On the Distribution of Scattering Coefficients in the Context of Image Processing

Alexander Braun

Born 23rd March 1995 in Fürth, Germany

10th September 2019

Master's Thesis Mathematics

Advisor: Prof. Dr. Massimiliano Gubinelli

Second Advisor: Dr. Martin Lenz

Institut für Angewandte Mathematik

Mathematisch-Naturwissenschaftliche Fakultät der

Rheinischen Friedrich-Wilhelms-Universität Bonn

Abstract

Image generation is a task of primary importance in the area of signal processing. Given a signal representation Φx, the aim is to exploit its statistical properties by drawing a sample $(\Phi x)'$ and then create an image using an inversion of Φ. Consequently, gaining a deeper insight into the statistical behavior of the embedding is highly attractive.

We begin by getting access to a signal representation which produces remarkable results in digit classification tasks, called the scattering transform, invented by S. Mallat and his team. During the development in chapter 2, we will pass by the Fourier transform as well as wavelets which build the foundation of the scattering transform. The latter is computed by iterating on wavelet convolutions, followed by non-linear modulus operators and finalized by an averaging. In chapter 3, we will summarize desireable characteristics of the scattering transform such as contractivity, translation invariance or stability with respect to the action of diffeomorphisms. Further, the structural similarities of the scattering transform to deep convolutional neural networks are discussed. Turning our view towards the statistical properties of the scattering coefficients in chapter 4, we compare their distribution to its dependence on the transformed image. We generate images as realizations of random variables and evaluate numerical experiments by making use of the Kolmogorov-Smirnov test as well as a discussion of the empirical moments. Since taming a distribution by Gaussianization is a common goal in signal processing, we compare the distribution of scattering representations to the one of a Gaussian in several environments. Therefore, we allow a whitening of the scattering vector by its covariance as well as a normalization after rotation.

It turns out that each scattering coefficient follows a particular law, no matter what kind of image is used as input. Further, the hypothesis, that the distribution is Gaussian, can be rejected since scattering coefficients show a slightly right-skewed behavior. But still the law is reasonably close to a normal distribution, by far closer than the one of a Fourier-modulus or wavelet representation. This allows to remove the skewness for the majority of scattering coefficients by rotation.

Acknowledgements

I would first like to thank Prof. Dr. Massimiliano Gubinelli for providing the topic and advising my thesis. Whenever I ran into trouble, he pushed my work into the right directions and gave inspiring ideas on possible proceedings without restricting my own progression.

Further, I would like to acknowledge and thank Dr. Martin Lenz as the second reader of my thesis.

Additionally, I very gratefully thank Benjamin Zaslansky, Maximilian Gläser, Frederik Brüning and Melina Schwabach for the valuable discussions and comments concerning my work.

Last but not least, I would like to thank Prof. Stephane Mallat and his team for their work on the scattering transform, which allowed me to gain a magnificent insight into the world of signal processing.

Contents

List of Tables

List of Figures

1 Image Classification Tasks

A fundamental topic in the area of image processing is the task of classification. Consider one-channel images as vectors $x \in \Gamma \subseteq \mathbb{R}^D$. Assume a set of images $\{x_1, ..., x_T\}$ and a map f connecting each image with its class index $c \in \mathcal{C}$ to be given. The aim is to approximate f for images x which are not assigned to a class yet, given the set of training data $\{x_i, f(x_i)\}_{i \leq T}$. The case of handwritten digits which should be classified can be considered as an illustrating example. Usually, Γ is a high-dimensional subset of $\mathbb{R}^D$, often of dimension larger than 10^6, cf. [39], whereas there is no hope for having training data in the same order, calling directly the curse of dimensionality. Hence, classification of images in high dimensional spaces demands a sensitive understanding of the signals as well as an appropriate monitoring of variability in the data. A dimension reduction in suitable directions can further help to overcome the curse of dimensionality. This motivates the search of a variable $\Phi(x)$ which is supposed to simplify the classification task. A naturally arising question is if and how a suitable signal representation $\Phi(x)$ can be found, such that for all images $x, x' \in \Gamma$ the implication

$$f(x) \neq f(x') \Rightarrow \Phi(x) \neq \Phi(x')$$

holds. As a first remark, having a *margin condition* on Φ of the form

$$\exists \epsilon > 0 \ \forall x, x' \in \Gamma \text{ s.t. } f(x) \neq f(x') \ : \|\Phi(x) - \Phi(x')\| \geq \epsilon$$

ensures the distinguishability, cf. [39]. Considering images x as discretizations of functions $x(t)$ for a spatial coordinate $t \in \mathbb{R}^2$, the problem of finding a suitable representation $\Phi(x)$ extends to the task of finding an operator Φ acting on $x(t)$ with the aim of gaining valuable information concerning the classification.

When determining reasonable requirements on the operator Φ, first, in order to avoid small perturbations on images to create a large discrepancy in the target representation, the operator should be non-expansive, i.e.

$$\|\Phi x - \Phi x'\| \leq \|x - x'\| \ .$$

Further, since translation does usually not affect the class index, the operator should be translation invariant. Referring to [11, 38], we define translation invariance as follows: Given an operator L_c such that $L_c x(t) = x(t - c)$ for $c \in \mathbb{R}^d$ and $x \in L^2(\mathbb{R}^d)$, then Φx is translation invariant if

$$\Phi L_c x = \Phi x$$

for all $c \in \mathbb{R}^d$ and all $x \in L^2(\mathbb{R}^d)$. This becomes more transparent when taking the example of the position of a digit in an image as illustrated in

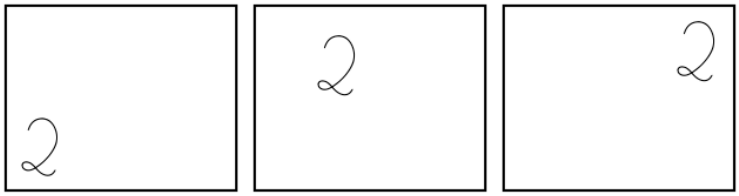

Figure 1: Translation of a handwritten digit 'two'

figure 1 into account, where all images show the exactly same object and should consequently be assigned to one class.

Additionally, the class index of an image is usually also uneffected by small deformations as displayed in figure 2. Therefore, the operator is supposed to satisfy a stability criterion with respect to the action of diffeomorphisms.

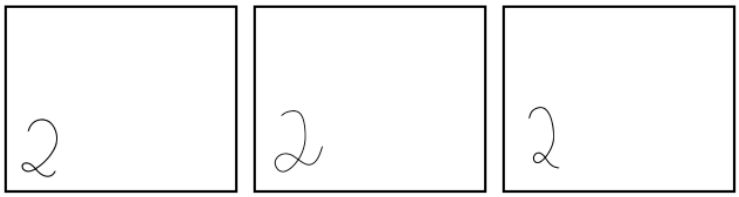

Figure 2: Deformation of a handwritten digit 'two'

We consider diffeomorphisms $\tau \in \mathrm{Diff}(\mathbb{R}^d)$. For a function $x \in L^2(\mathbb{R}^d)$, the operator L_τ acts on x by $L_\tau x(t) = x(t - \tau(t))$. Further, let $\|\tau\|$ be a metric on the space of diffeomorphisms. Referring to [11, 38], stability with respect to these deformations is given by a Lipschitz continuity condition with respect to this metric, i.e.

$$\|\Phi x - \Phi L_\tau x\| \leq C \, \|x\| \, \|\tau\|$$

for all $x \in L^2(\mathbb{R}^d)$ and all $\tau \in \mathrm{Diff}(\mathbb{R}^d)$. When regarding C^2-diffeomorphisms τ, we denote by $|\tau(t)|$ the Euclidean norm on $\mathbb{R}^d$ and by $|\nabla\tau(t)|$ be the supremum norm of the Jacobian $\nabla\tau(t)$. For the Hessian tensor $H\tau(t)$, the supremum norm is denoted by $|H\tau(t)|$. The norm $\|\tau\|$ to measure the deformation over a compact $\Omega \subseteq \mathbb{R}^d$ is then given by

$$\|\tau\| = \sup_{t\in\Omega} |\tau(t)| + \sup_{t\in\Omega} |\nabla\tau(t)| + \sup_{t\in\Omega} |H\tau(t)| \ .$$

Given a translation-invariant operator Φ, we define Lipschitz-continuity to the action of C^2-diffeomorphisms as follows: For any compact $\Omega \subseteq \mathbb{R}^d$ there

is a constant C such that for any function $x \in L^2(\mathbb{R}^d)$ supported in Ω and any diffeomorphism $\tau \in C^2(\mathbb{R}^d)$ the following holds:

$$\|\Phi x - \Phi L_\tau x\| \leq C \|x\| \left(\sup_{t \in \Omega} |\nabla \tau(t)| + \sup_{t \in \Omega} |H\tau(t)| \right) .$$

As a side remark, due to the translation invariance, the supremum norm of τ is not occuring in the bound on the right-hand side. Further, also stability with respect to scaling phenomena and rotation can be taken into account.

All these conditions lead to the consequence that encoding the important information of images in a representation Φx on the one hand and getting rid of superfluous variability on the other hand is one of the main difficulties when tackling the task of image classification. Therefore, in chapter 2, a suitable representation Φx is developed whose characteristics are regarded in chapter 3. Statistical properties of this representation are investigated in chapter 4 to form a basis for a possible application in image generation environments as in [2].

2 From Fourier to Scattering

In order to obtain a transform that satisfies the desired characteristics as translation invariance or stability with respect to deformations, a reasonable approach can be found in tackling the problem from a frequency point of view. Therefore, we begin with a short review of the most important properties of the Fourier transform in this chapter and progress by introducing wavelets as the corner-stone of the wavelet transform. Further, we use wavelets as the foundation of the scattering transform at the end.
Let us start by fixing some notation, cf. [19]. We use $\int$ for the integral over a whole domain Ω (in our cases mostly $\mathbb{R}^d$) rather than for the indefinite integral. For functions $x, y \in \mathcal{L}^1(\mathbb{R}^d)$ we denote by $x \star y(t) = \int_{\mathbb{R}^d} x(u)y(t-u)du$ the convolution of x and y. For complex-valued $x, y \in \mathcal{L}^2(\mathbb{R}^d)$, the scalar product is denoted as $< x, y >= \int \overline{x(t)}y(t)dt$.

2.1 Frequency Analysis via Fourier Transform

When talking about frequency analysis, a natural approach can be found in the Fourier transform. Given the canonical scalar product defined by $< s, t > := \sum_{k=1}^{d} s_k t_k \quad =: s \cdot t$ for vectors $s, t \in \mathbb{R}^d$, we know that for all functions $x \in \mathcal{L}^1(\mathbb{R}^d)$ and all $\xi \in \mathbb{R}^d$, the map $t \mapsto x(t)e^{-i<t,\xi>}$ also is in $\mathcal{L}^1(\mathbb{R}^d)$. This holds true since $|x(t)| = |x(t)\exp(-i < t, \xi >)|$. Hence the Fourier transform can be defined as follows.

Definition 1. (Fourier Transform) [19]
Given a function $x \in \mathcal{L}^1(\mathbb{R}^d)$ and $\xi \in \mathbb{R}^d$, the Fourier transform of x is given by

$$\widehat{x}(\xi) := \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} x(t)e^{-i<t,\xi>}dt$$

where $\widehat{x} : \mathbb{R}^d \to \mathbb{C}$.

Referring to [18, 19, 43, 49], the Fourier transform satisfies a well-known class of beneficial properties some of which are quickly reviewed in the following useful theorems.

At first, there exists the Riemann-Lebesgue-Lemma for the Fourier transform.

Theorem 2. (Riemann-Lebesgue-Lemma)
Let $x \in L^1(\mathbb{R}^d)$. Then

$$\lim_{|\xi| \to \infty} \widehat{x}(\xi) = 0 \ .$$

Further, we would like to summarize some useful properties.

Theorem 3. (Properties of the Fourier Transform)
Let $\widehat{x}, \widehat{y}$ describe the Fourier transforms of $x, y \in \mathcal{L}^1(\mathbb{R}^d)$. Then the following statements hold:

- $\xi \mapsto \widehat{x}(\xi)$ *is continuous*
- $\widehat{x}$ *is uniformly bounded for all* ξ *in the sense that* $|\widehat{x}(\xi)| \leq \frac{1}{(2\pi)^{d/2}} \|x\|_1$
- $\widehat{x \star y} = (2\pi)^{d/2} \widehat{x}\widehat{y}$
- *the functions* $\widehat{x}y$ *and* $x\widehat{y}$ *are integrable and* $\int_{\mathbb{R}^d} x(t)\widehat{y}(t)dt = \int_{\mathbb{R}^d} \widehat{x}(\xi)y(\xi)d\xi$
- *if* $\frac{\partial x}{\partial t_j}$ *exists and is integrable, then* $\widehat{\frac{\partial x}{\partial t_j}}(\xi) = i\xi_j \widehat{x}(\xi)$

As a consequence, the Fourier transform defines a map from $L^1(\mathbb{R}^d)$ to $C_0(\mathbb{R}^d)$, where $C_0(\mathbb{R}^d)$ denotes the set of continuous functions vanishing at infinity.

The construction of the extension of the Fourier transform to the space of square-integrable functions is skipped here in favor of mentioning theorems for the energy conservation and for the inversion of the Fourier transform.

Theorem 4. (Parseval's Theorem)
Let $x \in \mathcal{S}(\mathbb{R}^d)$, where $\mathcal{S}(\mathbb{R}^d)$ denotes the Schwartz space of rapidly decreasing functions, then

$$\int_{\mathbb{R}^d} |x(t)|^2 \, dt = \int_{\mathbb{R}^d} |\widehat{x}(\xi)|^2 \, d\xi \ .$$

Theorem 5. (Inversion of the Fourier Transform)
Let $x \in L^1(\mathbb{R}^d)$ such that $\widehat{x} \in L^1(\mathbb{R}^d)$. Then for all $t \in \mathbb{R}^d$, x can be reconstructed from its Fourier transform via

$$x(t) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \widehat{x}(\xi) e^{i<\xi,t>} d\xi \ .$$

Next, we analyze if the Fourier transform satisfies the desired properties introduced in section 1. First, consider the translation invariance property. Therefore, let L_c for $c \in \mathbb{R}^d$ be the operator that translates a function by c, i.e.

$$L_c x(t) = x(t - c)$$

for $x \in L^2(\mathbb{R}^d)$. When computing the Fourier transform of a function translated by c, applying a simple change of variables leads to the following equation, cf. [18, 19, 43]:

$$\begin{aligned}\widehat{L_c x}(\xi) &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} x(t-c) e^{-i<t,\xi>} dt = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} x(v) e^{-i<v+c,\xi>} dv = \\ &= e^{-i<c,\xi>} \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} x(v) e^{-i<v,\xi>} dv = e^{-i<c,\xi>} \widehat{x}(\xi)\end{aligned}$$

As a consequence, the Fourier transform as such is not invariant under translation. Taking the absolut value of the Fourier transform leads to a representation of the function x which is translation invariant: [1]

$$\left|\widehat{L_c x}(\xi)\right| = \left|e^{-i<c,\xi>}\widehat{x}(\xi)\right| = |\widehat{x}(\xi)|$$

Disadvantageously, the Fourier transform is accompanied by a couple of drawbacks, which are explained further now.

In the first place, the Fourier transform does not allow any spatial localization at all as illustrated in the following. To see this, consider the one dimensional case $d = 1$ and the corresponding time-frequency domain. Note that the notion of spatial domain and real space are used equivalently as well as frequency domain and Fourier space. Define by

$$D(x) := \int t^2 \, |x(t)|^2 \, dt$$

the second moment of the function $|x(t)|^2$ around 0 as in [21, 37, 43, 48]. Then we can state the following principle that relates a function with its Fourier transform.

Theorem 6. (Uncertainty principle) [43]
For an absolutly continuous, complex-valued function $x \in L^2(\mathbb{R})$ such that $t \cdot x(t)$ and $x'(t)$ are both in $L^2(\mathbb{R})$, the following uniform bound holds:

$$D(x) \cdot D(\widehat{x}) \geq \frac{1}{16\pi^2} \ .$$

Equality can only be achieved in the case of $x(t) = C_1 e^{-\pi \frac{t^2}{\sigma^2}}$ for some $\sigma > 0$ and $C_1 = \frac{\sqrt[4]{2}}{\sqrt{\sigma}}$. Then the Fourier transform of x can be computed to be $\widehat{x}(\xi) = \sigma C_1 e^{-\pi\sigma^2\xi^2}$. [43]

When generalizing to the d-dimensional case, the lower bound is determined by $\frac{d^2}{16\pi^2}$ and the integrals on the left hand side are taken over $\mathbb{R}^d$.

For any function x we would like to have highly accurate spatial and frequency information. As a consequence of the uncertainty principle, if we wanted to reduce the variability in the frequency domain, the accuracy in the spatial variable becomes worse. In the limit, where we have perfect localization in frequency, we cannot count on any localization in the spatial domain anymore. This causes severe problems in classification tasks, since here we are not only interested in the occuring frequencies, but also their

position of appearance. Consider the order of digits as an example, where 91 is different to 19.

To tackle another characteristic, the global behavior of the Fourier transform is highly dependent on the behavior of the function in small open sets, cf. [43]. Changing the function values in a small area might have a heavy impact on the Fourier representation.

Additionally, the Fourier transform provides instabilities with respect to deformation in high frequencies. Similar to the translation case, let L_τ be the operator that deforms a function by a small diffeomorphism τ satisfying $\|\nabla\tau\|_\infty = \sup_{t\in\mathbb{R}^d} |\nabla\tau(t)| < 1$, i.e.

$$L_\tau x(t) = x(t - \tau(t)) \ .$$

Using this, we want to have a look at the following example in order to illustrate the instabilities.

Example 7.
Let us consider the one dimensional case and let τ be given by $\tau(t) = \epsilon t$, where $0 < \epsilon \ll 1$. As a consequence, $\|\nabla\tau\|_\infty = \sup_{t\in\mathbb{R}} |\nabla\tau(t)| = \epsilon < 1$. Now, again by a change of variables, the Fourier transform of $L_\tau x$ turns out to satisfy the following:

$$\begin{aligned}\widehat{L_\tau x}(\xi) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x(t - \tau(t)) e^{-i\xi t} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x((1-\epsilon)t) e^{-i\xi t} dt = \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x(u) e^{-i\xi \frac{u}{1-\epsilon}} \frac{1}{1-\epsilon} du = \frac{1}{1-\epsilon}\ \widehat{x}\left(\frac{\xi}{1-\epsilon}\right)\end{aligned}$$

For low frequencies, i.e. ξ small, we have that $\frac{\xi}{1-\epsilon} \approx \xi$. Hence, the Fourier transform of the deformed function $\widehat{L_\tau x}(\xi)$ is close to the Fourier transform of the original function $\widehat{x}(\xi)$. In contrast, for high frequencies, i.e. ξ large, $\frac{\xi}{1-\epsilon} > \xi$. This leads to the observation that $\widehat{L_\tau x}(\xi)$ can be far apart from $\widehat{x}(\xi)$.

