Pontus Westermark

Wavelets, Scattering transforms and Convolutional neural networks

Tools for image processing

UPPSALA
UNIVERSITET

Contents

*

1. Introduction

Wavelets in their modern form has been studied since the 1980's [25] and has found many applications in signal processing such as in compression and image denoising [18] and the theoretical properties of the waveforms are well understood.

In the most recent years, wavelets have been extended to the domain of machine learning and neural networks, which provides a way to enhance a neural network with well-defined mathematical properties.

This thesis aims to present all the relevant theory to be able to understand wavelets and how they can be used to define the scattering transform. Then we cover the theoretical properties of the scattering transform and how it can be used in handwritten digit classification.

Moreover, the appendices contain several numerical examples which show both a multi-resolution analysis of, and how the scattering transform acts on, a signal. All figures in this thesis has been developed by the author.

2. Mathematics

2.1 Measure and integration

A lot of important theorems and results that we will use originate from Fourier analysis and functional analysis, which in turn rely on measure and integration theory. Thus we talk briefly about which measure spaces we consider and the notation that we use for them. However in practice we need mostly ordinary calculus to work with the integrals that we will encounter. For a complete overview of measure theory, see [1].

Our measure space will be $\mathbb{R}^n$ for some positive integer n, equipped with the associated Lesbegue measure, here denoted μ. We will allow for any common denotation of an integral as made precise below.

Notation 1. If f is an integrable function with $\mathbb{R}^n$ as its domain, and $x = (x_1, ..., x_n)$ an n-dimensional vector, we permit ourselves to write its Lesbegue integral over f's entire domain as follows.

$$\begin{aligned}\int_{\mathbb{R}^n} f d\mu &= \int f d\mu \\ &= \int f(x_1, ..., x_n) dx_1 ... dx_n \\ &= \int f(x) dx.\end{aligned}$$

Similarly, for integration of f over a subset $((a_1, b_1), ..., (a_n, b_n)) \subset \mathbb{R}^n$ we write

$$\begin{aligned}\int_A f d\mu &= \int_{a_1}^{b_1} ... \int_{a_n}^{b_n} f(x_1, ..., x_n) dx_1 ... dx_n \\ &= \int_{a_1}^{b_1} ... \int_{a_n}^{b_n} f(x) dx.\end{aligned}$$

Finally, for integration over $(a, b) \subseteq \mathbb{R}$ we write integration over the inverse orientation as $\int_b^a f(x)dx = -\int_a^b f(x)dx$.

This may seem like a redundant amount of ways to write an integral, but it will aid us later since we will work a lot with concepts based on different types of integration.

2.2 Function spaces

We begin by giving a suitable definition of an L^p-space, followed by a closer look at two important special cases, namely L^1 and L^2 over both $\mathbb{R}$ and $\mathbb{R}^2$. Then we will briefly look at C^k spaces of continuous and continuously differentiable functions, with which we assume that the reader is already familiar.

2.2.1 L^p-space

Definition 1. For a measure space $\mathbb{R}^n$ with associated Lesbegue measure μ, the function space $L^p(\mathbb{R}^n)$ is the collection of all functions $f : \mathbb{R}^n \to \mathbb{C}$ such that

$$\int_{\mathbb{R}^n} |f|^p d\mu < \infty. \tag{2.1}$$

In this definition, $\mathbb{R}^n$ signifies the domain of the functions of $L^p(\mathbb{R}^n)$. This motivates the following notation.

Notation 2. We write simply L^p when the domain is either obvious in the current context, or if we wish to speak about something which holds for $L^p(\mathbb{R}^n)$ for all n.

Definition 2. To each L^p space, we denote the L^p norm

$$||f||_p = (\int |f|^p d\mu)^{1/p}.$$

Notation 3. When it is obvious to which L^p space a function f belongs, we will not write the subscript of the norm, i.e. $||f||_p$ will be written $||f||$.

That all L^p spaces are vector spaces is apparent from their respective definitions. More information about their properties can be found in [1].

Square-integrable, $L^2(\mathbb{R})$ and $L^2(\mathbb{R}^2)$

A function $f \in L^2(\mathbb{R})$ or $f \in L^2(\mathbb{R})$ is said to be square-integrable if its squared modulus has a finite value. That is, $f \in L^2(\mathbb{R})$ is square integrable if

$$\int_{-\infty}^{\infty} |f(t)|^2 dt < \infty.$$

Similarly, $f \in L^2(\mathbb{R}^2)$ and is also said to be square integrable if

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |f(x,y)|^2 dxdy < \infty.$$

Absolutely integrable, $L^1(\mathbb{R})$ and $L^1(\mathbb{R}^2)$

Analogously, we say that a function f is absolutely integrable if $f \in L^1(\mathbb{R})$ or if $f \in L^1(\mathbb{R}^2)$ and

$$\int_{-\infty}^{\infty} |f(t)|dt < \infty, \text{ or if } \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |f(x,y)|dxdy < \infty \text{ respectively.}$$

Example 1. The function $f(x) = \frac{1}{1+|x|}$ is in $L^2(\mathbb{R})$ but not in $L^1(\mathbb{R})$, as can be seen by the following equations.

$$\frac{1}{1+|x|} = \frac{d}{dx}\log(1+|x|), \quad \frac{1}{(1+|x|)^2} = \frac{d}{dx}\frac{1}{1+|x|}.$$

The integral over $\mathbb{R}$ of the former clearly does not converge to a finite value while the latter does.

2.2.2 Continuous function spaces

Definition 3. For k and n positive integers, we define $C^k(\mathbb{R}^n)$ as the space of all k times continuously differentiable functions $f : \mathbb{R}^n \to \mathbb{C}$.

Notation 4. In particular, we write $C^0(\mathbb{R}^n) = C(\mathbb{R}^n)$, and if n is obvious from the context or if we wish to talk about all k times continuously differentiable functions with range $\mathbb{C}$, we simply write $C^k = C^k(\mathbb{R}^n)$ or $C = C^0$.

2.3 Hilbert spaces

Equipped with the inner product $\langle f, g \rangle = \int f \bar{g} d\mu$, where $\bar{g}$ denotes complex conjugation of g, a function space $L^2(\mathbb{R}^n)$ becomes a normed vector space. To be more specific, it becomes a Hilbert space.

Definition 4. A normed vector space V is said to be complete if for all sequences (x_n) in V, if (x_n) converges to x then x is also in V.

Definition 5. A *Hilbert space H* is a complete normed vector space equipped with an inner product.

For the Hilbert spaces $L^2(\mathbb{R}^n)$, we have already mentioned its inner product. For a function $f \in L^2(\mathbb{R}^n)$, we denote the norm from (2.1) without subscript,

$$||f|| = ||f||_2 = \sqrt{\int |f|^2 d\mu}.$$

For a proof that $L^2(\mathbb{R}^n)$ is in fact a Hilbert space, see [7].

2.4 Convolutions

Later on, we will define wavelet transforms through convolution, which is why we will cover some of its most relevant properties here.

Definition 6. The convolution operator $\star$ of two function f and g with shared domain is defined by

$$(f \star h)(x) = \int f(y) g(x-y) dy. \tag{2.2}$$

Example 2. For $f \in L^1(\mathbb{R}^2)$ and $g \in L^2(\mathbb{R}^2)$ we have

$$(f \star g)(y_1, y_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) g(y_1 - x_1, y_2 - x_2) dx_1 dx_2.$$

Proposition 1. For three functions f,g,h in $L^p(\mathbb{R}^n)$, the following equalities hold whenever both sides are finite almost everywhere.

1. $f \star g = g \star f$,
2. $f \star (g+h) = f \star g + f \star h$.
3. $(f \star g) \star h = f \star (g \star h)$.

Proof. 1. follows from the obvious change of variables $y = y - x$, and 2. follows from $\int (a+b)d\mu = \int a d\mu + \int b d\mu$. Finally, 3. follows from theorem 1 below. □

In particular, the equalities in proposition 1 are always true when $f,g,h \in L^1$. This follows from an important result found in Haim Brezis book on functional analysis [7], which will only be stated here.

Theorem 1. Let $f \in L^1(\mathbb{R}^n)$ and let $g \in L^p(\mathbb{R}^n)$ with $1 \leq p \leq \infty$. Then for almost every $x \in \mathbb{R}^n$ the function $y \mapsto f(x-y)g(y)$ is integrable on $\mathbb{R}^n$.

In addition, $f \star g \in L^p(\mathbb{R}^n)$ and $||f \star g||_p \leq ||f||_1 ||g||_p$.

Proof. See [7]. □

Proposition 2. Suppose $f \in L^p(\mathbb{R}^n)$ and $g \in L^q(\mathbb{R}^n)$, and that $f \star g$ is finite almost everywhere. Suppose also that g is continuous. Then the convolution $f \star g$ is continuous.

Proof.

$$\begin{aligned} &\lim_{\tau \to 0} |(f \star g)(x+\tau) - (f \star g)(x)| \\ &= \lim_{\tau \to 0} \left| \int f(y) g(x+\tau-y) dy - \int f(y) g(x-y) dy \right| \\ &= \lim_{\tau \to 0} \left| \int f(y) (g(x+\tau-y) - g(x-y)) dy \right| \\ &\leq \lim_{\tau \to 0} \int |f(y)| |g(x+\tau-y) - g(x-y)| dy = 0. \end{aligned} \tag{2.3}$$

□

Proposition 3. Suppose that f is continuous, $g \in C^1$, and that $f \star g$ is finite almost everywhere. Then $f \star g \in C^1$. In particular, we have that

$$\frac{d}{dx_i}(f \star g) = f \star (\frac{d}{dx_i} g) = (\frac{d}{dx_i} f) \star g.$$

Proof. $\frac{d}{dx_i}(f \star g) = f \star (\frac{d}{dx_i} g)$, since

$$\begin{aligned}
&|\lim_{h \to 0} \int \frac{f(y)g(x+hx_i-y) - f(y)g(x-y)}{h} dy - \int f(y) \frac{dg}{dx_i}(x-y) dy| \\
&= |\lim_{h \to 0} \int f(y) \left(\frac{g(x+hx_i-y) - g(x-y)}{h} - \frac{dg}{dx_i}(x-y) \right) dy| \\
&= |\int f(y) \left(\frac{dg}{dx_i}(x-y) - \frac{dg}{dx_i}(x-y) \right) dy| = 0.
\end{aligned}$$

The second equality follows by the obvious change of variables. □

***Corollary* 1.** Suppose f and g are absolutely integrable function and g is k times continuously differentiable. Then the convolution $f \star g$ is k times continuously differentiable.

Proof. Since $\frac{d}{dx_i}(f \star g) = f \star \frac{d}{dx_i} g$ is continuous by proposition 2, the corollary follows from induction over $\frac{d}{dx_i}$. □

***Corollary* 2.** Suppose f and g are absolutely integrable functions, f is k times continuously differentiable and g is l times continuously differentiable. Then the convolution $f \star g$ is $k+l$ times continuously differentiable.

Proof. Follows immediately from Corollary 1 and commutativity of the convolution operator. □

2.5 Frequency analysis

We begin by relating a function $f \in L^1(\mathbb{R}^n)$ to its Fourier transform $\hat{f}(\omega)$, which changes the domain of f from the spatial domain to the frequency domain and provides the coefficients of f's constituent waveforms $e^{i\omega \cdot x}$ in the latter. Then we generalize the Fourier transform to the function spaces $L^2(\mathbb{R}^n)$. Sequentially we define and refer to a proof of the inverse Fourier transform which provides an important link that allows us to go "to and from" the frequency domain. We finish this section by deriving some properties of the Fourier transform .

Notation 5. We will sometimes also refer to the Fourier transform of a function f by $\mathscr{F}(f)(\omega)$.

