mathrm{Aff}(\Omega)$, with dimension $O(d^2)$. A much stronger prior may be defined by specifying how the function f behaves under geometric perturbations of the domain which are 'nearby' these global transformation groups.

[1] Assuming periodic boundary conditions to ensure that the operation is well-defined over $L^2(\Omega)$.

Local deformations and scale separation: In particular, given a smooth vector field $\tau : \Omega \to \Omega$, a deformation by τ acts on $L^2(\Omega)$ as $\mathbf{x}_\tau(u) := \mathbf{x}(u - \tau(u))$. Deformations can model local translations, changes in point of view, rotations and frequency transpositions [BM13], and have been extensively used as models of image variability in computer vision [JZL96, FGMR10, GDDM14]. Most tasks studied in computer vision are not only translation invariant/equivariant, but also stable with respect to local deformations [Mal16, BM13]. In tasks that are translation invariant, this prior may be expressed informally as

$$|f(\mathbf{x}_\tau) - f(\mathbf{x})| \approx \|\tau\|, \tag{1}$$

for all $\mathbf{x}, \tau$. Here, $\|\tau\|$ measures the distance of the associated diffeomorphism $\varphi(u) := u - \tau(u)$ to the translation group; we will see in next section how to specify this metric in the space of diffeomorphisms. In other words, the target to be predicted does not change much if the input image is slightly deformed. In tasks that are translation equivariant, we have $|f(\mathbf{x}_\tau) - f_\tau(\mathbf{x})| \approx \|\tau\|$ instead. The deformation stability property is much stronger than the global invariance one, since the space of local deformations has high dimensionality, as opposed to the group of global invariants.

As we will see later, a key consequence of (1) is that long-range dependencies may be broken into multi-scale local interaction terms, leading to hierarchical models in which spatial resolution is progressively reduced. To illustrate this principle, denote by

$$q(z_1, z_2; v) = \text{Prob}(\mathbf{x}(u) = z_1 \text{ and } \mathbf{x}(u + v) = z_2) \tag{2}$$

the joint distribution of two image pixels at an offset v from each other, where we have assumed a stationary statistical model for natural images (hence q does not depend upon the location u). In presence of long-range dependencies, this joint distribution will not be separable for any v. However, the deformation stability prior states that $q(z_1, z_2; v) \approx q(z_1, z_2; v(1 + \epsilon))$ for small ϵ. In other words, whereas long-range dependencies indeed exist in natural images and are critical to object recognition, they can be captured and down-sampled at different scales. This principle of stability to local deformations has been exploited in the computer vision community in models other than CNNs, for instance, deformable parts models [FGMR10], as we will review next. In practice, the Euclidean domain Ω is discretized using a regular grid with n points; the translation and deformation operators are still well-defined so the above properties hold in the discrete setting.

2.2 Representations with Euclidean Geometric Stability

Motivated by the previous geometric stability prior, we are interested in building signal representations that are compatible with such a prior. Specifically, suppose our estimation for f, the target function, takes the form

$$\hat{f}(\mathbf{x}) := \langle \Phi(\mathbf{x}), \theta \rangle \ , \tag{3}$$

where $\Phi : L^2(\Omega) \to \mathbb{R}^K$ corresponds to the signal representation and $\theta \in \mathbb{R}^K$ the classification or regression coefficients, respectively. In a CNN, one would associate Φ with the operator that maps the input to the last hidden layer, and θ with the very last output layer of the network.

The linear relationship between $\Phi(\mathbf{x})$ and $\hat{f}(\mathbf{x})$ above implies that geometric stability in the representation is sufficient to guarantee a predictor which is also geometrically stable. Indeed, if we assume that

$$\forall \ \mathbf{x}, \tau \ , \ \|\Phi(\mathbf{x}) - \Phi(\mathbf{x}_\tau)\| \lesssim \|\mathbf{x}\| \|\tau\| \ , \tag{4}$$

then by Cauchy-Schwartz, it follows that

$$|\hat{f}(\mathbf{x}) - \hat{f}(\mathbf{x}_\tau)| \leq \|\theta\| \|\Phi(\mathbf{x}) - \Phi(\mathbf{x}_\tau)\| \lesssim \|\theta\| \|\mathbf{x}\| \|\tau\| \ .$$

This motivates the study of signal representations where one can certify (4), while ensuring that Φ captures enough information so that $\|\Phi(\mathbf{x}) - \Phi(\mathbf{x}')\|$ is large whenever $|f(\mathbf{x}) - f(\mathbf{x}')|$ is large. In this setting, a notorious challenge to achieving (4) while keeping enough discriminative power in $\Phi(\mathbf{x})$ is to transform the high-frequency content of $\mathbf{x}$ in such a way that it becomes stable.

In recognition tasks, one may not only want to consider geometric stability, but also stability with respect to the Euclidean metric in $L^2(\Omega)$:

$$\forall\, \mathbf{x}, \mathbf{x}' \in L^2(\Omega)\ ,\ \|\Phi(\mathbf{x}) - \Phi(\mathbf{x}')\| \lesssim \|\mathbf{x} - \mathbf{x}'\|\ . \tag{5}$$

This stability property ensures that additive noise in the input will not drastically change the feature representation.

The stability desiderata (4) and (5) may also be interpreted in terms of robustness to adversarial examples [SZS+13]. Indeed, the general setup of adversarial examples consists in producing small perturbations $\mathbf{x}'$ of a given input $\mathbf{x}$ (measured by appropriate norms) such that $|\langle \Phi(\mathbf{x}) - \Phi(\mathbf{x}'), \theta\rangle|$ is large. Stable representations certify that those adversarial examples cannot be obtained with small additive or geometric perturbations.

2.3 Non-Euclidean Geometric Stability

Whereas Euclidean domains may be used to model many signals of interest, such as images, videos or speech, a wide range of high-dimensional data across physical and social sciences is naturally defined on more general geometries. For example, signals measured on social networks have rich geometrical structure, encoding locality and multiscale properties, yet they on a non-Euclidean geometry. An important question is thus how to extend the notion of geometrical stability to more general domains.

Deformations provide the natural framework to describe geometric stability in Euclidean domains, but their generalization to non-Euclidean, non-smooth domains is not straightforward. Let $\mathbf{x} \in L^2(\mathcal{X})$ be a signal defined on a domain $\mathcal{X}$. If $\mathcal{X}$ is embedded into a low-dimension Euclidean space $\Omega \subset \mathbb{R}^d$, such as a 2-surface within a three-dimensional space, then one can still define meaningful deformations on $\mathcal{X}$ via *extrinsic* deformations of Ω. Indeed, if $\tau : \mathbb{R}^d \to \mathbb{R}^d$ is a smooth field and $\varphi(v) = v - \tau(v)$ the corresponding diffeomorphism (assuming $\|\tau\| < 1/2$), then we can define $\mathbf{x}_\tau \in L^2(\mathcal{X}_\tau)$ as

$$\mathbf{x}_\tau(u) := \mathbf{x}(\varphi^{-1}(u))\ , u \in \mathcal{X}\ .$$

Such deformation models have been studied in [KBPZ17] with applications in surface representation, in which the notion of geometric stability relies on its ambient Euclidean structure.

In more general applications, however, we may be interested in intrinsic notions of geometric stability, that do not necessarily rely on a pre-existent low-dimensional embedding of the domain. The change of variables $\varphi(u) = u - \tau(u)$ defining the deformation can be seen as a perturbation of the Euclidean metric in $L^2(\mathbb{R}^d)$. Indeed,

$$\langle \mathbf{x}_\tau, \mathbf{y}_\tau\rangle_{L^2(\mathbb{R}^d,\mu)} = \int_{\mathbb{R}^d} \mathbf{x}_\tau(u)\mathbf{y}_\tau(u)d\mu(u) = \int_{\mathbb{R}^d} \mathbf{x}(u)\mathbf{y}(u)|I - \nabla\tau(u)|d\mu(u) = \langle \mathbf{x}, \mathbf{y}\rangle_{L^2(\mathbb{R}^d,\tilde{\mu})}\ ,$$

with $d\tilde{\mu}(u) = |I - \nabla\tau(u)|d\mu(u)$, and $|I - \nabla\tau(u)| \approx 1$ if $\|\nabla\tau\|$ is small, where I is the identity. Therefore, a possible way to extend the notion of deformation stability to general domains $L^2(\mathcal{X})$ is to think of $\mathcal{X}$ as a metric space and reason in terms of stability of $\Phi : L^2(\mathcal{X}) \to \mathbb{R}^K$ to *metric changes* in $\mathcal{X}$. This requires a representation that can be defined on generic metric spaces, as well as a criteria to compare how close two metric spaces are. We will describe a general approach for discrete metric spaces based on diffusion operators in Section 5.3.

2.4 Examples

2.4.1 Kernel Methods

Kernel methods refer to a general theory in the machine learning framework, whose main purpose consists in embedding data in a high dimensional space, in order to express complex relationships in terms of linear scalar products.

For a generic input space $\mathcal{Z}$ (which can be thought of as $\mathcal{Z} = L^2(\mathcal{X})$ corresponding to the previous discussion), a *feature map* $\Phi : \mathcal{Z} \longrightarrow \mathcal{H}$ maps data into a Hilbert space $\mathcal{H}$ with the reproducing property: for each $f \in \mathcal{H}$ and $\mathbf{x} \in \mathcal{Z}$, $f(\mathbf{x}) = \langle f, \Phi(\mathbf{x}) \rangle$. Linear classification methods access the transformed data $\Phi(\mathbf{x})$ only through scalar products of the form [STC04]

$$\langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle \ .$$

Rather than building the mapping explicitly, the popular "Kernel Trick" exploits Mercer's theorem. It states that a continuous, symmetric and positive definite kernel $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ defines an integral operator of $L^2(\mathcal{Z})$, which diagonalizes in an orthonormal basis [MNY06] $\{\phi_n\}_n$ of $L^2(\mathcal{Z})$, with non-negative eigenvalues. As a result, $K(\mathbf{x}, \mathbf{x}')$ admits a representation

$$K(\mathbf{x}, \mathbf{x}') = \sum_{n \geq 1} \lambda_n \phi_n(\mathbf{x}) \phi_n(\mathbf{x}') \ ,$$

which yields

$$K(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle \ ,$$

with $\Phi(\mathbf{x}) = (\lambda_n^{1/2} \phi_n(\mathbf{x}))_n$. In Kernel methods it is thus sufficient to construct positive definite kernels K on $\mathcal{Z}^2$ in order to extend linear classification tools to more complex relationships.

Despite their success and effectiveness in a number of machine learning tasks, the high dimensional embeddings induced by kernel methods do not automatically enjoy the stability properties to additive noise or deformations. The kernel needs to be chosen accordingly. *Convolutional Kernels Networks* [MKHS14, BM17] have been developed to capture the geometric stability properties and offer competitive empirical performance to modern deep architectures. These kernels contrast with another recent family of *Neural Tangent Kernels* [JGH18], which linearize a generic deep architecture around its parameter initialization, and which do not offer the same amount of geometric stability [BM19].

2.4.2 Power Spectra, Autocorrelation and Registration Invariants

Translation invariant representations can be obtained from registration, auto-correlation or Fourier modulus operators. However, the resulting representations are not Lipschitz continuous to deformations.

A representation $\Phi(\mathbf{x})$ is translation invariant if it maps global translations $\mathbf{x}_c(u) = \mathbf{x}(u - c)$ by $c \in \mathbb{R}^d$ of any function $x \in \mathbf{L}^2(\mathbb{R}^d)$ to the same image:

$$\forall \, \mathbf{x} \in \mathbf{L}^2(\mathbb{R}^d) \, , \, \forall \, c \in \mathbb{R}^d \ , \ \Phi(\mathbf{x}_c) = \Phi(\mathbf{x}) \ . \tag{6}$$

The Fourier transform modulus is an example of a translation invariant representation. Let $\hat{\mathbf{x}}(\omega)$ be the Fourier transform of $\mathbf{x}(u) \in \mathbf{L}^2(\mathbb{R}^d)$. Since $\widehat{\mathbf{x}_c}(\omega) = e^{-ic.\omega} \, \hat{\mathbf{x}}(\omega)$, it follows that $|\widehat{\mathbf{x}_c}| = |\hat{\mathbf{x}}|$ does not depend upon c.

A Fourier modulus is translation invariant and stable to additive noise, but unstable to small deformations at high frequencies [Mal12], as illustrated with the following dilation example. Let $\tau(u) = su$ denote a linear displacement field where $|s|$ is small, and let $\mathbf{x}(u) = e^{i\xi u} \theta(u)$ be a modulated version of a lowpass window $\theta(u)$. Then the dilation $\mathbf{x}_\tau(u) = L[\tau]\mathbf{x}(u) = \mathbf{x}((1+s)u)$ moves the central frequency of $\hat{\mathbf{x}}$ from ξ to $(1+s)\xi$. If $\sigma_\theta^2 = \int |\omega|^2 |\hat{\theta}(\omega)|^2 d\omega$ measures the frequency spread of θ, then

$$\sigma_x^2 = \int |\omega - \xi|^2 |\hat{x}(\omega)|^2 d\omega = \sigma_\theta^2 \ ,$$

and

$$\begin{aligned} \sigma_{x_\tau}^2 &= (1+s)^{-d} \int (\omega - (1+s)\xi)^2 |\hat{x}((1+s)^{-1}\omega)|^2 d\omega \\ &= \int |(1+s)(\omega - \xi)|^2 |\hat{x}(\omega)|^2 d\omega = (1+s)^2 \sigma_x^2 \ . \end{aligned}$$

It follows that if the distance between the central frequencies of $\mathbf{x}$ and $\mathbf{x}_\tau$, $s\xi$, is large compared to their frequency spreads, $(2+s)\sigma_\theta$, then the frequency supports of $\mathbf{x}$ and $\mathbf{x}_\tau$ are nearly disjoint and hence

$$\| |\hat{\mathbf{x}}_\tau| - |\hat{\mathbf{x}}| \| \sim \|\mathbf{x}\| \ ,$$

which shows that $\Phi(\mathbf{x}) = |\hat{\mathbf{x}}|$ is not Lipschitz continuous to deformations, since ξ can be arbitrarily large.

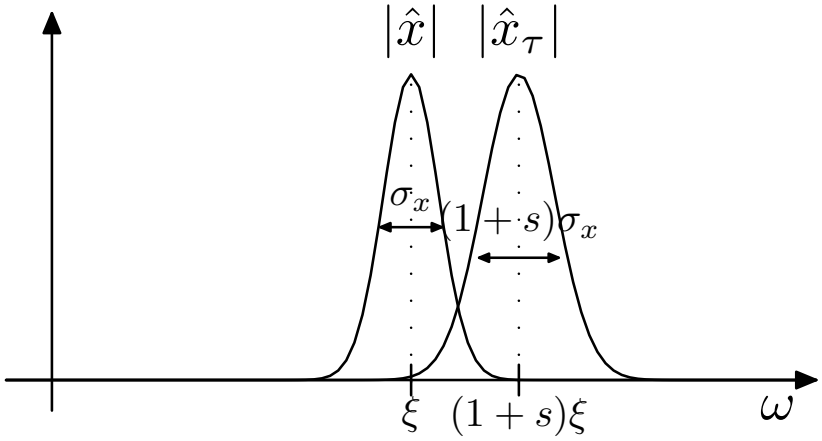


Figure 1: Dilation of a complex bandpass window. If $\xi \gg \sigma_x s^{-1}$, then the supports are nearly disjoint.

The autocorrelation of $\mathbf{x}$

$$R_{\mathbf{x}}(v) = \int \mathbf{x}(u)\mathbf{x}^*(u-v)du$$

is also translation invariant: $R_{\mathbf{x}} = R_{\mathbf{x}_c}$. Since $R_{\mathbf{x}}(v) = \mathbf{x} \star \overline{\mathbf{x}}(v)$, with $\overline{\mathbf{x}}(u) = \mathbf{x}^*(-u)$, it follows that $\widehat{R_x}(\omega) = |\hat{x}(\omega)|^2$. The Plancherel formula thus proves that it has the same instabilities as a Fourier transform:

$$\|R_{\mathbf{x}} - R_{\mathbf{x}_\tau}\| = (2\pi)^{-1} \| |\hat{\mathbf{x}}|^2 - |\hat{\mathbf{x}}_\tau|^2 \| \ .$$

Besides deformation instabilities, the Fourier modulus and the autocorrelation lose too much information. For example, a Dirac $\delta(u)$ and a linear chirp e^{iu^2} are two signals having Fourier transforms whose moduli are equal and constant. Very different signals may not be discriminated from their Fourier modulus.

A canonical invariant [KDGH07, Soa09] $\Phi(\mathbf{x}) = \mathbf{x}(u - a(\mathbf{x}))$ registers $\mathbf{x} \in \mathbf{L}^2(\mathbb{R}^d)$ with an anchor point $a(\mathbf{x})$, which is translated when $\mathbf{x}$ is translated:

$$a(\mathbf{x}_c) = a(\mathbf{x}) + c \ .$$

It thus defines a translation invariant representation: $\Phi\mathbf{x}_c = \Phi\mathbf{x}$. For example, the anchor point may be a filtered maximum $a(\mathbf{x}) = \arg\max_u |\mathbf{x} \star h(u)|$, for some filter $h(u)$. A canonical invariant $\Phi\mathbf{x}(u) = \mathbf{x}(u - a(\mathbf{x}))$ carries more information than a Fourier modulus, and characterizes $\mathbf{x}$ up to a global absolute position information [Soa09]. However, it has the same high-frequency instability as a Fourier modulus transform. Indeed, for any choice of anchor point $a(\mathbf{x})$, applying the Plancherel formula proves that

$$\|\mathbf{x}(u - a(\mathbf{x})) - \mathbf{x}'(u - a(\mathbf{x}'))\| \geq (2\pi)^{-1} \| |\hat{\mathbf{x}}(\omega)| - |\hat{\mathbf{x}}'(\omega)| \| \ .$$

If $\mathbf{x}' = \mathbf{x}_\tau$, the Fourier transform instability at high frequencies implies that $\Phi\mathbf{x} = \mathbf{x}(u - a(\mathbf{x}))$ is also unstable with respect to deformations.

3 Scattering on the Translation Group

This section reviews the Scattering transform on the translation group and its mathematical properties. Section 3.1 reviews windowed scattering transforms and its construction from Littlewood-Paley wavelet decompositions. Section 3.2 introduces the scattering metric and reviews the scattering energy conservation property, and Section 3.3 reviews the Lipschitz continuity property of scattering transforms with respect to deformations. Section 3.4 describes algorithmic aspects and implementation, and finally Section 3.5 illustrates scattering properties in computer vision applications.

3.1 Windowed Scattering transform

A wavelet transform is defined by dilating a mother wavelet $\psi \in \mathbf{L}^2(\mathbb{R}^d)$ with scale factors $\{a^j\}_{j\in\mathbb{Z}}$ for $a > 1$. In image processing applications one usually sets $a = 2$, whereas audio applications need smaller dilation factors, typically $a \leq 2^{1/8}$. Wavelets are not only dilated but also rotated along a discrete rotation group G of $\mathbb{R}^d$. As a result, a dilation by a^j and a rotation by $r \in G$ of ψ produce

$$\psi_{a^j r}(u) = a^{-dj}\psi(a^{-j}r^{-1}u) \ . \tag{7}$$

Wavelets are thus normalized in $\mathbf{L}^1(\mathbb{R}^d)$, such that $\|\psi_{a^j r}\|_1 = \|\psi\|_1$, which means that their Fourier transforms satisfy $\hat{\psi}_{a^j r}(\omega) = \hat{\psi}(a^j r\omega)$. In order to simplify notations, we denote $\lambda = a^j r \in a^{\mathbb{Z}} \times G$ and $|\lambda| = a^j$, and define $\psi_\lambda(u) = a^{-dj}\psi(\lambda^{-1}u)$. This notation will be used throughout the rest of the Chapter.

Scattering operators can be defined for general mother wavelets, but of particular interest are the complex wavelets that can be written as

$$\psi(u) = e^{i\eta u}\theta(u) \ ,$$

where θ is a lowpass window whose Fourier transform is real and has a bandwidth of the order of π. As a result, after a dilation and a rotation, $\hat{\psi}_\lambda(\omega) = \hat{\theta}(\lambda\omega - \eta)$ is centered at $\lambda^{-1}\eta$ and has a support size proportional to $|\lambda|^{-1}$. In Section 3.4.1 we shall specify the wavelet families used along all numerical experiments.

A Littlewood-Paley wavelet transform is a redundant representation which computes the following filter bank, without subsampling:

$$\forall u \in \mathbb{R}^d, \forall \lambda \in a^{\mathbb{Z}} \times G \ , W_\lambda \mathbf{x}(u) = \mathbf{x} \star \psi_\lambda(u) = \int \mathbf{x}(v)\psi_\lambda(u - v)\mathrm{d}v \ . \tag{8}$$

If $\mathbf{x}$ is real and the wavelet is chosen such that $\hat{\psi}$ is also real, then $W_{-\lambda}\mathbf{x} = W_\lambda \mathbf{x}^*$, which implies that in that case one can assimilate a rotation r with its negative version $-r$ into an equivalence class of positive rotations $G^+ = G/\{\pm 1\}$.

A wavelet transform with a finite scale 2^J only considers the subbands λ satisfying $|\lambda| \leq 2^J$. The low frequencies which are not captured by these wavelets are recovered by a lowpass filter ϕ_J whose spatial support is proportional to 2^J: $\phi_J(u) = 2^{-dJ}\phi(2^{-J}u)$. The wavelet transform at scale 2^J thus consists in the filter bank

$$\mathcal{W}_J \mathbf{x} = \{\mathbf{x} \star \phi_J, (W_\lambda \mathbf{x})_{\lambda\in\Lambda_J}\} \ ,$$

where $\Lambda_J = \{a^j r : r \in G^+, |\lambda| \leq 2^J\}$. Its norm is defined as

$$\|\mathcal{W}_J \mathbf{x}\|^2 = \|\mathbf{x} \star \phi_J\|^2 + \sum_{\lambda\in\Lambda_J} \|W_\lambda \mathbf{x}\|^2 \ .$$

$\mathcal{W}_J$ is thus a linear operator from $\mathbf{L}^2(\mathbb{R}^d)$ to a product space generated by copies of $\mathbf{L}^2(\mathbb{R}^d)$. It defines a frame of $\mathbf{L}^2(\mathbb{R}^d)$, whose bounds are characterized by the following Littlewood-Paley condition:

Proposition 3.1. *If there exists $\epsilon > 0$ such that for almost all $\omega \in \mathbb{R}^d$ and all $J \in \mathbb{Z}$*

$$1 - \epsilon \leq |\hat{\phi}(2^J\omega)|^2 + \frac{1}{2}\sum_{j \leq J}\sum_{r \in G} |\hat{\psi}(2^j r\omega)|^2 \leq 1 \ ,$$

then $\mathcal{W}_J$ is a frame with bounds given by $1-\epsilon$ and 1:

$$(1-\epsilon)\|\mathbf{x}\|^2 \leq \|\mathcal{W}_J\mathbf{x}\|^2 \leq \|\mathbf{x}\|^2 \quad , \mathbf{x} \in \mathbf{L}^2(\mathbb{R}^d) \ . \tag{9}$$

In particular, this Littlewood-Paley condition implies that $\hat{\psi}(0) = 0$ and hence that the wavelet must have at least a vanishing moment. When $\epsilon = 0$, the wavelet decomposition preserves the Euclidean norm and we say that it is unitary.

Wavelet coefficients are not translation invariant but translate as the input is translated, and their average $\int W_\lambda \mathbf{x}(u)\mathrm{d}u$ does not produce any information since wavelets have zero mean. A translation invariant measure which is also stable to the action of diffeomorphisms can be extracted out of each wavelet sub-band λ, by introducing a non-linearity which restores a non-zero, informative average value. This is for instance achieved by computing the complex modulus and averaging the result

$$\int |\mathbf{x} \star \psi_\lambda|(u)\mathrm{d}u \ . \tag{10}$$

Although many other choices of non-linearity are algorithmically possible, the complex modulus preserves the signal energy and enables overall energy conservation; see next Section. We will discuss in Section **??** how the choice of non-linearity is informed by geometric stability, and finally in Section 7 how half-rectified alternatives provide further insights into the signal through the Phase Harmonics.

The information lost by the averaging in (10) is recovered by a new wavelet decomposition $\{|\mathbf{x} \star \psi_\lambda| \star \psi_{\lambda'}\}_{\lambda' \in \Lambda_J}$ of $|\mathbf{x} \star \psi_\lambda|$, which produces new invariants by iterating the same procedure. Let $U[\lambda]\mathbf{x} = |\mathbf{x} \star \psi_\lambda|$ denote the wavelet modulus operator corresponding to the subband λ. Any sequence $p = (\lambda_1, \lambda_2, ..., \lambda_m)$ defines a *path*, i.e, the ordered product of non-linear and non-commuting operators

$$U[p]\mathbf{x} = U[\lambda_m] \, ... \, U[\lambda_2] \, U[\lambda_1]\mathbf{x} = |\ ||\mathbf{x} \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \, ... \, | \star \psi_{\lambda_m}| \ ,$$

with $U[\emptyset]\mathbf{x} = \mathbf{x}$.

Similarly as with frequency variables, one can manipulate path variables $p = (\lambda_1, \ldots, \lambda_m)$ in a number of ways. The scaling and rotation by $a^l g \in a^{\mathbb{Z}} \times G^+$ of a path p is denoted $a^l g p = (a^l g\lambda_1, \ldots, a^l g\lambda_m)$, and the concatenation of two paths is written $p + p' = (\lambda_1, \ldots, \lambda_m, \lambda'_1, \ldots, \lambda'_{m'})$.