Summing up, the deformation with a small diffeomorphism can introduce a large deviation in the corresponding Fourier representations. In other words, the Fourier transform is very unstable in high frequencies. [1]

But equivalently to the localization discussion before, in classification problems, a small deformation does usually not really affect the class index of the input. Consider for example handwritten digits. Even when written by the same person, these digits will never look identical, but rather slightly modified and deformed. Nonetheless, the same digit still belongs to the same class and should therefore be assigned to the same class index. A lack

of control over these deformations might have a heavy impact on the class index here. Hence, satisfying the desired stability conditions with respect to these small deformations is mandatory for the representation Φx.

2.2 Short Time Fourier Transform

Naturally, the question arises how the mentioned problems or disadvantages can be eradicated. Concerning the lack of spatial localization, we will next give a quick intuition-gaining overview of the so-called Short Time Fourier Transform (STFT), similar to [17, 21, 29, 37], focusing on its decomposition of the time-frequency domain in order to control the localizaton lack introduced by the uncertainty principle. Note that we restrict to the case of one dimension.

The main idea is to establish a window function w of compact support of length T. Further, the window function should be normalized such that $\int w(u)du = 1$. Having a look at a function $x \in L^2(\mathbb{R})$ through the window function w allows to compute

$$\widetilde{x}(\xi, t_0) \coloneqq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} w(t - t_0)x(t)e^{-i\xi t}dt \ .$$

This is what can be called a Short Time Fourier Transform. The support of the window function determines the accuracy of the STFT in the spatial domain, the uncertainty principle specifies the corresponding resolution in the frequencies.

Now, consider again the translation operator L_c. Then by a short computation, we can compare the STFT of the translated version of a function to the STFT of the original one.

$$\begin{aligned}
\sqrt{2\pi}\left|\widetilde{L_c x}(\xi, t_0)\right| &= \left|\int x(t-c)w(t-t_0)e^{-i\xi t}dt\right| = \\
&= \left|\int x(v)w(v-(t_0-c))e^{-i\xi(v+c)}dv\right| = \\
&= \left|e^{-i\xi c}\int x(v)w(v-(t_0-c))e^{-i\xi v}dv\right| = \sqrt{2\pi}\,|\widetilde{x}(\xi, t_0-c)|
\end{aligned}$$

Hence, shifting a function x by some c is equivalent to shifting the window function w by the same amount. Even further, for c satisfying $|c| \ll T$, this shift does not heavily affect the STFT of the function, i.e. $\left|\widetilde{L_c x}(\xi, t_0)\right| \approx |\widetilde{x}(\xi, t_0)|$, cf. [21].
The discussion on the inversion of the STFT is skipped here and the reader is referred to [21, 29].

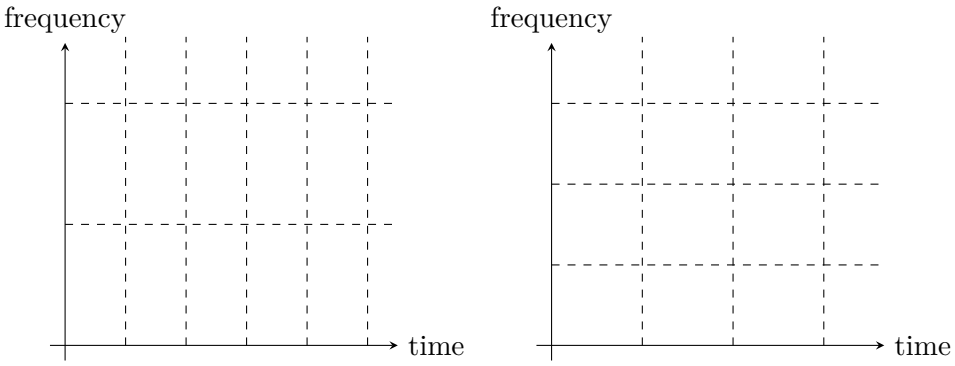

Figure 3: Decomposition of the time-frequency domain using the STFT, cf. [21]

Taking a closer look on the STFT, we can recognize that the STFT introduces a fixed window size via the window function w. Due to the uncertainty principle, there is the mentioned lower bound for the corresponding accuracy in the frequency domain which leads to a grid decomposition of the time-frequency domain as illustrated in figure 3.

This fixed resolution is responsible for some disadvantages of the STFT. Transforming a signal that has important frequency details which are too small or too large with respect to the window size T can lead to severe problems in detection. As an example, consider a window size that is spatially highly inaccurate, but therefore gains accuracy in the frequency domain. Assume further a signal, that has one high peak which is only occuring on a very small domain. A proper detection of this becomes more and more difficult, since there is only one STFT coefficient for a spatial domain which may be much larger than the neighborhood of the peak.

Consequently, using the STFT instead of the Fourier transform allows better time and spatial localization properties. But nonetheless, due to the fixed window size, there are problems remaining. Gaining detailed information on high frequencies while not losing the possiblity to detect low frequencies properly is still a heavy task for the STFT.

2.3 Wavelets and the Wavelet Transform

In contrast to the STFT where the Fourier transform was modulated to construct a transform with more powerful properties suiting our task, we are now trying to replace the sine and cosine waves that are used as basis functions in the Fourier transform. Instead, we would like to introduce basis functions that are compactly supported. Further, we still want to keep time

and frequency localization as a key characteristic of our transform.

During the last decades, wavelets and the wavelet transform arose in the context of frequency analysis. Before going into detail, we first want to give some examples and possible approaches for constructing wavelets in order to gain intuition on their shape and behavior.

As in [8, 30], wavelets can be constructed by solving dilation equations of the form

$$h(t) = \sum_{k=0}^{N-1} c_k h(at - k) \ ,$$

which is called a factor-of-2-reduction if $a = 2$. A solution to the dilation equation is called a scaling function ϕ. Shifting and scaling this function leads to a family of functions $\phi_{j,k}(t) = \phi(2^j t - k)$.

Since the dilation equation creates a natural form of dependence among the different $\phi_{j,k}$, we create the so-called mother wavelet ψ as a linear combination of suitable $\phi_{j,k}$ by

$$\psi(t) := \sum_{k=0}^{N-1} b_k \phi(2t - k)$$

in order to reduce dependencies. We illustrate this with some examples.

Example 8. (Haar wavelets)
Given the dilation equation

$$h(t) = h(2t) + h(2t - 1)$$

with the solution

$$\mathbb{1}_{[0,1)}(t) =: \phi(t) \ ,$$

the Haar wavelet can be constructed by

$$\psi(t) = \phi(2t) - \phi(2t - 1) \ .$$

The mother wavelet (see figure 4) is consequently determined by the equation $\psi(t) = \mathbb{1}_{[0,1/2)}(t) - \mathbb{1}_{[1/2,1)}(t)$, cf. [16].

Example 9. (Daubechies wavelets)
Another family of wavelets of considerable practical importance is named after the Belgian physicist and mathematician Ingrid Daubechies. Given a dilation equation of the form

$$h(t) = \frac{1+\sqrt{3}}{4} h(2t) + \frac{3+\sqrt{3}}{4} h(2t-1) + \frac{3-\sqrt{3}}{4} h(2t-2) + \frac{1-\sqrt{3}}{4} h(2t-3) \ ,$$

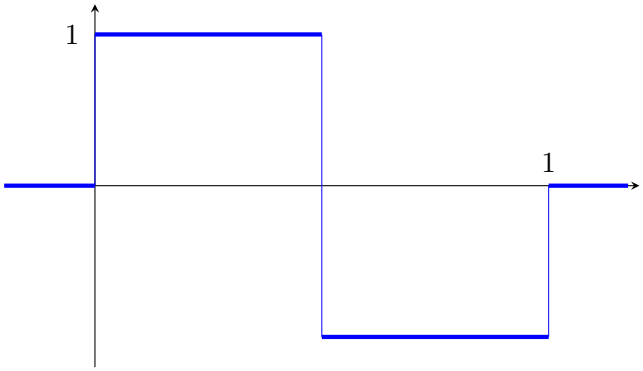

Figure 4: Haar wavelet $\psi(t)$

the mother wavelet is supposed to be constructed in an equivalent way as the Haar wavelet shown before, cf. [8]. Note that this dilation equation is not known to have an analytic solution so far. Therefore, wavelets are computed by approximating this solution, see figure 5.

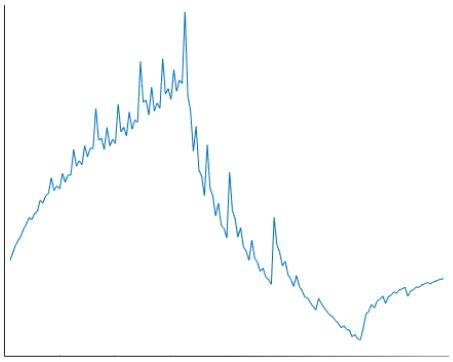

Figure 5: Scaling function for Daubechies wavelet, cf. [8]

Another possible approach to create wavelets is simply by modifying a sine wave to compactify its support, e.g. by multiplication with a Gaussian. This leads to the characteristic local structure. Further, wavelets can be obtained in many other ways, as for example the so-called Mexican hat wavelet which is simply the normalized and negative second derivative of a Gaussian, i.e. given by $\psi(t) = \frac{2}{\sqrt{3\sigma^2}\sqrt[4]{\pi}}\left(1-\left(\frac{t}{2}\right)^2\right)e^{-\frac{t^2}{2\sigma^2}}$, illustrated in figure 6.

As discussed in [38], for the ongoing development towards the scattering

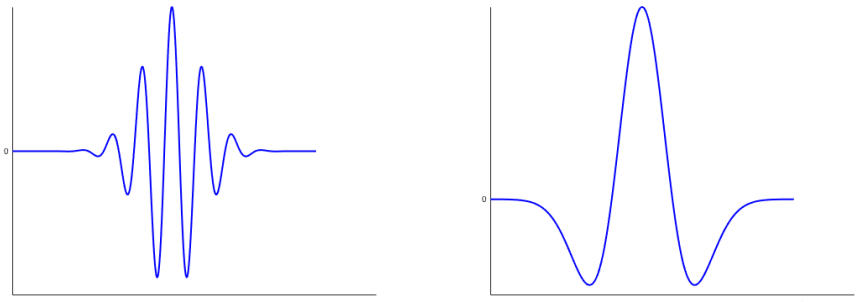

Figure 6: Gaussian Wavelet on the left and Mexican Hat Wavelet on the right, cf. [17]

transform, complex wavelets of the form

$$\psi(t) = e^{i\eta \cdot t}\Theta(t)$$

are used, where $\widehat{\psi}(\xi) = \widehat{\Theta}(\xi - \eta)$ and $\widehat{\Theta}$ is real-valued in a low-frequency ball centered at 0. As a consequence, $\widehat{\psi}$ has its center at η, i.e. we allow a shift in the frequency domain.

For the moment we interrupt the discussion concerning the creation of wavelets and turn towards the wavelet transform of an input signal x which leads us one step closer to the scattering transform.

Following the procedure from [12, 38, 53], the wavelet transform is constructed in the pursuing way.

The mother wavelet $\psi \in L^1 \cap L^2(\mathbb{R}^d)$ plays the key role in the transform combined with a sequence of scaling values $(a^j)_{j\in\mathbb{Z}}$ for some $a > 1$. In image processing, usually $a = 2$, so without loss of generality we set $a = 2$ to simplify notation. We get a family of wavelets of the form $\psi_{2^j}(t) = 2^{dj}\psi(2^j t)$. Note that the dilation of the mother wavelet is rescaled such that $\|\psi_{2^j}\|_2 = \|\psi\|_2$, cf. [17, 43].

In the case of dimensions $d \geq 2$, in addition to the scaling parameter, also a rotation parameter r is introduced in order to rotate the dilated wavelet with elements r of a finite discrete rotation group $G \subset SO(d)$ of $\mathbb{R}^d$, cf. [11]. This leads to dilated and rotated wavelets of the form

$$\psi_{2^j r}(t) = 2^{dj}\psi(2^j r^{-1} t) \ ,$$

i.e. $\psi_{2^j r}$ is the 2^j-dilated and r-rotated version of ψ. Further, denote by $\lambda := 2^j r \in 2^{\mathbb{Z}} \times G =: \Lambda$ with $|\lambda| = 2^j$ the parameter of dilation and rotation

of the wavelet.

This way, we create a family of wavelets $\{\psi_\lambda\}_{\lambda\in\Lambda}$ which we can now use to express our signal x with. Note that $\|\psi_\lambda\|_1 = \|\psi\|_1$, cf. [38]. When calculating the wavelet transform of a signal x, we compute the convolution of the signal with translated versions of dilated wavelets, cf. [17, 43].

Using the notation of [38], a so-called Littlewood-Paley wavelet transform is defined as

$$\forall\, t \in \mathbb{R}^d \;:\; W[\lambda]x(t) \coloneqq x \star \psi_\lambda(t) = \int x(u)\psi_\lambda(t-u)du$$

for all choices of $\lambda \in \Lambda$.

When focusing on frequencies, we can continue in the following way. First, note that in the Fourier domain by theorem 3 and a simple change of variables $\widehat{\psi_\lambda}(\xi) = \widehat{\psi}(\lambda^{-1}\xi)$ and hence, up to a constant of $\gamma = (2\pi)^{d/2}$,

$$\widehat{W[\lambda]x}(\xi) = \gamma\widehat{x}(\xi)\cdot\widehat{\psi_\lambda}(\xi) = \gamma\widehat{x}(\xi)\cdot\widehat{\psi}(\lambda^{-1}\xi)$$

which simply combines the frequencies of the wavelet transform as the product of the frequency of the input and the used wavelet.

Further, following [38], if we only consider wavelets of frequencies $2^j > 2^{-J}$ for $J \in \mathbb{Z}$, we can compute a wavelet transform at a scale 2^J. Therefore, denote by $\Lambda_J \coloneqq \{\lambda = 2^j r : r \in G, 2^j > 2^{-J}\}$ and compute $W[\lambda]x$ for $\lambda \in \Lambda_J$. Note that the low frequencies are not covered when taking only wavelets of higher frequencies into account. To fill this gap, we can compute an averaging with a kernel ϕ to cover the low frequencies, i.e.

$$A_J x \coloneqq x \star \phi_{2^J}$$

for $\phi_{2^J}(t) = 2^{-dJ}\phi(2^{-J}t)$ which averages over a spatial domain proportional to 2^J. This leads to a wavelet transform given by

$$W_J x \coloneqq \{A_J x, (W[\lambda]x)_{\lambda\in\Lambda_J}\} \tag{1}$$

with norm $\|W_J x\|^2 = \|A_J x\|^2 + \sum_{\lambda\in\Lambda_J}\|W[\lambda]x\|^2$. Hence the wavelet transform defines a linear operator on $L^2(\mathbb{R}^d)$, cf. [11]. Note that for $J = \infty$, $\Lambda_\infty = 2^{\mathbb{Z}} \times G = \Lambda$, so we are back in the previous setup such that $W_\infty x = \{W[\lambda]x\}_{\lambda\in\Lambda}$ with norm $\|W_\infty x\|^2 = \sum_{\lambda\in\Lambda}\|W[\lambda]x\|^2$.

If the domain of the input x is scaled and rotated, the wavelet transform is modified in a corresponding way. Assume $2^l g \in \Lambda = 2^{\mathbb{Z}} \times G$ and consider the scaled and rotated version of x, i.e. $2^l g \circ x(t) = x(2^l g t)$. By the application of a simple change of variables in the wavelet transform, we can see that

$$W[\lambda]\left(2^l g \circ x\right) = 2^l g \circ W[2^{-l}g\lambda]x \text{ , cf. [38].} \tag{2}$$

At a glance, computing the wavelet transform of a signal x gives a representation of wavelet coefficients of the form $\{W[\lambda]x\}_{\lambda\in\Lambda}$. Hereby, we filter the input x by means of the wavelet family $\{\psi_\lambda\}_{\lambda\in\Lambda}$ which consists of dilated and rotated versions of the mother wavelet ψ.

Since our motivation for following up the wavelet transform was to overcome the problems of the (Short Time) Fourier transform, we now compare the two in the upcoming paragraphs.

Let us start with the comparison of the decomposition of the time-frequency domain of STFT and the wavelet transform in the one dimensional case. When having a look at the scaling of wavelets, we can recognize that in each iteration, wavelets are doubling their height and halven their width in contrast, cf. [21]. Wavelets therefore have the same number of oscillations, but are scaled or dilated in the respective length and amplitude of the oscillation, cf. [5], as illustrated in figure 7.

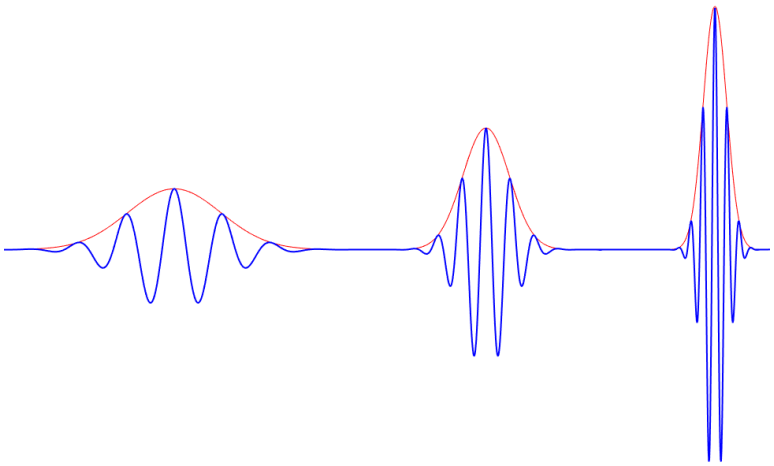

Figure 7: Envelope of wavelets, cf. [5]

In contrast, when having a look at the windowed sine waves of the STFT, the number of oscillations is increasing in a fixed support range, see figure 8 for an example with a Gaussian window function.

These observations lead to a different decomposition of the time-frequency domain for wavelets than for the STFT.