2.5.1 The $L^1(\mathbb{R})$ Fourier transform

Given $f \in L^1(\mathbb{R})$, we define the Fourier transform of f as

$$\hat{f}(\omega) = \mathscr{F} f(\omega) = \int_{-\infty}^{\infty} f(x) e^{-i\omega x} dx. \tag{2.4}$$

2.5.2 The $L^1(\mathbb{R}^2)$ Fourier transform

Analogously, given $f \in L^1(\mathbb{R}^2)$ we define the Fourier transform of f as

$$\hat{f}(\omega_1, \omega_2) = \mathscr{F} f(\omega_1, \omega_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) e^{-i(\omega_1 x + \omega_2 y)} dx dy. \tag{2.5}$$

If we let $\omega = (\omega_1, \omega_2)$ and $z = (x,y)$, we can write (2.5) as

$$\hat{f}(\omega) = \mathscr{F} f(\omega) = \int f(z) e^{-i\omega \cdot z} dz. \tag{2.6}$$

2.5.3 The $L^1(\mathbb{R}^n)$ Fourier Transform

The striking similarity between (2.6) and (2.4) allows us to define the general n-dimensional Fourier transform as follows.

Definition 7. For $f \in L^1(\mathbb{R}^n)$, we define the n-dimensional Fourier transform of f, for frequencies $\omega = (\omega_1, ..., \omega_n)$ as

$$\hat{f}(\omega) = \mathscr{F} f(\omega) = \int f(x) e^{-i\omega \cdot x} dx. \tag{2.7}$$

Proposition 4. If $f \in L^1(\mathbb{R}^n)$ and $\hat{f} \in L^1(\mathbb{R}^n)$, then the following equation is valid and is called the inverse Fourier transform,

$$f(x) = \frac{1}{(2\pi)^n} \int e^{i\omega \cdot x} \hat{f}(\omega) d\omega.$$

References to proofs can be found in [23]. A proof for the one-dimensional case is found in [18].

2.5.4 The $L^2(\mathbb{R})$ Fourier transform

Note that if $f \in L^1(\mathbb{R})$, equation 2.7 always converges to some value as can be seen by observing that $|\hat{f}(\omega)| \leq \int |f| d\mu$. For $f \in L^2(\mathbb{R}^n)$, we can not come to the same conclusion.

We need to deal with the remark above by considering functions in $f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. How this is done is thoroughly explained in [18] and will not be repeated here. We only repeat two important theorems found therein and note that there is a natural extension of the Fourier transform to any $f \in L^2(\mathbb{R})$.

Theorem 2. *Parseval's formula in one dimension.* If f and h are in $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, then

$$\int_{-\infty}^{\infty} f(t) \bar{h}(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega) \bar{\hat{h}}(\omega) d\omega. \tag{2.8}$$

Plancherel's formula in one dimension. For $h = f$ it follows that

$$\int_{-\infty}^{\infty} |f(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\hat{f}(\omega)|^2 d\omega. \tag{2.9}$$

2.5.5 The $L^2(\mathbb{R}^n)$ Fourier Transform

A similar extension can be done for $L^2(\mathbb{R}^n)$ by considering $L^1(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$. By a density argument, one gets the general n-dimensional Parseval's formula and the n-dimensional Plancherel's formula.

Theorem 3. *Parseval's formula.* If f and h are in $L^2(\mathbb{R}^n)$, then

$$\int f(x)\bar{h}(x)dx = \frac{1}{2\pi}\int \hat{f}(\omega)\bar{\hat{h}}(\omega)d\omega. \tag{2.10}$$

Plancherel's formula. When $h = f$ it follows that

$$\int |f(x)|^2 dx = \frac{1}{2\pi}\int |\hat{f}(\omega)|^2 d\omega. \tag{2.11}$$

Proof. For proof of (2.10), see [17]. (2.11) follows as an immediate corollary. □

From the Parceval and Plancherel formulas, we see that for f and g in L^2, $||f|| = ||\hat{f}||$ and $\langle f, g\rangle = \langle \hat{f}, \hat{g}\rangle$.

Proposition 5. If $f \in L^2(\mathbb{R}^n)$, then the Fourier transform $\hat{f}$ is in $L^2(\mathbb{R}^n)$ and is invertible, with its inverse given by the reconstruction formula

$$f(x) = \mathscr{F}^{-1}(\hat{f}) = \frac{1}{(2\pi)^n}\int \hat{f}(\omega)e^{i\omega\cdot x}d\omega. \tag{2.12}$$

Proof. See [17]. □

Observation 1. Any reader that sets out to verify the propositions introduced will encounter that in some works, the frequency ω in $\hat{f}(\omega)$ is multiplied by 2π in the integral of the transform. That is, $\hat{f} = \int f(x)e^{i2\pi\omega\cdot x}dx$. In such cases, some fractions $1/(2\pi)^n$ necessary in our propositions will vanish. This can be seen by looking at the n-dimensional reconstruction formula with $\omega = (\omega_1, ..., \omega_n)$ and $x = (x_1, ..., x_n)$.

$$\begin{aligned}
&\int \hat{f}(\omega)e^{i2\pi\omega\cdot x}dx = \int f(z)e^{-i2\pi z\omega}e^{i2\pi\omega\cdot x}dzd\omega \\
&= \text{ change of variables } (\omega = 2\pi\omega), \\
&= \frac{1}{(2\pi)^n}\int f(z)e^{-i\omega\cdot x}e^{i\omega\cdot x}dxd\omega.
\end{aligned}$$

From the equation above, it is clear that both definitions coincide. We have chosen to follow the style of [18], and thus we will keep the fractions.

2.5.6 Fourier transform properties

There are several useful properties of the Fourier transform. The ones we introduce below are essential for some interpretations of wavelets. Most of these properties are proven by a change of variables in the Fourier transform or in the reconstruction formula.

We will only state the propositions for $f \in L^2(\mathbb{R}^2)$, since the general case of $f \in L^2(\mathbb{R}^n)$ can be shown almost identically.

We denote function variables as x, y over the spatial domain and as ω_1, ω_2 over the frequency domain. So for example if we write $f(ax, by)$, we mean that we consider the function $f'(x,y) = f(ax, by)$, as will become apparent below.

Scaling property

$$\text{For } (a,b), a \neq 0, b \neq 0, \quad f(ax, by) \leftrightarrow \frac{1}{|a||b|} \hat{f}(\frac{\omega_1}{a}, \frac{\omega_2}{b}). \tag{2.13}$$

Derivation: If $a > 0$, $b = 1$, consider the change of variables $x = ax, y = y$. From evaluation of the Fourier transform we see that

$$\begin{aligned}
\hat{f}(\omega_1, \omega_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(ax, by) e^{-i(\omega_1 x + \omega_2 y)} dxdy \\
&= \text{ by the change of variables} \\
&= \frac{1}{a} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, by) e^{-i(\omega_1 \frac{x}{a} + \omega_2 y)} dxdy \\
&= \frac{1}{a} \hat{f}(\frac{\omega_1}{a}, \omega_2).
\end{aligned}$$

Similarly if $a < 0$, say $a = -c$, and $b = 1$, we have a change of limits of integration which leads us to the modulus in 2.13, namely

$$\begin{aligned}
\hat{f}(\omega_1, \omega_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(ax, by) e^{-i(\omega_1 x + \omega_2 y)} dxdy \\
&= \text{ by the change of variables} \\
&= \frac{-1}{c} \int_{-\infty}^{\infty} \int_{\infty}^{-\infty} f(x, by) e^{-i(\omega_2 y - \omega_1 \frac{x}{c})} dxdy \\
&= \frac{1}{c} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, by) e^{-i(\omega_2 y - \omega_1 \frac{x}{c})} dxdy \\
&= \frac{1}{c} \hat{f}(\frac{-\omega_1}{c}, \omega_2) = \frac{1}{|a|} \hat{f}(\frac{\omega_1}{a}, \omega_2)
\end{aligned}$$

Setting $a \neq 0$, $b \neq 0$, there will be four cases to consider, it should be clear from the changes of variables how the full derivation is then performed.

Frequency-shift property

$$e^{i\xi_1 x} e^{\xi_2 y} f(x,y) \leftrightarrow \hat{f}(\omega_1 - \xi_1, \omega_2 - \xi_2) \tag{2.14}$$

Derivation. Just as for the scaling property, this follows from a suitable change of variables and writing out the Fourier transform defined by (2.7).

Convolution in the time domain

$$(f \star g)(x,y) \leftrightarrow \frac{1}{2\pi} \hat{f}(\omega_1, \omega_2) \hat{g}(\omega_1, \omega_2). \tag{2.15}$$

See [17] for proof.

Conjugate symmetry

Suppose f is a real-valued function in $L^2(\mathbb{R}^n)$. Then $\hat{f}(\omega) = \overline{\hat{f}(-\omega)}$.

Derivation.

$$\begin{aligned}
\overline{\hat{f}(-\omega)} &= \overline{\int f(x) e^{-i(-\omega)\cdot x} dx} \\
&= \overline{\int f(x) e^{i(\omega)\cdot x} dx} \\
&= \overline{\int f(x) cos(i\omega \cdot x) dx + i \int f(x) sin(i\omega \cdot x) dx} \\
&= \int f(x) cos(i\omega \cdot x) dx - i \int f(x) sin(i\omega \cdot x) dx \\
&= \hat{f}(\omega).
\end{aligned} \tag{2.16}$$

2.5.7 On time and frequency

It is important to understand the connection between the spatial domain and the frequency domain associated to a function f. What we have introduced before should be familiar to most readers but we should emphasize certain points.

Functions have frequencies. This is confirmed from the theory introduced above, but it is not merely some esoteric mathematics that deliver useful theoretical results. It is something practical and witnessed in nature.

Example 3. A ray of white light can be divided into its frequency components using a prism, and later on reconstructed from these frequency components.

Motivated by the example, we can understand that there is a spectrum of frequencies which range from low to high oscillations. This is important from the interpretation that if we eliminate high frequency components of a signal, we obtain a signal which has a lower frequency spectrum, more slowly oscillating frequency components and is in a sense less detailed.

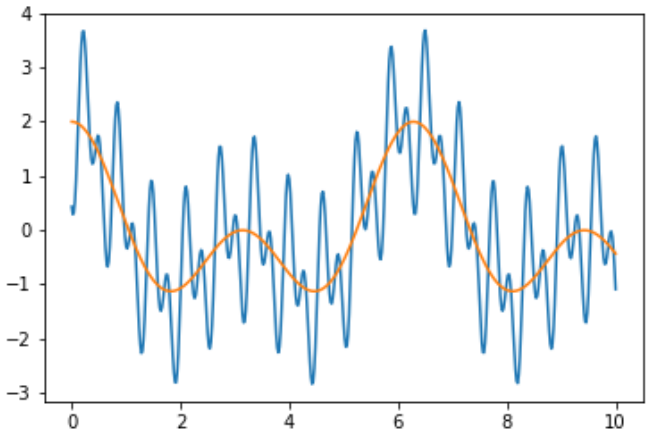

Figure 2.1. Functions f and f_l from example 4.

Example 4. Consider $f(t) = cos(x) + cos(2x) + cos(10x + 2 + \frac{\pi}{2}) + cos(20x + 7 + \frac{\pi}{2})$, and its lower frequency companion $f_l(t) = cos(x) + cos(2x)$. If f is the original signal of the two, $f_l(t)$ is a less detailed representation of f.

2.6 Frames

A well-known basis for functions f which are absolutely integrable over the unit circle is $\{e^{int}\}_{n\in\mathbb{Z}}$. The expansion of f in this basis is achieved with a Fourier transform of f over the unit circle, which can be defined similarly to (2.7).

Using this as motivation, we will now define frames, which will allow us to create an analysis of a function based on its frequency components.

Definition 8. For a Hilbert space H, and an arbitrary index set Γ, we say that the family $\{\phi_\gamma\}_{\gamma\in\Gamma}$ is a frame of H if three exists two constants $0 < A \leq B$ such that

$$\forall f \in H,\ A||f||^2 \leq \sum_{n\in\Gamma} |\langle f, \phi_n\rangle|^2 \leq B||f||^2. \tag{2.17}$$

Furthermore, if a frame $\{\phi_\gamma\}_{\gamma\in\Gamma}$ is linearly independent, then we call it a Riesz basis [12].

Frames are important in the study of wavelets and in the domain of signal processing. The interested reader should consult [18], from which the definition is borrowed, for more information. For us it is sufficient to have a some notion of what it means for a family of vectors $\{\phi_\gamma\}_{\gamma\in\Gamma}$ to localize most of a functions total energy, which is what (2.17) does.