Many applications in image and audio recognition require locally translation invariant representations, but which keep spatial or temporal information beyond a certain scale 2^J. A windowed scattering transform computes a locally translation invariant representation by applying a lowpass filter at scale 2^J with $\phi_{2^J}(u) = 2^{-2J}\phi(2^{-J}u)$.

Definition 3.2. *For each path $p = (\lambda_1, \ldots, \lambda_m)$ with $\lambda_i \in \Lambda_J$ and $\mathbf{x} \in \mathbf{L}^1(\mathbb{R}^d)$ we define the windowed scattering transform as*

$$S_J[p]\mathbf{x}(u) = U[p]\mathbf{x} \star \phi_{2^J}(u) = \int U[p]\mathbf{x}(v)\phi_{2^J}(u-v)\,\mathrm{d}v \ ,$$

A Scattering transform has the structure of a convolutional network, but its filters are given by wavelets instead of being learnt. Thanks to this structure, the resulting transform is locally translation invariant and stable to deformations, as will be discussed in 3.3. The scattering representation enjoys several appealing properties described in the following sections.

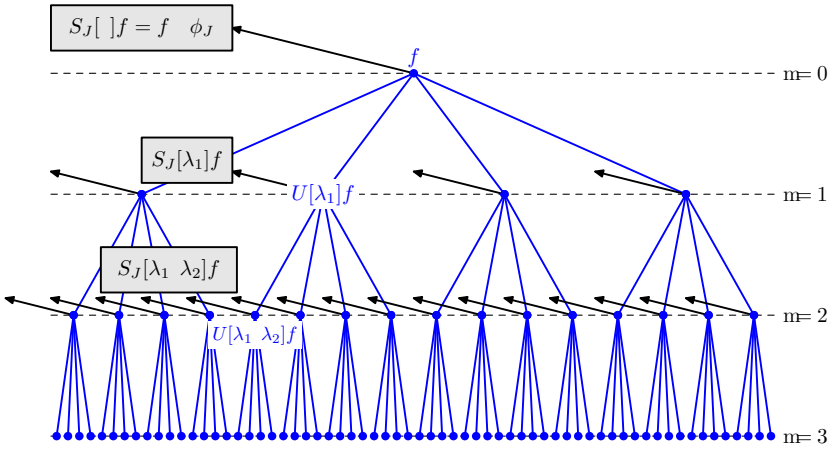


Figure 2: Convolutional structure of the windowed scattering transform. Each layer is computed from the previous by applying a wavelet modulus decomposition U on each envelope $U[p]\mathbf{x}$. The outputs of each layer are obtained via a lowpass filter ϕ_J.

3.2 Scattering metric and Energy Conservation

The windowed scattering representation is obtained by cascading a basic propagator operator,

$$\mathcal{U}_J\mathbf{x} = \{\mathbf{x} \star \phi_J, (U[\lambda]\mathbf{x})_{\lambda\in\Lambda_J}\} \ . \tag{11}$$

The first layer of the representation applies $\mathcal{U}_J$ to the input function, whereas successive layers are obtained by applying $\mathcal{U}_J$ to each output $U[p]\mathbf{x}$. Since $U[\lambda]U[p] = U[p+\lambda]$ and $U[p]\mathbf{x} \star \phi_J = S_J[p]\mathbf{x}$, it follows that

$$\mathcal{U}_J U[p]\mathbf{x} = \{S_J[p]\mathbf{x}, (U[p+\lambda]\mathbf{x})_{\lambda\in\Lambda_J}\} \ . \tag{12}$$

If Λ_J^m denotes the set of paths of length or *order* m, it follows from (12) that the $(m+1)$-th layer given by Λ_J^{m+1} is obtained from the previous layer via the propagator $\mathcal{U}_J$. We denote $\mathcal{P}_J$ the set of paths of any order up to scale 2^J, $\mathcal{P}_J = \cup_m \Lambda_J^m$.

The propagator $\mathcal{U}_J$ is non-expansive, since the wavelet decomposition $\mathcal{W}_J$ is non-expansive from (9) and the modulus is also non-expansive. As a result,

$$\|\mathcal{U}_J\mathbf{x} - \mathcal{U}_J\mathbf{x}'\|^2 = \|\mathbf{x}\star\phi_J - \mathbf{x}'\star\phi_J\|^2 + \sum_{\lambda\in\Lambda_J} \| |W_\lambda\mathbf{x}| - |W_\lambda\mathbf{x}'| \|^2 \leq \|\mathbf{x}-\mathbf{x}'\|^2 \ .$$

Moreover, if the wavelet decomposition is unitary, then the propagator $\mathcal{U}_J$ is also unitary.

For any path set Ω, the Euclidean norm defined by the scattering coefficients $S_J[p]$, $p \in \Omega$ is

$$\|S_J[\Omega]\mathbf{x}\|^2 = \sum_{p\in\Omega} \|S_J[p]\mathbf{x}\|^2 \ .$$

Since $S_J[\mathcal{P}_J]$ is constructed by cascading the non-expansive operator $\mathcal{U}_J$, it follows that $S_J[\mathcal{P}_J]$ is also non-expansive:

Proposition 3.3. *The windowed scattering transform is non-expansive:*

$$\forall \mathbf{x}, \mathbf{x}' \in \mathbf{L}^2(\mathbb{R}^d) \ , \quad \|S_J[\mathcal{P}_J]\mathbf{x} - S_J[\mathcal{P}_J]\mathbf{x}'\| \leq \|\mathbf{x}-\mathbf{x}'\| \ . \tag{13}$$

The windowed scattering thus defines a metric which is continuous with respect to the $\mathbf{L}^2(\mathbb{R}^d)$ euclidean metric, and thus it is stable to additive noise.

Let us now consider the case where the wavelet decomposition is unitary, ie $\epsilon = 0$ in (9). One can easily verify by induction on the path order $m = |p|$ that

$$\forall\, m\ ,\ \|\mathbf{x}\|^2 = \sum_{|p|<m} \|S_J[p]\mathbf{x}\|^2 + \sum_{|p|=m} \|U[p]\mathbf{x}\|^2\ .$$

This decomposition expresses the signal energy $\|\mathbf{x}\|^2$ in terms of coefficients captured by the first m layers of the scattering network, and a residual energy $\mathcal{R}_{J,\mathbf{x}}(m) := \sum_{p\in\mathcal{P}_J;|p|=m} \|U[p]\mathbf{x}\|^2$. An important question with practical implications is to understand the energy decay $\mathcal{R}_{J,\mathbf{x}}(m)$ as m grows, since this determines how many layers of processing are effectively needed to represent the input. In particular, the Scattering representation is energy-preserving if $\lim_{m\to\infty} \mathcal{R}_{J,\mathbf{x}}(m) = 0$.

This is established under mild assumptions on the wavelet decomposition for the univariate case $\mathbf{x} \in L^2(\mathbb{R})$ in [Wal17]:

Theorem 3.4 ([Wal17], Theorem 3.1)**.** *Let $\{\psi_j\}_{j\in\mathbb{Z}}$ be a family of wavelets satisfying the Littlewood-Paley condition (9), and such that*

$$\forall\, j, \omega > 0\ ,\ |\hat{\psi}_j(-\omega)| \leq |\hat{\psi}_j(\omega)|\ ,$$

with strict inequality for each ω for at least one scale. Finally, we assume for some $\epsilon > 0$

$$\hat{\psi}(\omega) = O(|\omega|^{1+\epsilon})\ .$$

Then for any $J \in \mathbb{Z}$, there exists $r > 0, a > 1$ such that for all $m \geq 2$ and $f \in L^2(\mathbb{R})$ it holds

$$\mathcal{R}_{J,\mathbf{x}}(m) \leq \|\mathbf{x}\|^2 - \|\mathbf{x} \star \chi_{ra^m}\|^2\ , \tag{14}$$

where χ_s is the Gaussian window $\chi_s(t) = \sqrt{\pi} s \exp(-(\pi s t)^2)$.

This result establishes in particular the energy conservation, owing to the square integrability of $\hat{\mathbf{x}} \in L^2(\mathbb{R})$. But, importantly, it also provides a quantitative rate in which the energy decays within the network: the energy in the input signal carried by frequencies around 2^k disappears after $O(k)$ layers, leading to exponential energy decay. An earlier version of the energy conservation was established in [Mal12] for general input dimensions, but under more restrictive admissibility conditions for the wavelet, and without the rate of convergence.

A similar energy conservation result with also exponential convergence rate has been established for extensions of the scattering transform, where the wavelet decomposition is replaced by other frames. [CL17] studies energy conservation for *uniform covering* frames, obtaining exponential convergence too. [WGB17] generalise this result to more general frames that are also allowed to vary from one layer to the next.

3.3 Local Translation Invariance and Lipschitz Continuity to Deformations

The windowed scattering metric defined in the previous section is non-expansive, which gives stability to additive perturbations. In this Section we review its geometric stability to the action of deformations, and its asymptotic translation invariance, as the localization scale 2^J increases.

Each choice of such localization scale defines a metric $d_J(\mathbf{x}, \mathbf{x}') := \|S_J[\mathcal{P}_J]\mathbf{x} - S_J[\mathcal{P}_J]\mathbf{x}'\|$. An induction argument over the non-expansive Littlewood-Paley property (9) shows that the limit of d_J as $J \to \infty$ is well defined thanks to the following non-expansive property:

Proposition 3.5 ([Mal12], Prop 2.9)**.** *For all $\mathbf{x}, \mathbf{x}' \in \mathbf{L}^2(\mathbb{R}^d)$ and $J \in \mathbb{Z}$,*

$$\|S_{J+1}[\mathcal{P}_{J+1}]\mathbf{x} - S_{J+1}[\mathcal{P}_{J+1}]\mathbf{x}'\| \leq \|S_J[\mathcal{P}_J]\mathbf{x} - S_J[\mathcal{P}_J]\mathbf{x}'\|\ .$$

As a result, the sequence $(\|S_J[\mathcal{P}_J]\mathbf{x} - S_J[\mathcal{P}_J]\mathbf{x}'\|)_J$ is positive and non-increasing as J increases, and hence it converges.

In fact, under mild assumptions, this limit metric is translation invariant:

Theorem 3.6 ([Mal12], Theorem 2.10)**.** *Let $x_v(u) = x(u-v)$. Then for admissible scattering wavelets satisfying the assumptions of Theorem (3.4) it holds*

$$\forall \mathbf{x} \in \mathbf{L}^2(\mathbb{R}^d)\ , \forall c \in \mathbb{R}^d\ ,\ \lim_{J\to\infty} \|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x_v\| = 0\ . \tag{15}$$

for $d = 1$.

Remark 3.7. *This result is proven in [Mal12] for general dimensions d under stronger assumptions on the wavelets (admissibility condition (2.28) in [Mal12]). However, these stronger assumptions may not be necessary, by extending the result in [Wal17] to arbitrary d.*

Remark 3.8. *[WB17] describes an interesting extension of Theorem 3.6 which holds for more general decomposition frames beyond wavelets, based on the notion of* vertical *translation invariance. This refers to the asymptotic translation invariance enjoyed by m-th layer coefficients of the network, as m grows.*

The translation invariance of the overall representation is based on two fundamental properties: (i) the equivariance of wavelet modulus decomposition operators with respect to translation, $\mathcal{U}_J\mathcal{T}_v\mathbf{x} = \mathcal{T}_v\mathcal{U}_J\mathbf{x}$, and (ii) the invariance provided by the local averaging operator $A_J\mathbf{x} := \mathbf{x} \star \phi_J$. Indeed, scattering coefficients up to order m are obtained by composing $\mathcal{U}_J$ up to m times followed by A_J. It follows that the translation invariance measured at order m is expressed as

$$\|S_J[\Lambda_J^m]\mathcal{T}_v\mathbf{x} - S_J[\Lambda_J^m]\mathbf{x}\| = \|A_J\mathcal{T}_vU[\Lambda_J^m]\mathbf{x} - A_JU[\Lambda_J^m]\mathbf{x}\| \leq \|U[\Lambda_J^m]\mathbf{x}\|\|A_J\mathcal{T}_v - A_J\|\ .$$

Besides asymptotic translation invariance, the windowed scattering transform defines a stable metric with respect to the action of diffeomorphisms, which can model non-rigid deformations. A diffeomorphism maps a point $u \in \mathbb{R}^d$ to $u - \tau(u)$, where $\tau(u)$ is a vector displacement field satisfying $\|\nabla\tau\|_\infty < 1$, where $\|\nabla\tau\|$ is the operator norm. As described in Section 2.1, it acts on functions $\mathbf{x} \in \mathbf{L}^2(\mathbb{R}^d)$ by composition: $\mathbf{x}_\tau(u) = \mathbf{x}(u - \tau(u))$. The following central theorem computes an upper bound of $\|S_J[\mathcal{P}_J]\mathbf{x}_\tau - S_J[\mathcal{P}_J]\mathbf{x}\|$. For that purpose, we assume an admissible scattering wavelet[2], and we define the auxiliary norm

$$\|U[\mathcal{P}_J]\mathbf{x}\|_1 = \sum_{m\geq 0} \|U[\Lambda_J^m]\mathbf{x}\|\ .$$

Theorem 3.9 ([Mal12], Theorem 2.12)**.** *There exists C such that every $\mathbf{x} \in \mathbf{L}^2(\mathbb{R}^d)$ with $\|U[\mathcal{P}_J]\mathbf{x}\|_1 < \infty$ and $\tau \in C^2(\mathbb{R}^d)$ with $\|\nabla\tau\|_\infty \leq 1/2$ satisfy*

$$\|S_J[\mathcal{P}_J]\mathbf{x}_\tau - S_J[\mathcal{P}_J]\mathbf{x}\| \leq C\|U[\mathcal{P}_J]\mathbf{x}\|_1 K(\tau)\ , \tag{16}$$

with

$$K(\tau) = 2^{-J}\|\tau\|_\infty + \|\nabla\tau\|_\infty \max\left(1, \log\frac{\sup_{u,u'}|\tau(u) - \tau(u')|}{\|\nabla\tau\|_\infty}\right) + \|H\tau\|_\infty\ ,$$

and for all $m \geq 0$, if $\mathcal{P}_{J,m} = \cup_{n<m}\Lambda_J^n$, then

$$\|S_J[\mathcal{P}_{J,m}]\mathbf{x}_\tau - S_J[\mathcal{P}_{J,m}]\mathbf{x}\| \leq Cm\|\mathbf{x}\|K(\tau)\ . \tag{17}$$

[2]Again, as mentioned in Remark 3.7, such admissible wavelet conditions can be relaxed by extending the energy conservation results from [Wal17].

This theorem shows that a diffeomorphism produces in the scattering domain an error bounded by a term proportional to $2^{-J}\|\tau\|_\infty$, which corresponds to the local translation invariance, plus a deformation error proportional to $\|\nabla\tau\|_\infty$. Whereas rigid translations $\mathcal{T}_v$ commute with all the convolutional or point-wise operators defining the scattering representation, non-rigid deformations no longer commute with convolutions. The essence of the proof is thus to control the *commutation error* between the wavelet decomposition and the deformation. If $\mathcal{L}_\tau$ denotes the deformation operator $\mathcal{L}_\tau \mathbf{x} = \mathbf{x}_\tau$, [Mal12] proves that

$$\|[\mathcal{W}_J, \mathcal{L}_\tau]\| = \|\mathcal{W}_J\mathcal{L}_\tau - \mathcal{L}_\tau\mathcal{W}_J\| \lesssim \|\nabla\tau\| \ ,$$

thanks to the scale separation properties of wavelet decompositions.

The norm $\|U[\mathcal{P}_J]\mathbf{x}\|_1$ measures the decay of the scattering energy across depth. Again, in the univariate case it is shown in [Wal17] that

$$\forall\, m \ , \ \|U[\Lambda_J^m]\mathbf{x}\| \leq \left(\int |\hat{\mathbf{x}}(\omega)|^2 h_m(\omega)\mathrm{d}\omega\right)^{1/2} \ ,$$

with $h_m(\omega) = 1 - \exp(-2(\omega/(ra^m))^2)$ and $a > 1$. Denote by

$$\mathcal{F} = \left\{\mathbf{x}; \int |\hat{\mathbf{x}}(\omega)|^2 \log(1+|\omega|)\mathrm{d}\omega < \infty\right\}$$

the space of functions whose Fourier transform is square integrable against a logarithmic scaling. This corresponds to a logarithmic Sobolev class of functions having an average modulus of continuity in $\mathbf{L}^2(\mathbb{R}^d)$. In that case, for $\mathbf{x} \in \mathcal{F}$, we verify that

Proposition 3.10. *If $\mathbf{x} \in \mathcal{F}$, then $\|U[\mathcal{P}_J]\mathbf{x}\|_1 < \infty$.*

This implies that the geometric stability bound from Theorem 3.9 applies to such functions, with an upper bound that does not blow up with depth. When $\mathbf{x}$ has compact support, the following corollary shows that the windowed scattering metric is Lipschitz continuous to the action of diffeomorphisms:

Corollary 3.11 ([Mal12], Corollary 2.15)**.** *For any compact set $\Omega \subset \mathbb{R}^d$ there exists C such that for all $\mathbf{x} \in \mathbf{L}^2(\mathbb{R}^d)$ supported in Ω with $\|U[\mathcal{P}_J]\mathbf{x}\|_1 < \infty$ and for all $\tau \in C^2(\mathbb{R}^d)$ with $\|\nabla\tau\|_\infty \leq 1/2$, then*

$$\|S_J[\mathcal{P}_{J,m}]\mathbf{x}_\tau - S_J[\mathcal{P}_{J,m}]\mathbf{x}\| \leq C\|U[\mathcal{P}_J]\mathbf{x}\|_1 \left(2^{-J}\|\tau\|_\infty + \|\nabla\tau\|_\infty + \|H\tau\|_\infty\right) \ . \tag{18}$$

The translation error term, proportional to $2^{-J}\|\tau\|_\infty$, can be reduced to a second-order error term, $2^{-2J}\|\tau\|_\infty^2$, by considering a first order Taylor approximation of each $S_J[p]\mathbf{x}$ [Mal12].

As mentioned earlier, [CL17] and [WB17] developed extensions of scattering representations by replacing scattering wavelets with other decomposition frames, also establishing deformation stability bounds. However, an important difference between these results and Theorem 3.9 is that no bandlimited assumption is made on the input signal $\mathbf{x}$, but rather the weaker condition that $\|U[\mathcal{P}_J]\mathbf{x}\|_1 < \infty$. For appropriate wavelets leading to exponential energy decay, such quantity is bounded for $\mathbf{x} \in L^1 \cap L^2$. Finally, another relevant work that connected the above geometric stability results with kernel methods is [BM17], in which a Convolutional Kernel is constructed that enjoys provable deformation stability.

3.4 Algorithms

We now describe algorithmic aspects of the scattering representation, in particular the choice of scattering wavelets and the overall implementation as a specific CNN architecture.

3.4.1 Scattering Wavelets

The Littlewood-Paley wavelet transform of $\mathbf{x}$, $\{\mathbf{x} \star \psi_\lambda(u)\}_\lambda$, defined in (8), is a redundant transform with no orthogonality property. It is stable and invertible if the wavelet filters $\hat{\psi}_\lambda(\omega)$ cover the whole frequency plane. On discrete images, to avoid aliasing, one may only capture frequencies in the circle $|\omega| \leq \pi$ inscribed in the image frequency square. Most camera images have negligible energy outside this frequency circle.

As mentioned in Section 3.1, one typically considers near-analytic wavelets, meaning that $|\hat{\psi}(-\omega)| \ll |\hat{\psi}(\omega)|$ for ω lying on a predefined half-space of $\mathbb{R}^2$. The reason is hinted in Theorem 3.4, namely the complex envelop of analytic wavelets is smoother than that of a real wavelet, and therefore more energy will be captured at earlier layers of the scattering representation.

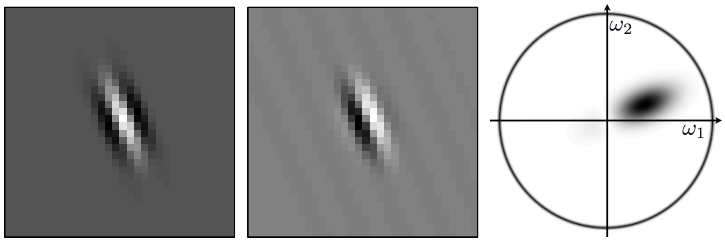


Figure 3: Complex Morlet wavelet. (a): Real part of $\psi(u)$. (b): Imaginary part of $\psi(u)$. (c): Fourier modulus $|\hat{\psi}(\omega)|$.

Let $u.u'$ and $|u|$ denote the inner product and norm in $\mathbb{R}^2$. A Morlet wavelet ψ is an example of complex wavelet given by

$$\psi(u) = \alpha \left(e^{iu.\xi} - \beta\right) e^{-|u|^2/(2\sigma^2)} ,$$

where $\beta \ll 1$ is adjusted so that $\int \psi(u)\, du = 0$. Its real and imaginary parts are nearly quadrature phase filters. Figure 3 shows the Morlet wavelet with $\sigma = 0.85$ and $\xi = 3\pi/4$, used in all classification experiments. The Morlet wavelet ψ shown in Figure 3 together with $\phi(u) = \exp(-|u|^2/(2\sigma^2))/(2\pi\sigma^2)$ for $\sigma = 0.7$ satisfy (9) with $\epsilon = 0.25$.

Cubic spline wavelets are an important family of unitary wavelets satisfying the Littlewood-Paley condition (9) with $\epsilon = 0$. They are obtained from a cubic-spline orthogonal Battle-Lemairé wavelet, defined from the conjugate mirror filter [Mal08]

$$\hat{h}(\omega) = \sqrt{\frac{S_8(\omega)}{2^8 S_8(2\omega)}} \text{ , with } S_n(\omega) = \sum_{k=-\infty}^{\infty} \frac{1}{(\omega + 2k\pi)^n} ,$$

which in the case $n = 8$ simplifies to the expression

$$S_8(2\omega) = \frac{5 + 30\cos^2(\omega) + 30\sin^2(\omega)\cos^2(\omega)}{1052^8 \sin^8(\omega)} + \frac{70\cos^4(\omega) + 2\sin^4(\omega)\cos^2(\omega) + 2/3\sin^6(\omega)}{1052^8 \sin^8(\omega)} .$$

In two dimensions, $\hat{\psi}$ is defined as a separable product in frequency polar coordinates $\omega = |\omega|\eta$, where η is a unit vector:

$$\forall |\omega|,\, \eta \in \mathbb{R}^+ \times \mathbf{S}^1 \ ,\ \hat{\psi}(\omega) = \hat{\psi}_1(|\omega|)\gamma(\eta) ,$$

with γ designed such that

$$\forall \eta \ ,\ \sum_{r \in G^+} |\gamma(r^{-1}\eta)|^2 = 1 .$$

Figure 4 shows the corresponding two-dimensional filters obtained with spline wavelets, by setting both $\hat{\psi}_1$ and γ to be cubic splines.

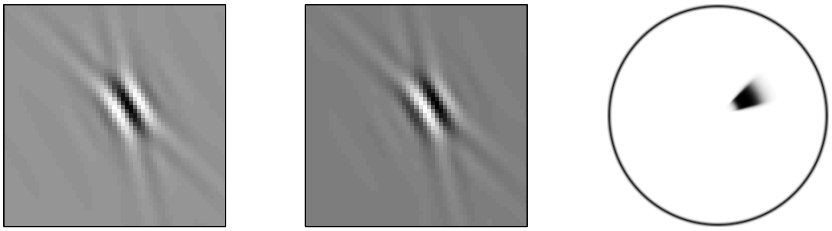


Figure 4: Complex cubic Spline wavelet. (a): Real part of $\psi(u)$. (b): Imaginary part of $\psi(u)$. (c): Fourier modulus $|\hat{\psi}(\omega)|$.

3.4.2 Fast Scattering Computations with Scattering Convolutional Network

A scattering representation is implemented with a CNN having a very specific architecture. As opposed to standard CNNs, output scattering coefficients are produced by each layer as opposed to the last layer.Filters are not learned from data but are predefined wavelets. If $p = (\lambda_1, ..., \lambda_m)$ is a path of length m then the windowed scattering coefficients $S_J[p]\mathbf{x}(u)$ of order m are computed at the layer m of a convolution network which is specified.

We describe a fast scattering implementation over frequency decreasing paths, where most of the scattering energy is concentrated. A frequency decreasing path $p = (2^{-j_1} r_1, ..., 2^{-j_m} r_m)$ satisfies $0 < j_k \leq j_{k+1} \leq J$. If the wavelet transform is computed over K rotation angles then the total number of frequency-decreasing paths of length m is $K^m \binom{J}{m}$. Let N be the number of pixels of the image x. Since ϕ_{2^J} is a low-pass filter scaled by 2^J, $S_J[p]\mathbf{x}(u) = U[p]\mathbf{x} \star \phi_{2^J}(u)$ is uniformly sampled at intervals $\alpha 2^J$, with $\alpha = 1$ or $\alpha = 1/2$. Each $S_J[p]\mathbf{x}$ is an image with $\alpha^{-2} 2^{-2J} N$ coefficients. The total number of coefficients in a scattering network of maximum depth $\overline{m}$ is thus

$$P = N\, \alpha^{-2}\, 2^{-2J} \sum_{m=0}^{\overline{m}} K^m \binom{J}{m}. \tag{19}$$

If $\overline{m} = 2$ then $P \simeq \alpha^{-2}\, N 2^{-2J} K^2 J^2/2$. It decreases exponentially when the scale 2^J increases.