According to [21], the wavelet transform decomposes the time-frequency domain as shown in figure 9. Due to the scaling of the mother wavelet, the wavelet transform is very accurate in low frequencies, but has a lack of accurancy in the time or spatial component. This does not really matter since low frequencies occur over larger regions, so the focus in low frequencies is

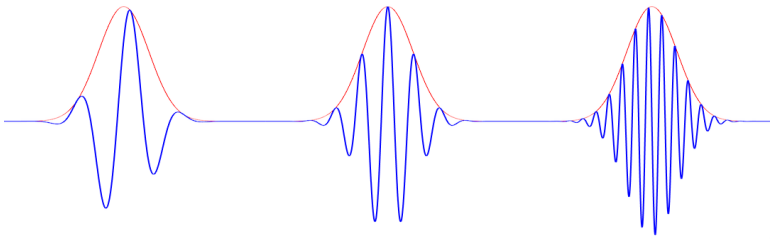

Figure 8: Envelope of the STFT, cf. [5]

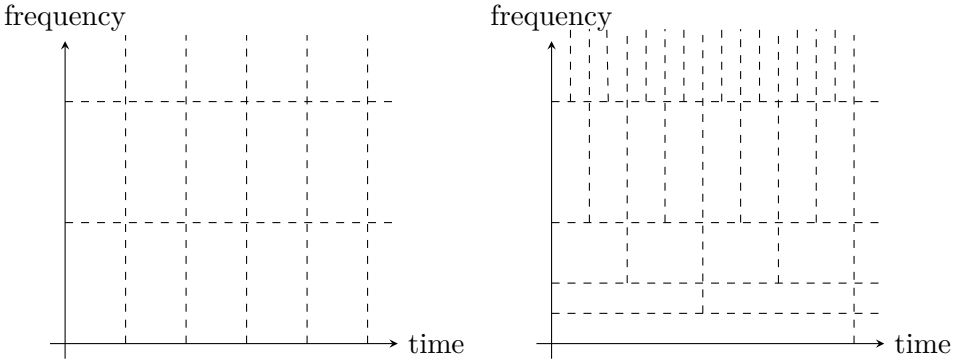

Figure 9: Comparing the decomposition of the time-frequency domain using the STFT on the left and the wavelet transform on the right

more on the existence than on the spatial localization. In contrast, in the high frequencies the wavelet transform provides a highly accurate representation of the signal in the time or spatial domain which allows a more precise localization of the considered frequency band.

After having talked about wavelets and the wavelet transform, we next would like to check if the wavelet transform fulfills the desired properties as defined at the beginning and hence evaluate, if the wavelet transform is a candidate for the representation in the classification task.

Therefore, consider again the translation operator L_c and the translated input $L_c x$. This allows the following computation:

$$W[\lambda]L_c x(t) = L_c x \star \psi_\lambda(t) = \int L_c x(u)\psi_\lambda(t-u)du = \int x(u-c)\psi_\lambda(t-u)du =$$
$$= \int x(v)\psi_\lambda((t-c)-v)dv = x \star \psi_\lambda(t-c) = L_c W[\lambda]x(t)$$

Hence, the wavelet transform is not invariant with respect to our definition of translation invariance, cf. [38], but instead commutes with translation. Consequently, when translating the input x, its wavelet representation

translates as well.

To overcome this, recall that the Lebesgue measure is translation invariant, i.e. $\mu(A) = \mu(A + c)$ for all $c \in \mathbb{R}^d$, all Borel sets $A \subseteq \mathbb{R}^d$ and $A + c = \{a + c : a \in A\}$. We can make use of this observation to create a translation invariant operator Φ. Therefore, consider an operator U on $L^2(\mathbb{R}^d)$ that commutes with translation. In the event that the integral $\int Ux(t)dt$ exists, it is translation invariant with respect to translations of the input function x since

$$\int U(L_c x)(t)dt = \int L_c(Ux)(t)dt = \int Ux(t)dt \ . \tag{3}$$

Note that the operator U is not necessarily linear here. [38]

Consequently, averaging the wavelet transform creates the desired invariance, i.e.

$$\int W[\lambda]x(t)dt = \int x \star \psi_\lambda(t)dt$$

is a translation invariant representation of the input x.

Unfortunately, wavelets are accompanied by some severe drawbacks when trying to average. Since in the literature wavelets are defined in many different ways, we interrupt and focus on this characteristic of wavelets for a short moment.

Referring to [16, 36, 43], wavelets are usually required to satisfy a so-called admissibility condition. It was originally introduced by Calderon, Grossmann and Morlet in [13, 23]. In the one dimensional case, this condition is given by

$$C_\psi := \int_0^{+\infty} \frac{\left|\widehat{\psi}(\xi)\right|^2}{\xi} d\xi < +\infty \ ,$$

where in some references the integral is taken over the whole real line with the denominator containing also a modulus. In any of the two, this characteristic enables the wavelet transform to be invertible and hence allows the recovery of $x(t)$ from its wavelet represenation.
As a side effect, in order to avoid the intergal in C_ψ diverging, the singularity at 0 needs to be removed. Hence, the Fourier transform in the numerator of C_ψ also needs to vanish. In the case of ψ integrable, the Fourier transform is continuous which implies that $\widehat{\psi}(0) = 0$, cf. [17, 43]. This then leads to

$$\int \psi(t)dt = \int \psi(t)e^{-i<t,0>}dt = (2\pi)^{d/2}\widehat{\psi}(0) = 0 \ .$$

In [38], instead of an admissibility condition for wavelets, the following unitary condition for the wavelet transform is introduced, given by a condition on the Fourier transform of the mother wavelet.

Lemma 10. (Unitary wavelets)
Let $J \in \mathbb{Z}$ or $J = \infty$, W_J as described in equation 1. Then the following holds: W_J is unitary in the space of real-valued functions in $L^2(\mathbb{R}^d)$ if and only if

$$\frac{1}{2} \sum_{j=-\infty}^{\infty} \sum_{r \in G} \left| \widehat{\psi}(2^{-j} r^{-1} \xi) \right|^2 = 1$$

and

$$\left| \widehat{\phi}(\xi) \right|^2 = \frac{1}{2} \sum_{j=-\infty}^{0} \sum_{r \in G} \left| \widehat{\psi}(2^{-j} r^{-1} \xi) \right|^2$$

for almost every $\xi \in \mathbb{R}^d$.

The proof is skipped here and the reader is referred to proposition 2.1. in [38]. Nonetheless, also this unitary condition leads to the fact that $\widehat{\psi}(0) = 0$ as before implying the same zero-average for wavelets.

Summing up, any wavelet is accompanied by a zero average, which will play an important role in the ongoings. Additionally, we impose that ψ is twice differentiable and that its decay and the decay of the first and second partial derivatives is of order $\mathcal{O}((1 - |t|)^{-(d+2)})$.

Now, knowing that wavelets integrate to 0, we can compute the following by a change of the order of integration and using the fact that the integral over the whole space is uneffected by translation.

$$\begin{aligned} \int W[\lambda]x(t)\ dt &= \int x \star \psi_\lambda(t)\ dt = \int \int x(u)\psi_\lambda(t-u)\ du\ dt = \\ &= \int \int x(u)\psi_\lambda(t-u)\ dt\ du = \int x(u) \int \psi_\lambda(t-u)\ dt\ du = \\ &= \int x(u) \cdot 0\ du = 0 \end{aligned}$$

Consequently, averaging the wavelet transform of a function leads to a zero representation. Even further, any linear operator applied to $W[\lambda]x$ that is translation invariant is zero, cf. [15, 38]. As a consequence, when trying to create a translation invariant representation by a simple average of the wavelet transform or any other linear operator, all information of the input function x is lost. This problem is tackled in the next section. For further reading concerning wavelets, see [17, 26, 37, 41].

2.4 From Wavelets to the Scattering Transform

Since the wavelet transform as such is not really appropriate for the image classification task due to the discussion of the previous section, we are trying to extend or modify the wavelet transform in a suitable way to gain the desired stabilites and invariances.

Making use of equation (3) for creating a translation invariant representation, we would like to modulate the wavelet transform $W[\lambda]x$ by joining an operator $M[\lambda]$ such that $U \coloneqq M[\lambda]W[\lambda]$ satisfies the translation invariant property from (3). We first argue what kind of characteristics this operator $M[\lambda]$ should fulfill.

Initially, in order to avoid small perturbations to create a large discrepancy in the target domain, we would like the operator to be non-expansive. Further, to satisfy a Lipschitz condition, the operator should also be commutative with respect to the actions of diffeomorphisms. This implies stability with respect to small deformations. And finally, the properties discussed in the introduction should also be fulfilled. The following theorem helps us to satisfy these requirements.

Theorem 11. *[11, 12]*
Let M be an operator acting on $L^2(\mathbb{R}^d)$ satisfying

- *non-expansiveness, i.e. $\|Mx - My\| \leq \|x - y\|$ and*
- *commutativity with respect to diffeomorphisms,*

then M is a pointwise operator almost everywhere.

Proof. [11, 12]
The proof mainly follows the idea of a monotone class argument. We start proving the statement for indicator functions $\mathbb{1}_\Omega(t)$ for a compact $\Omega \subseteq \mathbb{R}^d$, extend this to the case of C^∞-functions with compact support Ω and conclude by the density of C^∞ with compact support in $L^2(\mathbb{R}^d)$.

So let $\Omega \subseteq \mathbb{R}^d$ be compact, $x \in L^2(\mathbb{R}^d)$ and $\tau \in \mathrm{Diff}(\mathbb{R}^d)$ be a diffeomorphism where $\mathcal{L}_\tau x = x \circ \tau$. Further, by $G(x) = \{\tau \in \mathrm{Diff}(\mathbb{R}^d) : \mathcal{L}_\tau x = x \text{ a.e.}\}$ denote the isotropy group of x. Then

$$\tau \in G(x) \Rightarrow \tau \in G(M(x)) \ , \tag{4}$$

since $\|Mx - \mathcal{L}_\tau Mx\| = \|Mx - M\mathcal{L}_\tau x\| \leq \|x - \mathcal{L}_\tau x\| = 0$ exploiting the commutativity and the non-expansive property. Hence, any diffeomorphism which leaves x unchanged also leaves Mx unchanged. Now look at $x = c\mathbb{1}_\Omega$. Then $G(x)$ contains all τ such that

$$\tau(\Omega) = \Omega \text{ and } \tau(\Omega^c) = \Omega^c \ , \tag{5}$$

where $\Omega^c = \mathbb{R}^d \setminus \Omega$ is the complement of Ω, since otherwise, τ would affect values of x. Combining 4 and 5, Mx also needs to be constant on Ω and Ω^c which implies that $Mx(t) = 0$ for all $t \in \Omega^c$ since Mx is square-integrable. This leads to

$$Mx = M(c(\mathbb{1})_\Omega) = (Mc\mathbb{1}_\Omega)(t_0)\mathbb{1}_\Omega$$

for any $t_0 \in \Omega$. Consequently, M is pointwise in this case.

So let now $x \in C^\infty$ be compactly supported in Ω and $t_0 \in \Omega$. Further consider $(\tau_n)_{n\in\mathbb{N}} \subseteq \mathrm{Diff}(\mathbb{R}^d)$ such that

$$\lim_{n\to\infty} \|\mathcal{L}_{\tau_n} x - x(t_0)\mathbb{1}_\Omega\| = 0 \ , \tag{6}$$

i.e. the diffeomorphisms are constructed by leaving $t \in \Omega^c$ unchanged and extending a ball $B_{t_0}(2^{-n})$ of radius 2^{-n} around t_0 to the whole domain Ω, e.g. by taking straight lines through t_0 that are more and more contracted the closer they approach the boundary $\delta\Omega$. This way,

$$\tau_n : B_{t_0}\left(2^{-n}\right) \to \{t \in \Omega : \mathrm{dist}(t, \Omega^c) \geq 2^{-n}\} \ .$$

Since $x \in C^\infty$ and has compact support, it is bounded which allows the following computation:

$$\|\mathcal{L}_{\tau_n}(Mx) - M(x(t_0))\mathbb{1}_\Omega\| = \|M(\mathcal{L}_{\tau_n} x) - M(x(t_0))\mathbb{1}_\Omega\| \leq \|\mathcal{L}_{\tau_n} x - x(t_0)\mathbb{1}_\Omega\|$$

Using equation 6, this implies L^2-convergence of $\mathcal{L}_{\tau_n}(Mx)$ to $M(x(t_0))\mathbb{1}_\Omega$. Since M is pointwise on constant functions $x(t_0)\mathbb{1}_\Omega$ by the first part of the proof, and the sequence of diffeomorphism τ_n expands the neighborhood of t_0 to the whole domain Ω, we obtain that M is pointwise on this class of functions, too.

Concluding by the density of compactly supported C^∞-functions in L^2 and M being Lipschitz continuous, we can conclude that M is pointwise a.e. for all $x \in L^2$.

□

After having proved that $M[\lambda]$ needs to be a pointwise operator, following [38], we can further require a norm preserving property for the operator $M[\lambda]$, i.e.

$$\|M[\lambda]x\|_2 = \|x\|_2 \ \text{ for all } x \in L^2(\mathbb{R}^d) \ .$$

This implies that $|M[\lambda]x| = |x|$ leading to the fact that $M[\lambda]$ is necessarily a modulus operator, i.e.

$$M[\lambda]x = |x| \ ,$$

so consequently all possible phase variation is eliminated by the non-linear operator.

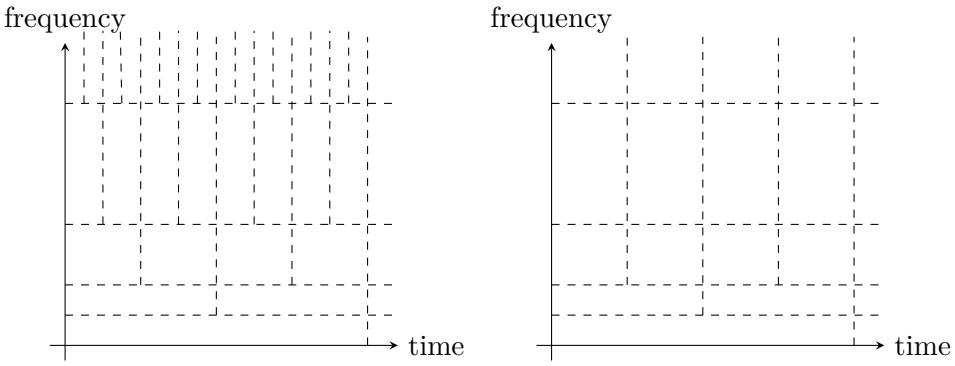

Figure 10: Decomposition of the time-frequency domain via wavelets on the left and averaged on the right

In order to create our desired representation of the input signal x, we now integrate $Ux \coloneqq M[\lambda]W[\lambda]x = |x \star \psi_\lambda|$ to establish translation invariance:

$$\int Ux(t)dt = \int |x \star \psi_\lambda(t)| \, dt \ ,$$

where instead of averaging over the whole domain we could also use an averaging kernel ϕ here which only involves a certain region.

When thinking back to figure 9, this averaging creates a huge loss of information in the high frequency region. But when distinguishing different kinds of signal characters, access to high-frequency information plays a key role. The wavelet transform provided a representation which was very accurate in the spatial domain for high frequency bands. This accuracy is lost in exchange for gaining translation invariance. For illustration purposes, consider the time-frequency domain as in figure 10, where the averaging is done with an averaging kernel only over an interval of length T instead of the whole domain.

The question arises how to recover the lost information and accuracy in the high frequencies. A possible answer can be found by iterating on computing wavelet coefficients, modulus operation and averaging - which is then called a scattering transform.

3 Scattering Transform

Iterating on this wavelet-modulus-averaging procedure, we gain what is called a scattering transform, invented by S. Mallat and his team. This representation of an input signal x is supposed to satisfy the characteristics developed in chapter 1. In the following, we start by its definition, having a closer look at its deep convolutional network structure, its properties as translation invariance or stability to small deformations and finish by some discussion concerning the invertibility of this transform.

As mentioned at the end of the previous section, the lack of spatial accuracy in the high frequency region of the averaged wavelet-modulus representation of the input x can be overcome by iterating on applying the wavelet-modulus operator U. To clarify notation, denote by $U[\lambda] \coloneqq |W[\lambda]|$ the wavelet operator followed by a modulus. Now, for any scale λ_1 we first compute $|W[\lambda_1]x| = |x \star \psi_{\lambda_1}|$. On the one hand, we average to create the translation invariance, i.e. by calculating the convolution with an averaging kernel $|x \star \psi_{\lambda_1}| \star \phi$. On the other hand, we want to recover the accuracy at high frequencies which was lost by the averaging, so we take a second parameter λ_2 and compute the wavelet convolution for all those scales, i.e.

$$||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|$$

what needs to be averaged to $||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi$ to gain translation invariance.

Iterating on this procedure for all scales $\lambda_1, ..., \lambda_m$ leads to the definition of the scattering transform. But let us start in a more formal way and continue step by step, referring to [12, 27, 38, 39].

Definition 12. (Path) [38]
Let $\lambda_k \in 2^{\mathbb{Z}} \times G$ for $k \in \mathbb{N}$ be the scaling and rotation parameter of a wavelet, where G describes a finite rotation group. Then we define the following:

- *An ordered sequence $p = (\lambda_1, ..., \lambda_m)$ is a path.*
- *The empty path is denoted by $p = \emptyset$.*

$\Rightarrow$ *Iterative application of the non-commutative operators $U[\lambda_k]$ is called scattering propagator denoted by $U[p] = U[\lambda_m]...U[\lambda_2]U[\lambda_1]$ with $U[\emptyset] = Id$.*

Note that the operator $U[p]$ is well-defined on $L^2(\mathbb{R}^d)$. To see this, apply Young's Theorem, cf. [10], iteratively on $U[\lambda]x$ which leads to

$$\|U[\lambda]x\|_2 \leq \|\psi_\lambda\|_1 \|x\|_2$$

for all $\lambda \in \Lambda$. Further, to fix some notation, let $\mathcal{P}_\infty$ be the set of all paths of finite length. For $p = (\lambda_1, ..., \lambda_m) \in \mathcal{P}_\infty$, $\lambda \in \Lambda$ we denote by $p + \lambda$ the path $(\lambda_1, ..., \lambda_m, \lambda) \in \mathcal{P}_\infty$.

Having this scattering propagator $U[\lambda]$, we can define the scattering transform.

Definition 13. (Scattering Transform) [11, 12, 38]
Let $x \in L^1(\mathbb{R}^d)$. Then for any $p \in \mathcal{P}_\infty$, the scattering transform of x along p is defined as

$$Sx(p) = \frac{1}{\mu_p} \int U[p]x(t)dt \ ,$$

where the normalization constant is given by $\mu_p = \int U[p]\delta(t)dt$ for a Dirac $\delta(t)$.