2.7 Multi-resolution analysis

Here we introduce a type of analysis that serves to create a direct correspondence between one-dimensional wavelets and discrete filter banks [18]. While we will briefly discuss some filters later on, discrete filters are important for numerical implementations but falls outside the scope of this document.

The goal is to arrive at a decomposition of a given function f into some type of averaging and a collection of different details. This can be achieved by a sequence of successive approximation spaces V_j [11]. In particular, following the definition in [18] and generalization to n dimensions in [21], we say that a sequence $\{V_j\}$ of closed subspaces of $L^2(\mathbb{R}^n)$ is a multiresolution approximation if the six properties presented below are satisfied.

2.7.1 Properties 1, 2, 3

$$\forall j \in \mathbb{Z}, k \in \mathbb{Z}^n, f(x) \in V_j \Leftrightarrow f(x - 2^j k) \in V_j, \tag{2.18}$$

$$\forall j \in \mathbb{Z}, V_{j+1} \subset V_j, \tag{2.19}$$

$$\forall j \in \mathbb{Z}, f(x) \in V_j \Leftrightarrow f(\frac{x}{2}) \in V_{j+1}, \tag{2.20}$$

We see that by (2.19), ... $\subset V_3 \subset V_2 \subset V_1 \subset V_0$, and (2.20) requires that for a given function f_0 in V_0, there are smoothed out version f_i of f_0 for $i < 0$ such that f_i is in V_i. An illustration will be given when we introduce the one-dimensional wavelet.

2.7.2 Properties 4, 5

$$\lim_{j \to +\infty} V_j = \bigcap_{j=-\infty}^{+\infty} V_j = \{0\}, \tag{2.21}$$

$$\lim_{j \to -\infty} V_j = \text{Closure}\Big(\bigcup_{j=-\infty}^{+\infty} V_j \Big) = L^2(\mathbb{R}^n) \tag{2.22}$$

(2.21) shows that as j tends to infinity, the component of a function f in V_j goes to 0. Or, V_j will eventually contain no details of f at all.

Similarly, (2.22) shows that as j tends to the negative infinity, for a given function $f \in L^2(\mathbb{R}^n)$, V_j eventually contains all the information of f.

For our purposes, (2.22) is not of very practical since our functions will be signals which have been sampled (e.g. sounds and images). In general, we will have our samples of f in V_0 and successively deal with decomposing f into subspaces of V_0, i.e. as j increases.

2.7.3 Property 6

There exists a θ such that $\{\theta(x-n)\}_{n\in\mathbb{Z}^n}$ is a Riesz basis of V_0. (2.23)

Proposition 6. For θ such that $\{\theta(x-n)\}_{n\in\mathbb{Z}^n}$ is a Riesz basis of V_0, it follows that $\{2^{-(nj/2)}\theta(2^{-j}x-m)\}_{m\in\mathbb{Z}^n}$ is a Riesz basis for V_j.

Proof. Let $\theta_{j,m}(t) = 2^{-(mj/2)}\theta(2^{-j}x-m)$, and $\theta_m = \theta_{0,m} = \theta(x-m)$.

Suppose that $f \in V_j$. The proposition follows from the change of variables $x = 2^{-j}x$ in the following equation.

$$\begin{aligned}\langle f, \theta_{j,m}\rangle &= \frac{1}{2^{mj/2}}\int f(x)\bar{\theta}(2^{-j}x-n)dx \\ &= \text{ by the change of variables} \\ &= 2^{mj/2}\int f(2^j x)\bar{\theta}(x-m)dx \\ &= \langle f(2^j\cdot), 2^{mj/2}\theta_m\rangle\end{aligned} \tag{2.24}$$

For $f = \theta_{j,m}$, we get orthogonality. Since by (2.20), $f(2^j x) \in V_0$ and $\{\theta_m\}_{m\in\mathbb{Z}^n}$ is a Riesz basis for V_0, then $\{\theta_{j,m}\}_{m\in\mathbb{Z}^n}$ is a Riesz basis for V_j. □

Notation 6. The function θ is also called the scaling function associated with the multi-resolution analysis.

Proposition 6 provides us with an important interpretation of the spaces $\{V_j\}_{j>0}$ as successively lower scale approximations of a function f. More precisely, for $f \in V_0$ there exists a lower resolution approximation of f in V_j which we will be able to access through the projection of f onto V_j by a low-pass filter.

2.8 Wavelets

Wavelets are oscillating wave-like forms that provide both time and frequency localization of a signal. This time-frequency localization property of the wavelet transform has great theoretical and practical value, and it is the foundation for the scattering transform, which is our primary study of interest.

We will cover wavelets in both one and two dimensions but have decided to treat each dimension in its own section because the one-dimensional case is much easier to develop in a thorough manner.

For the two-dimensional case, we will be more brief in the technicalities as they would otherwise occlude what we want to express, using the one-dimensional case as motivation. The following definitions have been assembled from [4, 11, 18].

Definition 9. A one-dimensional wavelet ψ is a function in $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ with unit norm with respect to the $L^2(\mathbb{R})$ norm and which satisfies the admissibility condition

$$C_\psi = \int_0^\infty \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty. \tag{2.25}$$

Definition 10. A two-dimensional wavelet ψ is a function in $L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$ with unit norm with respect to the $L^2(\mathbb{R}^2)$ norm and which satisfies the admissibility condition

$$C_\psi = \int_{-\infty}^\infty \int_{-\infty}^\infty \frac{|\hat{\psi}(x,y)|^2}{|(x,y)|^2} dxdy < \infty.(\text{[4]}) \tag{2.26}$$

Notation 7. For either a one- or two-dimensional wavelet, we write just wavelet if it is clear from the context which dimension we are talking about.

The above definitions implies that $\hat{\psi}(0) = \int \psi dx = 0$ since otherwise neither defining equation could possibly converge to a finite value. This provides motivation for the following notation.

Notation 8. We say that a function f is oscillating when it takes on both positive and negative values.

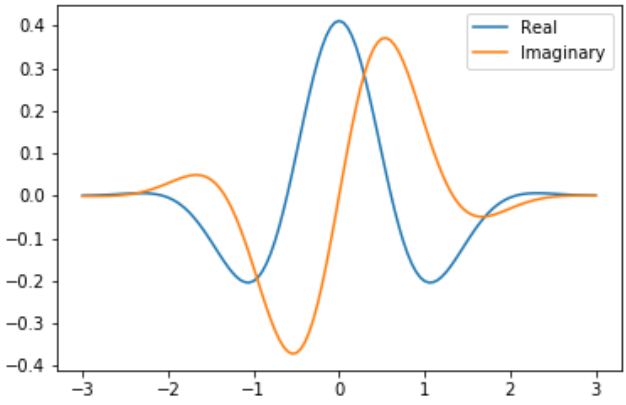

Figure 2.2. Real and imaginary part of a one-dimensional Morlet wavelet

Definition 11. Consider a one-dimensional wavelet ψ. For any function f in $L^2(\mathbb{R})$, we define the one-dimensional wavelet transform through the convolution $f \star \psi$. Analogously, we define the two-dimensional wavelet transform for a two-dimensional wavelet ψ and function $f \in L^2(\mathbb{R}^2)$ as $f \star \psi$.

Notation 9. We call both the one- and two-dimensional wavelet transforms just a wavelet transform, when it is clear which one is intended, or when the discussion is related to both of them.

Example 5. The one-dimensional Morlet wavelet is defined as

$$\psi(x) = \alpha(e^{i\xi \cdot x} - \beta)e^{-x^2/(2\sigma^2)},$$

where ξ, σ are parameters and α, β are calculated such that $\int \psi d\mu = 0$, and $\int |\psi|^2 d\mu = 1$ [9]. For $\xi = 3\pi/4, \sigma = 0.85$, we have

$$\psi(x) \approx 0.476(e^{i\frac{3}{4}\pi x} - 0.135)e^{-x^2/(1.445)}$$

which has be plotted in fig. 2.2.

2.9 One-dimensional wavelets

Here we will look at one-dimensional wavelets in more detail, which provide a clear view of what we are trying to accomplish through the lens of multi-resolution analysis. The concepts that we introduce here are also what guides our decomposition of functions onto wavelets in two dimensions.

2.9.1 Connection to multi-resolution analysis

In one dimension, [11] proposes that whenever we have a multiresolution analysis (fulfilling the 6 properties previously covered), then there is a wavelet ψ that we can construct explicitly, such that $\{\psi_{j,k} \mid j,k \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathbb{R})$ where each $\psi_{j,k}$ is defined as we did for $\theta_{j,n}$ in the proof for proposition 6, that is,

$$\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k),$$

such that the following hold for all $f \in L^2(\mathbb{R})$,

$$P_{j-1}f = P_j f + \sum_{k\in\mathbb{Z}} < f, \psi_{j,k} > \psi_{j,k}, \tag{2.27}$$

where $P_j f$ is the projection of f onto V_j.

Furthermore, suppose that we have a function $f \in V_0$, and that this is the highest resolution approximation that we have access to. Then by (2.27), we can write

$$f = P_1 f + \sum_{k\in\mathbb{Z}} < f, \psi_{1,k} > \psi_{1,k}.$$

Repeating this construction, we obtain a full representation of $f \in V_0$ either by

$$f = \sum_{i\in\mathbb{Z}, i\leq 1} \sum_{k\in\mathbb{Z}} < f, \psi_{j,k} > \psi_{j,k}, \tag{2.28}$$

or by

$$f = P_{V_i} f + \sum_{j=1}^{i} \sum_{k\in\mathbb{Z}} < f, \psi_{j,k} > \psi_{j,k} \tag{2.29}$$

We see that any function $f \in L^2(\mathbb{R})$ can be decomposed either fully onto wavelets or onto a low resolution approximation and several detail components.

2.9.2 Probability theory for wavelets

This will be a very brief introduction to some concepts from probability theory which applies to one-dimensional wavelets and can be extended to the two-dimensional case with little effort. We will follow the discourse in [18] and begin by recalling (2.13) which relates a dilation in the time domain to a contraction in the frequency domain

$$f(at) \leftrightarrow \frac{1}{|a|}\hat{f}(\frac{\omega}{a}),$$

which hints that we will not be able to localize a function both in time and frequency, an argument which will be made precise in upcoming sections.

We will defer any rigorous definition of a continuous random variable to [14], here denoted X, and simply note that X is associated with a density function $f : \mathbb{R} \to \mathbb{R}$ such that $\int f d\mu = 1$ and the probability that $X \leq x$ is given by

$$\mathbf{P}(X \leq x) = \int_{-\infty}^{x} f(x)dx.$$

Furthermore, recall that for a wavelet ψ, $\int |\psi|^2 d\mu = 1$, so we can interpret $|\psi|^2$ as a density function for a continuous random variable.

Definition 12. The expected value of a continuous random variable X with density function f is defined as:

$$\mathbb{E}(X) = \int xf(x)dx \tag{2.30}$$

Definition 13. The variance of a continuous random variable X with density function f is defined as:

$$\sigma^2 = \mathbf{V}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \int (x - \mathbb{E}(X))^2 f(x)dx \tag{2.31}$$

2.9.3 Time-frequency localization

Suppose that we have a given family of wavelets $\mathscr{W} = \{\psi_{j,k}\}_{j,k\in\mathbb{Z}}$. For $\varphi \in \mathscr{W}$, we know from the definition of a wavelet that $\int |\varphi|^2 d\mu = 1$, and similarly by Plancherel's formula (2.11), $(2\pi)^{-1}\int |\hat{\varphi}|^2 d\mu = 1$. Thus φ and $\hat{\varphi}$ can be interpreted as distribution functions which define random variables.