Algorithm 1 describes the computations of scattering coefficients on sets $\mathcal{P}_{\downarrow}^m$ of frequency decreasing paths of length $m \leq \overline{m}$. The initial set $\mathcal{P}_{\downarrow}^0 = \{\emptyset\}$ corresponds to the original image $U[\emptyset]\mathbf{x} = \mathbf{x}$. Let $p + \lambda$ be the path which begins by p and ends with $\lambda \in \mathcal{P}$. If $\lambda = 2^{-j} r$ then $U[p+\lambda]\mathbf{x}(u) = |U[p]\mathbf{x} \star \psi_\lambda(u)|$ has energy at frequencies mostly below $2^{-j}\pi$. To reduce computations we can thus subsample this convolution at intervals $\alpha 2^j$, with $\alpha = 1$ or $\alpha = 1/2$ to avoid aliasing.

At the layer m there are $K^m \binom{J}{m}$ propagated signals $U[p]\mathbf{x}$ with $p \in \mathcal{P}_{\downarrow}^m$. They are sampled at intervals $\alpha 2^{j_m}$ which depend on p. One can verify by induction on m that the layer m has a total number of samples equal to $\alpha^{-2}\, (K/3)^m\, N$. There are also $K^m \binom{J}{m}$ scattering signals $S[p]\mathbf{x}$ but they are subsampled by 2^J and thus have much less coefficients. The number of operation to compute each layer is therefore driven by the $O((K/3)^m\, N \log N)$ operations needed to compute the internal propagated coefficients with FFT's. For $K > 3$, the overall computational complexity is thus $O((K/3)^{\overline{m}}\, N \log N)$.

The package Kymatio [AAE+18] provides a modern implementation of scattering transforms leveraging efficient GPU-optimized routines.

3.5 Empirical Analysis of Scattering Properties

To illustrate the properties of scattering representations, let us describe a visualization procedure. For a fixed position u, windowed scattering coefficients $S_J[p]\mathbf{x}(u)$ of order $m = 1, 2$ are displayed as

Algorithm 1 Fast Scattering Transform

for $m = 1$ to $\overline{m}$ **do**
  for all $p \in \mathcal{P}_{\downarrow}^{m-1}$ **do**
    Output $S_J[p]\mathbf{x}(\alpha 2^J n) = U[p]\mathbf{x} \star \phi_{2^J}(\alpha 2^J n)$
  end for
  for all $p + \lambda_m \in \mathcal{P}_{\downarrow}^{m}$ with $\lambda_m = 2^{-j_m} r_m$ **do**
    Compute

$$U[p + \lambda_m]\mathbf{x}(\alpha 2^{j_m} n) = |U[p]\mathbf{x} \star \psi_{\lambda_m}(\alpha 2^{j_m} n)|$$

  end for
end for
for all $p \in \mathcal{P}_{\downarrow}^{\max}$ **do**
  Output $S_J[p]\mathbf{x}(\alpha 2^J n) = U[p]\mathbf{x} \star \phi_{2^J}(\alpha 2^J n)$
end for

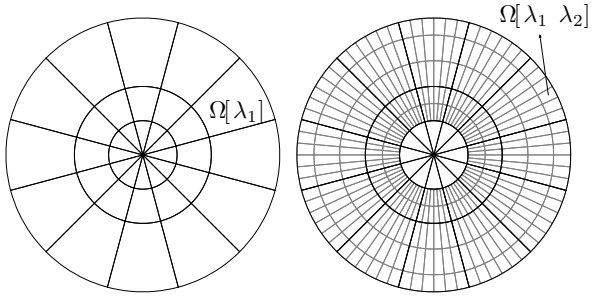


Figure 5: To display scattering coefficients, the disk covering the image frequency support is partitioned into sectors $\Omega[p]$, which depend upon the path p. (a): For $m = 1$, each $\Omega[\lambda_1]$ is a sector rotated by r_1 which approximates the frequency support of $\hat{\psi}_{\lambda_1}$. (b): For $m = 2$, all $\Omega[\lambda_1, \lambda_2]$ are obtained by subdividing each $\Omega[\lambda_1]$.

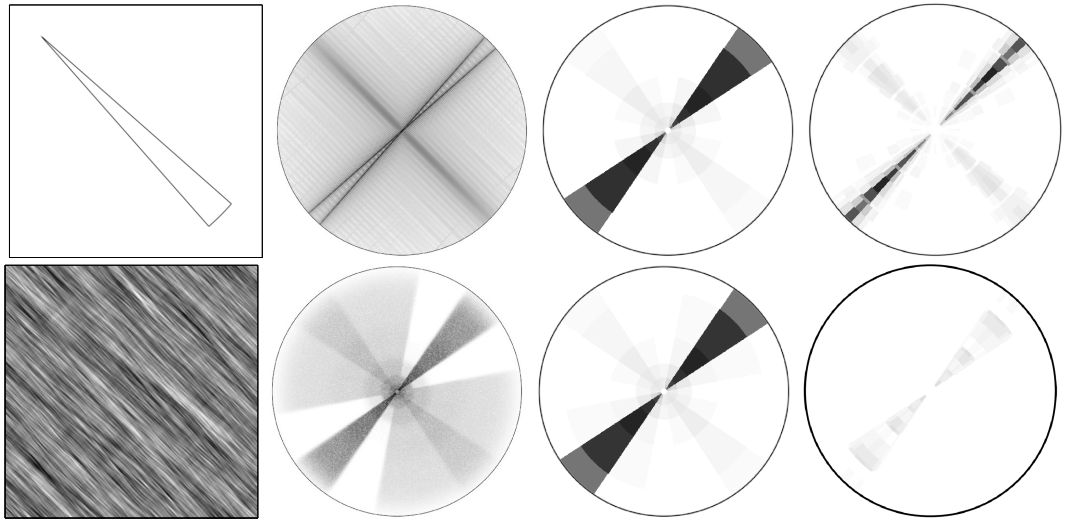


Figure 6: (a) Two images $x(u)$. (b) Fourier modulus $|\hat{x}(\omega)|$. (c) First order scattering coefficients $S_J x[\lambda_1]$ displayed over the frequency sectors of Figure 5(a). They are the same for both images. (d) Second order scattering coefficients $S_J x[\lambda_1, \lambda_2]$ over the frequency sectors of Figure 5(b). They are different for each image.

piecewise constant images over a disk representing the Fourier support of the image $\mathbf{x}$. This frequency disk is partitioned into sectors $\{\Omega[p]\}_{p \in \mathcal{P}^m}$ indexed by the path p. The image value is $S_J[p]\mathbf{x}(u)$ on the frequency sectors $\Omega[p]$, shown in Figure 5.

For $m = 1$, a scattering coefficient $S_J[\lambda_1]\mathbf{x}(u)$ depends upon the local Fourier transform energy of $\mathbf{x}$ over the support of $\hat{\psi}_{\lambda_1}$. Its value is displayed over a sector $\Omega[\lambda_1]$ which approximates the frequency support of $\hat{\psi}_{\lambda_1}$. For $\lambda_1 = 2^{-j_1} r_1$, there are K rotated sectors located in an annulus of scale 2^{-j_1}, corresponding to each $r_1 \in G$, as shown by Figure 5(a). Their area are proportional to $\|\psi_{\lambda_1}\|^2 \sim K^{-1}\, 2^{-j_1}$.

Second order scattering coefficients $S_J[\lambda_1, \lambda_2]\mathbf{x}(u)$ are computed with a second wavelet transform which performs a second frequency subdivision. These coefficients are displayed over frequency sectors $\Omega[\lambda_1, \lambda_2]$ which subdivide the sectors $\Omega[\lambda_1]$ of the first wavelets $\hat{\psi}_{\lambda_1}$, as illustrated in Figure 5(b). For $\lambda_2 = 2^{-j_2} r_2$, the scale 2^{j_2} divides the radial axis and the resulting sectors are subdivided into K angular sectors corresponding to the different r_2. The scale and angular subdivisions are adjusted so that the area of each $\Omega[\lambda_1, \lambda_2]$ is proportional to $\| |\psi_{\lambda_1}| \star \psi_{\lambda_2} \|^2$.

A windowed scattering S_J is computed with a cascade of wavelet modulus operators $\mathcal{U}$ defined in (11), and its properties thus depend upon the wavelet transform properties. Sections 3.3 and 3.2 gave conditions on wavelets to define a scattering transform which is non-expansive and preserves the signal norm. The scattering energy conservation shows that $\|S_J[p]\mathbf{x}\|$ decreases quickly as the length of p increases, and is non-negligible only over a particular subset of frequency-decreasing paths. Reducing computations to these paths defines a convolution network with much fewer internal and output coefficients.

Theorem 3.4 proves that the energy captured by the m-th layer of the scattering convolutional network, $\sum_{|p|=m} \|S_J[p]x\|^2$, converges to 0 as $m \to \infty$. The scattering energy conservation also proves that the more sparse the wavelet coefficients, the more energy propagates to deeper layers. Indeed, when 2^J increases, one can verify that at the first layer, $S_J[\lambda_1]\mathbf{x} = |\mathbf{x} \star \psi_{\lambda_1}| \star \phi_{2^J}$ converges to $\|\phi\|^2 \, \|\mathbf{x} \star \psi_\lambda\|_1^2$. The more sparse $\mathbf{x} \star \psi_\lambda$, the smaller $\|\mathbf{x} \star \psi_\lambda\|_1$ and hence the more energy is propagated to deeper layers to satisfy the global energy conservation of Theorem 3.4.

Figure 6 shows two images having same first order scattering coefficients, but the top image is piecewise regular and hence has wavelet coefficients which are much more sparse than the uniform

Table 1: Percentage of energy $\sum_{p \in \mathcal{P}^m_\downarrow} \|S_J[p]\mathbf{x}\|^2 / \|\mathbf{x}\|^2$ of scattering coefficients on frequency-decreasing paths of length m, depending upon J. These average values are computed on the Caltech-101 database, with zero mean and unit variance images.

J	$m=0$	$m=1$	$m=2$	$m=3$	$m=4$	$m \leq 3$
1	95.1	4.86	-	-	-	99.96
2	87.56	11.97	0.35	-	-	99.89
3	76.29	21.92	1.54	0.02	-	99.78
4	61.52	33.87	4.05	0.16	0	99.61
5	44.6	45.26	8.9	0.61	0.01	99.37
6	26.15	57.02	14.4	1.54	0.07	99.1
7	0	73.37	21.98	3.56	0.25	98.91

texture at the bottom. As a result the top image has second order scattering coefficients of larger amplitude than at the bottom. Higher-order coefficients are not displayed because they have a negligible energy. For typical images, as in the CalTech101 dataset [FFFP04], Table 1 shows that the scattering energy has an exponential decay as a function of the path length m. Scattering coefficients are computed with cubic spline wavelets, which define a unitary wavelet transform and satisfy the scattering admissibility condition for energy conservation. As expected, the energy of scattering coefficients converges to 0 as m increases, and it is already below 1% for $m \geq 3$.

The propagated energy $\|U[p]\mathbf{x}\|^2$ decays because $U[p]\mathbf{x}$ is a progressively lower frequency signal as the path length increases. Indeed, each modulus computes a regular envelop of oscillating wavelet coefficients. The modulus can thus be interpreted as a non-linear "demodulator" which pushes the wavelet coefficient energy towards lower frequencies. As a result, an important portion of the energy of $U[p]\mathbf{x}$ is then captured by the low pass filter ϕ_{2^J} which outputs $S_J[p]\mathbf{x} = U[p]\mathbf{x} \star \phi_{2^J}$. Hence fewer energy is propagated to the next layer.

Another consequence is that the scattering energy propagates only along a subset of frequency decreasing paths. Since the envelope $|\mathbf{x} \star \psi_\lambda|$ is more regular than $\mathbf{x} \star \psi_\lambda$, it follows that $|\mathbf{x} \star \psi_\lambda(u)| \star \psi_{\lambda'}$ is non-negligible only if $\psi_{\lambda'}$ is located at lower frequencies than ψ_λ, and hence if $|\lambda'| < |\lambda|$. Iterating on wavelet modulus operators thus propagates the scattering energy along frequency-decreasing paths $p = (\lambda_1, ..., \lambda_m)$ where $|\lambda_k| < |\lambda_{k-1}|$ for $1 \leq k < m$. We denote by $\mathcal{P}^m_\downarrow$ the set of frequency decreasing (or equivalently scale increasing) paths of length m. Scattering coefficients along other paths have a negligible energy. This is verified by Table 1 which shows not only that the scattering energy is concentrated on low-order paths, but also that more than 99% of the energy is absorbed by frequency-decreasing paths of length $m \leq 3$. Numerically, it is therefore sufficient to compute the scattering transform along frequency-decreasing paths. It defines a much smaller convolution network. Section 3.4.2 shows that the resulting coefficients are computed with $O(N \log N)$ operations.

Signal Recovery versus Energy Conservation: Preserving energy does not imply that the signal information is preserved. Since a scattering transform is calculated by iteratively applying $\mathcal{U}$, inverting S_J requires to invert $\mathcal{U}$. The wavelet transform $\mathcal{W}$ is a linear invertible operator, so inverting $\mathcal{U}z = \{z \star \phi_{2^J}, |z \star \psi_\lambda|\}_{\lambda \in \mathcal{P}}$ amounts to recovering the complex phases of wavelet coefficients removed by the modulus. The phase of Fourier coefficients can not be recovered from their modulus, but wavelet coefficients are redundant, as opposed to Fourier coefficients. For particular wavelets, it has been proved that the phase of wavelet coefficients can be recovered from their modulus, and that $\mathcal{U}$ has a continuous inverse [Wal12].

Still, one can not exactly invert S_J because we discard information when computing the scattering coefficients $S_J[p]\mathbf{x} = U[p]\mathbf{x} \star \phi_{2^J}$ of the last layer $\mathcal{P}^{\overline{m}}$. Indeed, the propagated coefficients $|U[p]\mathbf{x} \star \psi_\lambda|$ of the next layer are eliminated, because they are not invariant and have a negligible total energy. The

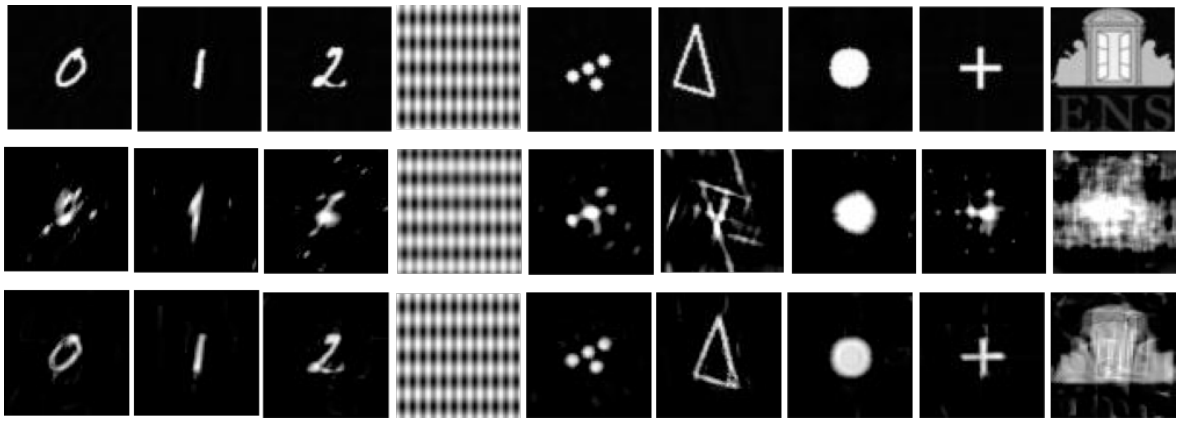


Figure 7: Signal Reconstruction from Scattering coefficients $S_J\mathbf{x}$ with $J = \log N$. Top: original images, Middle: reconstruction from only first-order coefficients. Bottom: reconstruction using first and second-order coefficients.

number of such coefficients is larger than the total number of scattering coefficients kept at previous layers. Initializing the inversion by considering that these small coefficients are zero produces an error. This error is further amplified as the inversion of $\mathcal{U}$ progresses across layers from $\overline{m}$ to 0.

Yet, under some structural assumptions on the signal $\mathbf{x}$, it is possible to recover the signal from its scattering coefficients $\mathbf{z} = S_J\mathbf{x}$. For instance, if $\mathbf{x}$ admits a sparse wavelet decomposition, [BM18] shows that important geometrical information of $\mathbf{x}$ is preserved in $S_J\mathbf{x}$. Figure 7 illustrates the signal recovery using either $m = 1$ or $m = 2$ with $J = \log N$. The recovery is obtained using a gradient-descent on the energy $E(\mathbf{x}) = \|S_J\mathbf{x} - \mathbf{z}\|^2$ described in Section 6. In this case, first-order scattering provides a collection of ℓ_1 norms $\{\|\mathbf{x} \star \psi_\lambda\|_1\}_\lambda$, which recover the overall regularity of the signal, but fail to reconstruct its geometry. Adding second-order coefficients results in $O(\log N^2)$ coefficients and substantially improves the reconstruction quality. In essence, the sparsity in these images creates no scale interactions on a large subset of scattering coefficients, which reduces the loss of information caused by the removal of the wavelet phases.

For natural images with weaker sparsity, Figure 8 shows reconstructions from second-order scattering coefficients for different values of J, using the same recovery algorithm. When the scale 2^J is such that the number of scattering coefficients is comparable with the dimensionality of $\mathbf{x}$, we observe good perceptual quality. When $\dim(S_J\mathbf{x}) \ll \dim(\mathbf{x})$, scattering coefficients define an underlying generative model based on a microcanonical maximum entropy principle, as described in Section 6.

3.6 Scattering in Modern Computer Vision

Thanks to their provable deformation stability and ability to preserve important geometric information, scattering representations are suitable as feature extractors in many computer vision pipelines.

First demonstrated in [BM13] on handwritten digit classification and texture recognition, scattering-based image classifcation models have been further developed in [OM15, OBZ17, OZH$^+$18], by extending the wavelet decomposition to other transformation groups (see Section 5) and by integrating them within CNN architectures as preprocessing stages.

In particular, the results from [OZH$^+$18] demonstrate that the geometric priors of scattering representations provide a better trade-off than data-driven models in the small-training regime, where large capacity CNNs tend to overfit. Even first-order scattering coefficients may be used to ease inference and learning within CNN pipelines, as demonstrated in [OBZV18].

Also, let us mention several works that considered 'hybrid' models between the fully-structured scattering networks and the fully-trainable CNNs. [JvGLS16] proposed to learn convolutional filters

Figure 8: Samples from $\Omega_{J,\epsilon}$ for different values of J using the gradient descent algorithm described in Section 6.3. Top row: original images, second row: $J = 3$, third row: $J = 4$, fourth row: $J = 5$, fifth row: $J = 6$. The visual quality of the reconstruction is nearly perfect for $J = 3$ and degrades for larger values of J.

in the wavelet domain, leveraging the benefits of multiscale decompositions. [CW16a, CW16b, KT18] added the structure of group convolution of the joint scattering representation of Section 5 into CNNs, significantly improving the sample complexity.

4 Scattering Representations of Stochastic Processes

This section reviews the definitions and basic properties of the expected scattering of random processes [Mal12, BMB$^+$15]. We prove a version of scattering mean-square consistency for orthogonal Haar scattering in Section 4.1.

4.1 Expected Scattering

If $(X(t))_{t\in\mathbb{R}}$ is a stationary process or has stationary increments, meaning that $\delta_s X(t) = X(t)-X(t-s)$ is stationary for all s, then $X \star \psi_\lambda$ is also stationary, and taking the modulus preserves stationarity. It follows that for any path $p = (\lambda_1, ..., \lambda_m) \in \mathcal{P}_\infty$, the process

$$U[p]X = |...|X \star \psi_{\lambda_1}| \star ...| \star \psi_{\lambda_m}|$$

is stationary, hence its expected value does not depend upon the spatial position t.

Definition 4.1. *Let $X(t)$ be a stochastic process with stationary increments. The expected scattering of X is defined for all $p \in \mathcal{P}_\infty$ by*

$$\overline{S}X(p) = \mathbb{E}(U[p]X) = \mathbb{E}(|...|X \star \psi_{\lambda_1}| \star ...| \star \psi_{\lambda_m}|) \ .$$

The expected scattering defines a representation for the process $X(t)$ which carries information on high order moments of $X(t)$, as we shall see in later sections. It also defines a metric between stationary processes, given by

$$\|\overline{S}X - \overline{S}Y\|^2 := \sum_{p\in\mathcal{P}_\infty} |\overline{S}X(p) - \overline{S}Y(p)|^2 \ .$$

The scattering representation of $X(t)$ is estimated by computing a windowed scattering transform of a realization $\mathbf{x}$ of $X(t)$. If $\Lambda_J = \{\lambda = 2^j \ ; \ 2^{-j} > 2^{-J}\}$ denotes the set of scales smaller than J, and $\mathcal{P}_J$ is the set of finite paths $p = (\lambda_1, .., \lambda_m)$ with $\lambda_k \in \Lambda_J \ \forall k$, then the windowed scattering at scale J of a realization $\mathbf{x}(t)$ is

$$S_J[\mathcal{P}_J]\mathbf{x} = \{U[p]\mathbf{x} \star \phi_J \ , \ p \in \mathcal{P}_J\} \ . \tag{20}$$

Since $\int \phi_J(u)du = 1$, we have $\mathbb{E}(S_J[\mathcal{P}_J]X) = \mathbb{E}(U[p]X) = \overline{S}X(p)$, so S_J is an unbiased estimator of the scattering coefficients contained in $\mathcal{P}_J$. When the wavelet ψ satisfies the Littlewood-Paley condition (9), the non-expansive nature of the operators defining the scattering transform implies that $\overline{S}$ and $S_J[\mathcal{P}_J]$ are also non-expansive, similarly as the deterministic case covered in Proposition 3.3:

Proposition 4.2. *If X and Y are finite second order stationary processes, then*

$$\mathbb{E}(\|S_J[\mathcal{P}_J]X - S_J[\mathcal{P}_J]Y\|^2) \ \leq \ \mathbb{E}(|X-Y|^2) \ , \tag{21}$$

$$\|\overline{S}X - \overline{S}Y\|^2 \ \leq \ \mathbb{E}(|X-Y|^2) \ , \tag{22}$$

in particular

$$\|\overline{S}X\|^2 \leq \mathbb{E}(|X|^2) \ . \tag{23}$$

The $\mathbf{L}^2(\mathbb{R}^d)$ energy conservation theorem (3.4) yields an equivalent energy conservation property for the mean squared power:

Theorem 4.3 ([Wal17], Theorem 5.1)**.** *Under the same assumptions on scattering wavelets as in Theorem 3.4, and if X is stationary, then*

$$\mathbb{E}(\|S_J[\mathcal{P}_J]X\|^2) = \mathbb{E}(|X|^2) \ . \tag{24}$$

Expected scattering coefficients are estimated with the windowed scattering $S_J[p]X = U[p]X \star \psi_J$ for each $p \in \mathcal{P}_J$. If $U[p]X$ is ergodic, $S_J[p]X$ converges in probability to $\overline{S}X(p) = \mathbb{E}(U[p]X)$ when $J \to \infty$. A process $X(t)$ with stationary increments is said to have a mean squared consistent scattering if the total variance of $S_J[\mathcal{P}_J]X$ converges to zero as J increases:

$$\lim_{J\to\infty} \mathbb{E}(\|S_J[\mathcal{P}_J]X - \overline{S}X\|^2) = \sum_{p\in\mathcal{P}_J} \mathbb{E}(|S_J[p]X - \overline{S}X(p)|^2) = 0 \ . \tag{25}$$

This condition implies that $S_J[\mathcal{P}_J]X$ converges to $\overline{S}X$ with probability 1. Mean square consistent scattering is observed numerically on a variety of processes, including Gaussian and non-Gaussian fractal processes. It was conjectured in [Mal12] that Gaussian stationary processes X whose autocorrelation R_X is in $\mathbf{L}^1$ have a mean squared consistent scattering.

Consistency of Orthogonal Haar Scattering: We show a partial affirmative answer of this conjecture, by considering a specific scattering representation built from discrete orthogonal real Haar wavelets. Consider $(X_n)_{n\in\mathbb{Z}}$ a stationary process defined over discrete time-steps. The orthogonal Haar scattering transform S_J^H maps 2^J samples of X_n into 2^J coefficients, defined recursively as

$$\begin{aligned}
x^{0,k} &= X_k,\ k = 0 \dots 2^J - 1 \\
x^{j,k} &= \frac{1}{2}(x^{j-1,2k} + x^{j-1,2k+1}) \ , \ x^{j,k+2^{J-j}} = \frac{1}{2}|x^{j-1,2k} - x^{j-1,2k+1}| \ , 0 < j \le J, k = 0 \dots 2^{J-1} - 1, \\
S_J^H X &:= (x^{J,k}; k = 0 \dots 2^J - 1) \, .
\end{aligned} \tag{26}$$

This representation thus follows a multiresolution analysis (MRA) [Mal99] but also decomposes the details at each scale, after applying the modulus non-linearity. It is easy to verify by induction that (26) defines an orthogonal transformation that preserves the energy: $\|S_J^H \mathbf{x}\| = \|\mathbf{x}\|$. However, contrary to the Littlewood-Paley wavelet decomposition, orthogonal wavelets are defined from downsampling operators, and therefore the resulting scattering representation S_J^H is not translation invariant when $J \to \infty$. We have the following consistency result:

Theorem 4.4 ([Bru19])**.** *The progressive Haar Scattering operator S_J^H is consistent in the class of compactly supported linear processes, in the sense that*

$$\lim_{J\to\infty} \mathbb{E}(\|S_J^H X - \mathbb{E}S_J^H X\|^2) = 0 \ , \tag{27}$$

for stationary processes X which can be represented as $X = W \star h$, where W is a white noise and h is compactly supported.