Note, that for the scattering propagator $U[p]x$, we get that $\|U[p]x\|_1 \leq \|x\|_1 \|\psi\|_1^m$, since wavelets were designed to satisfy $\|\psi\|_1 = \|\psi_\lambda\|_1$. As a consequence, the integral in the definition of the scattering transform is finite. For conditions to have $\mu_p \neq 0$ and hence to ensure the scattering transform to be well-defined, we refer to [38].

Due to the varying amount of iterative applications of wavelet convolutions and modulus operators, the scattering transform has a similar shape as a deep convolutional neural network (DCNN) which is described in the following.

3.1 Deep Convolutional Network Structure

Let us start with a short explanation of DCNNs, cf. [32, 39]. Given an input x, a DCNN applies at each step a composition of three layers:

- a filter bank layer
- a non-linearity layer (e.g. sigmoids, rectifiers, modulus,...)
- and a pooling layer (e.g. averaging each point over some neighborhood).

Thinking of the computation of a scattering transform, the filtering is done via a wavelet convolution and the modulus application afterwards corresponds to the non-linearity layer. Note that the pooling is left out in the scattering transform.

In classical DCNNs, the output is only influenced by the last layer. Hence there is a successive application of different linear and non-linear operators which all progress towards one single output layer. Additionally, the filters

are learned by backpropagation algorithms using a loss function and a set of training data on which this function is minimized, cf. [6, 47]. In comparison to this, the scattering transform averages at every level, hence creates output coefficients at each layer. Further, the filters of the scattering transform are fixed in advance, since the scattering transform filters the input x with the help of given wavelets, cf. [1, 32]. To get an intuition for this behavior of the scattering transform, figure 11 illustrates the interaction of the different operators.

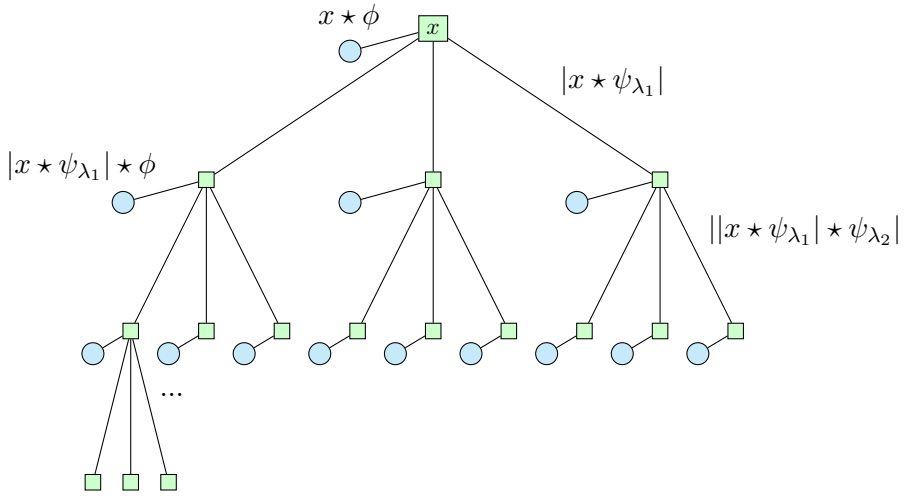

Figure 11: DCNN structure of the scattering transform, cf. [1, 11, 38]. The green squares represent the wavelet convolution modulus coefficients $U[\lambda_1, \lambda_2, ..., \lambda_m]x$. The blue nodes describe the output of the scattering transform: the averaged coefficients $Sx(\lambda_1, \lambda_2, ..., \lambda_m)$.

As a consequence, in contrast to DCNNs, the scattering transform needs to know the desired invariances a priori. In our case, we were aiming for a translation invariant, deformation stable representation of the input x which led to the given construction of the scattering transform.

3.2 Properties of Scattering Transform

After having briefly looked at the structure of a scattering transform, we will discuss its properties. Let us start with the modification of the amplitude of the input function x, i.e. consider $\mu \in \mathbb{R}$ and the scaled input μx. Out of this, we can state the following lemma.

Lemma 14. (Amplitude preservation) [38]
Let $\mu \in \mathbb{R}$. Then for any $p \in \mathcal{P}_\infty$ such that $p \neq \emptyset$, the scattering transform

of x along p preserves the amplitude modification, i.e.

$$S(\mu x)(p) = |\mu| \, Sx(p) \ .$$

Proof. First, we compute that

$$\begin{aligned} |\mu x \star \psi_\lambda(t)| &= \left| \int \mu x(u) \psi_\lambda(t-u) du \right| = \left| \mu \int x(u) \psi_\lambda(t-u) du \right| = \\ &= |\mu| \left| \int x(u) \psi_\lambda(t-u) du \right| = |\mu| \, |x \star \psi_\lambda(t)| \ . \end{aligned} \tag{7}$$

Applying this equation iteratively to each wavelet convolution within the scattering transform and using the linearity of the integral proves the statement. □

Hence, modifying the input function in its amplitude is carried through when applying the scattering operator.

Secondly, we would like to be able to control scaling and rotation of the input x. Therefore, consider the scaling and rotation parameter $2^l g \in 2^{\mathbb{Z}} \times G$ and the scaled and rotated version $2^l g \circ x(t) = x(2^l g t)$. From this, we can state the upcoming lemma, where we use the notation $2^l gp$ for a path $p = (\lambda_1, ..., \lambda_m) \in \mathcal{P}_\infty$ to abbreviate $(2^l g\lambda_1, ..., 2^l g\lambda_m) \in \mathcal{P}_\infty$.

Lemma 15. (Scaling and rotating) [38]
For a path $p \in \mathcal{P}_\infty$ and $2^l g \in 2^{\mathbb{Z}} \times G$, we have

$$S(2^l g \circ x)(p) = 2^{-dl} Sx(2^{-l} gp) \ .$$

Proof.
We have seen in equation 2 that wavelets satisfy $W[\lambda]\left(2^l g \circ x\right) = 2^l g \circ W[2^{-l} g\lambda]x$. Applying this to the wavelet-modulus operator $U[\lambda] = |W[\lambda]|$, we can see that $U[\lambda]\left(2^l g \circ x\right) = 2^l g \circ U[2^{-l} g\lambda]x$. Iterating on this over a path $p \in \mathcal{P}_\infty$, it follows that

$$U[p]\left(2^l g \circ x\right) = 2^l g \circ U[2^{-l} gp]x \ .$$

We conclude by applying the definition of the scattering transform which is simply integrating the last equation. □

In other words, rotating the input x identically rotates its scattering representation whereas scaling the input by a factor of 2^l forces a path scaling by 2^{-l}.

As further aspects, we would like to check the resistance to additive perturbations, norm preservation, translation invariance and the stability to

diffeomorphisms. Before turning to this, we would like to introduce what is called a windowed scattering transform (WST) in order to widen definition 13. Therefore, we consider an averaging kernel ϕ as in chapter 2.3. Let ϕ be real, symmetric, twice differentiable and invariant with respect to rotations, i.e. $\phi(rt) = \phi(t)$ for $r \in G$. Further, we impose that the decay of ϕ and its first and second partial derivatives is in $\mathcal{O}((1+|t|)^{-(d+2)})$. Then we can define the WST.

Definition 16. (Windowed Scattering Transform) [11, 38]
For $J \in \mathbb{Z}$, denote by $\Lambda_J = \{\lambda = 2^j r \in \Lambda \mid 2^j > 2^{-J}\}$ and $\mathcal{P}_J = \{p = (\lambda_1, ..., \lambda_m) \in \mathcal{P}_\infty \mid \lambda_k \in \Lambda_J \text{ for all } k\}$. Then for $x \in L^1(\mathbb{R}^d)$ the windowed scattering transform is defined as

$$S_J[p]x(t) = U[p]x \star \phi_{2^J}(t) = \int U[p]x(u)\phi_{2^J}(t-u)du$$

for all $p \in \mathcal{P}_J$, where $\phi_{2^J}(t) = 2^{-dJ}\phi(2^{-J}t)$.

The WST now allows local averaging instead of being forced to average over the whole domain. When talking about the relation between WST and the scattering transform, following [38], for $x \in L^1(\mathbb{R}^d)$, the WST converges pointwise to the scattering transform as 2^J goes towards infinity. To see this, take $t \in \mathbb{R}^d$ and consider

$$\lim_{J\to\infty} 2^{dJ} S_J[p]x(t) = \phi(0) \int U[p]x(u)du = \phi(0)\mu_p Sx(p) \ ,$$

where we exploit the continuity of ϕ at 0.

The WST can also be used to extend the scattering transform to the space of square-integrable functions $L^2(\mathbb{R}^d)$. A deeper discussion is skipped here and the interested reader is referred to [38]. To fix some notation, denote by $S_J[\mathcal{P}_J]x \coloneqq \{S_J[p]x\}_{p\in\mathcal{P}_J}$ and $S_J[\Omega]x \coloneqq \{S_J[p]x\}_{p\in\Omega}$ for a subset of paths Ω. In the same way, let $U[\Omega]x \coloneqq \{U[p]x\}_{p\in\Omega}$.

3.2.1 Non-Expansiveness

When trying to avoid small additive perturbations from having a heavy impact on the scattering representation, a non-expansiveness property for the WST needs to be introduced. To do so, for a subset of paths Ω, let $\|S_J[\Omega]x\|^2 \coloneqq \sum_{p\in\Omega} \|S_J[p]x\|^2$ and $\|U[\Omega]x\|^2 \coloneqq \sum_{p\in\Omega} \|U[p]x\|^2$, where $\|.\|$ denotes the L^2-norm in this context.

Theorem 17. (Non-expansiveness) [11, 38]
Let $x, x' \in L^2(\mathbb{R}^d)$. Then

$$\|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x'\| \leq \|x - x'\| \ .$$

Proof. [38]
Consider the one-step propagator

$$U_J x \coloneqq \{A_J x, (U[\lambda]x)_{\lambda\in\Lambda_J}\} \coloneqq \{x \star \phi_{2^J}, (|x \star \psi_\lambda|)_{\lambda\in\Lambda_J}\} \ .$$

Since $U[\lambda]U[p] = U[p+\lambda]$ and $A_J U[p] = S_J[p]$ we obtain

$$U_J U[p]x = \{S_J[p]x, (U[p+\lambda]x)_{\lambda\in\Lambda_J}\} \ .$$

As a side remark, note that U_J is non-expansive since W_J is unitary due to Lemma 10 and so is the modulus operator $|\ .\ |$, i.e. for $a, b \in \mathbb{C}$ we have $||a| - |b|| \leq |a-b|$. Let now $\|U_J x\|^2 = \|A_J x\|^2 + \sum_{\lambda\in\Lambda_J} \|U[\lambda]x\|^2$. This allows the following calculation:

$$\begin{aligned}
\left\|U_J x - U_J x'\right\|^2 &= \left\|A_J x - A_J x'\right\|^2 + \sum_{\lambda\in\Lambda_J} \left\|U[\lambda]x - U[\lambda]x'\right\|^2 = \\
&= \left\|A_J x - A_J x'\right\|^2 + \sum_{\lambda\in\Lambda_J} \left\||W[\lambda]x| - |W[\lambda]x'|\right\|^2 \leq \\
&\leq \left\|A_J x - A_J x'\right\|^2 + \sum_{\lambda\in\Lambda_J} \left\|W[\lambda]x - W[\lambda]x'\right\|^2 = \\
&= \left\|W_J x - W_J x'\right\|^2 \leq \left\|x - x'\right\|^2
\end{aligned}$$

Further, setting $x' = 0$ and performing the same calculation, we can see that $\|U_J x\|^2 = \|x\|^2$ as W_J is unitary.

Let now $\Lambda_J^m \coloneqq \{p \in \mathcal{P}_J \mid |p| = m\}$ where $\Lambda_J^0 = \{\emptyset\}$. Then:

$$U_J U[\Lambda_J^m]x = \{S_J[\Lambda_J^m]x, U[\Lambda_J^{m+1}]x\} \tag{8}$$

Using that $P_J = \cup_{m\in\mathbb{N}} \Lambda_J^m$, the computation of $S_J[P_J]x$ is possible via iteration on $U_J U[\Lambda_J^m]x$ for $m \geq 0$. As a side remark, to acknowledge is the correspondance to the DCNN structure mentioned in the previous chapter. Using equation 8 and exploiting the non-expansiveness of U_J, we obtain the following bound:

$$\begin{aligned}
&\left\|U[\Lambda_J^m]x - U[\Lambda_J^m]x'\right\|^2 \geq \left\|U_J U[\Lambda_J^m]x - U_J U[\Lambda_J^m]x'\right\|^2 = \\
&= \left\|S_J[\Lambda_J^m]x - S_J[\Lambda_J^m]x'\right\|^2 + \left\|U[\Lambda_J^{m+1}]x - U[\Lambda_J^{m+1}]x'\right\|^2
\end{aligned} \tag{9}$$

Combining all this, we can conclude with the help of a telescopic sum:

$$\begin{aligned}
\left\|S_J[P_J]x - S_J[P_J]x'\right\|^2 &= \sum_{m=0}^{\infty} \left\|S_J[\Lambda_J^m]x - S_J[\Lambda_J^m]x'\right\|^2 \leq \\
&\leq \sum_{m=0}^{\infty} \left(\left\|U[\Lambda_J^m]x - U[\Lambda_J^m]x'\right\|^2 - \left\|U[\Lambda_J^{m+1}]x - U[\Lambda_J^{m+1}]x'\right\|^2 \right) \leq \\
&\leq \left\|U[\Lambda_J^0]x - U[\Lambda_J^0]x'\right\|^2 = \left\|U[\emptyset]x - U[\emptyset]x'\right\|^2 = \left\|x - x'\right\|^2
\end{aligned}$$

□

Hence, the scattering representation satisfies a stability condition with respect to additive noise, cf. [1]. To illustrate this, consider $x \in L^2(\mathbb{R}^d)$ and a slightly perturbed version $x + h$. Then

$$\|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J](x+h)\| \leq \|x - (x+h)\| \leq \|h\| \ .$$

As a consequence, the metric introduced by $S_J[P_J]$ satisfies a Lipschitz continuity condition with respect to the Euclidean norm of the perturbing noise, cf. [11].

3.2.2 Norm Preservation

In order to determine the energy which is propagated through the consecutive levels of the scattering transform, the next aspect we are focusing on is a theorem concerning the preservation of the $L^2(\mathbb{R}^d)$ norm.

To do that, we need a technical feature for the used wavelet. Let $\eta \in \mathbb{R}^d$ and $\rho \in L^2(\mathbb{R}^d)$, where ρ is non-negative such that for the averaging kernel ϕ we have $|\widehat{\rho}(\xi)| \leq \left|\widehat{\phi}(2\xi)\right|$ and $\widehat{\rho}(0) = 1$. Further, let

$$\widehat{\Psi}(\xi) \coloneqq |\widehat{\rho}(\xi - \eta)|^2 - \sum_{k=1}^{\infty} k \left(1 - \left|\widehat{\rho}(2^{-k}(\xi - \eta))\right|^2\right)$$

satisfy

$$\alpha \coloneqq \inf_{1 \leq |\xi| \leq 2} \sum_{j=-\infty}^{\infty} \sum_{r \in G} \widehat{\Psi}(2^{-j} r^{-1} \xi) \left|\widehat{\psi}(2^{-j} r^{-1} \xi)\right|^2 > 0 \ . \tag{10}$$

If such an η and ρ exist for a wavelet ψ, then this wavelet is called an admissible scattering wavelet. Achieving this, we can state the desired theorem.

Theorem 18. (Norm preservation) [11, 38]
Let $x \in L^2(\mathbb{R}^d)$. If the used wavelet ψ is an admissible scattering wavelet and the corresponding wavelet transform W_J is unitary, i.e. $\|W_J x\|^2 = \|x\|^2$, then

$$\lim_{m \to \infty} \|U[\Lambda_J^m]x]\|^2 = \lim_{m \to \infty} \sum_{n \geq m} \|S_J[\Lambda_J^n]x\|^2 = 0$$

and further

$$\|S_J[P_J]x\| = \|x\| \ .$$

Proof. [38]
The proof consists of two parts. First we show equivalence of

(A) $\lim_{m \to \infty} \|U[\Lambda_J^m]x]\|^2 = 0$

(B) $\lim_{m\to\infty}\sum_{n\geq m}\|S_J[\Lambda_J^n]x\|^2 = 0$

(C) $\|S_J[P_J]x\| = \|x\|$.

The second part consists of the implication that an admissible scattering wavelet implies (A).

(A) $\Leftrightarrow$ (B)
In the proof of theorem 17 we demonstrated that $\|U_J x\|^2 = \|x\|^2$ for all $x \in L^2(\mathbb{R}^d)$. Further, recall that by definition of U_J we have

$$U_J U[\Lambda_J^n]x = \{S_J[\Lambda_J^n]x, U[\Lambda_J^{n+1}]x\}$$

and hence

$$\|U[\Lambda_J^n]x\|^2 = \|U_J U[\Lambda_J^n]x\|^2 = \|S_J[\Lambda_J^n]x\|^2 + \left\|U[\Lambda_J^{n+1}]x\right\|^2 \ . \qquad (11)$$

Equivalently, we get $\|S_J[\Lambda_J^n]x\|^2 = \|U[\Lambda_J^n]x\|^2 - \left\|U[\Lambda_J^{n+1}]x\right\|^2$ which allows the following when summing over $n \geq m$ for a fixed $m \in \mathbb{N}$ exploiting again the telescopic sum:

$$\sum_{n\geq m}\|S_J[\Lambda_J^n]x\|^2 = \sum_{n\geq m}\left(\|U[\Lambda_J^n]x\|^2 - \left\|U[\Lambda_J^{n+1}]x\right\|^2\right) = \|U[\Lambda_J^m]x\|^2$$

Taking the limit $m \to \infty$ on both sides proves the first equivalence.

(A) $\Leftrightarrow$ (C)
Using equation 11 and summing over $n \leq m-1$ leads to the following:

$$\sum_{n=0}^{m-1}\|S_J[\Lambda_J^n]x\|^2 = \sum_{n=0}^{m-1}\left(\|U[\Lambda_J^n]x\|^2 - \left\|U[\Lambda_J^{n+1}]x\right\|^2\right) = \left\|U[\Lambda_J^0]x\right\|^2 - \|U[\Lambda_J^m]x\|^2 =$$
$$= \|x\|^2 - \|U[\Lambda_J^m]x\|^2$$

Here we made use of the telescopic sum as well as of $x = U[\emptyset]x = U[\Lambda_J^0]x$.
As a consequence, we achieve

$$\|x\|^2 = \sum_{n=0}^{m-1}\|S_J[\Lambda_J^n]x\|^2 + \|U[\Lambda_J^m]x\|^2 \ .$$

For $m \to \infty$ we then get the second equivalence.