As per usual, we follow [18] and define for $\psi_{j,k} \in \mathscr{W}$

$$u = \int_{-\infty}^{\infty} x|\psi_{j,k}(t)|^2 dt, \tag{2.32}$$

$$\sigma_t^2 = \int_{-\infty}^{\infty} (t-u)^2|\psi_{j,k}(t)|^2 dt, \tag{2.33}$$

$$\xi = \frac{1}{2\pi}\int_{-\infty}^{\infty} \omega|\hat{\psi}_{j,k}(\omega)|^2 d\omega, \tag{2.34}$$

$$\sigma_\omega^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} (\omega-\xi)^2|\hat{\psi}_{j,k}(\omega)|^2 d\omega. \tag{2.35}$$

From the first pair of equations (2.32) and (2.33), we interpret u and σ_t^2 as the time-localization and time-spread respectively. The time-localization u gives the center of mass of $\psi_{j,k}$ in the time domain, while the time-spread σ_t^2 give the primary support (non-negligible magnitude of $\psi_{j,k}$ around u).

Similarly we get from the second pair of equations (2.34) and (2.35) an interpretation of ξ as the frequency-localization, and σ_ω^2 as the frequency-spread of $\hat{\psi}$, which provides the center of mass and the length of the primary support of $\hat{\psi}$ respectively.

With the following notation, we provide an important theorem from [18], whose general form is known as *Heisenberg's uncertainty principle*.

Theorem 4. The temporal variance σ_t^2 and the frequency variance σ_ω^2 of a wavelet ψ satisfy

$$\sigma_t^2\sigma_\omega^2 \geq \frac{1}{4}. \tag{2.36}$$

Proof. See [18]. □

As a consequence of (2.36), when we increase the scale j, the wavelets $\{\psi_{j,k}\}_{k\in\mathbb{Z}}$ covers a more narrow, lower band of frequencies.

2.9.4 Convolutions and filters

The last important link between time and frequency is to show how wavelets can be considered bandpass-filters, and how the scaling function can be considered a lowpass-filter. We will exploit the convolution properties of the Fourier transform, namely equations 2.15, to analyze the wavelet transform.

Definition 14. A lowpass filter is a functional F that keeps only low frequencies $|\omega| < \eta$ for some $\eta > 0$ from an input function f. Furthermore, θ is the lowpass function associated to F if and only if $\hat{\theta}(\omega) = 0$ for $|\omega| > \eta$ and we can write the filtering as $F(f) = (f \star \theta)(t)$.

Definition 15. A bandpass filter is a functional G that keeps only frequencies $\eta_1 < |\omega| < \eta_2$ for some $0 < \eta_1 < \eta_2$ from an input function f. Furthermore, ψ is the bandpass function associated to G if and only if $\hat{\psi}(\omega) = 0$ for $|\omega| < \eta_1$ or $\eta_2 < |\omega|$ and we can write the filtering as $G(f) = (f \star \psi)(t)$.

Notation 10. The filter functions θ and ψ are also known as transfer functions [24].

It makes sense to talk about the lowpass filter θ and the bandpass filter ψ, or just a filters ξ, and leave the associated functionals F and G unmentioned. There is also some vagueness to the meaning of keeps in the definition. It means that $|\xi(\omega)| << 1$ for ω outside the frequency band covered by a filter ξ.

By the convolution property 2.15, we know that $\mathscr{F}(f \star \psi)(\omega) = \hat{f}(\omega)\hat{\psi}(\omega)$. This means that filtering can be viewed as properties of $\hat{\psi}$ acting on a function f in its frequency-domain.

The choice of symbol θ for lowpass filter is intentional to emphasize a connection between the lowpass filter and the scaling function associated to a multiresolution analysis. Similarly, the choice of ψ for bandpass filter should then indicate a connection between the bandpass filter and an associated wavelet. Before we make this connection formal, we give an example of a filter.

Given the convolutional property of (2.15), we note that the filters are quite "simple" in the sense that they can be defined through their Fourier transforms $\hat{\theta}$ and $\hat{\psi}$ respectively, as non-zero over some frequency range.

Example 6. A lowpass filter θ such that $\hat{\theta} = 1$ for $\omega \in [-\pi/2, \pi/2]$ and 0 otherwise can be calculated by the inverse Fourier transform,

$$\begin{aligned} \theta(t) &= \frac{1}{2\pi} \int \hat{\theta}(\omega) e^{i\omega t} d\omega = \frac{1}{2\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} e^{i\omega t} d\omega \\ &= \left[\frac{1}{2\pi} \frac{1}{it} e^{i\omega t} \right]_{-\frac{\pi}{2}}^{\frac{\pi}{2}} = \frac{sin(\frac{\pi}{2}t)}{2\pi t}. \end{aligned} \tag{2.37}$$

The filter above is known as an ideal lowpass filter, ideal in the sense that it keeps all (and only) the frequencies $|\omega| < \frac{\pi}{2}$. However, it can not be represented by a rational transfer function and thus in practice other filters are constructed for applications [24].

2.9.5 Wavelets and filters

For a wavelet ψ, recall that the frequency support σ_ω^2 of (2.35) gives a measure of width of primary support of ψ. Thus writing the wavelet transform $f \star \psi$ for some function f, we see that a wavelet is a transfer function for some bandpass filter. In general, the wavelet transform for a given scale j is then a dilated bandpass filter. [18]

From the representations (2.22) and (2.29), namely

$$\lim_{j \to +\infty} V_j = \bigcap_{j=-\infty}^{+\infty} V_j = \{0\}, \text{ and}$$

$$f = P_{V_i} f + \sum_{j=1}^{i} \sum_{k \in \mathbb{Z}} < f, \psi_{j,k} > \psi_{j,k},$$

we note that the wavelet transform is a band-pass filter for a function f sampled in V_0, which captures successively lower frequency bands, up until the scaling function ϕ at scale i captures the remaining low frequencies.

Observation 2. A wavelet ψ captures high frequency components, and thus higher details, of a function f, while a scaling function θ captures low frequency components, and thus lower details of f.

2.9.6 Complex Wavelets

In this section, we will explore complex wavelets, introduce analytic wavelets, and explain why complex wavelets are important for the scattering transform.

Definition 16. An analytic wavelet is a wavelet $\psi(x)$ such that

$$\hat{\psi}(\omega) = 0 \text{ for } \omega < 0.$$

It follows from (2.16) that an analytic wavelet is necessarily complex, since if f is real, then $\hat{f}(-\omega) = 0 \Rightarrow \overline{\hat{f}(\omega)} = \overline{0} = 0$.

For the following propositions, suppose that $f \in L^2(\mathbb{R})$ such that f is real, and that ψ is a complex wavelet. Thus we can write $\psi(x) = u(x) + iv(x)$ for real-valued function u, v.

Proposition 7. The wavelet transform $f \star \psi$ is again complex.

Proof. We write $\psi = u + iv$, note that f is real, and calculate

$$f \star \psi = (f \star u) + i(f \star v).$$

□

Proposition 8. If ψ is analytic, then for the wavelet transform $f \star \psi$, we have that $\widehat{f \star \psi}(\omega) = 0$ for $\omega < 0$.

Proof. This follows immediately from the convolution property, equation 2.15,

$$(f \star \psi)(x) \leftrightarrow \hat{f}(\omega)\hat{\psi}(\omega), \tag{2.38}$$

since we know that $\hat{\psi}(\omega)$ is 0 for negative frequencies. □

Observation 3. The wavelet transform of f with $\psi = u + iv$ can be written as its complex polar representation

$$(f \star \psi)(x) = A(x)e^{i\phi(x)},$$

where $A(x)$, $\phi(x)$ are real-valued continuous positive functions. More precisely, we say that $A(x)$ is the amplitude and $\phi(x)$ is the phase of the wavelet transform $(f \star \psi)(x)$, and we write

$$A(x) = |(f \star \phi)(x)|, \tag{2.39}$$

$$\phi(x) = \arctan\left(\frac{(f \star u)(x)}{(f \star v)(x)}\right), \tag{2.40}$$

where (2.40) is well-defined for all $(f \star v)(x) \neq 0$, and extended to $\phi(x) = 0$ whenever $(f \star v)(x) = 0$.

This observation is incredibly important. It means that complex wavelets allows us to analyze the phase of a signal. In the scattering transform, we will remove the complex phase of a signal and thereby remove some of its oscillations. The leads to a lower resolution signal.

2.10 Two-dimensional wavelets

A two-dimensional wavelet $\psi(x,y)$ works similarly to a one-dimensional one. The only problem is that almost everything requires a more rigorous and sometimes non-informative treatment of theoretical results, and a lot of the required details fall outside the scope of this document. Therefore we will only go through the necessary results that provides a link to the more easily digested one-dimensional counterpart.

2.10.1 Time-frequency spread

Since a two-dimensional wavelet $\psi(x,y)$ is normalized w.r.t the $L^2(\mathbb{R}^2)$ norm, we note that

$$1 = ||\psi|| = \int |\psi(x,y)|^2 dx = 1.$$

This means that $|\psi(x,y)|^2$ can be viewed as a joint probability density function to the random variables X and Y [14]. Similarly to (2.32) and (2.34) for the one-dimensional wavelet, we can define the localization of X and Y respectively, for both time and frequency. That is, we let

$$u(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x|\psi(x,y)|^2 dxdy, \tag{2.41}$$

$$u(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y|\psi(x,y)|^2 dxdy, \text{ and} \tag{2.42}$$

$$\xi(X) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x|\hat{\psi}(x,y)|^2 dxdy, \tag{2.43}$$

$$\xi(Y) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y|\hat{\psi}(x,y)|^2 dxdy. \tag{2.44}$$

Similar to the one-dimensional case, the wavelet $\psi(x,y)$ is centered around $(u(X),u(Y))$ in space, and around $(\xi(X),\xi(Y))$ in frequency.

The spread of ψ in the xy-plane is given by the covariance matrix of X and Y which defines an ellipse within whose boundary most of the wavelet's energy is concentrated. In particular, the diagonal of the matrix defines the length of the major and minor axis of the ellipse, and its symmetrical values off the diagonal defines its rotation. By the same observation, we find an ellipse in the frequency plane.

How these covariance matrices are defined is outside the scope of this document, but can be found in e.g. [14].

There is an analog to the Heisenberg uncertainty principle in one dimension, which limits the possible localization of a wavelet in both time and frequency, by giving a lower bound on the area of the ellipse defined by the covariance matrix.

Theorem 5. *Two-dimensional Heisenberg's uncertainty principle.* Given a two-dimensional wavelet ψ, its time-frequency spread is bounded below by

$$1 \leq \int_{\mathbb{R}^2} |(x,y)|^2 |\psi(x,y)|^2 dxdy \int_{\mathbb{R}^2} |(x,y)|^2 |\hat{\psi}(x,y)|^2 dxdy \tag{2.45}$$

Proof. References to proofs are given in [5]. □

(2.45) is actually the square root of the trace of the covariance matrix in the time domain, multiplied with its counterpart in the frequency domain. Since the area of an ellipse can be determined by the length of its major and minor axis, (2.45) shows that the product of the two ellipses areas is bounded below by 1.

As with the one-dimensional wavelet transform, we can view two-dimensional wavelets as frequency filters based on their frequency localization and spread.

2.10.2 Complex wavelets in two dimensions

Similar to one-dimensional wavelets, the wavelet transform of a real-valued, two-dimensional signal $f(x,y)$ with a complex valued two-dimensional wavelet $\psi(x,y) = u(x,y) + iv(x,y)$ is also complex valued, and we can write its complex polar representation as

$$f \star \psi(x,y) = A(x,y)e^{i\phi(x,y)},$$

where $A(x,y)$ is a non-negative real-valued function, and $\phi(x,y)$ is the complex phase of $f \star \psi$ at (x,y). Furthermore, analogously to the one-dimensional case, we have that

$$A(x,y) = |(f \star \psi)(x,y)|, \text{ and}$$
$$\phi(x,y) = \arctan\left(\frac{(f \star u)(x,y)}{(f \star v)(x,y)}\right)$$

With sufficient dedication, one can even extend the concept of analytic wavelets to two and several dimensions (see [6]). However we will not.

We have mentioned already how a complex wavelet can be used to analyze, and suppress, the phase of a signal. An argument for this is made explicitly [4] who explains the Morlet wavelet in particular as "catching" the phase of a signal. As previously mentioned, this is sort of how the scattering transform works. We "catch" the phase, and then eliminate it.