As a consequence of Theorem 4.3, mean squared consistency implies an expected scattering energy conservation:

Corollary 4.5. *For admissible wavelets as in Theorem 4.3, $S_J[\mathcal{P}_J]X$ is mean squared consistent if and only if*

$$\|\overline{S}X\|^2 = \mathbb{E}(|X|^2) \ .$$

Expected scattering coefficients depend upon normalized high order moments of X. If one expresses $|U[p]X|^2$ as

$$|U[p]X(t)|^2 = \mathbb{E}(|U[p]X|^2)(1 + \epsilon(t)) \ ,$$

then, assuming $|\epsilon| \ll 1$, a first order approximation of

$$U[p]X(t) = \sqrt{|U[p]X(t)|^2} \approx \mathbb{E}(|U[p]X|^2)^{1/2}(1+\epsilon/2)$$

yields

$$U[p+\lambda]X = |U[p]X \star \psi_\lambda| \approx \frac{||U[p]X|^2 \star \psi_\lambda|}{2\mathbb{E}(|U[p]X|^2)^{1/2}} ,$$

thus showing that $\overline{S}X(p) = \mathbb{E}(U[p]X)$ for $p = (\lambda_1, \ldots, \lambda_m)$ depends upon normalized moments of X of order 2^m, determined by the cascade of wavelet sub-bands λ_k. As opposed to a direct estimation of high moments, scattering coefficients are computed with a non-expansive operator which allows consistent estimation with few realizations. This is a fundamental property which enables texture recognition and classification from scattering representations [Bru13].

The scattering representation is related to the sparsity of the process through the decay of its coefficients $\overline{S}X(p)$ as the order $|p|$ increases. Indeed, the ratio of the first two moments of X

$$\rho_X = \frac{\mathbb{E}(|X|)}{\mathbb{E}(|X|^2)^{1/2}}$$

gives a rough measure of the fatness of the tails of X.

For each p, the Littlewood-Paley unitarity condition satisfied by ψ gives

$$\mathbb{E}(|U[p]X|^2) = \mathbb{E}(U[p]X)^2 + \sum_\lambda \mathbb{E}(|U[p+\lambda]X|^2) ,$$

which yields

$$1 = \rho_{U[p]X} + \frac{1}{\mathbb{E}(|U[p]X|^2)} \sum_\lambda \mathbb{E}(|U[p+\lambda]X|^2) . \tag{28}$$

Thus, the fraction of energy that is trapped at a given path p is given by the relative sparsity $\rho_{U[p]X}$.

This relationship between sparsity and scattering decay across the orders is of particular importance for the study of point processes, which are sparse in the original spatial domain, and for regular image textures, which are sparse when decomposed in the first level UX of the transform. In particular, the scattering transform can easily discriminate between white noises of different sparsity, such as Bernouilli and Gaussian.

The autocovariance of a real stationary process X is denoted

$$RX(\tau) = \mathbb{E}\Big((X(x) - \mathbb{E}(X)) \, (X(x-\tau) - \mathbb{E}(X)) \Big) .$$

Its Fourier transform $\widehat{R}X(\omega)$ is the power spectrum of X. Replacing X by $X \star \psi_\lambda$ in the conservation energy formula (4.3) implies that

$$\sum_{p \in \mathcal{P}_J} \mathbb{E}(|S_J[p+\lambda]X|^2) = E(|X \star \psi_\lambda|^2) . \tag{29}$$

These expected squared wavelet coefficients can also be written as a filtered integration of the Fourier power spectrum $\widehat{R}X(\omega)$

$$\mathbb{E}(|X \star \psi_\lambda|^2) = \int \widehat{R}X(\omega) \, |\hat{\psi}(\lambda^{-1}\omega)|^2 \, d\omega . \tag{30}$$

These two equations prove that summing scattering coefficients recovers the power spectrum integral over each wavelet frequency support, which only depends upon second-order moments of X. However, scattering coefficients $\overline{S}X(p)$ depend upon moments of X up to the order 2^m if p has a length m. Scattering coefficients can thus discriminate textures having same second-order moments but different higher-order moments.

4.2 Analysis of stationary textures with scattering

Section 4.1 showed that the scattering representation can be used to describe stationary processes, in such a way that high order moments information is captured and estimated consistently with few realizations.

Image textures can be modeled as realizations of stationary processes $X(u)$. The Fourier spectrum $\widehat{R}_X(\omega)$ is the Fourier transform of the autocorrelation

$$R_X(\tau) = \mathbb{E}\Big([X(u) - \mathbb{E}(X)][X(u-\tau) - \mathbb{E}(X)]\Big) \ .$$

Despite the importance of spectral methods, the Fourier spectrum is often not sufficient to discriminate image textures because it does not take into account higher-order moments.

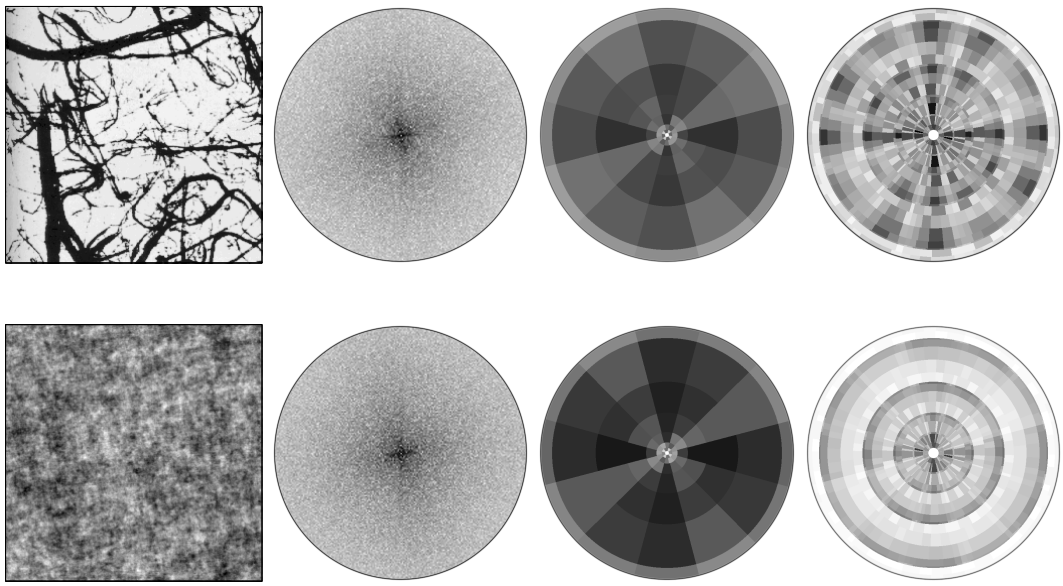


Figure 9: Two different textures having the same Fourier power spectrum. (a) Textures $X(u)$. Top: Brodatz texture. Bottom: Gaussian process. (b) Same estimated power spectrum $\widehat{R}X(\omega)$. (c) Nearly same scattering coefficients $S_J[p]X$ for $m = 1$ and 2^J equal to the image width. (d) Different scattering coefficients $S_J[p]X$ for $m = 2$.

The discriminative power of scattering representations is illustrated using the two textures in Figure 9, which have the same power spectrum and hence same second order moments. Scattering coefficients $S_J[p]X$ are shown for $m = 1$ and $m = 2$ with the frequency tiling illustrated in Figure 5. The ability to discriminate the top process X_1 from the bottom process X_2 is measured by a scattering distance normalized by the variance:

$$\rho(m) = \frac{\|S_J X_1[\Lambda_J^m] - \mathbb{E}(S_J X_2[\Lambda_J^m])\|^2}{\mathbb{E}(\|S_J X_2[\Lambda_J^m] - \mathbb{E}(S_J X_2[\Lambda_J^m])\|^2)} \ .$$

For $m = 1$, scattering coefficients mostly depend upon second-order moments and are thus nearly equal for both textures. One can indeed verify numerically that $\rho(1) = 1$ so both textures can not be distinguished using first order scattering coefficients. On the contrary, scattering coefficients of order 2 are highly dissimilar because they depend on moments up to order 4, and $\rho(2) = 5$. A scattering

Table 2: Decay of the total scattering variance $\sum_{p\in\mathcal{P}_J}\mathbb{E}(|S_J[p]X-\overline{S}X(p)|^2)/E(|X|^2)$ in percentage, as a function of J, averaged over the Brodatz dataset. Results obtained using cubic spline wavelets.

$J=1$	$J=2$	$J=3$	$J=4$	$J=5$	$J=6$	$J=7$
85	65	45	26	14	7	2.5

representation of stationary processes includes second order and higher-order moment descriptors of stationary processes, which discriminates between such textures.

The windowed scattering $S_J[\mathcal{P}_J]X$ estimates scattering coefficients by averaging wavelet modulus over a support of size proportional to 2^J. If X is a stationary process, Section 4.1 showed that the expected scattering transform $\overline{S}X$ is estimated with the windowed scattering

$$S_J[\mathcal{P}_J]X = \{U[p]X \star \phi_J \ , p\in\mathcal{P}_J\} \ .$$

This estimate is called mean-square consistent if its total variance over all paths converges:

$$\lim_{J\to\infty}\sum_{p\in\mathcal{P}_J}\mathbb{E}(|S_J[p]X-\overline{S}X(p)|^2)=0 \ .$$

Corollary 4.5 showed that mean-square consistency is equivalent to

$$\mathbb{E}(|X|^2)=\sum_{p\in\mathcal{P}_\infty}|\overline{S}X(p)|^2 \ ,$$

which in turn is equivalent to

$$\lim_{m\to\infty}\sum_{p\in\mathcal{P}_\infty \ , \ |p|=m}\mathbb{E}(|U[p]X|^2)=0 \ . \tag{31}$$

If a process $X(t)$ has a mean square consistent scattering, then one can recover the scaling law of its second moments with scattering coefficients:

Proposition 4.6. *Suppose that $X(t)$ is a process with stationary increments such that S_JX is mean square consistent. Then*

$$\mathbb{E}(|X\star\psi_j|^2)=\sum_{p\in\mathcal{P}_\infty}|\overline{S}X(j+p)|^2 \ . \tag{32}$$

For a large class of ergodic processes including most image textures, it is observed numerically that the total scattering variance $\sum_{p\in\mathcal{P}_J}\mathbb{E}(|S_J[p]X-\overline{S}X(p)|^2)$ decreases to zero when 2^J increases. Table 2 shows the decay of the total scattering variance, computed on average over the Brodatz texture dataset.

Corollary 4.5 showed that this variance decay then implies that

$$\|\overline{S}X\|^2=\sum_{m=0}^{\infty}\sum_{p\in\Lambda_\infty^m}|\overline{S}X(p)|^2=\mathbb{E}(|X|^2) \ .$$

Table 3 gives the percentage of expected scattering energy $\sum_{p\in\Lambda_\infty^m}|\overline{S}X(p)|^2$ carried by paths of length m, for textures in the Brodatz database. Most of the energy is concentrated in paths of length $m\leq 3$.

Table 3: Percentage of expected scattering energy $\sum_{p \in \Lambda_\infty^m} |\overline{S}X(p)|^2$, as a function of the scattering order m, computed with cubic spline wavelets, over the Brodatz dataset.

$m=0$	$m=1$	$m=2$	$m=3$	$m=4$
0	74	19	3	0.3

4.3 Multifractal Analysis with Scattering Moments

Many physical phenomena exhibit irregularities at all scales, as illustrated by the canonical example of turbulent flows or Brownian Motions. Fractals are mathematical models of stochastic processes that express such property through a scale-invariance symmetries of the form

$$\forall\ s > 0, \{X(st); t \in \mathbb{R}\} \overset{d}{=} A_s \cdot \{X(t); t \in \mathbb{R}\}\ . \tag{33}$$

In other words, the law of the stochastic process is invariant to time dilation, up to a scale factor. Here A_s denotes a random variable independent of X that controls the strength of irregularity of sample trajectories of $X(t)$. Fractional Brownian Motions are the only Gaussian Processes satisfying (33) with $A_s := s^H$, where $H = 0.5$ corresponds to the standard Wiener Process.

Fractals can be studied from wavelet coefficients through the distribution of point-wise Hölder exponents [PT02]. Moments of order q define a scaling exponent $\zeta(q)$ such that

$$\mathbb{E}[|X \star \psi_j|^q] \simeq 2^{j\zeta(q)}\ , (j \to -\infty)$$

This characteristic exponent provides rich information about the process, in particular the curvature of $\zeta(q)$ measures the presence of different Holder exponents within a realisation, and can be interpreted as a measure of *intermittency*. Intermittency is an ill-defined mathematical notion, which is used in physics to describe those irregular bursts of large amplitude variations, appearing for example in turbulent flows [YSO+11]. Multiscale intermittency appears in other domains such as network traffics, financial time series, geophysical and medical data.

Intermittency is created by heavy tail processes, such as Lévy processes. It produces large if not infinite polynomial moments of degree larger than two, and empirical estimations of second order moments have a large variance. These statistical instabilities can be reduced by calculating expected values of non-expansive operators in mean-square norm, which reduce the variance of empirical estimation. Scattering moments are computed with such a non-expansive operator.

In [BMB+15], it is shown that second-order scattering moments provide robust estimation of such intermittency through the following renormalisation scheme. In the univariate case, we consider for each $j, j_1, j_2 \in \mathbb{Z}$

$$\tilde{S}X(j) := \frac{\mathbb{E}[|X \star \psi_j|]}{\mathbb{E}[|X \star \psi_0|]}\ ,\ \tilde{S}X(j_1, j_2) = \frac{\mathbb{E}[||X \star \psi_{j_1}| \star \psi_{j_2}|]}{\mathbb{E}[|X \star \psi_{j_1}|]}\ . \tag{34}$$

This renormalised scattering can be estimated by plug-in of both numerator and denominator using the windowed scattering estimators (20). These renormalised scattering moments capture both self-similarity and intermittence, as illustrated by the following result.

Proposition 4.7 ([BMB+15], Proposition 3.1)**.** *Let $X(t)$ be a self-similar process (33) with stationary increments. Then for all $j_1 \in \mathbb{Z}$*

$$\tilde{S}X(j_1) = 2^{j_1 H}\ , \tag{35}$$

and for all $(j_1, j_2) \in \mathbb{Z}^2$

$$\tilde{S}X(j_1, j_2) = \overline{S}\tilde{X}(j_2 - j_1) \quad \textit{with} \quad \tilde{X}(t) = \frac{|X \star \psi(t)|}{\mathbb{E}(|X \star \psi|)}\ . \tag{36}$$

Moreover, the discrete curvature $\zeta(2) - 2\zeta(1)$ *satisfies*

$$2^{j(\zeta(2)-2\zeta(1))} \simeq \frac{\mathbb{E}(|X \star \psi_j|^2)}{\mathbb{E}(|X \star \psi_j|)^2} \geq 1 + \sum_{j_2=-\infty}^{+\infty} |\tilde{S}X(j, j_2)|^2 \ . \tag{37}$$

This proposition illustrates that second-order scattering coefficients $\tilde{S}X(j_1, j_2)$ of self-similar processes are only function of the difference $j_2 - j_1$, which can be interpreted as a stationarity property across scales. Moreover, it follows from (37) that if $\sum_{j_2=-\infty}^{+\infty} \tilde{S}X(j, j_2)^2 \simeq 2^{j\beta}$ as $j \to -\infty$ with $\beta < 0$, then $\zeta(2) - 2\zeta(1) < 0$. Therefore, the decay of $\tilde{S}X(j, j+l)$ with l (or absence thereof) captures a rough measure of intermittency. Figure 10 illustrates the behavior of normalised scattering coefficients for three representative processes, Poisson point processes, Fractional Brownian Motions and Mandelbrot Cascades. The asymptotic decay of scattering moments clearly distinguishes the different intermittent behavior. [BMB+15] explores the applications of such scattering moments to perform model selection in real-world applications, such as turbulent flows and finantial time-series.

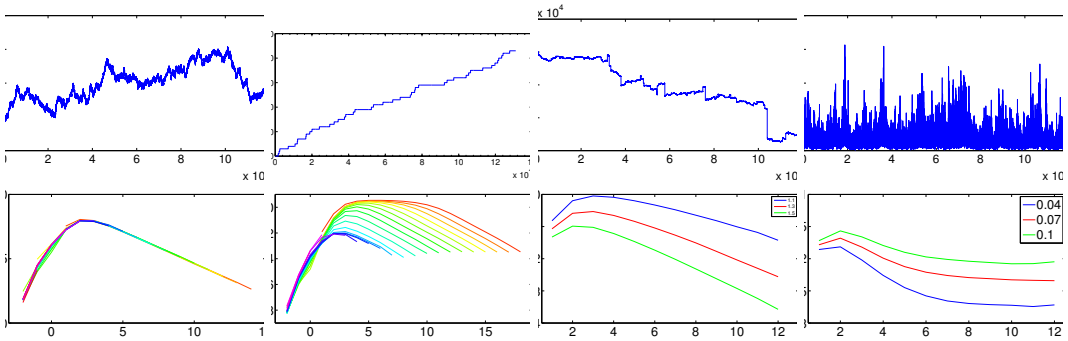


Figure 10: Top row: Realizations of a Brownian Motion, a Poisson point process, a Levy process and a Multifractal Random Cascade. Bottom Row: corresponding normalized second-order coefficietns.

5 Non-Euclidean Scattering

Scattering representations defined over the translation group are extended to other global transformation groups by defining Littlewood-Paley wavelet decompositions on non-Euclidean domains with group convolutions. Wavelet decompositions can also be defined on domains lacking global symmetries such as graphs and manifolds. In this section we present this formalism and discuss several applications.

5.1 Joint versus Separable Scattering

Let us consider the question of building a signal representation $\Phi(\mathbf{x})$ that is invariant to the action of a certain transformation group G acting on $\mathbf{L}^2(\mathbb{R}^d)$:

$$\begin{aligned} G \times \mathbf{L}^2(\mathbb{R}^d) &\to \mathbf{L}^2(\mathbb{R}^d) \\ (g, \mathbf{x}) &\mapsto \mathbf{x}_g \ . \end{aligned}$$

Φ is G-invariant if $\Phi(\mathbf{x}_g) = \Phi(\mathbf{x})$ for all $g \in G$, and G-equivariant if $\Phi(\mathbf{x}_g) = (\Phi(\mathbf{x}))_g$, that is, G acts on the image of Φ respecting the axioms of a group action.

Now, suppose that the group G admits a factorization as a *semidirect* product of two subgroups G_1, G_2:

$$G = G_1 \rtimes G_2 \ .$$

Figure 11: From [SM13]. The left and right textures are not discriminated by a separable invariant along rotations and translations, but can be discriminated by a joint invariant.

This means that G_1 is a normal subgroup of G and that each element $g \in G$ can be uniquely written as $g = g_1 g_2$, with $g_i \in G_i$. It is thus tempting to leverage group factorizations to build invariants to complex groups by combining simpler invariants and equivariants as building blocks.

Suppose Φ_1 is G_1-invariant and G_2-equivariant, and Φ_2 is G_2-invariant. Then $\bar{\Phi} := \Phi_2 \circ \Phi_1$ satisfies, for all $(g_1, g_2) \in G_1 \rtimes G_2$

$$\bar{\Phi}(\mathbf{x}_{g_1 g_2}) = \Phi_2((\Phi_1(\mathbf{x}))_{g_2}) = \Phi_2(\Phi_1(\mathbf{x})) = \bar{\Phi}(\mathbf{x}) \ ,$$

showing that we can effectively build larger invariants by composing simpler invariants and equivariants.

However, such compositional approach comes with a loss of discriminative power [SM13]. Indeed, whereas the group can be factorised into smaller groups, the group action that acts on the data is seldom separable, as illustrated in Figure 11. In the case of images $\mathbf{x} \in \mathbf{L}^2(\mathbb{R}^2)$, an important example comes from the action of general affine transformations of $\mathbb{R}^2$. This motivates the construction of joint scattering representations in the roto-translation group, discussed next.

5.2 Scattering on Global Symmetry Groups

We illustrate the ideas from Section 5.1 with the construction of a scattering representation over the roto-translation group for images, developed in [SM13, OM15], the Heisenberg group of frequency transpositions [AM14, ALM18], and SO(3) for quantum chemistry [HMP17, EEHM17]. In essence, these representation adapt the construction of Section 3 by defining appropriate wavelet decompositions over the roto-translation group.

Roto-Translation Group: The roto-translation group is formed by pairs $g = (v, \alpha) \in \mathbb{R}^2 \times \mathrm{SO}(2)$ acting on $u \in \Omega$ as follows:

$$(g, u) \mapsto g.u := v + R_\alpha u \ ,$$

where R_α is the rotation of the plane of angle α. One can easily verify that the set of all pairs (v, α) forms a group $G_{\mathrm{Rot}} \simeq \mathbb{R}^2 \rtimes \mathrm{SO}(2)$, with the multiplication defined as

$$(v_1, \alpha_1).(v_2, \alpha_2) = (v_1 + R_{\alpha_1} v_2, \alpha_1 + \alpha_2) \ .$$

The group acts on images $\mathbf{x}(u)$ by the usual composition: $\mathbf{x}_g := \mathbf{x}(g^{-1}.u)$.

Wavelet decompositions over a compact group are obtained from group convolutions, defined as weighted averages over the group. Specifically, if $\tilde{\mathbf{x}} \in L^2(G)$ and $h \in L^1(G)$, the *group convolution* of $\mathbf{x}$ with the filter h is

$$\tilde{\mathbf{x}} \star_G h(g) := \int_G \mathbf{x}_g h(g^{-1}) \mathrm{d}\mu(g) \ . \tag{38}$$

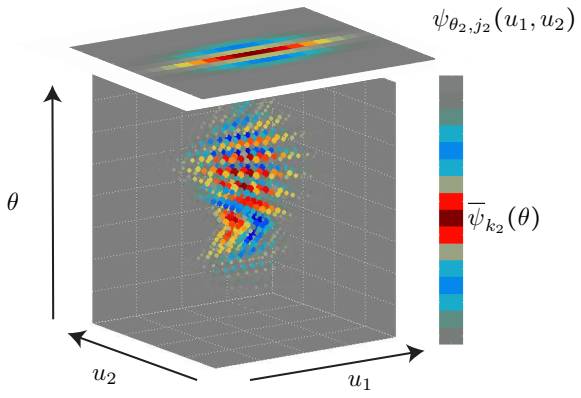


Figure 12: From [SM13]. A Wavelet defined on the Roto-Translation group, displayed in the 3D domain defined by positions u_1, u_2 and angles θ.

Here μ is the uniform Haar measure over G. One can immediately verify that group convolutions are the only linear operators which are equivariant with respect to the group action: $\tilde{\mathbf{x}}_{g'} \star_G h(g) = \tilde{\mathbf{x}} \star_G h((g')^{-1}.g)$ for all $g, g' \in G$.

Given an input $\mathbf{x}(u)$, $u \in \Omega \subset \mathbb{R}^2$, we consider first a wavelet decomposition over the translation group $W_1 = \{\psi_{j,\theta}\}_{\theta \in SO(2), j \in \mathbb{Z}}$, with dilations and rotations of a given mother wavelet. The corresponding propagated wavelet modulus coefficients become

$$U_1(\mathbf{x})(p_1) = |\mathbf{x} \star \psi_{j_1,\theta_1}|(u) \ , \ \text{with } p_1 := (u, j_1, \theta_1) \ .$$

This vector of coefficients is equivariant with respect to translations, since it is defined through spatial convolutions and point-wise nonlinearities. We verify that it is also equivariant with respect to rotations, since

$$U_1(r_\alpha \mathbf{x})(u, j_1, \theta_1) = U_1(\mathbf{x})(r_{-\alpha} u, j_1, \theta_1 - \alpha) \ .$$

In summary, the first layer U_1 is G_{Rot}-equivariant, $U_1(\mathbf{x}_g) = [U_1(\mathbf{x})]_g$, with group action on the coefficients $g.p_1 = (g.u, j_1, \theta_1 - \alpha)$, for $g = (v, \alpha) \in G_{\text{Rot}}$.

While the original Scattering operator from Section 3 would now propagate each sub-band of $U_1\mathbf{x}$ independently using the same wavelet decomposition operator, roto-translation scattering now considers a joint wavelet decomposition W_2 defined over functions of G_{Rot} in that case. Specifically, $W_2 = \{\Psi_\gamma\}_\gamma$ is a collection of wavelets defined in $L^1(G_{\text{Rot}})$. In [SM13, OM15] these wavelets are defined as separable products of spatial wavelets defined in $\Omega \subset \mathbb{R}^2$ with 1d wavelets defined in $SO(2)$. Figure 12 illustrates one such Ψ_γ.

Importantly, the geometric stability and energy conservation properties described in Sections 3.2 and 3.3 carry over the roto-translation scattering [Mal12, OM15]. As discussed earlier, addressing the invariants jointly or separately gives different discriminability tradeoffs. Some numerical applications greatly benefit from the joint representatation, in particular texture recognition under large point of view variability [SM13].