The second - rather technical - part of the proof is skipped here, so the reader is referred to Appendix A of [38]. □

Note that as soon as the level index m, i.e. the number of computed wavelet convolutions, gets large enough, the remaining energy in the last network layers $\sum_{n \geq m} \|S_J[\Lambda_J^n]x\|^2$ converges to 0, cf. [12]. Even further, due to numerical simulations, the energy $\|U[\Lambda_J^m]x\|^2$ shows an exponential decay for $m \to \infty$ and hence the energy of the scattering transform converges to the energy of the input x exponentially, cf. [1]. Having this property is mandatory in numerical applications since the network depth can consequently be limited to a finite number of levels without running into the loss of too much information, cf. [12].

3.2.3 Translation Invariance

Initially, one of our main objectives was the creation of a translation invariant representation of the input x. So next, we want to turn our attention towards this property. To achieve this, we use the following proposition.

Proposition 19. [38]
Let $x, x' \in L^2(\mathbb{R}^d)$ and $J \in \mathbb{Z}$, then

$$\|S_{J+1}[P_{J+1}]x - S_{J+1}[P_{J+1}]x'\| \leq \|S_J[P_J]x - S_J[P_J]x'\| \ .$$

Note, that as $\|S_J[P_J]x - S_J[P_J]x'\|$ is non-negative and non-expansive, it converges as $J \to \infty$. Combining this with the non-expansiveness property shown in theorem 17 leads to a non-expansive metric in the limiting case, i.e.

$$\lim_{J \to \infty} \|S_J[P_J]x - S_J[P_J]x'\| \leq \|x - x'\|$$

and also

$$\lim_{J \to \infty} \|S_J[P_J]x\| = \|x\| \ .$$

Exploiting these properties also for the limiting case, we can state the following theorem concerning translation invariance. Therefore, recall the translation operator L_c for $c \in \mathbb{R}^d$ which translates a function x by c, i.e. $L_c x(t) = x(t - c)$.

Theorem 20. [11, 38]
Let $x \in L^2(\mathbb{R}^d)$ and $c \in \mathbb{R}^d$. Further, assume the scattering transform to be computed with admissible scattering wavelets, then

$$\lim_{J \to \infty} \|S_J[P_J]x - S_J[P_J]L_c x\| = 0 \ .$$

Proof. (Sketch) [38]
At first, note that the windowed scattering transform commutes with translations. To see this, let $p \in \mathcal{P}_\infty$ be an arbitrary path. A simple change of

variables allows the following computation:

$$\begin{aligned} S_J[p]L_c x(t) = U[p]L_c x \star \phi_{2^J}(t) &= \int U[p]L_c x(u)\phi_{2^J}(t-u)du = \\ &= \int U[p]x(u-c)\phi_{2^J}(t-u)du = \int U[p]x(v)\phi_{2^J}((t-c)-v)dv = \\ &= U[p]x \star \phi_{2^J}(t-c) = S_J[p]x(t-c) = L_c S_J[p]x(t) \end{aligned}$$

As a consequence, $S_J[P_J]L_c = L_c S_J[P_J]$ and since further $S_J[P_J]x = A_J U[P_J]x$, we can find the following bound using that A_J and $L_c A_J$ are bounded linear operators:

$$\begin{aligned} \|S_J[P_J]L_c x - S_J[P_J]x\| = \|L_c A_J U[P_J]x - A_J U[P_J]x\| \leq \\ \leq \|L_c A_J - A_J\| \, \|U[P_J]x\| \ , \end{aligned} \tag{12}$$

where the operator norm is the usual sup-norm for linear operators.

Secondly, we bound the first factor $\|L_c A_J - A_J\|$. To do this, we use an application of Schur's Lemma for an operator $Kx(t) = \int x(u)k(t,u)du$ by following Appendix B of [38], which states:

$$\int |x(u)k(t,u)| \, du \leq C \ \text{ and } \ \int |x(u)k(t,u)| \, dt \leq C \ \Rightarrow \ \|K\| \leq C$$

This can be used to show that there exists a C such that

$$\|L_c A_J - A_J\| \leq C 2^{-J} |c| \ . \tag{13}$$

As a third part, we aim to bound the second factor $\|U[P_J]x\|$. To do this, we use the following statement, proven in Appendix A of [38]. Let

$$\|x\|_w^2 \coloneqq \sum_{j \geq 0} \sum_{r \in G} j \left\| W[2^j r]x \right\|^2 < \infty$$

and let the wavelets be admissible scattering wavelets, then

$$\frac{\alpha}{2} \|U[P_J]x\|^2 \leq \max(J+1, 1) \|x\|^2 + \|x\|_w^2 \ ,$$

where α is defined as in equation 10.

Applying all this to equation 12, leads to the following bound in the case of $\|x\|_w^2 < \infty$:

$$\|S_J[P_J]L_c x - S_J[P_J]x\|^2 \leq \left((J+1) \|x\|^2 + \|x\|_w^2 \right) \frac{C^2}{\alpha 2^{2J-1}} |c|^2$$

Hence, $\lim_{J \to \infty} \|S_J[P_J]L_c x - S_J[P_J]x\| = 0$.

To extend this case of functions satisfying $\|x\|_w^2 < \infty$ to all $x \in L^2(\mathbb{R}^d)$, we use a density argument and write x as the limit of a sequence of $(x_n)_{n\in\mathbb{N}}$ with $\|x_n\|_w^2 < \infty$ for all $n \in \mathbb{N}$. Exploiting the unitary property of L_c and the non-expansiveness of $S_J[P_J]$, we can conclude by

$$\|S_J[P_J]L_c x - S_J[P_J]x\| \leq \|S_J[P_J]L_c x_n - S_J[P_J]x_n\| + 2\,\|x - x_n\| \;\longrightarrow\; 0$$

as $n \to \infty$.

□

As a consequence, when enlarging the window size of the averaging kernel to the limit, the scattering transform provides a translation invariant representation of the input x.

3.2.4 Lipschitz Continuity with respect to Diffeomorphisms

Now, as we have shown that the windowed scattering transform satisfies a translation invariant criterion, we would like to give a statement concerning a Lipschitz condition with respect to the action of diffeomorphisms. Recall the operator $L_\tau x(t) = x(t - \tau(t))$ for a diffeomorphism τ with $\|\nabla\tau\|_\infty < 1$ and $x \in L^2(\mathbb{R}^d)$. Let $\|\Delta\tau\|_\infty \coloneqq \sup_{u,v\in\mathbb{R}^d} |\tau(u) - \tau(v)|$. Further, $\|U[P_J]x\|_1 = \sum_{m\geq 0} \|U[\Lambda_J^m]x\|$ and $P_{J,m} = \{p \in P_J \mid |p| < m\} = \cup_{n<m}\Lambda_J^n \subseteq P_J$. Using this, we can state the following theorem.

Theorem 21. (Lipschitz condition) [11, 38]
There is a constant C such that for all $x \in L^2(\mathbb{R}^d)$ satisfying $\|U[P_J]x\|_1 < \infty$ and for all $\tau \in C^2(\mathbb{R}^d)$ with $\|\nabla\tau\|_\infty \leq \frac{1}{2}$ the following holds:

$$\|S_J[P_J]L_\tau x - S_J[P_J]x\| \leq C\,\|U[P_J]x\|_1\,\kappa(\tau) \tag{14}$$

Further, for all $m \geq 0$, it holds that

$$\|S_J[P_{J,m}]L_\tau x - S_J[P_{J,m}]x\| \leq Cm\,\|x\|\,\kappa(\tau) \tag{15}$$

for $\kappa(\tau) \coloneqq 2^{-J}\,\|\tau\|_\infty + \|\nabla\tau\|_\infty \max\left(\log\left(\frac{\|\Delta\tau\|_\infty}{\|\nabla\tau\|_\infty}\right), 1\right) + \|H\tau\|_\infty$, where $H\tau$ denotes the Hessian of τ.

Proof. (Sketch) [38]
At first, consider the commutator $[S_J[P_J], L_\tau] \coloneqq S_J[P_J]L_\tau - L_\tau S_J[P_J]$ and start with the following bounds:

$$\|S_J[P_J]L_\tau x - S_J[P_J]x\| \leq \|L_\tau S_J[P_J]x - S_J[P_J]x\| + \|[S_J[P_J], L_\tau]\,x\| \tag{16}$$

Now, we bound the first summand in an equivalent way to equation 12 with the use of $\|U[P_J]x\| = \left(\sum_{m\geq 0} \|U\left[\Lambda_J^m\right]x\|^2\right)^{1/2} \leq \sum_{m\geq 0} \|U\left[\Lambda_J^m\right]x\| = \|U[P_J]x\|_1$. Hence:

$$\|L_\tau S_J[P_J]x - S_J[P_J]x\| \leq \|L_\tau A_J - A_J\|\,\|U[P_J]x\| \leq \|L_\tau A_J - A_J\|\,\|U[P_J]x\|_1$$

Next we want to use a lemma which states for any operator L on $L^2(\mathbb{R}^d)$ the following bound:

$$\|[S_J[P_J], L]\, x\| \leq \|U[P_J]x\|_1 \, \|[U_J, L]\|$$

A proof of this can be found in Appendix D of [38]. Combining this with the fact that $U_J = |W_J|$ and exploiting the non-expansiveness property of $|\,.\,|$, we get

$$\|[U_J, L_\tau]\| \leq \|[W_J, L_\tau]\| \ .$$

Plugging all these together leads to the following:

$$\|S_J[P_J]L_\tau x - S_J[P_J]x\| \leq \|U[P_J]x\|_1 \left(\|L_\tau A_J - A_J\| + \|[W_J, L_\tau]\|\right) \ ,$$

where the term $\|L_\tau A_J - A_J\|$ can be bounded by $C2^{-J}\|\tau\|_\infty$ using the corresponding result of equation 13. A proof can be found in Appendix B of [38].

So the remaining task is to find a suitable bound for the commutator term $\|[W_J, L_\tau]\|$. Therefore we use a lemma proven in Appendix E of [38].

Lemma. *There is $c > 0$ such that for all $J \in \mathbb{Z}$ and for all $\tau \in C^2(\mathbb{R}^d)$ satisfying $\|\nabla\tau\|_\infty \leq \frac{1}{2}$ the following holds:*

$$\|[W_J, L_\tau]\| \leq c\left(\|\nabla\tau\|_\infty \max\left(\log\left(\frac{\|\Delta\tau\|_\infty}{\|\nabla\tau\|_\infty}\right), 1\right) + \|H\tau\|_\infty\right)$$

Again, combining all these leads to the desired bound 14.

In order to extent this to equation 15, note that equation 14 still holds when replacing P_J by $P_{J,m} \coloneqq \cup_{n<m}\Lambda_J^n$, if we substitute $\|U[P_J]x\|_1$ with $\|U[P_{J,m}]x\|_1$. Then we can compute

$$\|U[\Lambda_J^n]x\| \leq \left\|U[\Lambda_J^{n-1}]x\right\| \leq \|x\| \ ,$$

since $U[\Lambda_J^n]x$ is calculated by applying U_J to $U[\Lambda_J^{n-1}]x$ and as discussed in the previous section, U_J is norm-preserving. As a consequence, we obtain:

$$\|U[P_{J,m}]x\|_1 = \sum_{n=0}^{m-1} \|U[\Lambda_J^n]x\| \leq m\,\|x\|$$

This leads to the second bound which concludes the sketch. □

Obtaining this theorem, with the help of the following corollary, we can establish Lipschitz continuity.

Corollary 22. [11, 38]
Let $\Omega \subseteq \mathbb{R}^d$ be compact. Then there is C such that for all $x \in L^2(\mathbb{R}^d)$ with support in Ω satisfying $\|U[P_J]x\|_1 < \infty$ and for all $\tau \in C^2(\mathbb{R}^d)$ with $\|\nabla\tau\|_\infty \leq \frac{1}{2}$, if $2^J \geq \frac{\|\tau\|_\infty}{\|\nabla\tau\|_\infty}$, then

$$\|S_J[P_J]L_\tau x - S_J[P_J]x\| \leq C \|U[P_J]x\|_1 (\|\nabla\tau\|_\infty + \|H\tau\|_\infty) \ .$$

For the case of numerical applications of the scattering transform where only a finite number of levels $m_{\max}$ is calculated, we can use that the Hessian term can be neglected for regular τ, cf. [12], and bound

$$\|S_J[P_J]L_\tau x - S_J[P_J]x\| \leq C m_{\max} \|x\| \|\nabla\tau\|_\infty \ ,$$

which leads to a Lipschitz continuous representation of the input.

3.3 Invertibility and Image Generation

Now, a question that arises naturally relates the scattering operator with its invertibility characteristics. When trying to invert, we are facing the following problem. The scattering transform only outputs the averaged coefficients and not every wavelet convolution coefficient, see figure 11. When iterating through the levels, we create an output at every level and in the same moment calculate the wavelet convolution coefficients of the next level. When terminating this procedure in applications, we lose the last wavelet convolution coefficients. This implies a loss of information which does not allow a direct inversion of the scattering transform, but just enables possible approximations. Further reading on this can be found in [53].

Since the scattering transform has a structure related to a DCNN, we can try to find an inverting operator similar to convolutional network generators, which are used to invert DCNNs. For further reading concerning the inversion of DCNNs, cf. [9, 22, 35].

Therefore, following [2], let Φ be a fixed operator and assume a set of training data $\{x_i\}_{i \leq T}$ of size T to be given. The goal is to find an operator which best approximates the inversion of the embedding $\{\Phi(x_i)\}_{i \leq T}$ on this data set. To do so, let $\mathcal{G}$ be the set of convolutional network generators. We aim to find an inversion operator $G \in \mathcal{G}$ that minimizes a L^1-loss function on the training data, i.e.

$$\widehat{G} := \ \text{argmin}_{G \in \mathcal{G}} \frac{1}{T} \sum_{i=1}^{T} \|x_i - G(\Phi(x_i))\|_1 \ . \tag{17}$$

The inversion operator is consequently created to ensure $\widehat{G}(\Phi(x_i)) \approx x_i$ for all $i \leq T$. An essential feature of this lies in the distribution of the output coefficients of the DCNNs. There is a lot of research done on this, e.g.

[3, 4] for further reading, but still, a lot of questions are not well understood.

Let us now turn towards the scattering transform to which we try to adapt the DCNN inversion task. In order to create flexibility concerning the statistical properties of the scattering transform we allow a whitening of scattering coefficients by a normalization. Following the notation from the preceding procedure, let $\Phi = A\bar{\Phi}$ be this normalization of the windowed scattering transform $\bar{\Phi} = S_J[P_J]$. The created embedding is now supposed to be inverted by an operator G, see figure 12.

$$x \xrightarrow{A\ S_J[P_J]} \Phi x$$

$$x' \xleftarrow{G} (\Phi x)'$$

Figure 12: Inverting the scattering operator $S_J[P_J]$

In order to create the foundation of a suitable approach for determining a valuable inverting operator, we want to gain access to the distribution and the statistical properties of scattering coefficients which the discussion in the upcoming chapter focuses on.

Note that a possible approach to create an approximate inversion operator can be found in [1] or [53].

Further, we can embed this into the context of image generation where we want to exploit the statistical structure of the representation Φx in order to create images. Therefore, assume we possess detailed information concerning the distribution of coefficients in the scattering vector. Knowing the statistical details of representations for a class of images in the scattering domain, we can easily create a sample having the same statistical properties. Applying an inversion operator to this sample, e.g. as determined by equation 17, will create an image out of a sample in the scattering domain.

4 Discussion on Statistics of the Scattering Transform

In order to obtain a deeper insight into the characteristics of scattering coefficients, we evaluate their distribution in several environments. The overall setup is as follows: We start by creating images as realizations of random variables, then we compute the scattering transform for these inputs. Holding the scattering representation of the input images, statistical tests are used for comparisons. The attached code to run the experiments with is implemented in MATLAB. In the process, we compare the scattering coefficients for different classes of input realizations such as Bernoulli and (jointly) Gaussian random variables or realizations of the Ising model. We begin by testing a universality property in order to establish a canonical model for scattering coefficients. Further, since there are several reasons to aim for a Gaussian distribution of scattering coefficients, we evaluate simulations testing the Gaussianization properties of the scattering transform. Therefore, we allow a whitening of scattering coefficients. During this procedure, we compare the results to the Fourier-modulus and wavelet transform, which were both introduced during the derivation of the scattering operator in chapter 2.

4.1 Methods for Numerical Experiments

In the following, we start working through the used methods for the generation of test images as well as the implemented version of the scattering transform, followed by a quick review of suitable statistical methodology and tests.

4.1.1 Image Generation

The position of a pixel in the image is determined by its spatial coordinates t. We concentrate on black-and-white-images $x(t)$ for $t = (t_1, t_2)$, where only one channel is needed in order to determine the pixel's color. As a side remark, note that colored red-green-blue images can be treated in a similar way by introducing a channel variable v for $v \in \{1, 2, 3\}$ so that an image can be denoted by $x(t, v)$.

In the first case, we create images with pixels being realizations of a Bernoulli distribution. Hence, we simply toss a coin independently for each t which determines the color as 0 or 1, i.e.

$$x(t) \sim \text{Bern}\left(\frac{1}{2}\right) \text{ for all } t \ .$$

We focus on images of size $N \times N$ for $N = 64$ and $1 \leq t_1, t_2 \leq N$. Examples are illustrated in figure 13.

Figure 13: Images as realizations of an i.i.d. Bernoulli distribution

The same procedure is run for independent and identically distributed (i.i.d.) standard Gaussian random variables, where the color is now in $\mathbb{R}$ instead of $\{0, 1\}$. Hence,

$$x(t) \sim \mathcal{N}(0, 1) \text{ for all } t = (t_1, t_2)$$

and again $1 \leq t_1, t_2 \leq N$. Examples of the resulting images are presented in figure 14. Note that although printed in colors, we still work with one-channel images, where yellow colors indicate highly positive while blue colors state highly negative values. In what follows, when talking about a Gaussian distribution without further parameters, we always mean the standard normal distribution.