2.10.3 Rotations

Definition 17. A wavelet ψ is said to be rotation invariant if $\psi(x,y) = \psi(x',y')$ whenever (x,y) and (x',y') lies on the same circle of radius r from the origin, or when $\psi(x) = \psi(-x)$ in the one-dimensional case.

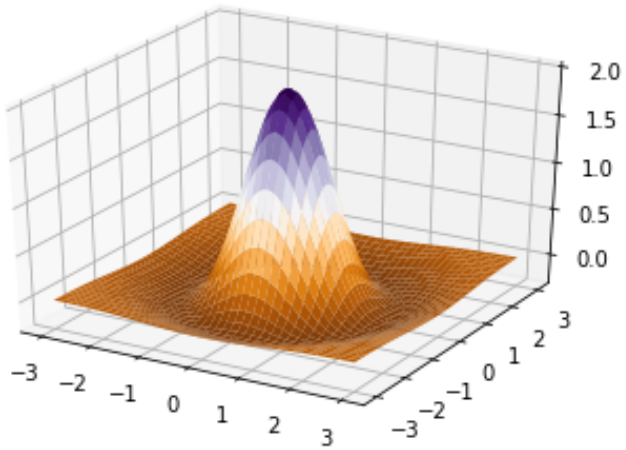

Figure 2.3. The rotation invariant Mexican hat wavelet.

Example 7. The two-dimensional Mexican hat wavelet defined by

$$\psi(x,y) = (2 - |(x,y)|^2)e^{-(1/2)|(x,y)|^2}$$

can be easily seen to be rotation invariant through its graph [4].

Definition 18. A wavelet is *directional* if it is not rotation invariant.

More precisely, a two-dimensional wavelet is directional if its covariance matrix has non-zero entries off its diagonal.

Example 8. The two-dimensional Morlet wavelet with parameters $\xi = (\frac{3\pi}{4}, 0)$, $\sigma = 0.85$, and $\beta \approx 0.13, a \approx 0.22$, given by

$$\psi(x,y) = \alpha(e^{iu\cdot\xi} - \beta)e^{-|(x,y)|^2/(2\sigma^2)} \tag{2.46}$$

is directional, where α, β are numerical approximations which normalizes and creates a zero average of its integral [9]. See fig. 2.4.

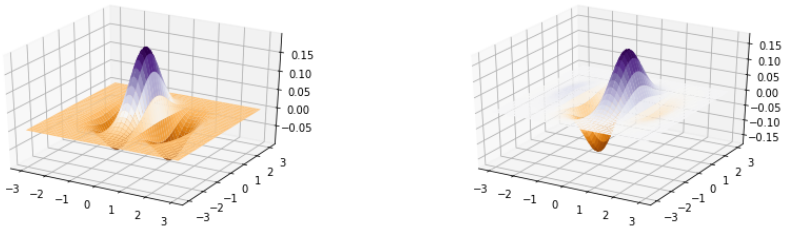

Figure 2.4. Real and imaginary parts of the 2-dimensional Morlet wavelet.

Notation 11. Let R_n denote the set of rotations of $\mathbb{R}^n$. For $r \in R_n$ and $x \in \mathbb{R}^n$, we write $rx = r(x)$. Similarly, we denote r^{-1} as the inverse rotation of r. For R_2, we identify a counter-clockwise rotation of angle θ with r_θ.

A directional wavelet can thus be rotated. A rotation in $\mathbb{R}$ is just a flip of a function f's argument, $f(x) \to f(-x)$. For $\mathbb{R}^2$, a wavelet ψ can be rotated by the argument θ and we can write

$$\psi_\theta(x,y) = \psi(r_\theta(x,y)).$$

The graph of ψ_θ will be the same as for the wavelet ψ but with a clockwise rotation in the plane.

2.11 Scattering wavelets

We will be looking for a specific set of wavelets that can help us achieve translation invariance and linearization of small diffeomorphisms. These properties are very natural for the classification of digits as will be explained in the following sections.

Notation 12. We denote L_c the translation functional in arbitrary dimension.

$$L_c f(x) = f(x-c).$$

Definition 19. An operator $\Phi : L^2(\mathbb{R}^d) \to H$ where H is a Hilbert space is said to be translation invariant if $\Phi(L_c f) = \Phi(f)$ for all $f \in L^2(\mathbb{R}^d)$, $c \in \mathbb{R}^d$. [19]

To see why translation invariance is an important property, consider an arbitrary object in an image. The presence of this object is independent of its location in the image. So for example searching for images which contains a certain object requires some translation invariance.

Another example is that of an image of a hand-written digit. The written digit five is still a five, even if it is written in the center of an image or towards the top left corner of the image.

Definition 20. An operator $\Phi : L^2(\mathbb{R}^d) \to H$ is said to be stable if

$$||\Phi(f) - \Phi(g)||_H \leq ||f - g||.$$

Stability is an important property because it guarantees that the operator Φ does not increase the distance between small deformations of a function f in the Hilbert space H if the differences in $L^2(\mathbb{R}^d)$ is small. This may seem almost trivial, but it is not, as example 9 will show.

Definition 21. A translation-invariant operator $\Phi : L^2(\mathbb{R}^d) \to H$ is said to be Lipschitz-continuous to the action of C^2-diffeomorphisms if for any compact $\Omega \subset \mathbb{R}$ there exists $C \in \mathbb{R}$ such that for all $\tau \subset C^2(\mathbb{R})$,

$$||\Phi(f) - \Phi(L_\tau f)||_H \leq C||f||(\sup_{x\in\mathbb{R}} |\nabla\tau(x)| + \sup_{x\in\mathbb{R}} |H\tau(x)|). \text{ [19]} \qquad (2.47)$$

Where H in $|H\tau(x)|$ denotes the Hessian of τ at x. Thus equation 2.47 depends on the first- and second-order terms of τ.

Lipschitz-continuity to the action of diffeomorphisms gives a notion of similarity between two objects of the same class. For digits that are classified in the obvious manner of 0 through 9, consider two samples $5_1, 5_2$ of a hand-written five. An operator Φ such as above then takes these samples to $\Phi(5_1), \Phi(5_2)$ into another Hilbert space such that in that space, the difference between the samples are small, i.e (2.47) can be written

$$||\Phi(5_1) - \Phi(5_2)|| \leq \min_{i \in \{1,2\}} C||5_i||(\sup_{x \in \mathbb{R}} |\nabla \tau_i(x)| + \sup_{x \in \mathbb{R}} |H\tau_i(x)|). \tag{2.48}$$

This means that for a classification problem, finding a good translation-invariant operator Φ means that almost all objects in a given class are "close together" in the associated Hilbert space H, bounded by the largest possible diffeomorphism between same-class objects such that the objects maintains the same class property. Ideally then, two classes are completely disjoint in H and can be linearly separated there.

Example 9. This is a motivating example for why it is necessary to develop the scattering transform, presented in both [19] and [10].

The modulus of the Fourier transform is translation invariant but unstable to small deformations at high frequencies. While translation invariance is trivial by the definition of the Fourier transform, instability takes a bit more work. A more detailed presentation can be found in [2].

In essence, suppose that $f(x) = ag(x)cos(n\xi x)$. Then $f(t) = \frac{a}{2} g(x)(e^{-i\xi x} + e^{i\xi x})$ with Fourier transform

$$\begin{aligned}
\frac{a}{2} \int g(x)(e^{-i\xi x} + e^{i\xi x}) e^{-i\omega x} dx = & \\
&= \frac{a}{2} \int_{-\infty}^{\infty} g(x) e^{-i\xi x} e^{-i\omega x} dx + \frac{a}{2} \int_{-\infty}^{\infty} g(x) e^{i\xi x} e^{-i\omega x} dx \\
&= \text{ by (2.14)} \\
&= \frac{a}{2} (\hat{g}(\omega + \xi) + \hat{g}(\omega - \xi)).
\end{aligned}$$

Now, a scaling of $f(x)$ by $\tau(x) = sx$, $0 < s < 1$, that is we set $f' = f(x - \tau(x)) = f((1-s)x)$, is not Lipschitz-continuous to the action of small diffeomorphisms. To see why, we calculate $\hat{f}'(\omega)$,

$$\begin{aligned}
\hat{f}'(\omega) &= \int_{-\infty}^{\infty} f((1-s)x)e^{-i\omega x}dx \\
&= \text{ by a change of variables} \\
&= \frac{1}{1-s}\int_{\infty}^{-\infty} f(x)e^{-i\frac{\omega}{1-s}x}dx \\
&= \frac{1}{1-s}\hat{f}(\frac{\omega}{1-s}) \\
&= \frac{a}{2(1-s)}(\hat{g}(\frac{\omega}{1-s}+\xi)+\hat{g}(\frac{\omega}{1-s}-\xi) \\
&= \frac{a}{2(1-s)}(\hat{g}(\frac{\omega+\xi}{1-s}-\frac{s\xi}{1-s})+\hat{g}(\frac{\omega-\xi}{1-s}+\frac{s\xi}{1-s})).
\end{aligned}$$

For high frequencies ξ, the support of $\hat{g}$ will be smaller than $\frac{s\xi}{1-s}$ and thus the difference $|||\hat{f}|-|\hat{f}'|||$ will not be proportional to s. This has been graphically illustrated in [2].

We have managed to identify what the scattering transform sets out to accomplish, what type of problems it can solve and why this could be important. Now we will show the construction of the scattering transform and its properties.

Definition 22. A scattering wavelet is a wavelet that can be written

$$\psi(x) = e^{i\eta\cdot x}\theta(x), \tag{2.49}$$

with $\hat{\theta}(\omega)$ a real-valued function with primary support in a low-frequency ball with radius π, centered around $\omega = 0$, and such that ψ fulfills the admissibility criterion in [19].

The admissibility condition in [19] is an inequality to ensure that the scattering transform using ψ is norm preserving. Reproducing it here would not be constructive.

By (2.14) we see that $\hat{\psi}(x) = \hat{\theta}(\omega-\eta)$ and thus $\hat{\psi}(\omega)$ has the same properties as $\hat{\theta}$ but centered around $\omega = \eta$. Now we define the time-frequency localization of a scaled and rotated wavelet, following the notation in [9].

Definition 23. For a scattering wavelet, a rotation $r \in R_n$ and a scaling of 2^j in the time-domain, we define the scaled and rotated wavelet as

$$\begin{aligned}
\psi_{2^j r}(x) &= 2^{-nj}\psi(2^{-j}r^{-1}x), \text{ by which it follows that} \\
\hat{\psi}_{2^j r}(\omega) &= 2^{-nj}\int \psi(2^{-j}r^{-1}x)e^{-i\omega\cdot x}dx \\
&= \text{ change of variables } x = 2^{-j}x \\
&= \int \psi(r^{-1}x)e^{-i2^j\omega\cdot x}dx \\
&= \text{ change of variables } x = r^{-1}x \\
&= \hat{\psi}(2^j r\omega).
\end{aligned} \tag{2.50}$$

When $j > 0$ then $\psi_{2^j r}$ has a lower frequency localization than ψ. Expanding the Fourier transform of $\psi_{2^j r}$ in terms of θ,

$$\hat{\psi}(2^j r\omega) = \hat{\theta}(2^j r\omega - \eta). \tag{2.51}$$

Since θ is localized around 0, we see that $\psi_{2^j r}$ is localized around $2^{-j}\eta$. Similarly, if $\hat{\theta}(\omega)$ covers frequencies for $0 \leq |\omega| < c$, the scaled and translated scattering wavelet $\psi_{2^j r}$ covers the frequency band $2^{-j}\eta \leq |\omega| < 2^{-j}\eta + 2^{-j}c$.

Finally we need a lowpass filter θ associated to the scattering wavelet ψ, which is scaled analogously to definition 23,

$$\theta_{2^j}(x) = 2^{-j}\theta(2^{-j}x).$$

In the case of using a Morlet wavelet as a scattering wavelet, its Gaussian scaling function works well.

Observation 4. Contrary to the scaling and rotation of an ordinary wavelet which is normalized in L^2, the scaling and rotation of a scattering wavelet preserves its L^1 norm but is not necessarily normalized in L^1.