Time-Frequency Scattering: Joint scattering transforms also appear naturally in speech and audio processing, to leverage interactions of the signal energy at different time-frequency scales. Successful recognition of audio signals requires stability to small time-warps as well as frequency transpositions. Similarly as in the previous example, where the input $\mathbf{x}(u)$ was 'lifted' to a function over the roto-translation group with appropriate equivariance properties, in the case of audio signals this initial lifting is carried out by the so-called *scalogram*, which computes a Littlewood-Paley wavelet decomposition mapping a time-series $\mathbf{x}(t)$ to a two-dimensional function $\mathbf{z}(t, \lambda) = |\mathbf{x} \star \psi_\lambda(t)|$ [AM14].

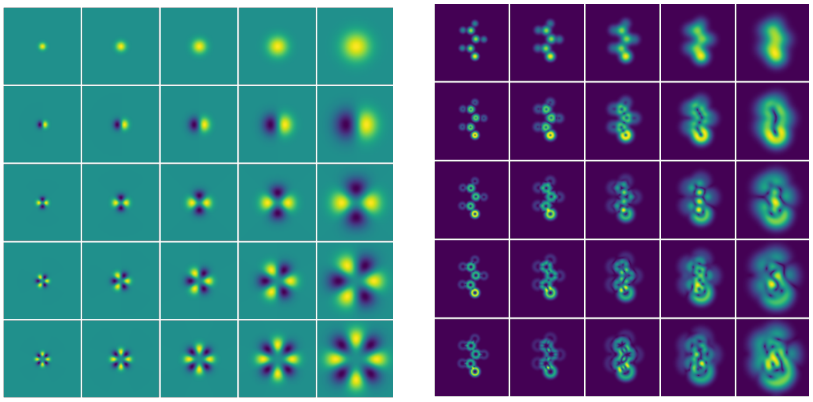


Figure 13: From [EEHM17]. Left: Real parts of 2D solid harmonic wavelets. Cartesian slices of 3D spherical harmonic wavelets yield similar patterns. Right: Solid harmonic wavelet moduli $S[j,l,1](\rho x)(u) = |\rho x \star \psi_j|(u)$ of a molecule ρx. The interference patterns at the different scales are reminiscent of molecular orbitals obtained in e.g. density functional theory.

The time-frequency interactions in $\mathbf{z}$ can be captured by a joint wavelet decomposition frame, leading to state-of-the-art classification and synthesis on several benchmarks [ALM18].

Solid Harmonic Scattering for Quantum Chemistry Building representations of physical systems with rotational and translational invariance and stability to deformations is of fundamental importance across many domains, since these symmetries are present in many physical systems. Specifically, [HMP17, EEHM17] study scattering representations for quantum chemistry, by considering a wavelet decomposition over SO(3). Such wavelet decomposition is constructed in the spectral domain, given by spherical harmonics. The resulting scattering representation enjoys provable roto-translational invariance and stability to small deformations, and leads to state-of-the-art performance in the regression of molecular energies [EEHM17]. Figure 13 illustrates the 'harmonic' wavelets as well as the resulting scattering coefficients for some molecules.

5.3 Graph Scattering

In Section 5.2 we described invariant representations of functions defined over a *fixed* domain with *global* symmetries. Despite being of fundamental importance in physics, global symmetries are lacking in many systems in other areas of science, such as networks, surface meshes, or proteins. In those areas, one is rather interested in local symmetries, and often the domain is variable, as well as the measurements over that domain.

5.3.1 Invariance and Stability in Graphs

In this context, graphs are flexible data structures that enable general metric structures and modeling non-Euclidean domains. The main ingredients of the scattering transform can be generalized using tools from computational harmonic analysis on graphs. As described in Section 2.3, the Euclidean treatment of deformations as changes of variables in the signal domain $\Omega \subset \mathbb{R}^d$, $u \mapsto \varphi_\tau(u) = u - \tau(u)$, can now be seen more generally as a change of metric, from an original metric domain $\mathcal{X}$ to a deformed metric domain $\mathcal{X}_\tau$.

We shall thus focus on deformations on the underlying graph domain, while keeping the same function mapping, i.e. we model deformations as a change of the underlying graph support and analyze how this affects the interaction between the function mapping and the graph. Similarly as

with the Group Scattering constructions of Section 5.2, defining scattering representations for graphs amounts to defining wavelet decompositions with appropriate equivariance and stability, and averaging operators providing the invariance.

Consider a weighted, undirected graph $G = (V, E, W)$ with $|V| = n$ nodes, edge set E and adjacency matrix $W \in \mathbb{R}^{n\times n}$, with $W_{i,j} > 0$ iff $(i,j) \in E$. In this context, the natural notion of invariance is given by permutations acting simultaneously on nodes and edges. Let us define $G_\pi = (\tilde{V}, \tilde{E}, \tilde{W})$ such that there exists a permutation $\pi \in S_n$ with $\tilde{v}_i = v_{\pi(i)}$, $(\tilde{i}, \tilde{j}) \in \tilde{E}$ iff $(\pi(i), \pi(j)) \in E$ and $\tilde{W} = \Pi W \Pi^\top$, where $\Pi \in \{0,1\}^{n\times n}$ is the permutation matrix associated with π. Many applications require a representation Φ such that $\Phi(\mathbf{x}; G) = \Phi(\mathbf{x}_\pi, G_\pi) = \Phi(\mathbf{x}, G)$ for all π.

Previously, Littlewood-Paley wavelets were designed as a non-expansive operator $\|\mathcal{W}\| \leq 1$ with small commutation error with respect to deformations: $\|[\mathcal{W}, \mathcal{L}_\tau]\| \lesssim \|\nabla\tau\|$. The first task is to quantify metric perturbations $\mathcal{X}_\tau$ induced by deforming the graph.

5.3.2 Diffusion Metric Distances

A weighted, undirected graph $G = (V, E, W)$ with $|V| = n$ nodes, edge set E and adjacency matrix $W \in \mathbb{R}^{n\times n}$ defines a diffusion process A on its nodes, given in its symmetric form by the normalized adjacency

$$\overline{W} := D^{-1/2} W D^{-1/2} \ , \text{ with } D = \mathrm{diag}(d_1, \ldots, d_n) \, , \tag{39}$$

where $d_i = \sum_{(i,j)\in E} W_{i,j}$ denotes the degree of node i. Denote by $\mathbf{d} = W\mathbf{1}$ the degree vector containing d_i in the i-th element. By construction, $\overline{W}$ is well-localized in space (it is nonzero only where there is an edge connecting nodes), it is self-adjoint and satisfies $\|\overline{W}\| \leq 1$, where $\|\overline{W}\|$ is the operator norm. It is convenient to assume that the spectrum of A (which is real and discrete since $\overline{W}$ is self-adjoint and in finite-dimensions) is non-negative. Since we shall be taking powers of $\overline{W}$, this will avoid folding negative eigenvalues into positive ones. For that purpose, we adopt the so-called *lazy diffusion*, given by $T = \frac{1}{2}(I + \overline{W})$. We will use this diffusion operator to define both a multiscale wavelet filter bank and a low-pass average pooling, leading to the diffusion scattering representation.

This diffusion operator can be used to construct a metric on G. The so-called *diffusion maps* [CL06, NLCK06] measure distances between two nodes $x, x' \in V$ in terms of their associated diffusion at time s: $d_{G,s}(x, x') = \|T_G^s \delta_x - T_G^s \delta_{x'}\|$, where δ_x is a vector with all zeros except a 1 in position x. This diffusion metric can be now used to define a distance between two graphs G, G'. Assuming first that G and G' have the same size, the simplest formulation is to compare the diffusion metric generated by G and G' up to a node permutation:

Definition 5.1. *Let $G = (V, E, W)$, $G' = (V', E', W')$ have the same size $|V| = |V'| = n$. The normalized diffusion distance between graphs G, G' at time $s > 0$ is*

$$\mathrm{d}^s(G, G') := \inf_{\Pi\in\Pi_n} \|(T_G^s)^*(T_G^s) - \Pi^\top (T_{G'}^s)^*(T_{G'}^s)\Pi\| = \inf_{\Pi\in\Pi_n} \|T_G^{2s} - \Pi^\top T_{G'}^{2s}\Pi\| \ , \tag{40}$$

where Π_n is the space of $n \times n$ permutation matrices.

The diffusion distance is defined at a specific time s. As s increases, this distance becomes weaker[3], since it compares points at later stages of diffusion. The role of time is thus to select the smoothness of the 'graph deformation', similarly as $\|\nabla\tau\|$ measures the smoothness of the deformation in the Euclidean case. For convenience, we denote $\mathrm{d}(G, G') = \mathrm{d}^{1/2}(G, G')$ and use the distance at $s = 1/2$ as our main deformation measure. The quantity d defines a distance between graphs (seen as metric spaces) yielding a stronger topology than other alternatives such as the Gromov-Hausdorff distance, defined as

$$d^s_{\mathrm{GH}}(G, G') = \inf_{\Pi} \sup_{x,x'\in V} |d^s_G(x, x') - d^s_{G'}(\pi(x), \pi(x'))|$$

[3]In the sense that it defines a weaker topology, i.e., $\lim_{m\to\infty} \mathrm{d}^s(G, G_m) \to 0 \Rightarrow \lim_{m\to\infty} \mathrm{d}^{s'}(G, G_m) = 0$ for $s' > s$, but not vice-versa.

with $d_G^s(x,x') = \|T_G^t(\boldsymbol{\delta}_x - \boldsymbol{\delta}_{x'})\|_{L^2(G)}$. Finally, we consider for simplicity only the case where the sizes of G and G' are equal, but definition (5.1) can be naturally extended to compare variable-sized graphs by replacing permutations by soft-correspondences [see BBK$^+$10].

Our goal is to build a stable and rich representation $\Phi_G(\mathbf{x})$. The stability property is stated in terms of the diffusion metric above: For a chosen diffusion time s, $\forall\ \mathbf{x} \in \mathbb{R}^n$, $G = (V, E, W), G' = (V', E', W')$ with $|V| = |V'| = n$, we want

$$\|\Phi_G(\mathbf{x}) - \Phi_{G'}(\mathbf{x})\| \lesssim \|\mathbf{x}\| \mathrm{d}(G, G') \ . \tag{41}$$

This representation can be used to model both signals and domains, or just domains G, by considering a prespecified $\mathbf{x} = f(G)$, such as the degree, or by marginalizing from an exchangeable distribution, $\Phi_G = \mathbb{E}_{\mathbf{x}\sim Q}\Phi_G(\mathbf{x})$.

The motivation of (41) is two-fold: On the one hand, we are interested in applications where the signal of interest may be measured in dynamic environments that modify the domain, e.g. in measuring brain signals across different individuals. On the other hand, in other applications, such as building generative models for graphs, we may be interested in representing the domain G itself. A representation from the adjacency matrix of G needs to build invariance to node permutations, while capturing enough discriminative information to separate different graphs. In particular, and similarly as with Gromov-Hausdorff distances, the definition of $\mathrm{d}(G, G')$ involves a matching problem between two kernel matrices, which defines an NP-hard combinatorial problem. This further motivates the need for efficient representations of graphs Φ_G that can efficiently tell apart two graphs, and such that $\ell(\theta) = \|\Phi_G - \Phi_{G(\theta)}\|$ can be used as a differentiable loss for training generative models.

5.3.3 Diffusion Wavelets

Diffusion wavelets [CL06] provide a simple framework to define a multi-resolution analysis from powers of a diffusion operator defined on a graph, and they are stable to diffusion metric changes.

Let $\lambda_0 \geq \lambda_1 \geq \ldots \lambda_{n-1}$ denote the eigenvalues of A in decreasing order. Defining $\mathbf{d}^{1/2} = (\sqrt{d_1}, \ldots, \sqrt{d_n})$, one can easily verify that the normalized squared root degree vector $\boldsymbol{v} = \mathbf{d}^{1/2}/\|\mathbf{d}^{1/2}\|_2 = \mathbf{d}/\|\mathbf{d}\|_1$ is the eigenvector with associated eigenvalue $\lambda_0 = 1$. Also, note that $\lambda_{n-1} = -1$ if and only if G has a connected component that is non-trivial and bipartite [CG97].

Following [CL06], we construct a family of multiscale filters by exploiting the powers of the diffusion operator T^{2^j}. We define

$$\psi_0 := I - T \,,\ \psi_j := T^{2^{j-1}}(I - T^{2^{j-1}}) = T^{2^{j-1}} - T^{2^j} \ , (j > 0) \,. \tag{42}$$

This corresponds to a graph wavelet filter bank with optimal spatial localization. Graph diffusion wavelets are localized both in space and frequency, and favor a spatial localization, since they can be obtained with only two *filter coefficients*, namely $h_0 = 1$ for diffusion $T^{2^{j-1}}$ and $h_1 = -1$ for diffusion T^{2^j}. The finest scale ψ_0 corresponds to one half of the normalized Laplacian operator $\psi_0 = (1/2)\Delta = 1/2(I - D^{-1/2}WD^{-1/2})$, here seen as a temporal difference in a diffusion process, seeing each diffusion step (each multiplication by Δ) as a time step. The coarser scales ψ_j capture temporal differences at increasingly spaced diffusion times. For $j = 0, \ldots, J_n - 1$, we consider the linear operator

$$\begin{aligned} \mathcal{W} : L^2(G) &\to (L^2(G))^{J_n} \\ \mathbf{x} &\mapsto (\psi_j \mathbf{x})_{j=0,\ldots,J_n-1} \ , \end{aligned} \tag{43}$$

which is the analog of the wavelet filter bank in the Euclidean domain. Whereas several other options exist to define graph wavelet decompositions [RG13, GNC10], we consider here wavelets that can be expressed with few diffusion terms, favoring spatial over frequential localization, for stability reasons that will become apparent next. We choose dyadic scales for convenience, but the construction is analogous if one replaces scales 2^j by $\lceil \gamma^j \rceil$ for any $\gamma > 1$ in (42). If the graph G exhibits a *spectral gap*, i.e., $\beta_G = \sup_{i=1,\ldots n-1} |\lambda_i| < 1$, the linear operator $\mathcal{W}$ defines a stable frame.

Proposition 5.2 ([GRB18], Prop 4.1)**.** *For each n, let $\mathcal{W}$ define the diffusion wavelet decomposition (43) and assume $\beta_G < 1$. Then there exists a constant $0 < C(\beta)$ depending only on β such that for any $\mathbf{x} \in \mathbb{R}^n$ satisfying $\langle \mathbf{x}, \boldsymbol{v} \rangle = 0$,*

$$C(\beta)\|\mathbf{x}\|^2 \leq \sum_{j=0}^{J_n-1} \|\psi_j \mathbf{x}\|^2 \leq \|\mathbf{x}\|^2 \ . \tag{44}$$

This proposition thus provides the Littlewood-Paley bounds of $\mathcal{W}$, which control the ability of the filter bank to capture and amplify the signal $\mathbf{x}$ along each 'frequency'. We note that diffusion wavelets are neither unitary nor analytic and therefore do not preserve energy. However, the frame bounds in Proposition 5.2 provide lower bounds on the energy lost. They also inform us about how the spectral gap β determines the appropriate diffusion *scale* J: The maximum of $p(u) = (u^r - u^{2r})^2$ is at $u = 2^{-1/r}$, thus the cutoff r_* should align with β as $r_* = \frac{-1}{\log_2 \beta}$, since larger values of r capture energy in a spectral range where the graph has no information. Therefore, the maximum scale can be adjusted as $J = \lceil 1 + \log_2 r_* \rceil = 1 + \left\lceil \log_2 \left(\frac{-1}{\log_2 \beta} \right) \right\rceil$.

5.3.4 Diffusion Scattering

Recall that the Euclidean Scattering transform is constructed by cascading three building blocks: a wavelet decomposition operator, a pointwise modulus activation function, and an averaging operator. Following the Euclidean scattering, given a graph G and $\mathbf{x} \in L^2(G)$, we define an analogous Diffusion Scattering transform $S_G(\mathbf{x})$ by cascading three building blocks: the Wavelet decomposition operator $\mathcal{W}$, a pointwise activation function ρ, and an average operator A which extracts the average over the domain. The average over a domain can be interpreted as the diffusion at infinite time, thus $A\mathbf{x} = \lim_{t\to\infty} T^t \mathbf{x} = \langle \mathbf{v}, \mathbf{x} \rangle$. More specifically, we consider a first layer transformation given by

$$S_G[\Lambda_1]\mathbf{x}) = A\rho\mathcal{W}\mathbf{x} = \{A\rho\psi_j \mathbf{x}\}_{0 \leq j \leq J_n - 1} \ , , \tag{45}$$

followed by second order coefficients

$$S_G[\Lambda_2]\mathbf{x}) = A\rho\mathcal{W}\rho\mathcal{W}\mathbf{x} = \{A\rho\psi_{j_2}\rho\psi_{j_1}\mathbf{x}\}_{0 \leq j_1, j_2 \leq J_n - 1} \ , , \tag{46}$$

and so on. The representation obtained from m layers of such transformation is thus

$$S_{G,m}(\mathbf{x}) = \{A\mathbf{x}, S_G[\Lambda_1](\mathbf{x}), \ldots, S_G[\Lambda_m](\mathbf{x})\} = \{A(\rho\mathcal{W})^k \mathbf{x} \, ; \, k = 0, \ldots, m-1\} \ . \tag{47}$$

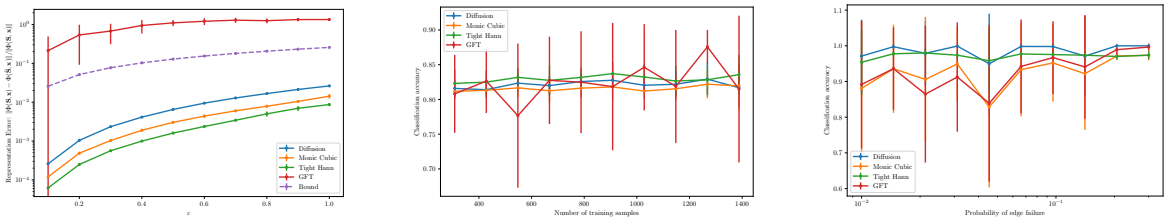
5.3.5 Stability and Invariance of Diffusion Scattering

The scattering transform coefficients $S_G(\mathbf{x})$ obtained after m layers are given by (47), for low-pass operator A such that $A\mathbf{x} = \langle \boldsymbol{v}, \mathbf{x} \rangle$.

The stability of diffusion wavelets with respect to small changes of the diffusion metric can be leveraged to obtain a resulting diffusion scattering representation with prescribed stability, as shown by the following Theorem.

Theorem 5.3 ([GRB18], Theorem 5.3)**.** *Let G, G' be two graphs and let $\mathrm{d}(G, G')$ be their distance measured as in (40). Let $\beta_- = \min(\beta_G, \beta_{G'})$, $\beta_+ = \max(\beta_G, \beta_{G'})$ and assume $\beta_+ < 1$. Then, we have that, for each $k = 0, \ldots, m-1$, the following holds*

$$\begin{aligned} \|S_{G,m}(\mathbf{x}) - S_{G',m}(\mathbf{x})\|^2 &\leq \sum_{k=0}^{m-1} \left[\left(\frac{2}{1-\beta_-} \mathrm{d}(G, G') \right)^{1/2} + k \sqrt{\frac{\beta_+^2(1+\beta_+^2)}{(1-\beta_+^2)^3}} \mathrm{d}(G, G') \right]^2 \|\mathbf{x}\|^2 \\ &\lesssim m \, \mathrm{d}(G, G') \|\mathbf{x}\|^2 \qquad \textit{if } \mathrm{d}(G, G') \ll 1. \end{aligned} \tag{48}$$



(a) Small world (b) Authorship attribution (c) Facebook graph

Figure 14: From [GBR19]. a: Difference in representation between the signal defined using the original graph scattering S_G and $S_{G'}$ corresponding to the deformed graph as a function of the perturbation size $\mathrm{d}(G, G')$. b-c Classification accuracy as a function of perturbation for the authorship attribution and the Facebook graph, respectively.

This result shows that the closer the graphs are in terms of the diffusion metric, the closer their scattering representations will be. The constant is given by topological properties, the spectral gaps of G and G', as well as design parameters, the number of layers m. We observe that the stability bound grows the smaller the spectral gap is and also as more layers are considered. The spectral gap is tightly linked with diffusion processes on graphs, and thus it does emerge from the choice of a diffusion metric. Graphs with values of β closer to 1 exhibit weaker diffusion paths, and thus a small perturbation on the edges of these graphs would lead to a larger diffusion distance. We also note that the spectral gap appears in our upper bounds, but it is not necessarily sharp. In particular, the spectral gap is a poor indication of stability in regular graphs, and we believe our bound can be improved by leveraging structural properties of regular domains.

Finally, we note that the size of the graphs impacts the stability result inasmuch as it impacts the distance measure $\mathrm{d}(G, G')$. A similar scattering construction was developed in [ZL18], where the authors established stability with respect to a graph measure that depends on the spectrum of the graph through both eigenvectors and eigenvalues. More specifically, it is required that the spectrum gets concentrated as the graphs grow. However, in general, it is not straightforward to relate the topological structure of the graph with its spectral properties.

As mentioned above, the stability is computed with a metric $\mathrm{d}(G, G')$ which is stronger than what could be hoped for. This metric is permutation-invariant, in analogy with the rigid translation invariance in the Euclidean case, and stable to small perturbations around permutations. Recently, [GBR19] extended the previous stability analysis to more general wavelet decompositions and using a *relative* notion of deformation. Figure 14 illustrates the performance of Graph Scattering operators on several graph signal processing tasks. Also, [GWH19] developed a similar scattering representation for graphs, achieving state-of-the-art results on several graph classification tasks. The extension of (48) to weaker metrics, using e.g. multiscale deformations, is an important open question.

5.3.6 Unsupervised Haar Scattering on Graphs

A particularly simple wavelet representation on graphs – that avoids any spectral decomposition – is given by Haar wavelets [GNC10]. Such wavelets were used in the first work that extended scattering representations to graphs in [CCM14]. Given an undirected graph $G = (V, E)$, an *orthogonal Haar Scattering* transform is obtained from a multiresolution approximation of G. Let $G_0 := G$. In the dyadic case $|V| = 2^J$, it is defined as a hierarchical partition $\{V_{j,n}\}_{j,n}$ of G of the form

$$V_{0,i} = \{i\}, i \leq 2^J \; V_{j+1,i} = V_{j,a_i} \sqcup V_{j,b_i} \;, i = 1 \leq 2^{J-j-1} \;,$$

where the pairings (a_i, b_i) are connected in the induced subsampled graph $G_j = (V_j, E_j)$ of size $|V_j| = 2^{J-j}$, whose vertices are precisely $V_j := \{V_{j,i}\}_i$ and its edges are inherited recursively from G_{j-1}: $(i, i') \in E_j$ iff there exists $\bar{e} = (\bar{i}, \bar{i}') \in E_{j-1}$ with $\bar{i} \in V_{j,i}$ and $\bar{i}' \in V_{j,i'}$.

Let $\mathbf{x} \in \mathrm{l}^2(G)$. By rearranging the pairings sequentially, the resulting orthogonal Haar Scattering representation $S_J\mathbf{x}$ is defined recursively as

$$\begin{aligned} S_0\mathbf{x}(i,0) &:= \mathbf{x}(i)\,,\, i = 1 \dots 2^J \\ S_{j+1}\mathbf{x}(i,2q) &:= S_j\mathbf{x}(a_i,q) + S_j\mathbf{x}(b_i,q), \\ S_{j+1}\mathbf{x}(i,2q+1) &:= |S_j(a_i,q) - S_j(b_i,q)|\,, i = 1 \dots 2^{J-j-1},\, q = 2^j. \end{aligned} \tag{49}$$

One easily verifies [CCM14, CCM16] that the resulting transformation preserves the number of coefficients (equal to 2^J), and is contractive and unitary up to a normalization factor $2^{J/2}$. However, since the multiresolution approximation defines an orthogonal transformation, the resulting orthogonal scatering coefficients are not permutation invariant. In order to recover an invariant representation, it is thus necessary to average an ensemble of orthogonal transforms using different multiresolution approximations. Nevertheless, the main motivation in [CCM14, CCM16] was to perform graph scattering on domains with unknown (but presumed) graph connectivity structure. In that case, the sparsity of scattering coefficients was used as a criteria to find the optimal multiresolution approximation, resulting in state-of-the-art performance on several graph classification datasets.

5.4 Manifold Scattering

In the previous sections we have seen some instances of extending scattering representations to non-Euclidean domains, including compact Lie Groups and graphs. Such extensions (which in fact also apply to the wider class of Convolutional Neural Network architectures; see [BBL$^+$16] for an in-depth review) can be understood from the lens of the spectrum of differential operators, in particular the Laplacian. Indeed, the Laplacian operator encapsulates the symmetries and stability requirements that we have been manipulating so far, and can be defined across many different domains.