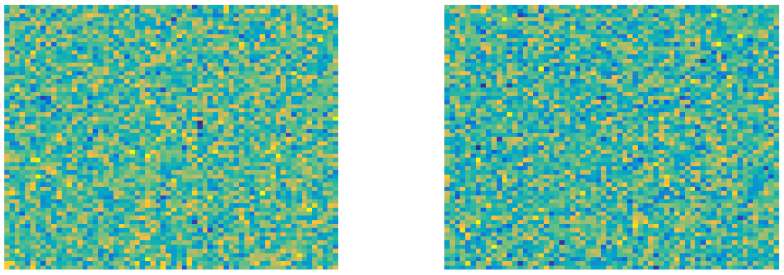

Figure 14: Images as realizations of an i.i.d. Gaussian distribution

Turning towards more dependencies among the pixels, we next implement images with a jointly Gaussian distribution. For two points $t = (t_1, t_2)$ and $u = (u_1, u_2)$ with $1 \leq t_1, t_2, u_1, u_2 \leq N$ we consider their periodic Euclidean distance in an image

$$\text{dist}(t, u) = \left(\min\{|t_1 - u_1|, N - |t_1 - u_1|\}^2 + \min\{|t_2 - u_2|, N - |t_2 - u_2|\}^2 \right)^{1/2} .$$

This allows to define a covariance matrix $\Sigma \in \mathbb{R}^{N^2 \times N^2}$ for all points in the image. The covariance between two points is set to be

$$\Sigma_{t,u} \coloneqq (1 + \operatorname{dist}(t,u))^{-\alpha}$$

for $\alpha = 1/2$ in our case. We then sample an image x due to a jointly Gaussian distribution with mean 0 and covariance matrix Σ, i.e.

$$x \sim \mathcal{N}(0, \Sigma) \ .$$

This leads to realizations as shown in figure 15, where again yellow colors indicate highly positive and blue highly negative values.

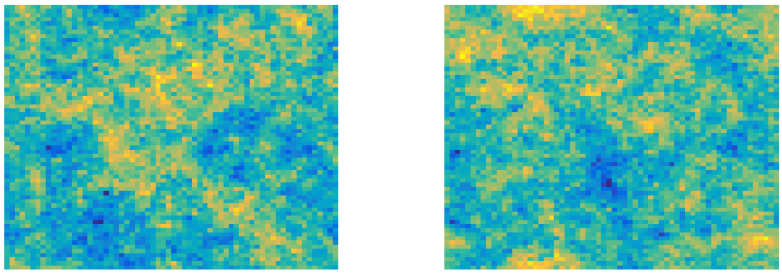

Figure 15: Images as realizations of a jointly Gaussian distribution

As a final example, we consider the Ising model, cf. [20, 34, 44], which plays a very important role in statistical mechanics. In dimensions $d \geq 2$, the Ising model becomes a very useful model showing phase transition phenomena. Since we would like to use the Ising model to create images, we limit our view to the two dimensional case. Therefore, we start with a two dimensional squared lattice $\Pi \subseteq \mathbb{Z}^2$. For each lattice point $t = (t_1, t_2)$, we assign a variable $x(t) \in \{\pm 1\}$ characterizing the spin at its position. This leads to a spin configuration $x = (x(t))_{t \in \Pi}$. Now, we want to represent the interaction of two adjacent points. As a side remark, note that we could extend this setup by representing the underlying structure of points by a graph. For any two points $t, t' \in \Pi$, we denote their interaction by $I_{t,t'}$. If t, t' are not nearest neighbors, $I_{t,t'} = 0$. When considering all interactions being equal, then for nearest neighbors $I_{t,t'} = I > 0$. Further, we allow an external magnetic field, denoted by h, which can be neglected for $h = 0$. Combining all this allows to define the Hamiltonian of a state as

$$H(x) = -\frac{1}{2} \sum_{t,t'} I_{t,t'} x(t) x(t') - h \sum_{t} x(t) \ ,$$

where the factor $1/2$ on the first sum is used to rescale the double counting of neighboring points. Given a temperature T and its inverse temperature

$\beta = T^{-1} \geq 0$, we can define the probability for each possible configuration as

$$P_\beta(x) = \frac{1}{Z} \exp(-\beta H(x))$$

for Z being the normalization constant.

In order to create realizations of the Ising model, there exist multiple algorithms with different advantages and drawbacks: Starting with a Metropolis algorithm approach dating back to the 1950s, cf. [7, 40], followed by the Swendsen-Wang algorithm, cf. [51, 54] or when turning towards probabilistic running time a coupling-from-the-past algorithm, cf. [45, 46]. In the following applications, we make use of a MATLAB implementation of the Swendsen-Wang algorithm[1]. Concerning the choice of the temperatur for the Ising model, note that for very low temperatures, the configurations are mainly dominated by spins in one direction. This would create images that constist of large areas of the same color. On the other hand, for high temperatures, neighboring spins become more and more independent which leads to images similar to the ones we created by the Bernoulli distribution. As a consequence, we focus on temperatures near the critical temperature of the Ising model generating images showing phase transition phenomena as displayed in figure 16.

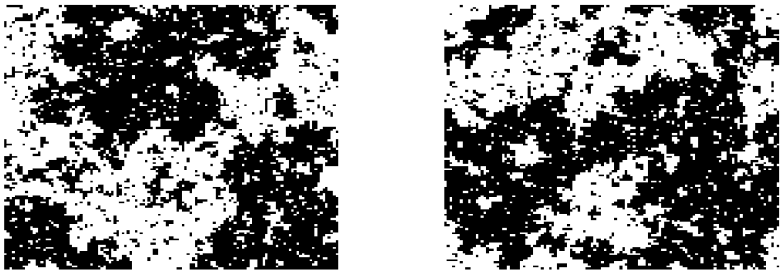

Figure 16: Images as realizations of the Ising model at critical temperature

We consider images of size $N \times N$ for $N = 64$, where the value of each pixel $x(t_1, t_2) \in \{\pm 1\}$ for coordinates $1 \leq t_1, t_2 \leq N$.

4.1.2 Implementation of the Scattering Transform

In order to compute the scattering transform for the images, we use the MATLAB library ScatNet 0.2 [2]. The wavelet convolutions are computed by the use of scaled and rotated versions of a two dimensional Morlet wavelet,

[1] https://github.com/jzavatoneveth/sw-ising - opened: 29th Mai 2019
[2] https://www.di.ens.fr/data/software/ - opened: 15th January 2019

scaled up to $J = 4$ and orientated in $|G| = 8$ different directions, see figure 17.

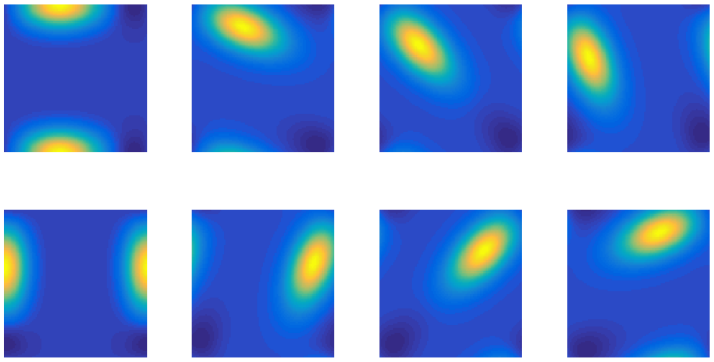

Figure 17: Two dimensional Morlet wavelets ψ_λ to compute the scattering transform. Displayed are all eight rotations for one choice of scale j. Again, yellow indicates positive, dark blue negatives values.

The averaging ϕ is done over the whole domain and the scattering transform is computed up to $M = 2$ levels of wavelet transformations. Hence, including the first averaging of the input, the resulting transform consists of three levels of output coefficients. Note that referring to [12], at least 99% of the energy of the input is captured within the first two layers of convolutions. For the first level, there is only one output coefficient, in level two there are 32 and the third level consists of 384 coefficients. In total, the computed scattering representation in the simulations consists of 417 coefficients.

4.1.3 Statistical Methodology and Tests

In order to evaluate patterns in the distribution of coefficients, we make use of the Kolmogorov-Smirnov test (KS-test), cf. [25, 33, 55], being constructed to test the null hypothesis of two distribution functions being equal against the alternative hypothesis of not being equal. We quickly review the structure of the KS-test starting with testing a set of data against a theoretical distribution. Given a set of realizations $\{x_1, ..., x_n\}$ for a random variable X, we consider its distribution function to be F_X and its empirical distribution function to be $F_{X,n}$. The distribution function of the theoretical distribution is denoted by F. This allows to define the null hypothesis as

$$H_0 : F_X(t) = F(t) \text{ for all } t,$$

which is tested against the alternative hypothesis

$$H_1 : F_X(t) \neq F(t) \text{ for at least one } t.$$

The KS-test statistic is given by

$$D = \sup_t \{|F_{X,n}(t) - F(t)|\}$$

and we reject the null hypothesis in the case that D is larger than a critical value c_α depending on the significance level α which is set a priori and well approximated by $\sqrt{\frac{-0.5\ln(\alpha/2)}{n}}$.

When comparing two sets of realizations $\{x_1, ..., x_n\}$ and $\{y_1, ..., y_m\}$ for random variables X and Y, we can replace the theoretical distribution function F in the above setting by the (empirical) distribution function of the second random variable Y. This leads to a null hypothesis

$$H_0 : F_X(t) = F_Y(t) \text{ for all } t,$$

which is tested against the alternative hypothesis

$$H_1 : F_X(t) \neq F_Y(t) \text{ for at least one } t.$$

The test statistic is adjusted to be

$$D = \sup_t \{|F_{X,n}(t) - F_{Y,m}(t)|\} \ .$$

Again, we reject H_0 if the value of D extends some critial value c_α, which can now be approximated by $c_\alpha = K_\alpha \sqrt{\frac{n+m}{nm}}$ for some constant K_α depending on the significance level α, e.g. $K_{0.05} \approx 1.36$.

Since we are not only interested if the null hypothesis is accepted or rejected, but would further like to measure the strength of the outcome, we shortly have a look at the p-value. Loosely speaking, the p-value is the probability to see the given data (and more extreme data) under the null hypothesis. Small p-values lead to the conclusion of rejecting the null hypothesis whereas high p-values (i.e. closer to 1) indicate an acceptance of the null hypothesis. Note that we usually reject the null hypothesis, if the given p-value is less than the significance level α of the test. For further reading concerning p-values, we refer to [25, 50].

In our application, we run the Kolmogorov-Smirnov test at a significance level of $\alpha = 0.05$. In the version of comparing data to a theoretical distribution, unless stated otherwise, we test against a standard normal distribution.

Concerning empirical moments of data, to fix some notation, let $\mathcal{Z} = \{z_1, ..., z_T\}$ be a set of data. The empirical mean of this data set is then given by

$$\mu_{\mathcal{Z}} = \frac{1}{T} \sum_{i=1}^{T} z_i$$

and the empirical variance can be defined to be

$$\sigma_{\mathcal{Z}} = \frac{1}{T-1}\sum_{i=1}^{T}(z_i - \mu_{\mathcal{Z}})^2 \ .$$

Further, to allow the comparison of the skewness, i.e. the third order moment of a random variable normalized by its mean and variance, we consider the empirical third order moment of the form

$$\text{skew}_{\mathcal{Z}} = \frac{\frac{1}{T}\sum_{i=1}^{T}(z_i - \mu_{\mathcal{Z}})^3}{\sqrt{\sigma_{\mathcal{Z}}}^3} = \frac{\frac{1}{T}\sum_{i=1}^{T}(z_i - \mu_{\mathcal{Z}})^3}{\left(\frac{1}{T-1}\sum_{i=1}^{T}(z_i - \mu_{\mathcal{Z}})^2\right)^{3/2}} \ .$$

Using this value, we can measure a possible asymmetric behavior of the data. Note that for a Gaussian distribution, the skewness is zero. Additionally, when considering fourth order moments of random variables normalized by mean and variance, we can define the empirical kurtosis as

$$\text{kurt}_{\mathcal{Z}} = \frac{\frac{1}{T}\sum_{i=1}^{T}(z_i - \mu_{\mathcal{Z}})^4}{\left(\frac{1}{T}\sum_{i=1}^{T}(z_i - \mu_{\mathcal{Z}})^2\right)^2} \ .$$

This measures how likely a distribution produces outliers or how heavy a distribution is dominated by its tails. A standard normal distribution has a kurtosis of three. The excess kurtosis is defined to be the kurtosis minus three in order to have a zero value for a standard Gaussian.

4.2 The Canonical Model

Aiming for a universality property for the scattering coefficients, we start by comparing the distribution of coefficients for different kinds of input. To fix some notation, for a set of images $\mathcal{Y} \coloneqq \{y_1, ..., y_T\}$, each y_i being a realization of Y, we denote by

$$S_J[p]\mathcal{Y} \coloneqq \{S_J[p]y_1, ..., S_J[p]y_T\}$$

the scattering transform along path p for all images in $\mathcal{Y}$.

As mentioned in the chapter before, we compare the scattering representations of images $\mathcal{X} \coloneqq \{x_1, ..., x_T\}$ created as realizations from X for Bernoulli random variables, jointly Gaussian random variables and realizations of the Ising model to samples $\mathcal{R} \coloneqq \{r_1, ..., r_T\}$ from R created by i.i.d. Gaussian random variables. To do so, we generate $T = 5000$ images, except for the Ising model where $T = 2000$. Afterwards we compute the scattering transform for each realization which leads to a sample of size T for every scattering coefficient. In order to make the distributions comparable, we

are normalizing each value in the set of coefficients $S_J[p]\mathcal{Y}$ by the empirical mean $\mu_{S_J[p]\mathcal{Y}}$ and empirical variance $\sigma_{S_J[p]\mathcal{Y}}$ for $\mathcal{Y} = \mathcal{X}, \mathcal{R}$, i.e.

$$(\sigma_{S_J[p]\mathcal{Y}})^{-1/2} \left(S_J[p]y - \mu_{S_J[p]\mathcal{Y}} \right)$$

leading to a normalized representation of $S_J[p]\mathcal{Y}$.

Exploiting this, we can now compare the distribution of the two normalized scattering representations for random inputs X and R by the use of the KS-test[3]. The results are shown in table 1. The first column denotes the method to create images for the test set $\mathcal{X}$ as described in chapter 4.1.1. We compare this images to our canonical form of images $\mathcal{R}$ created as realizations of R sampled from i.i.d. Gaussian random variables. In both cases, we apply the scattering transform to each of the images and normalize the coefficients by its empirical mean and variance. The number of rejected null hypotheses (the two distributions are equal) are displayed in the following three columns sorted by levels (recall that level one consists of one coefficient, level two of 32 and level three of 384). The last column gives the total percentage of null hypotheses which are rejected.

Family X	**Family** R	**Level** 1	**Level** 2	**Level** 3	**Total**
IID Gaussian	IID Gaussian	0	0	0	0.0%
Bernoulli	IID Gaussian	0	0	0	0.0%
Jointly Gaussian	IID Gaussian	0	0	0	0.0%
Ising Model $T > T_c$	IID Gaussian	0	0	1	0.2 %
Ising Model $T = T_c$	IID Gaussian	0	3	1	1.0 %
Ising Model $T < T_c$	IID Gaussian	0	1	0	0.2 %

Table 1: Comparing single scattering coefficients for different families of inputs. We display the **rejected** null hypotheses per level and a total percentage.

As a consequence of this KS-test results, a first idea of suggesting that the scattering coefficients always follow a particular law, modulated by mean and variance depending on the input's properties, does not seem far-fetched. Nonetheless, a deeper discussion is required due to a possibly high type II error, i.e. failing to reject the null hypothesis although it is incorrect. Instead of normalizing each entry of the scattering vector itself, another natural approach is to normalize the whole vector $S_J[P_J]x$ by its covariance matrix Σ, i.e. computing

$$\Sigma^{-1/2}(S_J[P_J]x - \mu) \ .$$

[3]MATLAB code to reproduce all experiments can be found attached

Doing so leads to equivalent results concerning the number of rejected null hypotheses of the KS-test as the ones in table 1.

To examine the proposed universality behavior, in figures 19 and 20 we illustrate examples of quantile-quantile plots for some coefficients. For purposes of comparison, we start by plotting quantiles of i.i.d. Gaussian input against the quantiles of different i.i.d. Gaussian realizations in figure 18.

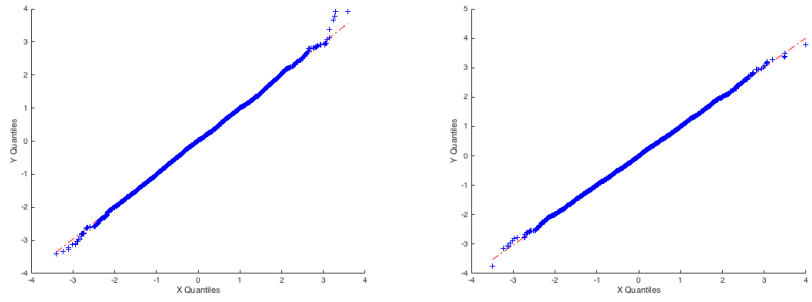

Figure 18: Quantile-quantile plots. Left: Plot the quantiles of a level two coefficient for two sets of i.i.d. Gaussian input against each other. Right: same for a level three coefficient.

Afterwards, in figure 19 we compare the normalized scattering coefficients for jointly Gaussian input images to the ones created by an i.i.d. Gaussian distribution. On the left, we plot the quantiles of a level two coefficient against each other. On the right, we do the same for a level three coefficient. The qq-plots show a straight line, indicating that the data sets follow the same distribution.

In the case of the Ising model input in figure 20, we plot the quantiles of two third level coefficients against the quantiles for an i.i.d. Gaussian input class. The outcome is the same as before: the qq-plots indicate straight lines proposing an identical distribution of coefficients. The same behavior is obtained for all other tested inputs as well as all other coefficients. Further plots can be found attached.

When having a closer look at empirical cumulative distribution functions (ECDF) of scattering coefficients, in nearly all cases of comparisons, i.e. for any coefficient as well as any input, we obtain congruent ECDFs underlining the hypothesis of an identical law for scattering coefficients again. This behavior can be used as a possible explanation of the results of the KS-tests, since thereby, we used the maximal distance between the two ECDFs as test statistic. Having nearly congruent ECDFs and hence very small dis-

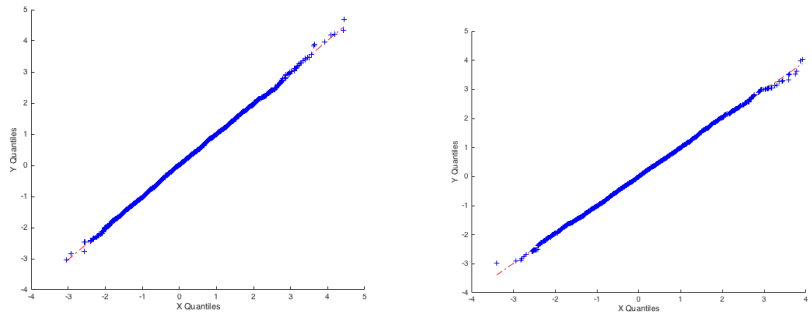

Figure 19: Quantile-quantile plots. Left: Plot the quantiles of a level two coefficient for a jointly Gaussian input against the quantiles of the same coefficient from an i.i.d. Gaussian input. Right: same for a level three coefficient.