2.11.1 The scattering transform

Notation 13. Let $2^{\mathbb{Z}} \times G = \{(2^j, r), j \in \mathbb{Z}, G \text{ a finite subset of } R_n\} = \Lambda_\infty$, and denote $\lambda = (s,r) \in 2^{\mathbb{Z}} \times G$. We write $\psi_\lambda = \psi_{sr}$ as the wavelet ψ scaled by s and rotated by r.

Notation 14. Let $\Lambda_j = \{(s,r) = \lambda \in \Lambda_\infty, s \leq 2^j\}$.

Definition 24. A path p is a finite sequence $p = (\lambda_1, ..., \lambda_i) \in \Lambda_x^i$ where x may equal ∞. The empty path is denoted $p = \emptyset$. Moreover we write $p + \lambda = (\lambda_1, ..., \lambda_n, \lambda) \in \Lambda_x^i$ for $\lambda \in \Lambda_x$.

For a scattering wavelet ψ, a function $f \in L^2(\mathbb{R}^n)$, and $\lambda \in \Lambda_\infty$, let $U[\lambda]f = |\psi_\lambda \star f|$, with $U[\emptyset]f = f$. That is, U is an operator that calculates the complex wavelet transform of a signal f and ψ_λ, which captures the signals phase and subsequently eliminates it. So $U[\lambda]f$ is a lower frequency, real-valued non-negative function which is more regular than f.

Definition 25. For a path $p = (\lambda_1, ..., \lambda_i) \in \Lambda_x^\infty = \{\lambda \in \Lambda_x^n, n \in \mathbb{N}\}$, we let the scattering propagator $U[p]$ be defined by

$$U[p] = U[\lambda_i]U[\lambda_{i-1}]...U[\lambda_1]$$

for U being the functional above.

Observation 5. Because of (2.16) and that we are taking the norm of the wavelet transforms when applying U to a function f, we only need to compute $U[p]f$ for paths with positive rotations in the frequency plane. For the one-dimensional case, we do not have to calculate any rotations at all.

Definition 26. For $f \in L^2(\mathbb{R}^n)$, and path $p \in \Lambda_j^\infty$, we define the windowed scattering transform of f as

$$S_j[p]f(u) = U[p] \star \phi_{2^j}(u) = \int U[p]f(v)\phi_{2^j}(u-v)dv. \tag{2.52}$$

So the windowed scattering transform performs a signal averaging of a scattering path p. In particular, note that $S_j[\emptyset]f = f \star \phi_{2^j}$, $S_j[\lambda_1]f = U[\lambda_1]f \star \phi_{2^j}$, and $S_j[\lambda_1, \lambda_2]f = U[\lambda_1, \lambda_2]f \star \phi_{2^j}$.

Definition 27. For $\Omega \subseteq \Lambda_j^\infty$, and $f \in L^2(\mathbb{R}^n)$, let

$$S_j[\Omega]f = \{S_j[\omega]f, \omega \in \Omega\},$$

with associated norm

$$||S_j[\Omega]f|| = \sum_{\omega \in \Omega} ||S_j[\omega]f||.$$

Notation 15. We call the operator $S_j[\Omega]$ defined as above the scattering transform at scale j.

2.11.2 Properties of the scattering transform

For the following propositions, proofs can be found in [19].

Proposition 9. For $f, g \in L^2(\mathbb{R}^d)$, the scattering transform is non-expansive,

$$||S_j[\Lambda_j^\infty]f - S_j[\Lambda_j^\infty]g|| \leq ||f - g||.$$

Proposition 10. For $f, g \in L^2(\mathbb{R}^d)$, the scattering distance is nonincreasing,

$$||S_j[\Lambda_j^{k+1}]f - S_j[\Lambda_j^{k+1}]g|| \leq ||S_j[\Lambda_j^k]f - S_j[\Lambda_j^k]g||.$$

Proposition 11. For $f \in L^2(\mathbb{R}^d)$ and L_c a translation operator,

$$\lim_{j \to \infty} ||S_j[\Lambda_j^\infty]f - S_j[\Lambda_j^\infty]L_c f|| = 0.$$

Proposition 12. There exists C such that for all $f \in L^2(\mathbb{R}^d)$ with $||U[\Lambda_j^\infty f|| < \infty$ and all $C^2(\mathbb{R}^d)$ diffeomorphisms τ such that $||\nabla \tau|| \leq 1/2$ satisfy

$$||S_j[\Lambda_j^\infty]f - S_j[\Lambda_j^\infty]L_\tau f|| \leq C||U[\Lambda_j^\infty]f||K(\tau), \text{ where}$$

$$K(\tau) = 2^{-j}||\tau||_\infty + ||\nabla \tau||_\infty \left(\min\{\log \frac{||\tau||_\infty}{||\nabla \tau||_\infty}, 1\} \right) + ||H\tau||_\infty,$$

and for all $m \geq 0$,

$$||S_j[\Lambda_j^m]f - S_j[\Lambda_j^m]L_\tau f|| \leq Cm||f||K(\tau). \tag{2.53}$$

Note that (2.53) of proposition 12 gives the Lipschitz continuity for finite paths of length m, which is important for applications since in practice we will only calculate paths with bounded length.

3. Convolutional Networks

Convolutional neural networks has a strong similarity to the structure of the scattering transform on finite paths which are frequency decreasing, to the extend where scattering transforms are used to develop CNN-like architectures which perform very well on image classification tasks [10, 22].

Notation 16. Convolutional neural networks, sometimes called deep convolutional networks or similarly, will be denoted CNNs for brevity.

Here we hope to give a brief but concise introduction to CNNs, mainly about how the composite nodes are individually constructed and how the network is then pieced together, but also how it is trained to learn its feature maps. We will also talk about some of the mathematical difficulties that presents themselves in understanding CNNs. This motivates the construction of networks such as the scattering convolution network which uses some of the mathematical properties of the scattering transform and manages to perform well on various classification tasks while answering some questions concerning the mathematics of CNNs.

3.1 Success of CNNs

In 2010, LeCun *et al.* published a paper which showed state of the art performance on a handwritten digit classification dataset called *MNIST* using a CNN [16]. Since then, classification tasks using deep learning, the use of networks such as CNNs which employs hidden layers, layers whose action is not governed by the class of the input, has flourished.

In 2017, Bronstein *et al.* explains that CNNs are "among the most successful deep learning architectures" and notes that breakthroughs in several fields by deep learning multilayer hierarchies has been made, which can be partly attributed to growing computational power and availability of large training data sets [8].

This serves as motivation for why we should concern ourselves with CNN-like architectures and why it is relevant today.

3.2 Classical convolutional neural networks

Influenced from the interconnections of neurons in the brain and the individual neuron's ability to "respond", of "fire", based on the response it in turn receives from its inputs, a CNN shares some of this neural structure through a tree of layered nodes which individually calculates a response function, which is a non-linearity composed with a convolution of its inputs with some kernel.

The non-linearity present in each node is crucial to the networks construction and is what turns a CNN into a universal approximator, a result established in 1989 by Hornik *et. al* [15]. Loosely speaking this means that the structure of a CNN provides the capability to represent "any" function, and leaves the open question of how to get it to actually do so.

The final layer of the created network must have a response function which corresponds to the type of classification that is performed [13], and may be the same used throughout the network.

We will describe the classical CNN from the perspective of image classification, in which case the input is an $n \times n$-dimensional matrix A who's value $A_{i,j}$ represents the color of the pixels at column i, row j. For simplicity, we have chosen to consider grayscale images which has only one color channel. The output of the CNN is then the class to which the image belongs.

Furthermore, we only consider neural networks for which each node at a layer k only receives input from layer $k-1$.

3.2.1 Structure of a neural node

For a node n^i in layer k of a neural network, its input is given as an $m_i \times n_i$ matrix A^i from outputs of nodes n^l for n^l in layer $k-1$. A new $m_i \times n_i$ matrix $A^{i'}$ is calculated using a discrete convolution

$$(A^{i'})_{i,j} = A^i \star K^i(i,j) = \sum_{n,m} A^i(n,m) K^i(i-n, j-m),$$

where $A^i(n,m) = (A^i)_{n,m}$ if n,m are within the bounds of the matrix A^i, and K^i is a convolutional kernel for layer i. Often $K^i(x,y) \neq 0$ only for x,y close to origo. Details about how to define $A^i(n,m)$ for m,n outside of the bounds of the matrix A^i can be found in [13].

After the computation of $A^{i'}$, one applies a non-linear function h^i on the matrix, usually pointwise such that

$$(h^i(A^{i'}))_{i,j} = h^i((A^{i'})_{i,j}).$$

Finally a pooling operation may be used to reduce the size of $A^{i'}$, e.g. by choosing the maximum value of each 2×2 square covering the matrix. Another way to reduce the size of $A^{i'}$ is to calculate only a subset of convolutions, known as using stride, such that

$$(A^{i'})_{i,j} = A^i \star K^i(ui+v, uj+v), \text{ where } v < u.$$

There are several details that needs to be worked out to give a complete formal description of how a node n^i works, e.g. how to deal with the sum of over the edges of an image and how to define the kernels K^i. These details has been treated in [13].

3.2.2 The interconnection of nodes

To create a CNN, then, is to combine several layers of nodes n^i, with the first layer being the input image A, and the final layer being a response of which class the image belongs to. This structure is a deep learning technique, which is really any type of machine learning method which employs hidden layers. In turn, a layer is hidden if its response is not governed directly by the desired response of the input image.

E.g. to classify an image as an image of a monkey, a hidden layer may search for the presence of a banana somewhere. While an image of a monkey in no way necessitates the presence of a banana, the network might have discovered that a characteristic of images of monkeys are that they like to surround themselves with the yellow fruit.

3.2.3 Training of CNNs

A CNN has to be trained. It has to learn what the convolutions employed at each node should be, which is done by classical cost functions widely employed in statistics and an algorithm known as back propagation. It is outside the scope of this document to explain how and why these methods work, so we will only cover what they need and what they do and refer to [13] for details.

First, to train a CNN we need labeled data. This corresponds to a set of images A_i, and corresponding classes c_i such that we preferably have at our disposal a large set $\{(A_i, c_i), 1 \leq i \leq M \text{ large}\}$ of labeled images.

Secondly, we need to run the back propagation algorithm over our labeled data set until we find some local minima of the cost function, which may be computationally costly. So we need either a lot of time or a lot of computational power.

Finally we hope that our network has learned "the correct" features of our data set, such that it will perform well on images that it has not seen before. It is curious to see that the features that are learned really do seem to correspond well to what we might say are features of the objects. See e.g. [26], where it is visualized how a layer has learned that features of a human is the presence of a head and a shoulder.

3.2.4 Mathematical challenges

There is little insight into "the internal operation and behavior of these complex models, or how they achieve such good performance" [26]. Among the insights that we do have, is that CNNs are "computing progressively more powerful invariants as depth increases" [20]. An invariant being for example the orbit O_I of an image I, i.e the class $\{gI, g \text{ a group acting on images } I\}$ [3].

Several questions are posed in [10], such as "Why use multiple layer?" and "How many layers?". These questions seem to remain unanswered for CNNs, but can be answered by scattering convolution networks presented next.

3.3 Scattering Convolution Networks

A subset P of Λ_J^∞ such that $(\lambda_1, ..., \lambda_{i-1}, \lambda_i) \in P \Rightarrow (\lambda_1, ..., \lambda_{i-1}) \in P$ has the structure of a convolutional neural network. Any such set P can be built up starting from $P_0 = \emptyset$ and defining each collection of paths of length no longer than $n+1$ by choosing $P_{n+1} \subseteq P_n \cup \{p + \lambda, p \in P_n, \lambda \in \Lambda_J\}$.

Definition 28. We call a set P constructed as above an inductive set of paths.

A scattering convolution network is a classification method applied to the scattering transforms along an inductive set of paths. This makes use of properties of the scattering transform, linearization of small deformations and translation invariance, and provides state of the art results for "hand written digit recognition and for texture discrimination" [10].

Definition 29. A path $p = (\lambda_1, ..., \lambda_i)$ is said to be frequency decreasing if $\lambda_k < \lambda_{k+1}$ for $1 \leq k \leq i-1$.