In particular, if $\mathcal{M}$ denotes a compact, smooth Riemannian manifold without boundary, one can define the Laplace-Beltrami operator Δ in $\mathcal{M}$ as the divergence of the manifold gradient. In these conditions $-\Delta$ is self-adjoint and positive semi-definite, therefore its eigenvectors define an orthonormal basis of $L^2(\mathcal{M}, \mu)$, where μ is the uniform measure on $\mathcal{M}$. Expressing any $f \in L^2(\mathcal{M})$ in this basis amounts to computing a 'Fourier transform' on $\mathcal{M}$. Indeed, the Laplacian operator in $\mathbb{R}^d$ is precisely diagonal in the standard Euclidean Fourier basis. Convolutions in the Euclidean case can be seen as linear operators that diagonalise in the Fourier basis, or equivalently that commute with the Laplacian operator. A natural generalisation of convolutions to non-Euclidean domains $\mathcal{M}$ is thus to formally see them as linear operators that commute with the Laplacian defined in $\mathcal{M}$ [BZSL13, BBL$^+$16]. Specifically, if $\{\varphi_k\}_k$ are the eigenvectors of Δ and $\Lambda := \{\lambda_k\}_k$ its eigenvalues, a function of *spectral multipliers* $\eta : \Lambda \to \mathbb{R}$ defines a kernel in $\mathcal{M}$:

$$K_\eta(u,v) = \sum_k \eta(\lambda_k)\varphi_k(u)\varphi_k(v)\ ,u,v \in \mathcal{M}\ ,$$

and a 'convolution' from its corresponding integral operator:

$$\begin{aligned} L^2(\mathcal{M}) &\to L^2(\mathcal{M}) \\ \mathbf{x} &\mapsto (T_\eta\mathbf{x})(u) = \int K_\eta(u,v)\mathbf{x}(v)\mu(\mathrm{d}v). \end{aligned} \tag{50}$$

In [PWH18], the authors use this formalism to build scattering representations on Riemannian manifolds, by defining Littlewood-Paley wavelet decompositions from appropriately chosen spectral multipliers $(\eta_j)_j$. The resulting scattering representation is shown to be stable to additive noise and to smooth diffeomorphisms of $\mathcal{M}$.

6 Generative Modeling with Scattering

In this Section we discuss applications of scattering representation to build high-dimensional generative models. Data priors defined from the scattering representation enjoy geometric stability and may be used as models for stationary processes or to regularize ill-posed inverse problems.

6.1 Scattering Sufficient Statistics

Defining probability distributions of signals $\mathbf{x}(u) \in \mathbf{L}^2(\mathbb{R}^d)$ is a challenging task due to the curse of dimensionality and the lack of tractable analytic models of "real" data. A powerful framework to approach this challenge is to rely on the principle of maximum entropy: construct probability models that are maximally regular while satisfying a number of constraints given by a vector $\Phi(\mathbf{x}) \in \mathbb{R}^K$ of sufficient statistics that is fit to the available data. When $\Phi(\mathbf{x}) = \mathbf{x}\mathbf{x}^\top$ consists in covariance measurements, the resulting maximum entropy model is a Gaussian Process, and when $\Phi(\mathbf{x})$ computes local potentials one obtains Markov Random Fields instead. In either case, one is quickly confronted with fundamental challenges, either statistical (exponential sample complexity for powerful statistical models, or large bias in small parametric ones) or computational, coming from the intractability of computing partition functions and sampling in high-dimensions.

Sufficient statistics in a maximum-entropy model capture our prior information about "what matters" in the input data. In this Section, we shall explore maximum entropy models where sufficient statistics are given by scattering representations. Depending on the localization scale 2^J, two distinct regimes emerge. For fixed and relatively small scale, windowed scattering representations provide local statistics that are nearly invertible, and help regularize ill-posed inverse problems (Section 6.4). As $J \to \infty$, expected scattering moments may be used to define models for stationary processes (Section 6.5).

Thanks to the scattering mean-squared consistency discussed in Section 4, we can circumvent the aforementioned challenges of maximum entropy models with the so-called *microcanonical* models from statistical physics, described in Section 6.2. In both regimes an important algorithmic component will be to solve a problem of the form $\min_{\mathbf{x}} \|S(\mathbf{x}) - y\|$. We first discuss a gradient descent strategy for that purpose in Section 6.3.

6.2 Microcanonical Scattering Models

Suppose first that we wish to characterize a probability distribution μ over input signals $\mathbf{x} \in \mathcal{D} = \mathbf{L}^2(\mathbb{R}^d)$, from the knowledge that $S_J(\mathbf{x}) \approx y$. In this setup, one could think of y as being an empirical average

$$y = \frac{1}{n}\sum_{i=1}^{n} S_J(\mathbf{x}_i),$$

where $\mathbf{x}_i$ are training samples that are conditionally independent and identically distributed.

Recall that the differential entropy of a probability distribution μ which admits a density $p(\mathbf{x})$ relatively to the Lebesgue measure is

$$H(\mu) := -\int p(\mathbf{x}) \, \log p(\mathbf{x}) \, d\mathbf{x} \, . \tag{51}$$

In absence of any other source of information, the classic macrocanonical model from Boltzmann and Gibbs μ^{ma} has density p_{ma} with maximum entropy, conditioned on $\mathbb{E}_{p_{\mathrm{ma}}}(S_J(\mathbf{x})) = y$. Instead, a microcanonical model replaces the expectation constraint with an empirical constraint of the form $\|S_J(\mathbf{x}) - y\| \leq \epsilon$ for small, appropriately chosen ϵ .

Despite being in appearance similar models, microcanonical and macrocanonical models have profound differences. On the one hand, under appropriate conditions, macrocanonical models may be

expressed as Gibbs distributions of the form

$$p_{\mathrm{ma}}(\mathbf{x}) = \frac{e^{\langle \theta, S_J(\mathbf{x}) \rangle}}{Z_\theta} \ ,$$

where Z_θ is the normalizing constant or partition function, and θ is a vector of Lagrange multipliers enforcing the expectation constraint. Unfortunately, this vector has no closed form expression in terms of estimable quantities in general, and needs to be adjusted using MCMC [WJ+08]. On the other hand, microcanonical models have compact support. However, under mild ergodicity assumptions on the underlying data-generating process, one can show that both models become asymptotically equivalent via the Boltzmann equivalence principle [DZ93] as $J \to \infty$, although microcanonical models may exist even when their macrocanonical equivalents do not [BM18, Cha17]. Also, estimating microcanonical models does not require the costly estimation of Lagrange multipliers.

The microcanonical set of width ϵ associated to y is

$$\Omega_{J,\epsilon} = \{x \in \mathcal{D} \ : \ \|S_J(\mathbf{x}) - y\| \leq \epsilon\} \ .$$

A maximum entropy microcanonical model $\mu^{\mathrm{mi}}(J, \epsilon, y)$ was defined by Boltzmann as the maximum entropy distribution supported in $\Omega_{J,\epsilon}$. If we assume the conditions that guarantee that S_J preserves energy (Section 3.2), then one can verify that $\Omega_{J,\epsilon}$ is a compact set. It follows that the maximum entropy distribution has a uniform density $p_{d,\epsilon}$:

$$p_{d,\epsilon}(x) := \frac{1_{\Omega_{J,\epsilon}}(x)}{\int_{\Omega_{J,\epsilon}} dx} \ . \tag{52}$$

Its entropy is therefore the logarithm of the volume of $\Omega_{J,\epsilon}$:

$$H(p_{d,\epsilon}) = -\int p_{d,\epsilon}(x) \, \log p_{d,\epsilon}(x) \, dx = \log \Big(\int_{\Omega_{J,\epsilon}} dx \Big) \ . \tag{53}$$

The scale J plays an important tradeoff in this model, as illustrated in the case where $y = S_J(\bar{\mathbf{x}})$ are measurements coming from a single realisation. When J is small, as explained in Section 3.4.2, the number of scattering coefficients is larger than the input dimension, and thus one may expect $\Omega_{J,\epsilon}$ to converge to a single point $\bar{\mathbf{x}}$ as $\epsilon \to 0$. As J increases, the system of equations $S_J(\mathbf{x}) = S_J(\bar{\mathbf{x}})$ becomes under-constrained, and thus $\Omega_{J,\epsilon}$ will be a non-singular set. Figure 8 illustrates this fact on a collection of input images. The entropy of the microcanonical model thus grows with J. It is proved in [BM18] that under mild assumptions the entropy is an *extensive* quantity, meaning that its growth is of the same order as 2^J, the support of the representation.

The appropriate scale J needs to balance two opposing effects: On the one hand, we want S_J to satisfy a concentration property to ensure that typical samples from the unknown data distribution μ are included in $\Omega_{J,\epsilon}$ with high probability, and hence typical for the microcanonical measure μ^{mi}. On the other hand, the sets $\Omega_{J,\epsilon}$ must not be too large to avoid having elements of $\Omega_{J,\epsilon}$ — and hence typical samples of μ^{mi}— which are not typical for μ. To obtain an accurate microcanonical model, the scale J must define microcanonical sets of minimum volume, while satisfying the concentration (**??**). In particular, this implies that the only data distributions that admit a valid microcanonical model as J increases need to be ergodic, stationary textures, where spatial averages converge to the expectation. [BM18] developed microcanonical models built from scattering represenations, showing their ability to model complex stationary phenomena such as Ising models, point processes or natural textures with tractable sample complexity. We illustrate scattering microcanonical models for such textures in Section 6.5. In essence, these models need to approximately sample from the uniform measure of sets of the form $\{\mathbf{x}; \|S(\mathbf{x}) - y\| \leq \epsilon\}$. We describe next how to efficiently solve this using gradient descent.

6.3 Gradient Descent Scattering Reconstruction

Computing samples of a maximum entropy microcanonical model is typically done with MCMC algorithms or Langevin Dynamics [Cre83], which is computationally expensive. Microcanonical models computed with alternative projections and gradient descents have been implemented to sample texture synthesis models [HB95, PS00, GEB15].

We consider microcanonical gradient descent models obtained by transporting an initial measure towards a microcanonical set, using gradient descent with respect to the distance to the microcanincal ensemble. Although this gradient descent sampling algorithm does not in general correspond to the maximum entropy microcanonical model, it preserves many symmetries of the maximum entropy microcanonical measure, and is shown to converge to the microcanonical set for appropriate choices of energy vector [BM18].

We transport an initial measure μ_0 towards a measure supported in a microcanonical set $\Omega_{J,\epsilon}$, by iteratively minimising

$$E(\mathbf{x}) = \frac{1}{2}\|S_J(\mathbf{x}) - y\|^2 \tag{54}$$

with mappings of the form

$$\varphi_n(\mathbf{x}) = \mathbf{x} - \kappa_n \nabla E(\mathbf{x}) = \mathbf{x} - \kappa_n \partial S_J(\mathbf{x})^t op(S_J(\mathbf{x}) - y) \ , \tag{55}$$

where κ_n is the gradient step at each iteration n.

Given an initial measure μ_0, the measure update is

$$\mu_{n+1} := \varphi_{n,\#}\mu_n, \tag{56}$$

with the standard pushforward measure $f_\#(\mu)[\mathcal{A}] = \mu[f^{-1}(\mathcal{A})]$ for any μ-measurable set $\mathcal{A}$, where $f^{-1}(\mathcal{A}) = \{x; f(x) \in \mathcal{A}\}$.

Samples from μ_n are thus obtained by transforming samples $\mathbf{x}_0$ from μ_0 with the mapping $\bar{\varphi} = \varphi_n \circ \varphi_{n-1} \cdots \circ \varphi_1$. It corresponds to n steps of a gradient descent initialized with $\mathbf{x}_0 \sim \mu_0$:

$$\mathbf{x}_{l+1} = \mathbf{x}_l - \kappa_l \partial S_J(\mathbf{x}_l)^t op(S_J(\mathbf{x}_l) - y) \ .$$

[BM18] studies the convergence of the gradient descent measures μ_n for general choices of sufficient statistics inculding scattering vectors. Even if they converge to a measure supported in a microcanonical set $\Omega_{J,\epsilon}$, in general they do not converge to a maximum entropy measure on this set. However, the next theorem proves that if μ_0 is a Gaussian measure of i.i.d Gaussian random variables then they have a large class of common symmetries with the maximum entropy measure. Let us recall that a symmetry of a measure μ is a linear invertible operator L such that for any measurable set $\mathcal{A}$, $\mu[L^{-1}(\mathcal{A})] = \mu[\mathcal{A}]$. A linear invertible operator L is a symmetry of Φ_d if for all $\mathbf{x} \in \mathcal{D}$, $S_J(L^{-1}\mathbf{x}) = S_J(\mathbf{x})$. It preserves volumes if its determinant satisfies $|\det L| = 1$. It is orthogonal if $L^tL = LL^t = I$ and we say that it preserves a stationary mean if $L\mathbf{1} = \mathbf{1}$ for $\mathbf{1} = (1, ..., 1) \in \mathbb{R}^\ell$.

Theorem 6.1 ([BM18], Theorem 3.4)**.** *(i) If L is a symmetry of S_J which preserves volumes then it is a symmetry of the maximum entropy microcanonical measure.*
(ii) If L is a symmetry of S_J and of μ_0 then it is a symmetry of μ_n for any $n \geq 0$.
(iii) Suppose that μ_0 is a Gaussian white noise measure of d i.i.d Gaussian random variables. Then, if L is a symmetry of Φ_d which is orthogonal and preserves a stationary mean then it is a symmetry of μ_n for any $n \geq 0$.

The initial measure μ_0 is chosen so that it has many symmetries in common with Φ_d and hence the gradient descent measures have many symmetries in common with a maximum entropy measure. A Gaussian measure of i.i.d Gaussian variables of mean m_0 and σ_0 is a maximum entropy measure conditioned by a stationary mean and variance. It is uniform over spheres which guarantees that it has a large group of symmetries.

Observe that periodic shifts are linear orthogonal operators and preserve a stationary mean. The following corollary applies property (iii) of Theorem 6.1 to prove that μ_n are circular-stationary.

Corollary 6.2 ([BM18], Corollary 3.5)**.** *When $J \to \infty$ then S_J is invariant to periodic shift. Therefore if μ_0 is a Gaussian white noise then μ_n is circular-stationary for $n \geq 0$.*

6.4 Regularising Inverse Problems with Scattering

Ill-posed inverse problems attempt to estimate an unknown signal $\mathbf{x}$ from noisy, possibly non-linear and under-determined measurements $\mathbf{y} = \mathcal{G}\mathbf{x} + w$, where w models additive noise. A natural Bayesian perspective is to consider the maximum-a-posteriori (MAP) estimate, given by

$$\hat{\mathbf{x}} \in \arg\max p(\mathbf{x}|\mathbf{y}) = \arg\max p(\mathbf{x}) \cdot p(\mathbf{y}|\mathbf{x}) = \arg\max \log p(\mathbf{x}) + \log p(\mathbf{y}|\mathbf{x}) \ .$$

Under a Gaussian noise assumption, $-\log p(\mathbf{y}|\mathbf{x})$ takes the familiar form $C\|\mathbf{y} - \mathcal{G}\mathbf{x}\|^2$. Regularising inverse problems using microcanonical scattering generative models thus amounts to choosing a prior $\log p(\mathbf{x})$ of the form

$$\|S_J(\mathbf{x}) - \mathbf{z}\|^2 \ ,$$

where $\mathbf{z}$ can be adjusted using a training set.

If μ denotes the underlying data-generating distribution of signals $\mathbf{x}$, such a prior implicitly assumes that scattering coefficients $S_J(\mathbf{x}), \mathbf{x} \sim \mu$ concentrate. In some applications, however, μ may not enjoy such ergodocity properties, in which case one can also consider a microcanonical 'amortised' prior that is allowed to depend on scattering coefficients of the measurements. The resulting estimator thus becomes

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}} \|\mathcal{G}\mathbf{x} - \mathbf{y}\|^2 + \beta\|S_J\mathbf{x} - MS_J\mathbf{y}\|^2 \ , \tag{57}$$

where M is a linear operator learnt by solving a linear regression of pairs $(S_J\mathbf{x}_i, S_J\mathbf{y}_i)_i$ in the scattering domain, where $\{\mathbf{x}_i, \mathbf{y}_i\}_i$ is a training set of input-output pairs.

This estimator differs from typical data-driven estimators that leverage supervised training in inverse problems using CNNs. More specifically, given a trainable model $\mathbf{x}_\theta = \Phi(\mathbf{y}; \theta)$, one considers

$$\hat{\mathbf{x}}_{\text{CNN}} = \mathbf{x}_{\theta^*} \ , \text{where } \theta^* \in \arg\min_\theta \sum_i \|\mathbf{x}_i - \Phi(\mathbf{y}_i; \theta)\|^2 \ . \tag{58}$$

See [AÖ18, ZZGZ17, JMFU17] for recent surveys on data-driven models for imaging inverse problems. Despite their phenomenal success across many inverse problems, such estimators suffer from the so-called 'Regression-to-the-mean' phenomena, in which the model is asked to predict a specific input $\mathbf{x}_i$ from potentially many plausible signals explaining the same observations $\mathbf{y}_i$ – leading to an estimator that averages all such plausible solutions, thus losing high-frequency and texture information. Instead, the scattering mircocanonical estimator (57) learns a linear operator using the scattering metric, which leverages the stability of the scattering transform to small deformations to avoid the regression to the mean phenomena of baseline estimators.

The estimator (57) was studied in [BSL15] using localized scattering, in the context of single-image super-resolution, and in [DBMdH16] for other imaging inverse problems such as tomography. In all cases, the gradient descent algorithm from Section 6.3 was employed. Figure 15 compares the resulting estimates with spline interpolation and with estimators of the form of (58).

Generative Networks as Inverse Problems with Scattering Transforms: In [AM18b], the authors consider a variant of the microcanonical scattering model, by replacing the gradient descent sampling scheme of Section 6.3 with a learnt deep convolutional network *generator*, that learns to map a vector of scattering coefficients $\mathbf{z} = S_J(\mathbf{x})$ back to $\mathbf{x}$. Deep generative models such as Variational Autoencoders [KW13, RMW14] or GANs [GPAM$^+$14] consider two networks, an *encoder* and a *decoder*. The encoder maps the data to a latent space with prescribed probability density, e.g. a standard Gaussian distribution, and the decoder maps it back to reconstruct the input. In this context, the scattering transform S_J may be used as an encoder on appropriate data distributions, thanks to its ability to linearize small deformations and 'Gaussianize' the input distribution [AM18b].

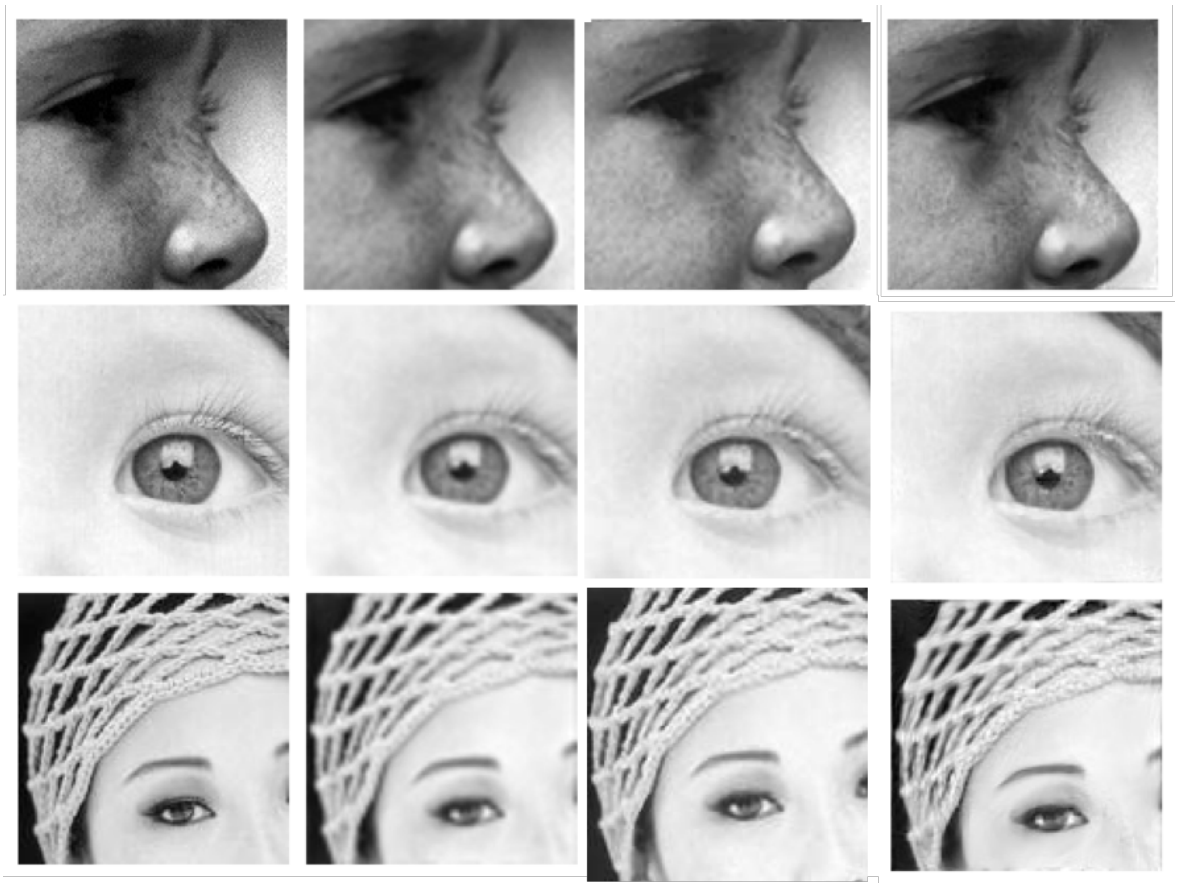


Figure 15: Comparison of single-image Super-Resolution using scattering microcanonical prior and pure data-driven models, using a linear model (leading to spline interpolation) and a CNN model from [DLHT14]. From left to right: original, linear model, CNN model, and scattering model.

Finally, in [AM18a] the authors used the time-frequency joint scattering transform of Section 5.2 and the learnt decoder from [AM18b] for generation and transformation of musical sounds.

6.5 Texture Synthesis with Microcanonical Scattering

An image or an audio texture is usually modeled as the realization of a stationary process. A texture model computes an approximation of this stationary process given a single realization, and texture synthesis then consists in calculating new realizations from this stochastic model.

Since in general the original stochastic process is not known, perceptual comparisons are the only criteria used to evaluate a texture synthesis algorithm. Microcanonical models can be considered as texture models computed from an energy function $S_J(\mathbf{x})$ which concentrate close to its mean.

[GG84] introduced macrocanonical models based on Markov random fields. They provide good texture models as long as these textures are realizations of random processes having no long range correlations. Several approaches have then been introduced to incorporate long range correlations. [HB95] capture texture statistics through the marginal distributions obtained by filtering images with oriented wavelets. This approach has been generalized by the macrocanonical Frame model of [ZWM98], based on marginal distributions of filtered images. The filters are optimized by trying to minimize the maximum entropy conditioned by the marginal distributions. Although the Cramer-Wold theorem proves that enough marginal probability distributions characterize any random vector defined over $\mathbb{R}^d$ the number of such marginals is typically intractable, which limits this approach. [PS00] made important improvements to these texture models, with wavelet transforms. They capture the correlation of the modulus of wavelet coefficients with a covariance matrix which defines an energy vector $\Phi_d(x)$. Although they use a macrocanonical maximum entropy formalism, their algorithm computes a microcanonical estimation from a single realization, with alternate projections as opposed to a gradient descent.

Excellent texture synthesis have recently been obtained with deep convolutional neural networks. In [GEB15], the authors consider a deep VGG convolutional network, trained on a large-scale image classification task. The energy vector is defined as the spatial cross-correlation values of feature maps at every layer of the VGG networks. This energy vector is calculated on a particular texture image. Texture syntheses of very good perceptual quality are calculated with a gradient descent microcanonical algorithm initialized on random noise. However, the dimension of this energy vector is larger than the dimension of the input image. These estimators are therefore not statistically consistent and have no asymptotic limit.

Figure 16 displays examples of textures from the Brodatz dataset synthesized using the scattering microcanonical model from [BM18], and compares the effect of using only first-order scattering coefficients or only covariance information. Although qualitatively better than these alternatives, deep convolutional networks reproduce image and audio textures of even better perceptual quality than scattering coefficients [GEB15], but use over 100 times more parameters. Much smaller models providing similar perceptual quality can be constructed with wavelet phase harmonics for audio signals [MZR18] or images [ZM19], which capture alignment of phases across scales. However, understanding how to construct low-dimensional multiscale energy vectors to approximate random processes remains mostly an open problem.

7 Final Remarks

This chapter aimed at providing a comprehensive overview of Scattering Representations, and specifically at motivating their role in the puzzle of understanding the effectiveness of deep learning.