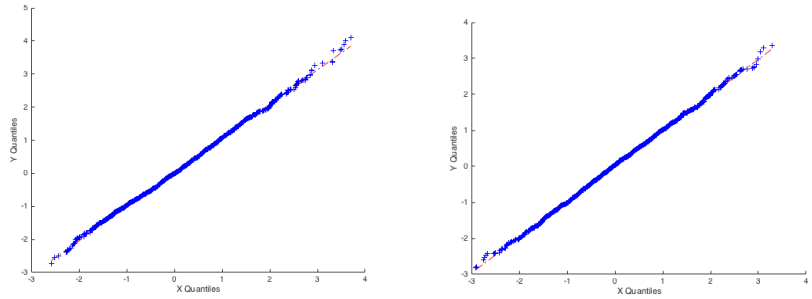

Figure 20: Quantile-quantile plots. Plot the quantiles of two level three coefficient for an Ising model input against the quantiles of the same coefficient from an i.i.d. Gaussian input.

tances between them does not allow a rejection of the null hypothesis in the KS-test. For illustration of examples, see figure 21, where we plot the empirical cumulative distribution function of a third level coefficient for the Ising model input class together with the ECDF of the same coefficient for i.i.d. Gaussian images.

During our development of the scattering transform in chapter 2, we encountered the Fourier-modulus transformation of functions as well as the wavelet transform as possible candidates for suitable representations of images. We want to compare the recently described behavior of the scattering transform to the Fourier-modulus and the wavelet transform.
To do so, as before, we take different families of input images, i.e. Bernoulli, jointly Gaussian and Ising model realizations. We compute the Fourier-modulus and wavelet transform, normalize by empirical mean and variance

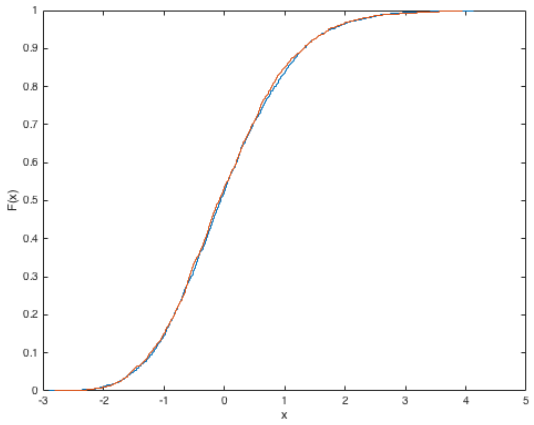

Figure 21: Plot the empirical distribution functions of a level two coefficient for an Ising model input and the ECDF of the same coefficient from an i.i.d. Gaussian input.

and compare the behavior to the one of i.i.d. Gaussian input images being transformed in the same way. Note that the wavelet transform is computed with Morlet wavelets. KS-tests are again evaluted as before, the results are shown in table 2.

Family X	**Family** R	**Fourier-Modulus**	**Wavelet**
Bernoulli	IID Gaussian	22.0 %	11.4 %
Jointly Gaussian	IID Gaussian	30.1 %	10.1 %
Ising Model $T = T_c$	IID Gaussian	54.4 %	98.5 %

Table 2: Comparing the Fourier-modulus and wavelet transform of different families of inputs to the transformation of an i.i.d. Gaussian class. We display the percentage of coefficients for which the null hypothesis can be **rejected**.

When trying to adapt the approach for the scattering transform to Fourier-modulus or wavelet transform, we can see while focusing on the Ising model that more than 50% in the one and 98% of coefficients in the other case do allow a rejection of the null hypothesis of following the same distribution at a significance level of $\alpha = 0.05$. Even in the case of independent Bernoulli images, we can reject the null hypothesis in every fifth or tenth trial. Comparing the results of table 2 to the ones of the scattering transform in table 1 clearly indicates that the universality property suggested for the scattering transform cannot be adapted to the cases of Fourier-modulus or wavelet transform.

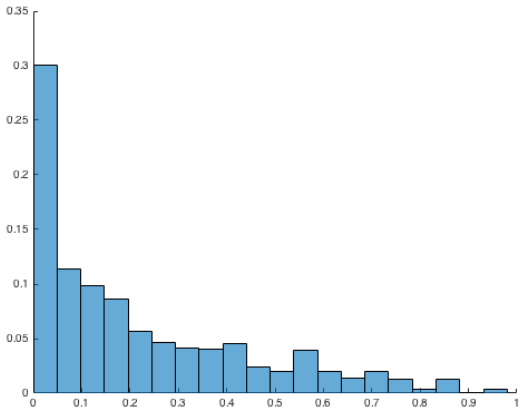

Figure 22: Histogram with relative frequencies of p-values for KS-test of Fourier-modulus for jointly Gaussian images.

Illustrating this, we can have a look at the p-values of the KS-test in the case of jointly Gaussian input images being transformed with the Fourier-modulus operation. Figure 22 shows the tendency of the p-values towards zero which also indicates a contradiction of the data with the null hypothesis.

Compared to Fourier-modulus and wavelet transform, the scattering vector shows a much more identically distributed behavior, inspiring the idea of exploiting this universality property for the scattering transform. It could be used to define a canonical model for the distribution of scattering coefficients $S_J[p]x$. Starting with an i.i.d. Gaussian random image r, we compute the scattering transform $\bar{\Phi}r \coloneqq S_J[P_J]r$. Further, we normalize each coefficient $S_J[p]r$ to get a whitened representation Φr. Running the same procedure for another input image x, we obtain two representations Φr and Φx which follow the same distribution. We can illustrate the canonical model as in figure 23, where the normalization of the scattering vector $\bar{\Phi}x$ is meant componentwise.

$$\begin{array}{ccccc} x & \xrightarrow{S_J[P_J]} & \bar{\Phi}x & \xrightarrow{\sigma^{-1/2}(Id-\mu)} & \Phi x \\ & & & & \wr \\ r & \xrightarrow{S_J[P_J]} & \bar{\Phi}r & \xrightarrow{\sigma^{-1/2}(Id-\mu)} & \Phi r \end{array}$$

Figure 23: The canonical model

4.3 Gaussianization with Scattering Coefficients

Since many methods in image processing share as a common goal to Gaussianize the output values in order to tame their distributions, cf. [14, 31], we want to extend our simulations to test a possible Gaussian behavior of scattering coefficients. As in chapter 3.3, we allow a whitening A of the coefficients which leads to an approach as illustrated in figure 26. In the upcoming subsections, we are testing different transformations A and compare their results.

4.3.1 Whitening each Coefficient

Starting with the same setup as for the canonical model, we are comparing normalized scattering coefficients for different classes of images $\mathcal{X} = \{x_1, .., x_T\}$ to a standard normal distribution using the Kolmogorov-Smirnov test. To do so, we whiten each coefficient by its empirical mean and variance. Sample sizes are again $T = 5000$, except for the Ising model, where $T = 2000$. The results are displayed in table 3 following a similar logic as table 1. The given values indicate the number of rejected null hypothesis of the KS-test.

Family X	**Level** 1	**Level** 2	**Level** 3	**Total**
IID Gaussian	0	12	289	72.2 %
Bernoulli	0	12	285	71.2 %
Jointly Gaussian	0	12	291	72.7 %
Ising Model $T > T_c$	0	6	194	48.0 %
Ising Model $T = T_c$	1	3	192	47.0 %
Ising Model $T < T_c$	0	5	180	44.4 %

Table 3: Comparing single scattering coefficients for different families of inputs to a Gaussian distribution after whitening. We display **rejected** null hypotheses.

As a result, the null hypothesis that the law of scattering coefficients follows a normal distribution can be rejected in at least 40% of all investigated cases, for some image classes even more than 70%. Note that the different sample sizes can cause the discrepancy in the percentages between the Ising model tests and the other inputs due to the occurance of the sample size in the critical values of the test statistic. Since there remain several coefficients, for which the null hypothesis is not rejected, we can hope for a behavior of coefficients, which is not highly different to the attitude of a Gaussian. Whitening the scattering vector $S_J[P_J]x$ by its covariance matrix instead of normalizing each single coefficient $S_J[p]x$ by its empirical mean and variance shows similar effects as the results in table 3. In order to get

a better intuition, we start having a look at empirical distribution functions of scattering coefficients and qq-plots against a normal distribution as illustrated in figure 24. Note that using different image classes as input as well as choosing other coefficients does not really affect the plots.

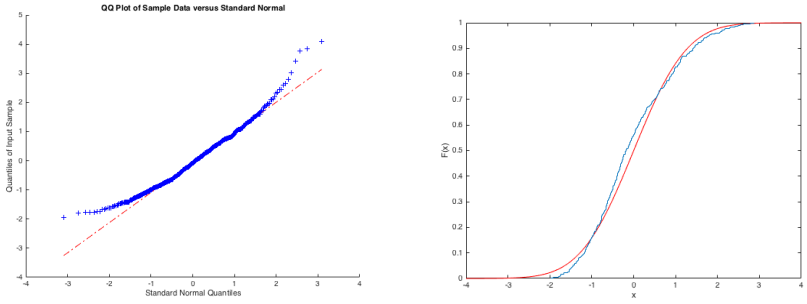

Figure 24: Left: Plot the quantiles of a level two coefficient for Ising model images against the quantiles of a standard normal distribution. Right: ECDF of a level two scattering coefficient using Ising model inputs in blue. The theoretical CDF of a standard Gaussian distribution is plotted in red.

Figure 24 indicates a right-skewed behavior of scattering coefficients when normalizing each coefficient by its empirical mean and variance. This seems reasonable while having a look on the definition of the scattering transform. Recall, that the scattering transform of x along a path $p = (\lambda_1, ..., \lambda_m)$ was defined to be

$$|||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| ... \star \psi_{\lambda_m}| \star \phi \ .$$

In every iteration, the modulus operator pushes all values onto the positive real line which introduces a lower bound on the domain of scattering coefficients causing a right-skewed distribution.

In order to measure the skewed behavior of the scattering transform, we turn our view towards its empirical skewness. Therefore, we compute the empirical third order moment of each scattering coefficient for different classes of input. In table 4, we show the minimal and maximal value for skewness as well as the median, i.e. the threshold which separates the lower half of the skewness values from the upper one.

Evaluating the empirical skewness of the scattering data, we obtain that at least 50% of coefficients show a positive skewed behavior of at least 0.22 up to 0.66 for the Ising case. For the other classes of input images we even obtain half of the skewness values in the interval $[0.30, 0.62]$. This noticeable deviation from a standard normal distribution can be used to underline and explain the rejection of the Gaussian hypothesis in the KS-tests. In order

Family X	**Min. skew**	**Max. skew**	**Median**
IID Gaussian	0.01	0.62	0.32
Bernoulli	-0.05	0.59	0.32
Jointly Gaussian	-0.02	0.53	0.30
Ising Model $T = T_c$	-0.04	0.66	0.22

Table 4: Empirical skewness of scattering coefficients for different classes of inputs. Shown are the minimal and maximal value for skewness among the 417 computed values as well as the median.

to visualize this, we plot the histrogram of a third level coefficient for the Ising model input class in figure 25 as an example.

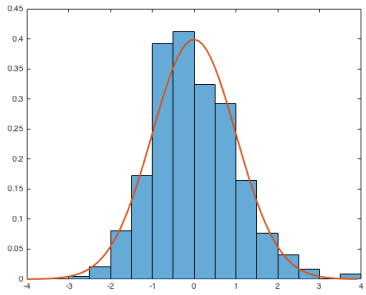

Figure 25: Histogram of the distribution of a third level scattering coefficient, computed for inputs of the Ising model at critical temperature. The red line shows the bell curve of a standard normal distribution.

As before, we further want to compare these results to the behavior of the Fourier-modulus and the wavelet transform and examine a possible Gaussian behavior in these representations. Therefore, we take input images from our classes of i.i.d. Gaussian, Bernoulli, jointly Gaussian or the Ising model, compute the transformation, normalize the coefficients by mean and variance and perform a KS-test for each coefficient comparing the data set to a standard normal distribution. The computation is done in an equivalent way as described in chapter 4.2. The results are illustrated in table 5. We can base ourselves on a percentage of at least 94.9% for null hypotheses being rejected in case of the Ising model, which indicates a highly non-Gaussian behavior of the corresponding coefficients.

Comparing this to the results for scattering coefficients, the latter seem to be closer to a Gaussian distribution than Fourier-modulus or wavelet coefficients. Nonetheless, scattering coefficients still show a slightly right-skewed behavior in their distributions after normalization with empirical mean and

Family X	**Fourier-Modulus**	**Wavelet**
IID Gaussian	92.3 %	38.0 %
Bernoulli	95.0 %	12.1 %
Jointly Gaussian	90.5 %	8.1 %
Ising Model $T = T_c$	94.9 %	99.9 %

Table 5: Comparing the Fourier-modulus and wavelet transform of different families of inputs to a Gaussian distribution after whitening. We display the percentage of **rejected** null hypotheses.

variance which leads to a rejection of the Gaussian hypothesis for at least two out of five scattering coefficients in the KS-test. Following a rule of thumb for skewness, only values larger than 0.5 indicate moderately skewed data, what suggests that the skewness may be removed for the majority of coefficients.

4.3.2 Rotation and Dimension Reduction

Since the distribution of scattering coefficients seems to be reasonably close to a Gaussian, we would like to test a principal component analysis (PCA) approach for the normalization. PCA was originally introduced by Pearson in [42] and has become a widely spread tool in statistics and data processing, cf. [24, 28, 56]. To get an idea of this, consider a set of data in some high dimensional space. Intuitively, the aim is to fit a lower dimensional ellipsoid, let us say of dimension d, to the data. Hence, when dealing with data generated from a Gaussian distribution, this fit will work perfectly, whereas there might occur some problems in the cases of data following distributions far away from a Gaussian.

We inspire ourselves by the whitening approach from [2] which mainly mimics the PCA concept. Therefore, as in the previous setup, we consider a set $\{x_1, ..., x_T\}$ of $T = 5000$ ($T = 2000$ in case of the Ising model) images created as explained in chapter 4.1.1. We compute the scattering transform for each image, leading to a set of T scattering vectors $\{\bar{\Phi}x_i\}_{i\leq T} \coloneqq \{S_J[P_J]x_i\}_{i\leq T}$. To perform the desired normalization, let μ be the empirical mean and Σ be the empirical covariance matrix for the scattering coefficients. In order to reduce the variability among the coefficients, instead of only normalizing with the covariance matrix, as described, we also want to project the scattering representation into a lower dimensional space of dimension d. Therefore, we compute the eigendecomposition of the covariance matrix

$$\Sigma = WDW^t \ ,$$

where D is a diagonal matrix consisting of eigenvalues of Σ and W contains the corresponding eigenvectors, the superscript t denotes the transposed of

the given matrix. To find a suitable subspace for the projection, we consider the d largest eigenvalues and the subspace spanned by their corresponding eigenvectors. In PCA approaches, these are the so-called principal components. Denote by W_d the truncated version of the matrix W to these d eigenvectors and by

$$\mathrm{Proj}_d = W_d W_d^t$$

the orthogonal projection onto the introduced subspace.
As a side remark, when choosing a value for d, we have a trade-off between the reduction of variability at the expense of reducing the distance of two points in the scattering domain. The latter can cause problems when trying to distinguish two different images and hence create trouble in classification tasks. In order to control this compromise, a bi-Lipschitz condition as in [2] of the form that there exists an $\alpha > 0$ satisfying

$$\frac{1}{\alpha} \|x_i - x_j\| \leq \left\| \mathrm{Proj}_d \bar{\Phi} x_i - \mathrm{Proj}_d \bar{\Phi} x_j \right\| \leq \|x_i - x_j\|$$

for all $i, j \leq T$ should be imposed.

Now, instead of normalizing by $\Sigma^{-1/2}$, we only rotate the scattering representations once and divide by the square root of the eigenvalues. Therefore, introduce the matrix

$$C_d^{-1/2} := D_d^{-1/2} W_d^t \,,$$

where $D_d^{-1/2}$ is the inversed square root of D projected to the space spanned by the d eigenvectors corresponding to the d largest eigenvalues of Σ. Hence, when multiplying by W_d^t we perform a change of basis which is not reversed later. Afterwards, we normalize the variance of the data in each direction of the new basis vectors. As a side remark, note that the data remains in the lower dimensional space with new basis vectors due to omitting the final multiplication with W_d. Concluding, normalizing the scattering vectors is done by a subtraction of the mean and whitening by C_d, i.e.

$$C_d^{-1/2}(\bar{\Phi} x - \mu) \,.$$

This leads to a representation as illustrated in figure 26 which we would like to compare to a Gaussian distribution.

$$x \xrightarrow{S_J[P_J]} \bar{\Phi} x \xrightarrow{C^{-\frac{1}{2}}(Id - \mu)} \Phi x \overset{?}{\sim} \mathcal{N}(0, \mathbb{1})$$

Figure 26: Normalized and truncated scattering coefficients

Therefore, we evaluate KS-tests in the similar setup as before by computing images as realizations of random variables, calculating the scattering

representation of those and finally normalize by mean and truncated covariance matrix projecting into the lower dimensional space with new basis vectors. For our simulations, we choose $d = 100$ in one case and d being equal to the dimension of the scattering domain in the other case (i.e. $d = 417$). In table 6 we show rejected null hypotheses of the KS-test, comparing the distribution of the normalized and truncated scattering coefficients to a standard normal distribution.

Family X	**Truncated to** $d = 100$	**Without truncation**
IID Gaussian	0.0 %	0.0 %
Bernoulli	0.0 %	0.0 %
Jointly Gaussian	0.0 %	0.0 %
Ising Model $T = T_c$	2.0 %	0.5 %

Table 6: Comparing dimensional truncated scattering coefficients for different families of inputs to a Gaussian distribution after whitening by the truncated covariance matrix $C_d^{-1/2}$. We display the percentage of coefficients for which the null hypothesis can be **rejected**.

For the case of truncating to $d = 100$, we obtain a number of at most 2.0% of cases where we can reject the null hypothesis that coefficients follow a standard Gaussian distribution. When comparing these results to the ones we generated before by a normalization of each coefficient itself or with the covariance matrix, we recognize a massive change in the number of non-rejected null hypotheses going now up to a hundred percent. Consequently, in contrast to the results of the preceding experiments, the hypothesis that the coefficients do follow a Gaussian distribution cannot be rejected in this new environment anymore. As possible explanations, we take either the reduction of dimension or the rotation during the process of normalization into account. When further considering the percentages for the case of $d = 417$, i.e. the case where we did not truncate at all, thus just normalized the data after a change of basis, we receive similar results as for $d = 100$. Consequently, it seems that the rotation is responsible for the different behavior concerning the rejection of the Gaussian hypothesis and, highly remarkable, it appears to be no difference whatever the dimension of the truncated space is chosen to be.