Depending on the scale j of the scattering transform S_j, numerical calculations performed on a specific image classification data set show that a large part ($\geq 99\%$) of the input signals energy will be contained in frequency decreasing paths on layers $l \leq 4$ when $j \leq 6$ [9]. This can be seen as giving a rather satisfactory answer to how deep a network should be, and why we use multiple layers (to capture most of a signals energy), which are the questions highlighted in section 3.2.4.

3.3.1 Hybrid networks

In [22], a hybrid network architecture is constructed where the first layers consists of a scattering network and later layers a CNN, which aims to capture more complex variabilities in the classification datasets. This hybrid network does very well on supervised image representation tasks and presents another way to introduce well-known mathematical properties into a CNN.

A. Multi-resolution analysis of a signal

Using the one-dimensional Morlet wavelet with associated scaling function (shown in fig. A.1) to create a multi-resolution analysis of an input signal (fig. A.2).

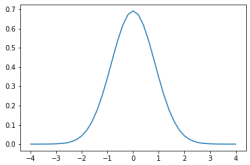

Figure A.1. Scaling function

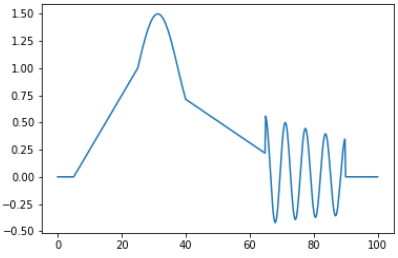

Figure A.2. Input signal

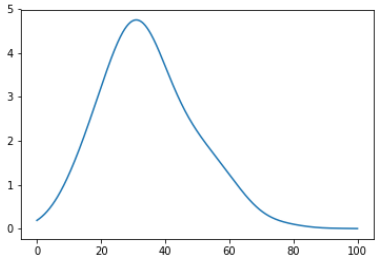

Figure A.3. Lowpass filtering at scale 2^3

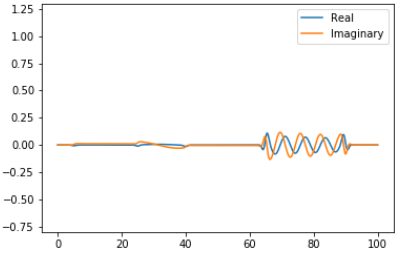

Figure A.4. Scale 1

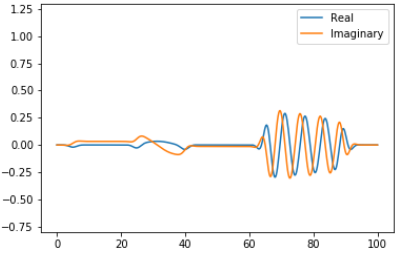

Figure A.5. Scale 2

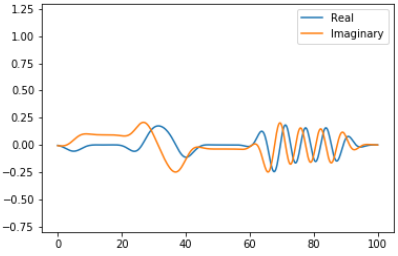

Figure A.6. Scale 2^2

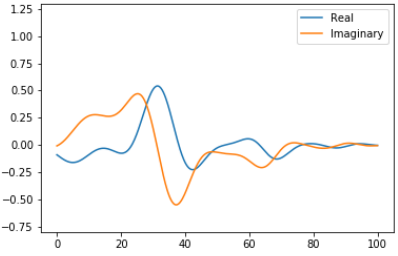

Figure A.7. Scale 2^3

B. Scattering transforms of signals

In this section, we consider scattering transforms using first a regular Morlet ψ and associated scaling function θ, and finally one example using a Morlet wavelet with associated scaling function both scaled by 2^{-4}, i.e., scaled as wavelets,