In the context of high-dimensional learning problems involving geometric data, beating the curse of dimensionality requires exploiting as many geometric priors as possible. In particular, good signal representations should be stable with respect to small metric perturbations of the domain, expressed as deformations in the case of natural images. Scattering representations, through their constructive

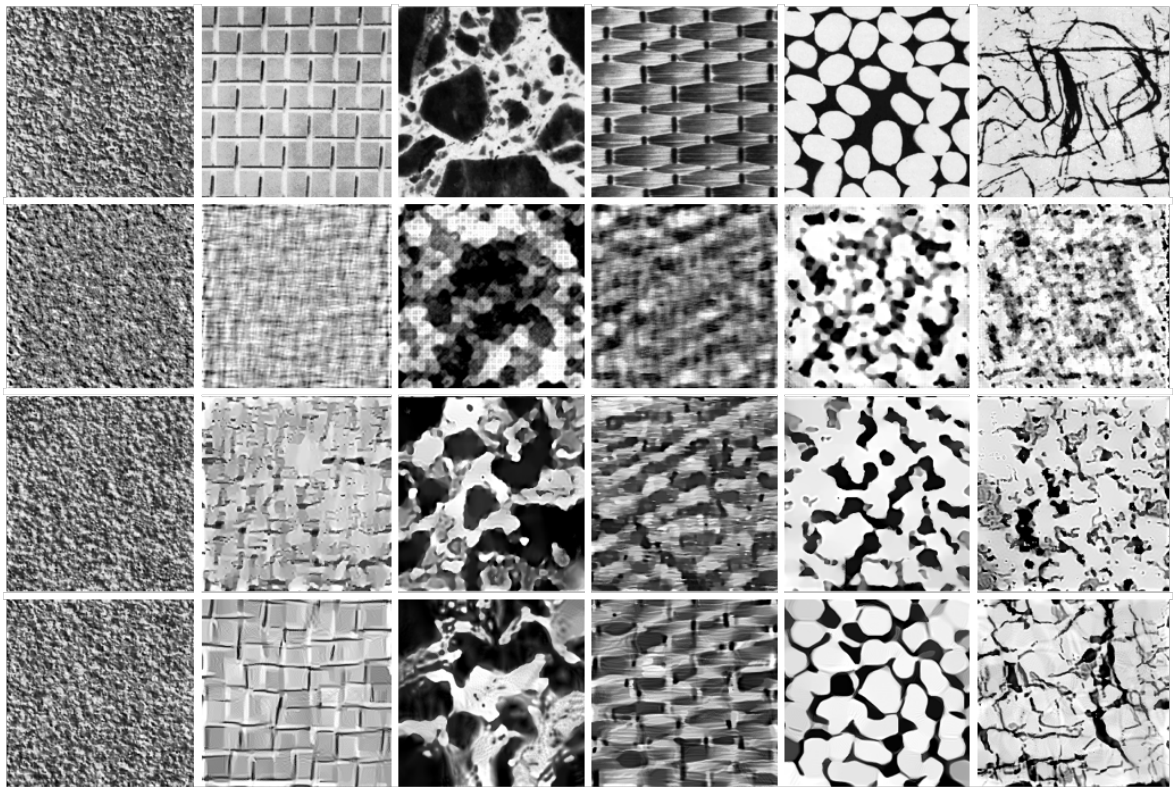


Figure 16: Examples of microcanonical texture synthesis using different vector of sufficient statistics. From top to bottom: original samples, gaussian model, first-order scattering and second-order scattering.

approach to build such stability, reveal the role of convolutions, depth and scale that underpins the success of CNN architectures.

We have mostly focused on the theoretical aspects of the scattering representation, and some of its ramifications beyond the context of computer vision and learning. That said, we logically could not cover all application areas nor some of the recent advances, especially the links with turbulence analysis and other non-linear PDEs in physics, applications to financial time-series [LRBM19], or video. Another important aspect that we did not address is the role of the non-linear activation function. All our discussion has focused on the complex modulus, but recent related work [MZR18] has considered the half-rectification case through the notion of 'phase harmonics', of which the modulus can be seen as the 'fundamental', complemented by higher harmonics.

Despite the above points, the inherent limitation of a scattering theory to explain deep learning is that precisely it does not consider the dynamical aspects of learning. Throughout numerous computer vision benchmarks, one systematically finds a performance gap between hand-designed scattering architectures and their fully trained counterparts, as soon as datasets become sufficiently large. The ability of CNNs to interpolate high-dimensional data while seemingly avoiding the curse of dimensionality remains an essential ability that scattering-based models currently lack. Hybrid approaches such as those outlined in [OZH+18] hold the promise of combining the interpretability and robustness of scattering models with the data-fitting power of large neural networks.

References

[AAE+18] Mathieu Andreux, Tomás Angles, Georgios Exarchakis, Roberto Leonarduzzi, Gaspar Rochette, Louis Thiry, John Zarka, Stéphane Mallat, Eugene Belilovsky, Joan Bruna, et al. Kymatio: Scattering transforms in python. *arXiv preprint arXiv:1812.11214*, 2018.

[ALM18] Joakim Andén, Vincent Lostanlen, and Stéphane Mallat. Classification with joint time-frequency scattering.(jul 2018). *arXiv preprint arXiv:1807.08869*, 2018.

[AM14] Joakim Andén and Stéphane Mallat. Deep scattering spectrum. *IEEE Transactions on Signal Processing*, 62(16):4114–4128, 2014.

[AM18a] Mathieu Andreux and Stéphane Mallat. Music generation and transformation with moment matching-scattering inverse networks. In *ISMIR*, pages 327–333, 2018.

[AM18b] Tomás Angles and Stéphane Mallat. Generative networks as inverse problems with scattering transforms. *arXiv preprint arXiv:1805.06621*, 2018.

[AÖ18] Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE transactions on medical imaging*, 37(6):1322–1332, 2018.

[BBK+10] Alexander M Bronstein, Michael M Bronstein, Ron Kimmel, Mona Mahmoudi, and Guillermo Sapiro. A gromov-hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching. *International Journal of Computer Vision*, 89(2-3):266–286, 2010.

[BBL+16] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *arXiv preprint arXiv:1611.08097*, 2016.

[BM13] J. Bruna and S. Mallat. Invariant scattering convolution networks. *Trans. PAMI*, 35(8):1872–1886, 2013.

[BM17] Alberto Bietti and Julien Mairal. Group invariance, stability to deformations, and complexity of deep convolutional representations. *arXiv preprint arXiv:1706.03078*, 2017.

[BM18] Joan Bruna and Stéphane Mallat. Multiscale sparse microcanonical models. *arXiv preprint arXiv:1801.02013*, 2018.

[BM19] Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. *arXiv preprint arXiv:1905.12173*, 2019.

[BMB+15] Joan Bruna, Stéphane Mallat, Emmanuel Bacry, Jean-François Muzy, et al. Intermittent process analysis with scattering moments. *The Annals of Statistics*, 43(1):323–351, 2015.

[Bru13] Joan Bruna. *Scattering representations for recognition*. PhD thesis, Ecole Polytechnique X, 2013.

[Bru19] Joan Bruna. Consistency of haar scattering. *arXiv preprint*, 2019.

[BSL15] Joan Bruna, Pablo Sprechmann, and Yann LeCun. Super-resolution with deep convolutional sufficient statistics. *arXiv preprint arXiv:1511.05666*, 2015.

[BZSL13] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *Proc. ICLR*, 2013.

[CCM14] Xu Chen, Xiuyuan Cheng, and Stéphane Mallat. Unsupervised deep haar scattering on graphs. In *Advances in Neural Information Processing Systems*, pages 1709–1717, 2014.

[CCM16] Xiuyuan Cheng, Xu Chen, and Stéphane Mallat. Deep Haar scattering networks. *Information and Inference*, 5:105–133, 2016.

[CG97] Fan RK Chung and Fan Chung Graham. *Spectral graph theory*. Number 92. American Mathematical Soc., 1997.

[Cha17] Sourav Chatterjee. A note about the uniform distribution on the intersection of a simplex and a sphere. *Journal of Topology and Analysis*, 9(04):717–738, 2017.

[CL06] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[CL17] Wojciech Czaja and Weilin Li. Analysis of time-frequency scattering transforms. *Applied and Computational Harmonic Analysis*, 2017.

[Cre83] Michael Creutz. Microcanonical monte carlo simulation. *Physical Review Letters*, 50(19):1411, 1983.

[CW16a] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999, 2016.

[CW16b] Taco S Cohen and Max Welling. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016.

[DBMdH16] Ivan Dokmanić, Joan Bruna, Stéphane Mallat, and Maarten de Hoop. Inverse problems with invariant multiscale statistics. *arXiv preprint arXiv:1609.05502*, 2016.

[DLHT14] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014.

[DZ93] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Johns and Bartett Publishers, Boston, 1993.

[EEHM17] Michael Eickenberg, Georgios Exarchakis, Matthew Hirn, and Stéphane Mallat. Solid harmonic wavelet scattering: Predicting quantum molecular energy from invariant descriptors of 3d electronic densities. In *Advances in Neural Information Processing Systems*, pages 6540–6549, 2017.

[FFFP04] L. Fei-Fei, R. Fergus, and P. Perona. Learning Generative Visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *IEEE CVPR*, 2004.

[FGMR10] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *Trans. PAMI*, 32(9):1627–1645, 2010.

[GBR19] Fernando Gama, Joan Bruna, and Alejandro Ribeiro. Stability of graph scattering transforms. *arXiv preprint arXiv:1906.04784*, 2019.

[GDDM14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[GEB15] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 262–270, 2015.

[GG84] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, pages 721–741, 1984.

[GNC10] Matan Gavish, Boaz Nadler, and Ronald R Coifman. Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning. In *Proc. ICML*, 2010.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[GRB18] Fernando Gama, Alejandro Ribeiro, and Joan Bruna. Diffusion scattering transforms on graphs. *arXiv preprint arXiv:1806.08829*, 2018.

[GWH19] Feng Gao, Guy Wolf, and Matthew Hirn. Geometric scattering for graph data analysis. In *International Conference on Machine Learning*, pages 2122–2131, 2019.

[HB95] David J Heeger and James R Bergen. Pyramid-based texture analysis/synthesis. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 229–238. ACM, 1995.

[HMP17] Matthew Hirn, Stéphane Mallat, and Nicolas Poilvert. Wavelet scattering regression of quantum chemical energies. *Multiscale Modeling & Simulation*, 15(2):827–863, 2017.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[JMFU17] Kyong Hwan Jin, Michael T McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

[JvGLS16] Jorn-Henrik Jacobsen, Jan van Gemert, Zhongyu Lou, and Arnold WM Smeulders. Structured receptive fields in cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2610–2619, 2016.

[JZL96] Anil K. Jain, Yu Zhong, and Sridhar Lakshmanan. Object matching using deformable templates. *IEEE Transactions on pattern analysis and machine intelligence*, 18(3):267–278, 1996.

[KBPZ17] Ilya Kostrikov, Joan Bruna, Daniele Panozzo, and Denis Zorin. Surface networks. *arXiv preprint arXiv:1705.10819*, 2017.

[KDGH07] D. Keysers, T. Deselaers, C. Gollan, and N. Hey. Deformation Models for Image Recognition. *IEEE trans of PAMI*, 2007.

[KT18] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. *arXiv preprint arXiv:1802.03690*, 2018.

[KW13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[LRBM19] Roberto Leonarduzzi, Gaspar Rochette, Jean-Phillipe Bouchaud, and Stéphane Mallat. Maximum-entropy scattering models for financial time series. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5496–5500. IEEE, 2019.

[Mal99] Stéphane Mallat. *A wavelet tour of signal processing*. Academic Press, 1999.

[Mal08] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, New York, 2008.

[Mal12] S. Mallat. Group Invariant Scattering. *Communications in Pure and Applied Mathematics (to appear)*, 2012.

[Mal16] Stéphane Mallat. Understanding deep convolutional networks. *Phil. Trans. R. Soc. A*, 374(2065), 2016.

[MKHS14] Julien Mairal, Piotr Koniusz, Zaid Harchaoui, and Cordelia Schmid. Convolutional kernel networks. In *Advances in neural information processing systems*, pages 2627–2635, 2014.

[MNY06] H. Minh, P. Niyogi, and Y. Yao. Mercer's Theorem, Feature Maps and Smoothing. *Proc. of Computational Learning Theory*, 2006.

[MZR18] Stéphane Mallat, Sixin Zhang, and Gaspar Rochette. Phase harmonics and correlation invariants in convolutional neural networks. *arXiv preprint arXiv:1810.12136*, 2018.

[NLCK06] Boaz Nadler, Stéphane Lafon, Ronald R Coifman, and Ioannis G Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.

[OBZ17] Edouard Oyallon, Eugene Belilovsky, and Sergey Zagoruyko. Scaling the scattering transform: Deep hybrid networks. *arXiv preprint arXiv:1703.08961*, 2017.

[OBZV18] 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.

[OM15] Edouard Oyallon and Stéphane Mallat. Deep roto-translation scattering for object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2865–2873, 2015.

[OZH+18] Edouard Oyallon, Sergey Zagoruyko, Gabriel Huang, Nikos Komodakis, Simon Lacoste-Julien, Matthew B Blaschko, and Eugene Belilovsky. Scattering networks for hybrid representation learning. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[PS00] Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70, 2000.

[PT02] G. Oppenheim P.Doukhan and M. Taqqu. *Theory and Applications of Long-Range Dependence*. Birkhauser, Boston, 2002.

[PWH18] Michael Perlmutter, Guy Wolf, and Matthew Hirn. Geometric scattering on manifolds. *arXiv preprint arXiv:1812.06968*, 2018.

[RG13] Raif Rustamov and Leonidas J Guibas. Wavelets on graphs via deep learning. In *Proc. NIPS*, 2013.

[RMW14] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and variational inference in deep latent gaussian models. *arXiv preprint arXiv:1401.4082*, 2014.

[SM13] L. Sifre and S. Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *Proc. CVPR*, 2013.

[Soa09] S. Soatto. Actionable Information in Vision. *ICCV*, 2009.

[STC04] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[SZS+13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[Wal12] I. Waldspurger. Recovering the phase of a complex wavelet transform, 2012.

[Wal17] Irène Waldspurger. Exponential decay of scattering coefficients. In *2017 international conference on sampling theory and applications (SampTA)*, pages 143–146. IEEE, 2017.

[WB17] Thomas Wiatowski and Helmut Bölcskei. A mathematical theory of deep convolutional neural networks for feature extraction. *IEEE Transactions on Information Theory*, 64(3):1845–1866, 2017.

[WGB17] Thomas Wiatowski, Philipp Grohs, and Helmut Bölcskei. Energy propagation in deep convolutional neural networks. *IEEE Transactions on Information Theory*, 64(7):4819–4842, 2017.

[WJ+08] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[YSO+11] K. Yoshimatsu, K. Schneider, N. Okamoto, Y. Kawahura, and M. Farge. Intermittency and geometrical statistics of three-dimensional homogeneous magnetohydrodynamic turbulence : A wavelet viewpoint. *Phys. Plasmas*, 2011.

[ZL18] D. Zou and G. Lerman. Graph convolutional neural networks via scattering. *arXiv:1804.00099v1 [cs.IT]*, 31 March 2018.

[ZM19] Sixin Zhang and Stephane Mallat. Wavelet phase harmonic covariance models of stationary processes. *arXiv preprint*, 2019.

[ZWM98] Song Chun Zhu, Yingnian Wu, and David Mumford. Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):107–126, 1998.

[ZZGZ17] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3929–3938, 2017.

A Proofs

A.1 Proof of Proposition 3.10

A.2 Proof of Theorem ??

Fix $p \in \mathcal{P}_\infty$, and let $q \in C_J(p) \cap \mathcal{P}_\infty$ be a path in the neighborhood of p. We can thus write $q = p + \tilde{q}$, with $\tilde{q} = \tilde{\lambda} + \check{q} \in \mathcal{P}_\infty$ satisfying $|\tilde{\lambda}|^{-1} \leq 2^{-J}$. We have

$$\begin{aligned}
\overline{S}x(q) &= \frac{\int U[q]x(u)du}{\int U[q]\delta(u)du} = \frac{\int U[\tilde{q}]U[p]x(u)du}{\int U[\tilde{q}]U[p]\delta(u)du} \\
&= \frac{\int U[\tilde{q}]U[p]x(u)du}{\int U[\tilde{q}]\delta(u)du} \cdot \frac{\int U[\tilde{q}]\delta(u)du}{\int U[\tilde{q}]U[p]\delta(u)du} \\
&= \overline{S}(U[p]x)(\tilde{q}) \cdot \left(\overline{S}(U[p]\delta)(\tilde{q})\right)^{-1} . \qquad (59)
\end{aligned}$$

The following lemma proves that if x is in $\mathbf{L}^1(\mathbb{R}^d)$ and is positive, then $\overline{S}x$ has a particularly simple form on "small" paths $\tilde{q} \in \mathcal{P}_\infty$ with finite order and finite excursion:

Lemma A.1. *Let m, $B \in \mathbb{N}$, and let*

$$\mathcal{A}_{J,m} = \{q \in \mathcal{P}_\infty \, ; \, q = (\lambda_1, \ldots, \lambda_m), \, |q| = m, \, |\lambda_1| = 2^J, \, \Delta(q) \leq M\} \, . \qquad (60)$$

If $x \in \mathbf{L}^1(\mathbb{R}^d)$, $x \geq 0$, then

$$\lim_{J \to \infty} \sup_{q \in \mathcal{A}_{J,m}} \left| \overline{S}x(q) - \int x(u)du \right| = 0 \, . \qquad (61)$$

If we apply Lemma A.1 to $f_1 = U[p]x$ and $f_2 = U[p]\delta$, then the identity (59) implies that for any $\epsilon > 0$ there exists $J > 0$ such that

$$\forall q \text{ s.t. } q \in C_J(p), \, |q| \leq m, \, \Delta(q) \leq B \, , \left| \overline{S}x(q) - \frac{\int U[p]x(u)du}{\int U[p]\delta(u)du} \right| \leq \epsilon \, ,$$

which implies (**??**) since $\overline{S}x(p) = (\int U[p]\delta(u)du)^{-1} \int U[p]f(u)du$.

We shall then prove (61). Fix $J > 0$, and let $q \in \mathcal{A}_{J,m}$. By definition (60), we can write $q = r2^J + \tilde{q}$, and without loss of generality, we can assume that $r = 1$. let $D_j x(u) = 2^{-jd} x(2^{-j}u)$ be a dilation

operator normalized in $\mathbf{L}^1(\mathbb{R}^d)$. A change of variables shows that

$$\begin{aligned} D_j x \star \psi_\lambda(u) &= 2^{-jd} \int x(2^{-j}v)\psi_\lambda(u-v)dv \\ &= \int x(v)\psi_\lambda(u-2^j v)dv = \int x(v)\psi_\lambda(2^j(2^{-j}u-v))dv \\ &= 2^{2^{-jd}} \int x(v)\psi_{2^{-j}\lambda}(2^{-j}u-v)dv \\ &= D_j(x \star \psi_{2^{-j}\lambda})(u) \ , \end{aligned} \tag{62}$$

and by cascading this property we obtain that

$$U[p]D_j x = D_j U[2^{-j}p]x \ ,$$

or equivalently $U[p]x = D_j U[2^{-j}p]D_{-j}x$. By setting $j = J$, we obtain

$$\overline{S}x(2^J + \tilde{q}) = \overline{S}D_{-J}x(1+\tilde{q}2^{-J}) = \frac{\int U[1+\tilde{q}2^{-J}]D_{-J}x(u)du}{\int U[1+\tilde{q}2^{-J}]\delta(u)du} \ , \tag{63}$$

since $D_j\delta = \delta \ \forall j$ with the $\mathbf{L}^1(\mathbb{R}^d)$ normalization. Now, if $\overline{x} = \int x(u)du$, (63) can be decomposed as

$$\begin{aligned} \overline{S}x(2^J+\tilde{q}) &= \\ &= \frac{\int \overline{x}U[1+\tilde{q}2^{-J}]\delta(u)du}{\int U[1+\tilde{q}2^{-J}]\delta(u)du} + \frac{\int \left(U[1+\tilde{q}2^{-J}]D_{-J}x(u) - \overline{x}U[1+\tilde{q}2^{-J}]\delta(u)\right)du}{\int U[1+\tilde{q}2^{-J}]\delta(u)du} \\ &= \overline{x} + \frac{\int \left(U[1+\tilde{q}2^{-J}]D_{-J}x(u) - U[1+\tilde{q}2^{-J}]\overline{x}\delta(u)\right)du}{\int U[1+\tilde{q}2^{-J}]\delta(u)du} \ , \end{aligned} \tag{64}$$

The path $2^{-J}q = 1+\tilde{q}2^{-J}$ is obtained by a translation in scale of q, and hence it satisfies $|2^{-J}q| = |q|$ and $\Delta(2^{-J}q) = \Delta(q)$. We will prove (61) by showing that

$$\inf_{q\in\mathcal{A}_{1,m}} \int U[q]\delta(u)du > 0 \ , \tag{65}$$

and

$$\lim_{J\to\infty} \sup_{q\in\mathcal{A}_{1,m}} \left| \int \left(U[q]D_{-J}x(u) - U[q]a\delta(u)\right)du \right| = 0 \ . \tag{66}$$

Let us first prove (65), by induction on the maximum path order m.

Let $m = 2$. In that case, the set $\mathcal{A}_{1,2}$ contains paths $q = (1,\lambda)$, where the scale of λ is lower bounded by $|\lambda|^{-1} \leq M$. We need to see that

$$\inf_{|\lambda|^{-1}\leq M} \int ||\psi| \star \psi_\lambda|(u)du = \||\psi| \star \psi_\lambda\|_1 > 0 \ .$$

From (62) we deduce that if $j = |\lambda|$, then

$$\||\psi| \star \psi_\lambda\|_1 = \|D_j(D_{-j}|\psi| \star \psi)\|_1 = \|D_{-j}|\psi| \star \psi\|_1 \ .$$

Since $|\psi| \in \mathbf{L}^1(\mathbb{R}^d)$ and $|\psi| \geq 0$, it follows that $D_{-j}|\psi|$ is an approximation of the identity in $\mathbf{L}^1(\mathbb{R}^d)$ as $j \to \infty$, with

$$\forall j \ , \int D_{-j}|\psi|(u)du = \|\psi\|_1 \ ,$$

and hence

$$\lim_{j\to\infty} \|D_{-j}|\psi| \star \psi - \|\psi\|_1\psi\|_1 = 0 \ . \tag{67}$$

But

$$\begin{aligned}\left|\int ||\psi| \star \psi_\lambda|(u)du - \|\psi\|_1 \int |\psi|(u)du\right| &= \left|\int |D_{-j}|\psi| \star \psi|(u)du - \|\psi\|_1 \int |\psi|(u)du\right| \\ &\leq \int ||D_{-j}|\psi| \star \psi|(u) - \|\psi\|_1 |\psi|(u)|\, du \\ &\leq \|D_{-j}|\psi| \star \psi - \|\psi\|_1 \psi\|_1 \ .\end{aligned}$$

As a result, $\forall \epsilon > 0$ there exists J such that if $|\lambda| > J$, then

$$\left|\int ||\psi| \star \psi_\lambda|(u)du - \|\psi\|_1^2\right| \leq \epsilon \ .$$

If ϵ is chosen such that $\epsilon < \|\psi\|_1^2/2$, and J_ϵ is the corresponding J, then the paths $q \in \mathcal{A}_{1,2}$, $q = (1, \lambda)$ with $|\lambda| > J_\epsilon$ satisfy

$$\forall q \in \mathcal{A} \ , q = (1, \lambda) \ , |\lambda| > J_\epsilon \ , \int U[q]\delta(u)du > \|\psi\|_1^2 - \epsilon = \frac{\|\psi\|_1^2}{2} > 0 \ . \tag{68}$$

On the other hand, there are only a finite number of paths $q \in \mathcal{A}_{1,2}$ with $|\lambda| \leq J_\epsilon$, since by definition $|\lambda| \geq M^{-1}$. As a result,

$$\inf_{\substack{q \in \mathcal{A} \\ q=(1,\lambda), |\lambda| \leq J_\epsilon}} \int U[q]\delta(u)du = \alpha_0 > 0 \ . \tag{69}$$

By combining (68) and (69) we obtain that

$$\inf_{q \in \mathcal{A}} \int U[q]\delta(u)du \geq \min(\alpha_0, \frac{\|\psi\|_1^2}{2}) = \alpha > 0 \ . \tag{70}$$

Let us now suppose the result true for $m = m_0 - 1$. We shall prove that it is also true for $m = m_0$. Let

$$\inf_{q \in \mathcal{A}_{1,m_0-1}} \int U[q]\delta(u)du = \alpha > 0 \ .$$

For each $l > 0$, we shall decompose the set $\mathcal{A}_{1,m_0}$ in terms of the maximum jump of the path:

$$\mathcal{A}_{1,m_0} = \mathcal{B}_l \cup (\mathcal{A}_{1,m_0} \setminus \mathcal{B}_l) \ ,$$

with

$$\mathcal{B}_l = \left\{ q \in \mathcal{A}_{1,m_0} \, , q = (\lambda_1, \ldots, \lambda_{m_0}); \ \chi(q) = \max_k \left(\frac{|\lambda_k|}{\sum_{k'<k} |\lambda_{k'}|} \right) \geq 2^l \right\} \ .$$

The maximum jump $\chi(q)$ of a path thus measures the largest decrease on the scale, with respect to the current cumulated support of $U[\lambda_1, \ldots, \lambda_k]$. Since the set $\mathcal{A}_{1,m}$ contains paths of finite order and finite slope, the maximum jump is lower bounded by a constant M_0 depending on M and the order m_0.