To gain a deeper insight, we focus on the case of the Ising model at critical temperature for the truncated scattering vector to $d = 100$ dimensions. The majority, i.e. all except the last entries in the scattering vector, shows marginals very similar to Gaussians, typified in the first two columns of figure 27. The distribution of the last coefficients slowly moves more and more away from a Gaussian, plotted in figure 27 in columns three and four.

At first, some slight right-skewness appears similar to the cases when normalizing each coefficient by its variance. Finally, the last coefficient shows a heavy-tailed behavior.

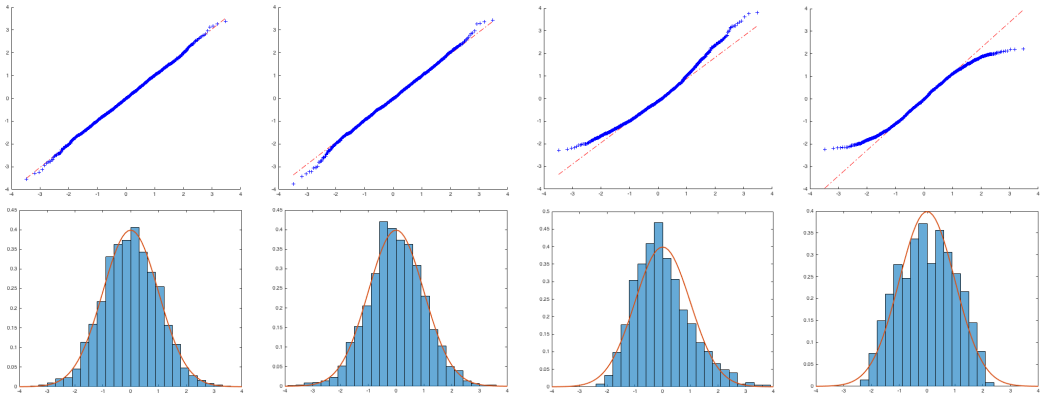

Figure 27: qq-plots and histograms for scattering coefficients of the 100-dimensional truncated scattering vector computed for inputs of the Ising model at critical temperature compared to a standard normal distribution. From left to right we plot the 29th, 91st, 99th and 100th coefficient, qq-plot against a standard Gaussian on top and its histogram below.

Note that due to the rotation, the entries in the scattering vector do no longer correspond to one fixed sequence of wavelet convolutions, modulus and averaging, but instead are linear combinations of these values. Still, since we normalized, all rotated coefficients do have zero mean and unit variance. When having a look at the qq-plots and histograms for $d = 417$, the same behavior occurs. All coefficients except a few seem to follow a Gaussian distribution whereas those show an equivalent behavior as the ones in the 100-dimensional case.

In order to compare this method of whitening to the intuitive normalization with the covariance matrix or the variance for each coefficient, we evaluate the empirical third order moments in the same way as before (compare to table 4) to get a measure for the skewness. In table 7, we plot the minimal and maximal skewness, as well as the 0.05 and 0.95-quantiles.

Hence, by the truncation and rotation of scattering vectors, we obtain marginal distributions a lot less skewed than before. In all cases of input images, the distribution of at least 90% of coefficients shows a skewness in the interval $[-0.1, 0.1]$ which can be read as approximately symmetric. Only for the Ising model at critical temperature, we obtain coefficients still taking skewness values of up to 0.63 indicating moderate skewness in their

Family X	**Min. skew**	**Max. skew**	0.05**-quant.**	0.95**-quant.**
IID Gaussian	-0.17	0.10	-0.07	0.05
Bernoulli	-0.17	0.11	-0.07	0.06
Jointly Gauss.	-0.09	0.18	-0.07	0.08
Ising Model	-0.25	0.63	-0.10	0.10

Table 7: Empirical skewness of truncated scattering coefficients for different classes of inputs. Shown are the minimal and maximal value for skewness among the 100 computed values as well as the 0.05 and 0.95-quantiles.

distribution. This coincides with the third column of figure 27 displaying a right-skewed coefficient of the scattering representation. When further evaluating fourth order moment to get access to the kurtosis of the distributions, we obtain results displayed in table 8. Note that we show the excess kurtosis, i.e. the kurtosis minus three in order to set the kurtosis for the standard normal distribution to zero. Again, for different classes of input we display the minimal and maximal values among the 100 kurtosis numbers for truncated scattering coefficients as well as the 0.05 and 0.95-quantiles.

Family X	**Min. kurt**	**Max. kurt**	0.05**-quant.**	0.95**-quant.**
IID Gaussian	-0.16	0.23	-0.12	0.13
Bernoulli	-0.24	0.19	-0.10	0.13
Jointly Gauss.	-0.14	0.31	-0.09	0.20
Ising Model	-0.83	0.41	-0.13	0.36

Table 8: Empirical excess kurtosis of truncated scattering coefficients for different classes of inputs. Shown are the minimal and maximal value for kurtosis among the 100 computed values as well as the 0.05 and 0.95-quantiles.

In the case of i.i.d. Gaussian, Bernoulli and jointly Gaussian input images, the range of the kurtosis is reasonably close to zero, indicating a very Gaussian-like behavior. For the Ising model at critical temperature, the majority of values, i.e. 90%, is also within a distance of at most 0.36 to zero. But there are a few outliers exhibiting a medium kurtosis. This corresponds to the intuition of the illustrations in figure 27, where the last coefficient shows a heavy-tailed behavior leading to an excess kurtosis different from zero. We obtain equivalent results in the case of 417 dimensions.

In summary, following a PCA approach by normalizing with a truncated covariance matrix which also rotates the scattering coefficients suggests a much more Gaussian behavior for the majority of coefficients. This Gaussianization for the major part of coefficients seems to be at the expense of a few remaining ones, showing a distribution still different from a Gaussian. Having a Gaussian distribution of coefficients would propose a very desirable

statistical behavior of the transformed scattering vectors.

5 Distribution of Scattering Coefficients in Comparison to Fourier-Modulus and Wavelet Transform

Finally, we would like to draw a conclusion from the numerical results of the performed experiments. Starting with an input image x and an image created as a sample from i.i.d. standard Gaussian random variables r, we can compute the scattering transform for both, denoted by $\bar{\Phi}x$ and $\bar{\Phi}r$. Afterwards, when performing a whitening by mean and variance for each single coefficient or normalizing the whole scattering vector by its covariance matrix, we obtain the same distributional behavior no matter which class of input image we use. This universality property may allow the introduction of a canonical model for the distribution of scattering coefficients. In the case of refining the normalization by a PCA approach, as described in chapter 4.3.2, we can further not reject the Gaussian hypothesis anymore. This leads to the assumption that the distribution of scattering coefficients is already close to a Gaussian distribution even if we could reject this in the case of whitening by variance or covariance and mean based on the results of the KS-tests. Nonetheless, a suitable rotation of the scattering representation Gaussianizes the majority of coefficients at the expense of a few coefficients whose distribution cannot be Gaussianized or is even further apart from a Gaussian than before. A different approach to get rid of the right-skewed behavior of scattering coefficients could consist of applying a transformation from the ladder of powers introduced in [52]. Matching results for this type of transformations to the PCA-inspired rotating modification could be a point of interest in future work.
Comparing the scattering representation to the cases of the Fourier-modulus or wavelet transform, we could not recognize a similar behavior for those two transformations: neither while aiming for a canonical model, nor comparing to a Gaussian distribution. In both cases, any of the equivalence hypotheses could be rejected at a significance level of $\alpha = 0.05$ for the majority of coefficients. For illustration purposes of this synthesis, see figure 28, where we denote by A the chosen whitening, either by mean and variance, covariance or truncated covariance matrix. The transforms Ψx and Ψr represent the Fourier-modulus or wavelet transform.

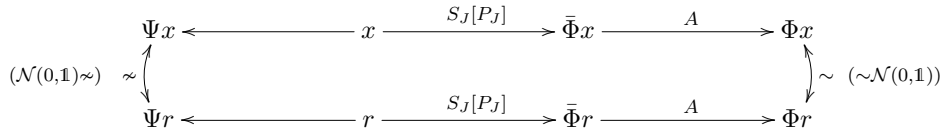

Figure 28: Comparing distributions of Fourier-modulus, wavelet and scattering transform

Concerning the used methods, in the cases where the null hypothesis of coefficients sharing a common (Gaussian) law could not be rejected, the issue emerges if the Kolmogorov-Smirnov test is the best possible way to compare the behavior in distributions.

When going back to the inversion task mentioned in chapter 3.3 and the corresponding image generation as displayed in figure 12, a naturally arising question is, how the different approaches of whitening would perform in this environment. Evaluating this could be a first step towards further indication on distributional characteristics of the scattering transform. Even further, inspired by the image generation with the scattering transform from Gaussian white noise in [2], we could possibly extend this by the canonical model proposed in chapter 4.2 and compare their behaviors. This could lead to an extended inverison model as illustrated in figure 29.

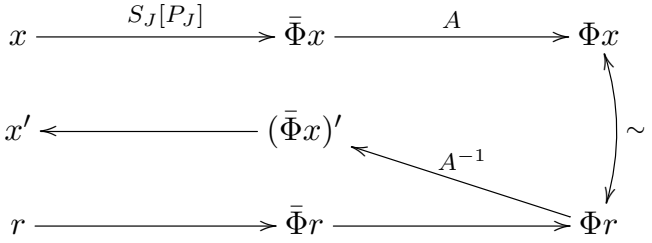

Figure 29: The extended inversion model

In conclusion, further research on this from the numerical point of view and more important from the analytical one is required to get a complete understanding of the underlying behavior of scattering coefficients.

References

[1] Anden, J. ; Mallat, S.: Deep Scattering Spectrum. In: *IEEE Transactions on Signal Processing* 62 (2013). http://dx.doi.org/10.1109/TSP.2014.2326991. – DOI 10.1109/TSP.2014.2326991

[2] Angles, T. ; Mallat, S.: Generative networks as inverse problems with scattering transforms. In: *International Conference on Learning Representations* (2018)

[3] Arjovsky, M. ; Chintala, S. ; Bottou, L.: Wasserstein Generative Adversarial Networks. In: *Proceedings of the 34th International Conference on Machine Learning* Bd. 70, p. 214-223, 2017 (Proceedings of Machine Learning Research)

[4] Arora, S. ; Ge, R. ; Liang, Y. ; Ma, T. ; Zhang, Y.: Generalization and Equilibrium in Generative Adversarial Nets (GANs). In: *CoRR* abs/1703.00573 (2017)

[5] Barford, L. A. ; Fazzio, R. S. ; Smith, D. R.: An Introduction to Wavelets. In: *HPL-92-124* (1992)

[6] Bengio, Y. ; Courville, A. ; Vincent, P.: Representation Learning: A Review and New Perspectives. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (2013)

[7] Binder, K. ; Heermann, D.: *Monte Carlo Simulation in Statistical Physics*. Springer Berlin Heidelberg, 2002. – ISBN 978–3–642–07746–3

[8] Blum, A. ; Hopcroft, J. ; Kannan, R.: *Foundations of Data Science*. 2018

[9] Bojanowski, P. ; Joulin, A. ; Lopez-Paz, D. ; Szlam, A.: *Optimizing the Latent Space of Generative Networks*. 2018

[10] Brezis, H.: *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, 2010. – ISBN 978–0–387–70913–0

[11] Bruna, J.: *Scattering Representations for Recognition*, Ecole Polytechnique, PhD thesis, 2012

[12] Bruna, J. ; Mallat, S.: Invariant Scattering Convolution Networks. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, no. 8, p. 1872-1886 (2013). http://dx.doi.org/10.1109/TPAMI.2012.230. – DOI 10.1109/TPAMI.2012.230. – ISSN 0162–8828

[13] Calderon, A.: Intermediate spaces and interpolation, the complex method. In: *Studia Mathematica* 24, no. 2, p. 113-190 (1964)

[14] Chen, S. ; Gopinath, R.: Gaussianization. In: *Advances in Neural Information Processing Systems* Bd. 13, p. 423-429. MIT Press, 2001

[15] Dales, H. G. ; Millinoton, A.: Translation-invariant linear operators. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 113, p. 161-172 (1993)

[16] Daubechies, I.: Orthonormal Bases of Compactly Supported Wavelets. In: *Communications on Pure and Applied Mathematics* XLI, p. 909-996 (1988)

[17] Daubechies, I.: *Ten Lectures on Wavelets*. 1992. – ISBN 978–0–898712–74–2

[18] Folland, G.: *Fourier Analysis and its Applications*. Pacific Grove, Calif. : Wadsworth + Brooks/Cole Advanced Books + Software, 1992 (Pure and Applied Undergraduate Texts, vol. 4). – ISBN 978–0–8218–4790–9

[19] Forster, O.: *Analysis 3*. Vieweg + Teubner Verlag, 2012. – ISBN 978–3–8348–2374–8

[20] Gallavotti, G.: *Statistical Mechanics - A Short Treatise*. Springer, 1999

[21] Gomes, J. ; Velho, L.: *From Fourier Analysis to Wavelets*. Springer International Publishing, 2015. – ISBN 978–3–319–22074–1

[22] Goodfellow, I. ; Pouget-Abadie, J. ; Mirza, M. ; Xu, B. ; Warde-Farley, D. ; Ozair, S. ; Courville, A. ; Bengio, Y.: Generative Adversarial Networks. arXiv:1406.2661 (2014)

[23] Grossmann, A. ; Morlet, J.: Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape. In: *SIAM Journal on Mathematical Analysis* 15, p. 723-736 (1984)

[24] Guan, Y. ; Dy, J.: Sparse Probabilistic Principal Component Analysis. In: *Journal of Machine Learning Research - Proceedings Track* 5, p. 185-192 (2009)

[25] Hedderich, J. ; Sachs, L.: *Angewandte Statistik*. 15. editon. Springer, 2016

[26] Hernandez, E. ; Weiss, G.: *A first course on wavelets*. Boca Raton: CRC Press, 1996. – ISBN 0–8493–8274–2

[27] Jacobsen, J.-H. ; Oyallon, E. ; Mallat, S. ; Smeulders, A.: Multiscale Hierarchical Convolutional Networks. In: *Proceedings of the 34th International Conference on Machine Learning* (2017)

[28] Jolliffe, I.: *Principal Component Analysis.* Springer Verlag, 1986

[29] Kaiser, G.: *A friendly guide to wavelets.* Birkhäuser, 1994. – ISBN 0–8176–3711–7

[30] Kessler, B. ; Payne, G. ; Polyzou, W.: Wavelet Notes. (2003). – arXiv:nucl-th/0305025

[31] Laparra, V. ; Camps-Valls, G. ; Malo, J.: Iterative Gaussianization: From ICA to Random Rotations. In: *Trans. Neur. Netw.* 22, no. 4, p. 537-549 (2011)

[32] LeCun, Y. ; Kavukcuoglu, K. ; Farabet, C.: Convolutional networks and applications in vision. p. 253-256 (2010). http://dx.doi.org/10.1109/ISCAS.2010.5537907. – DOI 10.1109/ISCAS.2010.5537907. – ISSN 0271–4302

[33] Lehmann, E.: *Testing Statistical Hypotheses.* 2. editon. Springer Texts in Statistics, 1997

[34] Ma, S.: *Statistical Mechanics.* 3. editon. World Scientifics, 2004

[35] Mahendran, A. ; Vedaldi, A.: Understanding Deep Image Representations by Inverting Them. In: *CoRR* abs/1412.0035 (2014)

[36] Mallat, S.: Mutlifrequency Channel Decomposition of Images and Wavelet Models. In: *IEEE Transactions on Acoustics. Speech and Signal Processing* 37, no. 12 (1989)

[37] Mallat, S.: *A wavelet tour of signal processing.* 2009

[38] Mallat, S.: Group Invariant Scattering. In: *Communications on Pure and Applied Mathematics* LXV, p. 1331-1398 (2012)

[39] Mallat, S.: Understanding deep convolutional networks. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374: 20150203 (2016). http://dx.doi.org/10.1098/rsta.2015.0203. – DOI 10.1098/rsta.2015.0203

[40] Metropolis, N. ; Rosenbluth, A. ; Rosenbluth, M. ; Teller, A. ; Teller, E.: Equation of State Calculations by Fast Computing Machines. In: *The Journal of Chemical Physics* 21, no. 6, p. 1087-1092 (1953)

[41] Meyer, Y.: *Wavelets and operators.* Cambridge studies in advanced mathematics, 1992

[42] Pearson, K.: On lines and planes of closest fit to systems of points in space. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, no. 11, p. 559-572 (1901)

[43] Pinsky, M.: *Introduction to Fourier Analysis and Wavelets.* Brooks Cole/Cengage Learning, 2002. – ISBN 978–0–534–37660–4

[44] Plischke, M.: *Equilibrium statistical physics.* 3. editon. World Scientifics, 2006

[45] Propp, J. ; Wilson, D.: Exact sampling with coupled Markov chains and applications to statistical mechanics. In: *Random Structures & Algorithms* 9, p. 223-252 (1996)

[46] Propp, J. ; Wilson, D.: *Coupling from the Past: a User's Guide.* 1997

[47] Ranzato, M. ; Huang, F. ; Boureau, Y. ; LeCun, Y.: Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition* (2007)

[48] Ricaud, B. ; Torresani, B.: A survey of uncertainty principles and some signal processing applications. In: *Advances in Computational Mathematics, Springer Verlag* 40 (3), p. 629-650 (2014)

[49] Rudin, W.: *Real and Complex Analysis.* McGraw-Hill Book Company, 1966

[50] Schervish, M.: *Theory of Statistics.* Springer, 1995

[51] Swendsen, R. ; Wang, J.: Nonuniversal critical dynamics in Monte Carlo simulations. In: *Phys. Rev. Lett.* 58, p. 86-88 (1987)

[52] Tukey, J.: *Exploratory Data Analysis.* Addison-Wesley, 1977

[53] Waldspurger, I.: *Wavelet Transform Modulus: Phase Retrieval and Scattering*, Ecole Normale Superieure, PhD thesis, 2015

[54] Wang, J. ; Swendsen, R.: Cluster Monte Carlo algorithms. In: *Physica A: Statistical Mechanics and its Applications* 167, no. 3, p. 565-579 (1990)

[55] Wilcox, R.: *Introduction to Robust Estimation and Hypothesis Testing.* US Academic Press, 2012

[56] Zou, H. ; Hastie, T. ; Tibshirani, R.: Sparse Principal Component Analysis. In: *Journal of Computational and Graphical Statistics* 15, no. 2, p. 265-286 (2006)