$$\psi_{2^{-4}}(x) = 2^2\psi(2^4x), \theta_{2^{-4}}(x) = 2^2\theta(2^4x).$$

B.1 Frequency decreasing paths over S_3, using ψ, θ

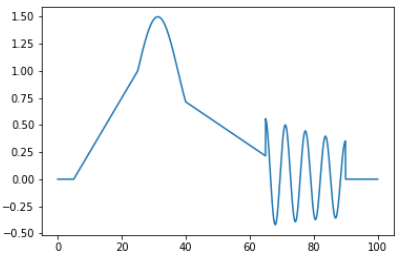

Figure B.1. Input signal f, or $U[\emptyset]$

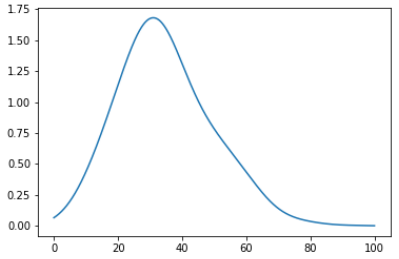

Figure B.2. $S_3[\emptyset]f$

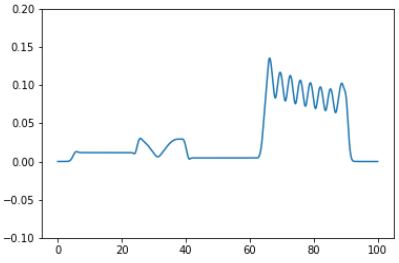

Figure B.3. $U[1]f = |\psi \star f|$

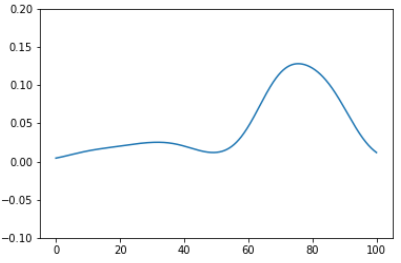

Figure B.4. $S_3[1]f$

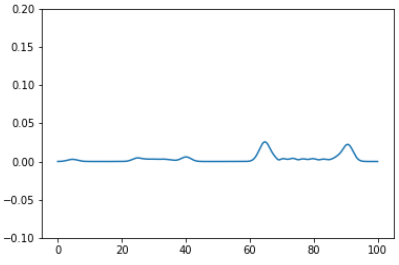

Figure B.5. $U[1,2]f = ||\psi \star f| \star \psi_2|$

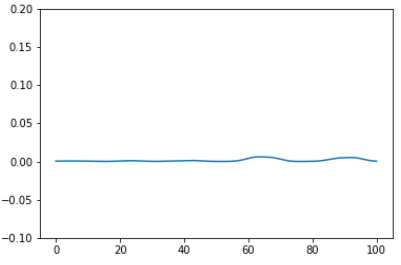

Figure B.6. $S_3[1,2,4]f$

B.2 Frequency decreasing paths over S_2, using ϕ, θ

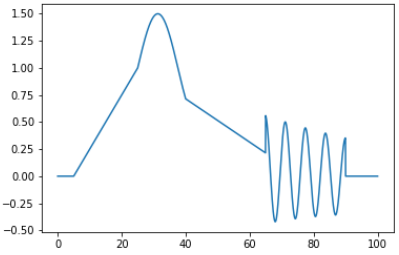

Figure B.7. Input signal f, or $U[\emptyset]$

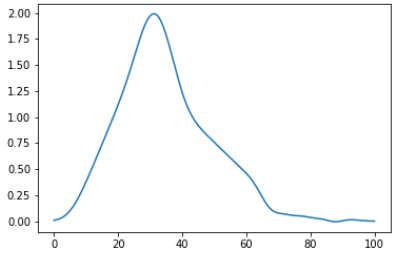

Figure B.8. $S_2[\emptyset]f$

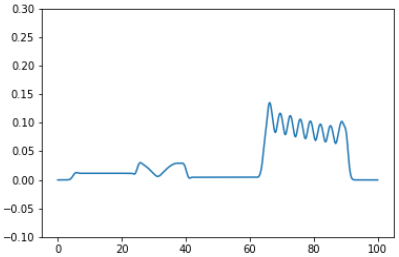

Figure B.9. $U[1]f = |\psi \star f|$

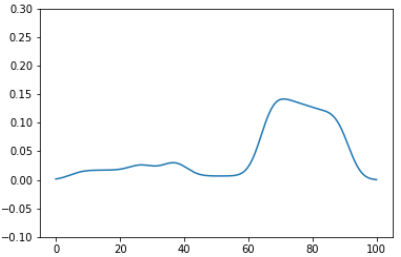

Figure B.10. $S_2[1]f$

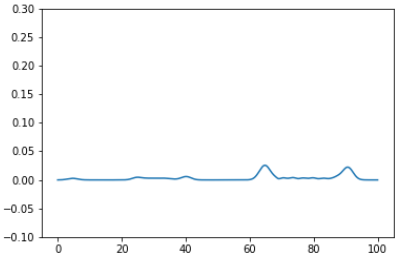

Figure B.11. $U[1,2]f = ||\psi \star f| \star \psi_2|$

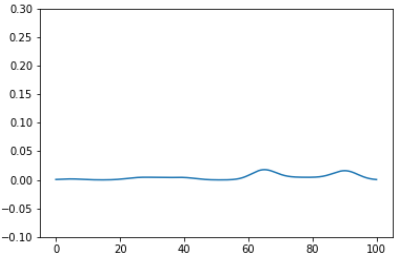

Figure B.12. $S_2[1,2]f$

B.3 Constant paths over S_3, using ψ, θ

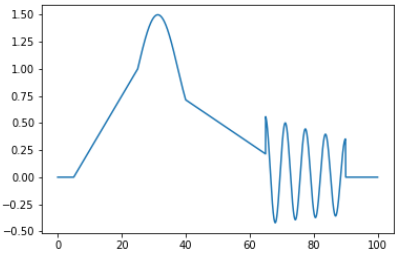

Figure B.13. Input signal f, or $U[\emptyset]$

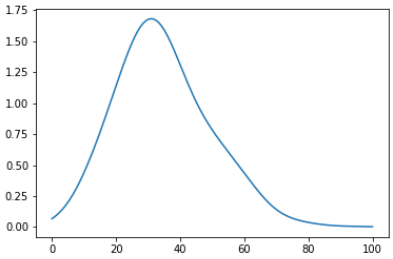

Figure B.14. $S_3[\emptyset]f$

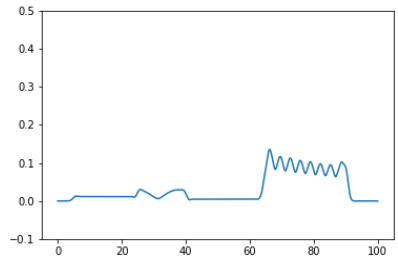

Figure B.15. $U[1]f = |\psi \star f|$

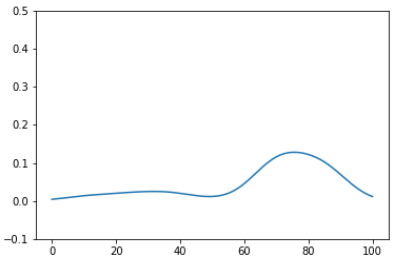

Figure B.16. $S_3[1]f$

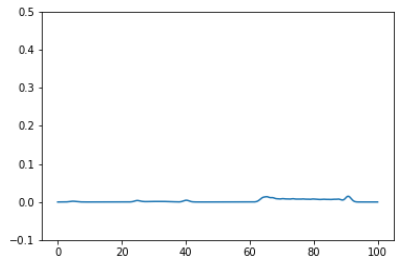

Figure B.17. $U[1,1]f = ||\psi \star f| \star \psi|$

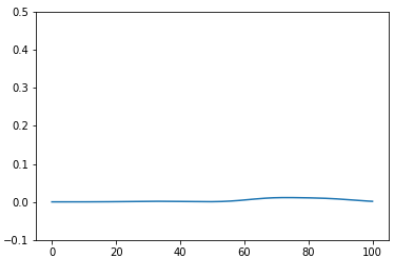

Figure B.18. $S_3[1,1]f$

B.4 Constant paths over S_3, using ψ, θ

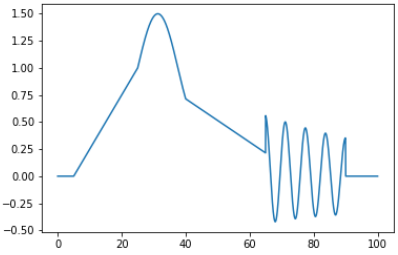

Figure B.19. Input signal f, or $U[\emptyset]$

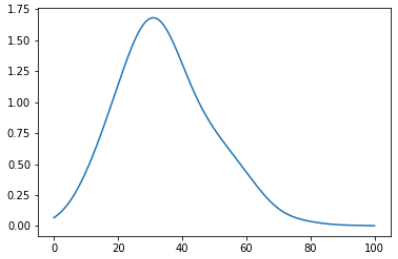

Figure B.20. $S_3[\emptyset]f$

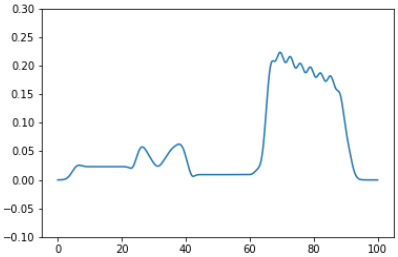

Figure B.21. $U[2]f = |\psi_2 \star f|$

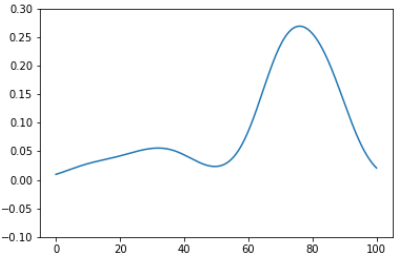

Figure B.22. $S_3[2]f$

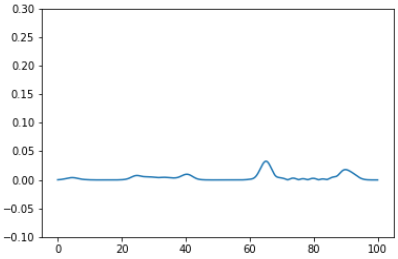

Figure B.23. $U[2,2]f = ||\psi_2 \star f| \star \psi_2|$

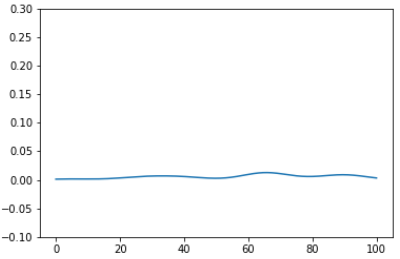

Figure B.24. $S_3[2,2]f$

B.5 Frequency decreasing path over S_5, using ψ, θ

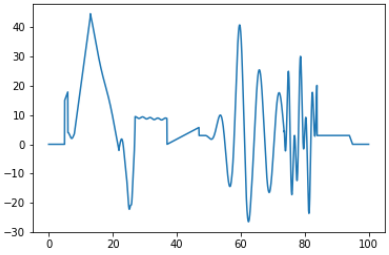

Figure B.25. Input signal f, or $U[\emptyset]$

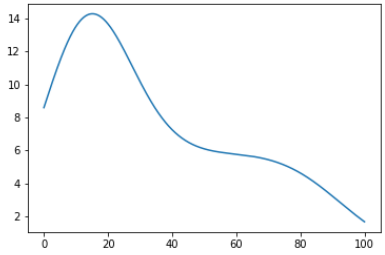

Figure B.26. $S_5[\emptyset]f = f \star \theta_{2^5}$

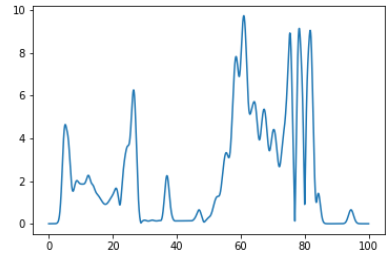

Figure B.27. $U[1]f = |\psi \star f|$

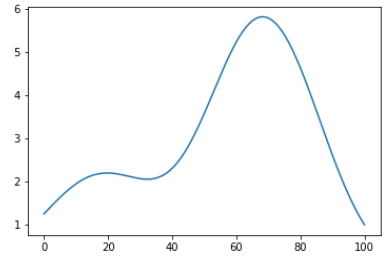

Figure B.28. $S_5[1]f$

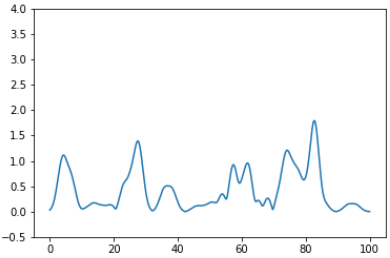

Figure B.29. $U[1,2]f$

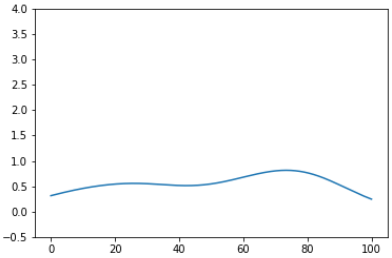

Figure B.30. $S_5[1,2]f$

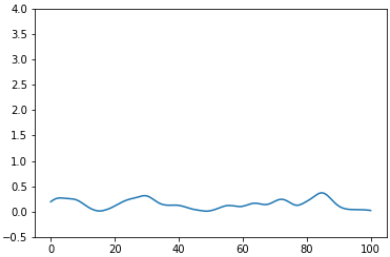

Figure B.31. $U[1,2,4]f$

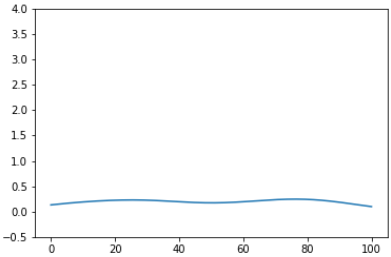

Figure B.32. $S_5[1,2,4]f$

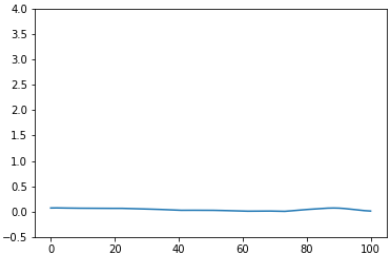

Figure B.33. $U[1,2,4,8]f$

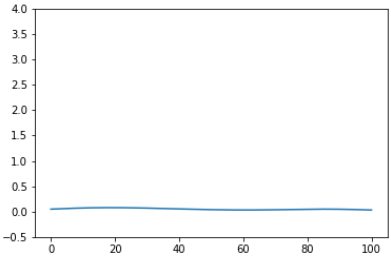

Figure B.34. $S_5[1,2,4,8]f$

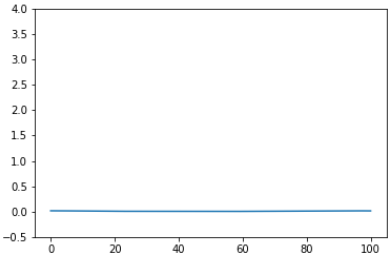

Figure B.35. $U[1,2,4,8,16]f$

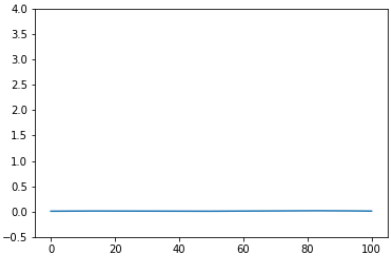

Figure B.36. $S_5[1,2,4,8,16]f$

B.6 Frequency decreasing path over S_5, using $\psi_{2^{-4}}, \theta_{2^{-4}}$

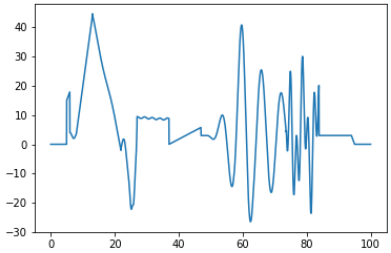

Figure B.37. Input signal f, or $U[\emptyset]$

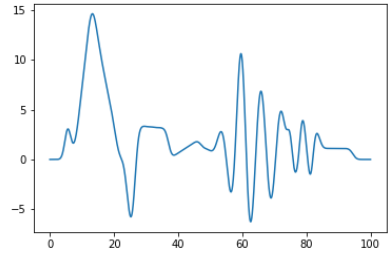

Figure B.38. $S_5[\emptyset]f$

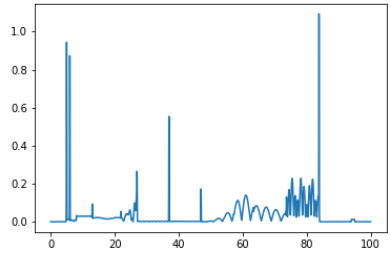

Figure B.39. $U[1]f$

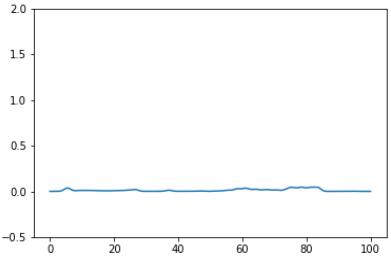

Figure B.40. $S_5[1]f$

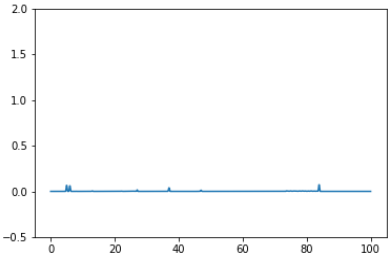

Figure B.41. $U[1,2]f$

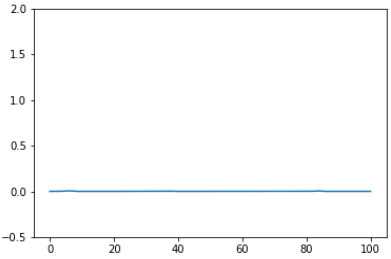

Figure B.42. $S_5[1,2]f$

4. References

[1] Malcolm Ritchie Adams and Victor Guillemin. *Measure theory and probability*. Springer, 1996.

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

[3] Fabio Anselmi, Joel Z Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso Poggio. Unsupervised learning of invariant representations in hierarchical architectures. *arXiv preprint arXiv:1311.4158*, 2013.

[4] Jean-Pierre Antoine, Romain Murenzi, Pierre Vandergheynst, and Syed Twareque Ali. *Two-dimensional wavelets and their relatives*. Cambridge University Press, 2008.

[5] Ashish Bansal and Ajay Kumar. Generalized analogs of the heisenberg uncertainty inequality. *Journal of Inequalities and Applications*, 2015(1):168, 2015.

[6] Swanhild Bernstein, Jean-Luc Bouchot, Martin Reinhardt, and Bettina Heise. Generalized analytic signals in image processing: comparison, theory and applications. In *Quaternion and Clifford Fourier Transforms and Wavelets*, pages 221–246. Springer, 2013.

[7] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.

[8] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

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

[10] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

[11] Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[12] Jonas Gomes and Luiz Velho. *From fourier analysis to wavelets*. Springer, 2015.

[13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[14] Geoffrey Grimmett and David Stirzaker. *Probability and random processes*. Oxford university press, 2001.

[15] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[16] Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. Convolutional networks and applications in vision. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pages 253–256. IEEE, 2010.

[17] Elliott H Lieb and Michael Loss. Analysis, volume 14 of graduate studies in mathematics. *American Mathematical Society, Providence, RI,*, 4, 2001.
[18] Stephane Mallat. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.
[19] Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.
[20] Stéphane Mallat. Understanding deep convolutional networks. *Phil. Trans. R. Soc. A*, 374(2065):20150203, 2016.
[21] Yves Meyer. *Wavelets and operators*, volume 1. Cambridge university press, 1995.
[22] Edouard Oyallon, Eugene Belilovsky, and Sergey Zagoruyko. Scaling the scattering transform: Deep hybrid networks. *CoRR*, abs/1703.08961, 2017.
[23] Ram Shankar Pathak. *The wavelet transform*, volume 4. Springer Science & Business Media, 2009.
[24] Paolo Prandoni and Martin Vetterli. *Signal processing for communications*. Collection le savoir suisse, 2008.
[25] Bruno Torrésani. *Analyse continue par ondelettes*. EDP Sciences, 1995.
[26] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.