Let $q \in \mathcal{B}_l$. We can write $q = q_0 + \lambda + q_1$, where $q_0 = (\lambda'_1, \ldots, \lambda'_{k'})$ satisfies

$$|\lambda| \geq (\sum_{i \leq k'} |\lambda'_i|) 2^l \ . \tag{71}$$

If $\lambda = 2^j r$, we have

$$\begin{aligned} U[q_0 + \lambda]\delta &= U[\lambda]U[q_0]\delta = |U[q_0]\delta \star \psi_\lambda| \\ &= D_j(D_{-j}U[q_0]\delta \star \psi_{2^{\mathrm{o}}r}) \end{aligned} \tag{72}$$

We will now exploit again the fact that $f_j(u) = D_{-j}U[q_0]\delta(u)$ is an approximation of the identity in $\mathbf{L}^1(\mathbb{R}^d)$. Let $\gamma = \int f_j(u)du$, which does not depend upon j. We have

$$\begin{aligned}\|D_j(D_{-j}U[q_0]\delta \star \psi_{2^0 r}) - \gamma D_j \psi_{2^0 r}\|_1 &= \|(f_j \star \psi_{2^0 r}) - \gamma\psi_{2^0 r}\|_1 \\ &= \int \left| \int (\psi_{2^0 r}(u-t) - \psi_{2^0 r}(u)) f_j(t)dt \right| du \\ &\leq \int \|T_t\psi_{2^0 r} - \psi_{2^0 r}\|_1 f_j(t)dt \ , \end{aligned} \tag{73}$$

where $T_t h(u) = h(u-t)$ is the translation operator. Since the translation operator $t \mapsto T_t h$ is continuous in $\mathbf{L}^1(\mathbb{R}^d)$ for any $h \in \mathbf{L}^1(\mathbb{R}^d)$, then for each $\epsilon > 0$, we can find $\eta > 0$ which only depends upon ψ such that

$$\forall |t| < \eta \ , \ \|T_t\psi_{2^0 r} - \psi_{2^0 r}\|_1 < \epsilon/2 \ . \tag{74}$$

On the other hand,

$$\begin{aligned}\int_{|t|>\eta} \|T_t\psi_{2^0 r} - \psi_{2^0 r}\|_1 f_j(t)dt &\leq 2\|\psi\|_1 \int_{|t|>\eta} f_j(t)dt \\ &= 2\|\psi\|_1 \int_{|t|>\eta} D_{-j}U[q_0]\delta(t)dt \\ &= 2\|\psi\|_1 \int_{|t|>2^j\eta} U[q_0]\delta(t)dt \ . \end{aligned} \tag{75}$$

By construction, the scale 2^j is such that

$$2^j \geq (\sum_{i \leq k'} |\lambda'_i|) \cdot 2^l \ ,$$

from (71). Since the wavelet ψ has fast decay, $U[q_0]\delta(t)$ satisfies

$$|U[q_0]\delta(t)| \leq C_1/(C_2 + (|t|/K))^n \ ,$$

where C_i and n only depend upon ψ and $K = \sum_{i \leq k'} |\lambda'_i|$ is proportional to the effective support of the cascade of convolutions given by $U[q_0]h = |||h \star \psi_{\lambda'_1}| \star \ldots| \star \psi_{\lambda'_{k'}}|$. As a result, the error in (75) can be bounded by

$$\begin{aligned}\int_{|t|>\eta} \|T_t\psi_{2^0 r} - \psi_{2^0 r}\|_1 f_j(t)dt &\leq C\|\psi\|_1 \epsilon(l) \int U[q_0]\delta(t)dt \\ &\leq C\|\psi\|_1 \gamma\epsilon(l) \ , \end{aligned} \tag{76}$$

where $\epsilon(l) \to 0$ as $l \to \infty$. By using (74) and (76) we can now bound (73) with

$$\begin{aligned}\|(f_j \star \psi_{2^0 r}) - \gamma\psi_{2^0 r}\|_1 &\leq \int \|T_t\psi_{2^0 r} - \psi_{2^0 r}\|_1 f_j(t)dt \\ &= \int_{|t|<\eta} \|T_t\psi_{2^0 r} - \psi_{2^0 r}\|_1 f_j(t)dt + \int_{|t|>\eta} \|T_t\psi_{2^0 r} - \psi_{2^0 r}\|_1 f_j(t)dt \\ &\leq \epsilon/2\gamma + C\|\psi\|_1 \gamma\epsilon(l) \\ &\leq \|\psi\|_1^{|q_0|}(\epsilon/2 + C\|\psi\|_1\epsilon(l)) \ , \end{aligned} \tag{77}$$

since $\gamma = \int U[q_0]\delta(u)du \leq \|\psi\|_1^{m_0}$ using the Young inequality $\|f \star g\|_1 \leq \|f\|_1\|g\|_1$.

Since

$$\begin{aligned}\|U[\lambda]f - U[\lambda]g\|_1 &= \||f \star \psi_\lambda| - |g \star \psi_\lambda|\|_1 \\ &\leq \|f \star \psi_\lambda - g \star \psi_\lambda\|_1 = \|(f-g) \star \psi_\lambda\|_1 \\ &\leq \|f-g\|_1\|\psi\|_1 \ , \end{aligned}$$

it follows that

$$\|U[p]f - U[p]g\|_1 \leq \|f-g\|_1 \|\psi\|_1^{|p|} \ . \tag{78}$$

As a result of (77), any path $q \in \mathcal{B}_l$, which was decomposed as $q = q_0 + \lambda + q_1$, satisfies

$$\begin{aligned}
\left| \int U[q]\delta(u)du - \gamma \int U[\lambda+q_1]\delta(u)du \right| &\leq \\
\|U[q]\delta - \gamma U[q_1]U[\lambda]\delta\|_1 &= \|U[q_1]U[\lambda]U[q_0]\delta - \gamma U[q_1]U[\lambda]\delta\|_1 \\
&\leq \|\psi\|_1^{|q_1|} \|U[q_0]\delta \star \psi_\lambda - \gamma\psi_\lambda\|_1 \\
&\leq \|\psi\|_1^{m_0}(\epsilon/2 + C\|\psi\|_1 \epsilon(l)) \ ,
\end{aligned} \tag{79}$$

by applying (78) on $U[q_1]$. (79) implies that for any $\epsilon > 0$ one can find sufficiently large l such that $\int U[q]\delta(u)du$ is at distance at most ϵ from $\gamma \int U[\tilde{q}]\delta(u)du$, where $|\tilde{q}| < |q|$ and $\alpha \leq \gamma \leq \|\psi\|_1^{m_0}$. By applying the induction hypothesis with $\epsilon = \alpha/2$, we conclude that

$$\forall q \in \mathcal{B}_l \ \ , \ \int U[q]\delta(u)du \geq \alpha^2/2 > 0 \ . \tag{80}$$

On the other hand, the set $\mathcal{A}_{1,m_0} \setminus \mathcal{B}_l$ contains only a finite number of paths, since their slope is bounded by $\Delta(q) \leq B$, and thus

$$\min_{q \in \mathcal{A}_{1,m_0} \setminus \mathcal{B}_l} \int U[q]\delta(u)du = \alpha_0 > 0 \ .$$

We conclude that

$$\forall q \in \mathcal{A}_{1,m_0} \ \ , \ \int U[q]\delta(u)du \geq \min(\alpha^2/2, \alpha_0) > 0 \ , \tag{81}$$

which proves (65).

Let us finally prove (66). Since $x \in \mathbf{L}^1(\mathbb{R}^d)$ and $x \geq 0$, $D_{-J}x$ is also an approximation of the identity, which, with $\overline{x} = \int x(u)du$, satisfies

$$\forall h \in \mathbf{L}^1(\mathbb{R}^d) \ , \quad \lim_{J \to \infty} \|D_{-J}x \star h - \overline{x}h\|_1 = 0 \ . \tag{82}$$

If $q \in \mathcal{A}$, $q = \lambda_1 + \tilde{q}$ with $\lambda_1 = 2^0 r$, and hence $U[q]D_{-J}x = U[\tilde{q}]|D_{-J}x \star \psi_{\lambda_1}|$. Then, by using again (78), it results that

$$\begin{aligned}
\left| \int U[q]D_{-J}x(u)du - \overline{x} \int U[q]\delta(u)du \right| &= \left| \int (U[q]D_{-J}x(u) - \overline{x}U[q]\delta(u))du \right| \\
&\leq \int |U[q]D_{-J}x(u) - \overline{x}U[q]\delta(u)| \, du \\
&= \|U[q]D_{-J}x - \overline{x}U[q]\delta\|_1 \\
&= \|U[\tilde{q}]|D_{-J}x \star \psi_{\lambda_1}| - \overline{x}U[\tilde{q}]|\psi_{\lambda_1}|\|_1 \\
&\leq \|\psi\|_1^{|\tilde{q}|} \| |D_{-J}x \star \psi_{\lambda_1}| - \overline{x}|\psi_{\lambda_1}| \|_1 \\
&\leq \|\psi\|_1^{|\tilde{q}|} \|D_{-J}x \star \psi_{\lambda_1} - \overline{x}\psi_{\lambda_1}\|_1 \ ,
\end{aligned} \tag{83}$$

which can be made arbitrarily small thanks to (82). This proves (66), which concludes the proof of Lemma A.1, and hence of (**??**) □.

A.3 Proof of Proposition ??

Let $(\mathbf{V}_k)_{k\in\mathbb{Z}}$ be a multiresolution analysis generated by a scaling function $\varphi \in \mathbf{L}^2(\mathbb{R}^d) \cap \mathbf{L}^1(\mathbb{R}^d)$, and fix $j \in \mathbb{Z}$. Let us first prove the result for $x_j = P_{\mathbf{V}_j} x$. x_j can be written using the orthonormal basis $\{\varphi_j(u - k2^j)\}_{k\in\mathbb{Z}}$, with $\varphi_j(u) = 2^{-jd/2}\varphi(2^{-j}u)$:

$$x_j(u) = \sum_k c_k \varphi_j(u - 2^j k) \ ,$$

with $c_k = \langle x(u), \varphi_j(u - 2^j k)\rangle$. But

$$Qx_j(u) = \sum_k c_k Q\varphi_j(u - 2^j k) = Q\varphi_j(u) \sum_k c_k \ , \tag{84}$$

thanks to the fact that Q is linear and $Q\varphi_j(u - 2^j k) = QT_{2^j k}\varphi_j = Q\varphi_j$. Moreover,

$$\begin{aligned} \int x_j(u)du &= \int \sum_k c_k \varphi_j(u - 2^j k)du \\ &= \sum_k c_k \left(\int \varphi_j(u - 2^j k)du \right) = \left(\sum_k c_k \right) \int \varphi_j(u)du \ , \end{aligned}$$

which implies, by substituting in (84), that

$$Qx_j = Q(\varphi_j) \left(\int \varphi_j(u)du \right)^{-1} \left(\int x_j(u)du \right) = C \left(\int x_j(u)du \right) \ ,$$

where C only depends upon the scaling function and the resolution. We finally extend the result to $\mathbf{L}^2(\mathbb{R}^d) \cap \mathbf{L}^1(\mathbb{R}^d)$ with a density argument. Given $x \in \mathbf{L}^2(\mathbb{R}^d) \cap \mathbf{L}^1(\mathbb{R}^d)$ and $\epsilon > 0$, there exists a resolution j such that $\|x - P_{V_j} x\| \leq \epsilon$. Let $x_j = P_{V_j} x$. Since $\int x(u)du = \int x_j(u)du$, it follows that

$$\begin{aligned} \|Qx - Qx_j\| &= \|Qx - C \int x(u)du\| \\ &= \|Q(x - x_j)\| \leq \|Q\|\|x - x_j\| \leq \|Q\|\epsilon \ , \end{aligned}$$

which concludes the proof since Q is a bounded operator $\square$.

A.4 Proof of Theorem ??

Proof: Let $\mathbf{1}_\Omega$ be the indicator of a compact ball $\Omega \subset \mathbb{R}^d$. Let us first show that $M\mathbf{1}_\Omega = \rho\mathbf{1}_\Omega$. Let $\phi \in Diff(\mathbb{R}^d)$ be a diffeomorphism of $\mathbb{R}^d$. For $f \in \mathbf{L}^2(\mathbb{R}^d)$, we denote $L_\phi f = f \circ \phi$. Given $f \in \mathbf{L}^2(\mathbb{R}^d)$, let

$$G(f) = \{\phi \in Diff(\mathbb{R}^d) \, , L_\phi f = f\}$$

denote the isotropy group of f, ie the subgroup of diffeomorphisms leaving f unchanged up to a set of zero measure. If $\phi \in G(f)$, then

$$\|Mf - L_\phi Mf\| = \|Mf - ML_\phi f\| \leq \|f - L_\phi f\| = 0 \ ,$$

which means that $\phi \in G(M(f))$ too.

If $f = c\mathbf{1}_\Omega$, then its isotropy group contains any diffeomorphism ϕ satisfying

$$\phi(\Omega) = \Omega \, , \quad \phi(\overline{\Omega}) = \overline{\Omega} \ ,$$

where $\overline{\Omega} = \mathbb{R}^d - \Omega$. Thus, Mf is also invariant to the action of any ϕ satisfying the above conditions. It results that Mf must also be constant within both Ω and $\overline{\Omega}$ up to a set of zero measure. Indeed,

otherwise we could find two subsets $I_1, I_2 \subset \Omega$ of strictly positive measure $\mu(I_1) = \mu(I_2) > 0$, such that

$$\int_{I_1} Mf(x)d\mu(x) \neq \int_{I_2} Mf(x)d\mu(x) \ ,$$

but then a diffeomorphism ϕ such that $\phi \in G(\mathbf{1}_\Omega)$ and mapping I_1 to I_2, does not satisfy $\|Mf - L_\phi Mf\|_2 = 0$, which is a contradiction.

Since Mf belongs to $\mathbf{L}^2(\mathbb{R}^d)$ and $\overline{\Omega}$ has infinite measure, it results that $Mf(x) = 0 \ \forall x \in \overline{\Omega}$, and hence

$$M(c\mathbf{1}_\Omega) = \rho(c, \Omega)\mathbf{1}_\Omega \ ,$$

with $\rho(c, \Omega) = (Mc\mathbf{1}_\Omega)(x_0)$ for any $x_0 \in \Omega$. Since the hypercube Ω can be obtained from the unit ball Ω_0 of $\mathbb{R}^d$ with a similarity transform T_Ω , $\Omega = T_\Omega\Omega_0$, we have $M(c\mathbf{1}_\Omega) = M(T_\Omega c\mathbf{1}_{\Omega_0}) = T_\Omega M(c\mathbf{1}_{\Omega_0})$, which shows that $\rho(c, \Omega)$ does not depend upon Ω, and we shall write it $\rho(c)$.

Let us now consider $f \in C^\infty$ with compact support Ω. Fix a point $x_0 \in \Omega$. We consider a sequence of diffeomorphisms $(\phi_n)_{n\in\mathbb{N}}$ which progressively warp f towards $f(x_0)\mathbf{1}_\Omega$:

$$\lim_{n\to\infty} \|L_{\phi_n} f - f(x_0)\mathbf{1}_\Omega\| = 0 \ , \tag{85}$$

For that purpose, we construct ϕ_n such that $\phi_n(x) = x$ for $x \in \overline{\Omega}$ for all n, and such that it maps a neighborhood of radius 2^{-n} of x_0 to the set $\Omega_n \subset \Omega$ defined as

$$\Omega_n = \{x \in \Omega \ , \mathrm{dist}(x, \overline{\Omega}) \geq 2^{-n}\} \ .$$

Thanks to the fact that the domain Ω is regular, such diffeormorphisms can be constructed for instance by expanding the rays departing from x_0 at the neighborhood of x_0 and contracting them as they approach the border $\partial\Omega$. Since f is C^∞ and it is compactly supported, it is bounded, and hence

$$\begin{aligned} \|L_{\phi_n}(Mf) - M(f(x_0)\mathbf{1}_\Omega)\| &= \|M(L_{\phi_n} f) - M(f(x_0)\mathbf{1}_\Omega)\| \\ &\leq \|L_{\phi_n} f - f(x_0)\mathbf{1}_\Omega\| \ , \end{aligned}$$

and it results from (85) that $\lim_{n\to\infty} L_{\phi_n}(Mf) = M(f(x_0)\mathbf{1}_\Omega)$ in $\mathbf{L}^2(\mathbb{R}^d)$. Since the diffeomorphisms ϕ_n expand the neighborhood of x_0 and $M(f(x_0)\mathbf{1}_\Omega) = \rho(f(x_0))\mathbf{1}_\Omega$, then necessarily $Mf(x_0) = M(f(x_0)\mathbf{1}_\Omega)(x_0)$, and hence $Mf(x_0) = \rho(f(x_0))$, which only depends upon the value of f at x_0.

Since C^∞, compact support functions are dense in $\mathbf{L}^2(\mathbb{R}^d)$ and M is Lipschitz continuous, for any $f \in \mathbf{L}^2(\mathbb{R}^d)$ and $\epsilon > 0$ we can find $f_0 \in C^\infty$ such that

$$\|Mf - Mf_0\| = \|f - f_0\| < \epsilon \ ,$$

and hence Mf can be approximated by a pointwise operator with arbitrary precision, and as a result $Mf(x) = \rho(f(x))$ almost everywhere for all $f \in \mathbf{L}^2(\mathbb{R}^d)$. □

B Proof of Theorem 4.4

B.1 Orthogonal Haar Scattering Consistency

We start by considering the case of discrete stationary processes with white autocorrelation and progressive Haar Scattering.

Let X be a random variable. We define the progressive Haar Scattering of X recursively as follows.

$$\begin{aligned} X_0 &\overset{d}{=} X \ , \\ X_{j+1,2k} &\overset{d}{=} \frac{Y+Z}{2} \ , \ X_{j+1,2k+1} \overset{d}{=} \frac{|Y-Z|}{2} \ , \text{ where } Y, \ Z \overset{d}{=} X_{j,k} \text{ independent.} \end{aligned} \tag{86}$$

We shall prove the following:

Theorem B.1. *If X is a random variable with finite energy $\mathrm{e}|X|^2$, then the progressive Haar scattering representation satisfies*

$$\lim_{J\to\infty} \sum_{k=0}^{2^J} \mathrm{var}\,[X_{J,k}] = 0 \ . \tag{87}$$

Proof: Let us suppose first that X is bounded, i.e $P(X > 2^M) = 0$ for some constant M. Without loss of generality we can assume that X is also positive.

We will base the proof of (87) on the Effron-Stein Inequality:

Lemma B.2. *Let Ω be a set and let $g : \Omega^n \to \mathbb{R}$ be a measurable function of n variables. Let $Z = g(X_1, \ldots, X_n)$ where the X_i are independent. Let $X'_1, \ldots, X'_n$ be an independent copy of the X_i and denote $Z'_i = g(X_1, \ldots, X'_i, X_{i+1}, \ldots, X_n)$. Then*

$$\mathrm{var}\,[Z] \leq \frac{1}{2} \sum_i \mathrm{e}(Z - Z'_i)^2 \ . \tag{88}$$

In particular, if a function $g : \Omega^n \to \mathbb{R}$ is uniformly bounded, in the sense that there exist constants $c_1, \ldots, c_n$ such that

$$\sup_{x_1,\ldots,x_n,x'_i\in\Omega} |g(x_1, \ldots, x_n) - g(x_1, \ldots, x'_i, x_{i+1}, \ldots, x_n)| \leq c_i \ ,$$

then it results from (88) that

$$\mathrm{var}\,[Z] \leq \frac{1}{2} \sum_{i=1}^{n} c_i^2 \ . \tag{89}$$

By construction, each of the random variables $X_{J,k}$ is a function of 2^J independent copies of X, via the nonlinear map

$$g_{J,k}(x_1, \ldots, x_{2^J}) = \frac{1}{2^J}(\ldots || \ldots x_k \pm x_{k+1}| \pm \ldots |x_l \pm x_{kl+1}| \ldots | \ldots) \ ,$$

with the particular choice of absolute values determined by the binary decomposition of k. Let us write the path $p(k) = (i_1, \ldots, i_J)$ using such binary decomposition. This suggests using the Effron-Stein inequality directly on the functions $g_{J,k}$, but the uniform bound is not effective in this case, since we obtain $c_i = \frac{M}{2^J}$ in that case, and therefore we do not exploit the contractive aspect of the nonlinearity in (86).

However, we can choose how to decompose each variable $X_{J,k}$ as a function of some of its ancestor variables in the recurrence tree. For that purpose, we represent each node of the recurrence tree of (86) as a path defined on the integer lattice

$$\Lambda = \{(j, m)\,;\, j, m \in \mathbb{Z}, 0 \leq m \leq j \leq J\} \ .$$

A node (j, m) intersects all scattering paths that at level j have gone through exactly m modulus nonlinearities, that is

$$p(k) \text{ intersects } (j, m) \Longleftrightarrow \sum_{j'\leq j} i_{j'} = m \ .$$

By observing that

$$x_1, x_2 \in \left[0, 2^M\right] \Rightarrow \frac{|x_1 - x_2|}{2} \in \left[0, \frac{2^M}{2}\right] \ ,$$

it results that a node (j, m) offers the possibility to represent any path that intersects it using 2^{J-j} independent variables taking values in $[0, 2^{M-m}]$. Applying (89) from node (j, m) produces constants $c_i = 2^{M-m+j-J}$ and therefore a bound

$$\sum_{i=1}^{2^{J-j}} 2^{2(M-J)+2(j-m)} = 2^{2M-J} 2^{j-2m} \ .$$

Thus, for each of the variables $X_{J,k}$ at the bottom layer J, we can select the best bound from all the ancestor nodes intersecting the corresponding path $p(k)$. The lines $\{(j,m) \in \Lambda,\, j-2m = \delta\}$ determine the regions where each bound is successively improved by 2^δ. The total variance $\sum_{k=0}^{2^J} \mathrm{var}\,[X_{J,k}]$ thus converges to

$$2^{2M} 2^J 2^{-J} \exp 2^{-\max r(k)} \ ,$$

where r is a random walk with no drift of J steps in the lattice Λ_J and the expectation is taken over the uniform distribution of random walks. Since the random walk has nonzero escape probability (since, for large J, the fluctuations of typical k are of the order of $\sqrt{J}/2$ by simple application of the Central Limit Theorem), we conclude that

$$\lim_{J\to\infty} \exp 2^{-\max r(k)} = 0 \ .$$

Finally, we extend the result for random variable X not necessarily bounded. Consider a sequence $X_n \to X$ of random variables defined as $X_n = X \cdot \mathbf{1}(X < c_n)$, where $c_n \to \infty$ It follows that $\exp |X_n - X|^2 \to 0$ as $n \to \infty$. We have

$$\mathrm{var}\,[X_{J,k}] \leq 2\mathrm{var}\,[(X_n)_{J,k}] + 2\exp |X_{J,k} - (X_n)_{J,k}|^2 \ ,$$

and

$$\sum_{k=0}^{2^J} \exp |X_{J,k} - Y_{J,k}|^2 \leq \exp |X - Y|^2$$

by induction over J and thanks to the non-expansive property of (86). Hence

$$\sum_{k=0}^{2^J} \mathrm{var}\,[X_{J,k}] \leq 2\sum_{k=0}^{2^J} \mathrm{var}\left[(X_n)_{J,k}\right] + 2\sum_{k=0}^{2^J} \exp |X_{J,k} - (X_n)_{J,k}|^2 \leq 2\sum_{k=0}^{2^J} \mathrm{var}\,[(X_n)_{J,k}] + 2\exp |X - X_n|^2$$

and both terms converge to 0 as J and n go to ∞ This concludes the proof. □

B.2 Extension to non-Orthogonal Haar

Another interpretation of the previous theorem is in terms of a white discrete process $Y[n]$, i.e. such that $Y[1], \ldots, Y[n], ..$ are iid. We now extend the previous proof to handle the case where either $Y[n]$ is a process not necessarily white, or the scattering is performed with oversampling.

The extension is based in the following simple modification of the Effron-Stein lemma.

Lemma B.3. *Suppose that $X_1, \ldots, X_n$ defined in Ω are random variables with the property that there exists $\Delta > 0$ such that X_i is independent of X_j for all $|j - i| > \Delta$. Let $Z = g(X_1, \ldots, X_n)$ with g measurable, and we assume that there exist constants $c_1, \ldots, c_n$ such that*

$$\sup_{x_1,\ldots,x_n,x_i' \in \Omega} |g(x_1, \ldots, x_n) - g(x_1, \ldots, x_i', x_{i+1}, \ldots, x_n)| \leq c_i \ ,$$

Then

$$\mathrm{var}\,[Z] \leq \frac{\Delta}{2} \sum_i^n c_i^2 \ . \tag{90}$$

In other words, the lemma is still valid if we replace the original n degrees of freedom by their effective number of degrees of freedom $\frac{n}{\Delta}$.

Proof: We use the same martingale argument as in [Lugosi, Bousquet, p??] and we use the same notation for convenience. We verify that the same definition of martinagles V_i also satisfies $\mathrm{e}V_iV_j = 0$ in our setting.

The only difference comes when bounding

$$\sum_i \mathrm{e}V_i^2$$

We verify that $\mathrm{e}V_i^2 \leq \Delta c_i^2$ by realizing that the probability space where the martingale takes values is a cartesian product of Δ copies of Ω, thus we can bound $\sup_{x,x' \in \Omega^\Delta} |g(x) - g(x')|^2$ with Δc_i^2 $\square$.

Using equation (90) we can extend the previous theorem to the case where Y is a linear stationary process of the form

$$Y = Y_0 * h \ ,$$

where Y_0 is a white process with finite energy and h is a compact support filter.

Similarly, introducing redundancy in the scattering becomes harmless in light of (90), since adding an oversampling of δ increases n to $n\delta$ but reduces c_i to c_i/δ, resulting in the same variance bound.

B.3 Proof of Proposition 4.6

If X is such that $S_J X$ is mean square consistent, then the process $X_j = |X \star \psi_j|$ also yields a mean square consistent scattering representation, since for each J

$$\begin{aligned} \sum_{p \in \mathcal{P}_J} \mathbb{E}(|S_J[p]X_j - \overline{S}X_j(p)|^2) &= \sum_{p \in \mathcal{P}_J} \mathbb{E}(|S_J[j+p]X - \overline{S}X(j+p)|^2) \\ &\leq \sum_{p \in \mathcal{P}_J} \mathbb{E}(|S_J[p]X - \overline{S}X(p)|^2) \ , \end{aligned}$$

which implies that $\lim_{J \to \infty} \mathbb{E}(\|S_J[\mathcal{P}_J]X_j - \overline{S}X_j\|^2) = 0$. As a result,

$$\mathbb{E}(|X \star \psi_j|^2) = \sum_{p \in \mathcal{P}_\infty} |\overline{S}X_j(p)|^2 = \sum_{p \in \mathcal{P}_\infty} |\overline{S}X(j+p)|^2 \ . \tag{91}$$

$\square$.