Mini-Course on Applied Harmonic Analysis

Philipp Christian Petersen*

PDE-CDT Summer School 2018
3-6 September 2018
at Ripon College

Abstract: This mini-course provides an introduction to applied harmonic analysis and its applications in the numerical analysis of partial differential equations. The standard Fourier transform exhibits various weaknesses, such as a lack of time-frequency localisation, which is quantified, for example, by the famous Heisenberg-Pauli-Weyl principle, and suboptimal approximation properties, leading to undesirable effects like the Gibbs phenomenon. For these reasons, a host of alternative transforms have been developed. Among these, the wavelet transform and the short-time Fourier transform are probably the most popular. We review properties of these transforms such as their time-frequency localisation and demonstrate how they lead to discrete representation systems. Then, we analyse the approximation properties of these systems and discuss their viability as tools in image processing and in the numerical analysis of partial differential equations. Finally, we will present further developments such as constructions and approximation properties of anisotropic directional systems including curvelets and shearlets.

*Email: Philipp.Petersen@maths.ox.ac.uk

Chapter 1

What is applied harmonic analysis?

Applied harmonic analysis a research area studying the efficient decomposition or representation, storage, and analysis of signals. In this framework, a signal will always be an element of a separable Hilbert space, and most of the time this will be $L^2(\mathbb{R}^d)$, $d \in \mathbb{N}$, and sometimes it will be ℓ^2.

We will start by presenting a prototype question in applied harmonic analysis. This part of the motivation is by no means a precise definition of what applied harmonic analysis is, but should rather give the participant of this workshop a feeling of the type of questions that an applied harmonic analyst is interested in. As a result, the presented question will be much less general than what we will study in the sequel and vague in parts where we lack the right terminology. We will sharpen and refine this setup more and more, once we have the right tools to do so.

Let $\mathcal{H}$ be a Hilbert space, $(\varphi_n)_{n\in\mathbb{N}} \subset \mathcal{H}$. Further assume, that for every $f \in \mathcal{H}$, there exist a unique sequence $(c_n(f))_{n\in\mathbb{N}} \subset \ell^2$ such that

$$f = \sum_{n\in\mathbb{N}} c_n(f)\varphi_n. \tag{1.0.1}$$

This is certainly possible if $(\varphi_n)_{n\in\mathbb{N}} \subset \mathcal{H}$ is an orthogonal basis, but also for much more general systems $(\varphi_n)_{n\in\mathbb{N}} \subset \mathcal{H}$. One of the main problems in applied harmonic analysis is to design the right bases $(\varphi_n)_{n\in\mathbb{N}}$ (or generalisations thereof) such that the representation of (1.0.1) satisfies the following criteria:

1. *Efficiency/Sparsity:* The representation (1.0.1) should be efficient in the sense that if one computes only partial sums over a carefully chosen subset $\Lambda \subset \mathbb{N}$ where $|\Lambda|$ is fairly small, then

   $$f \approx \sum_{n\in\Lambda} c_n(f)\varphi_n.$$

   Such an efficient representation is unlikely to exist for all $f \in \mathcal{H}$. Instead, we want such an efficient representation only for f in a subset of $\mathcal{H}$ containing functions of interest.

2. *Interpretability/Manipulation:* The representations should also serve as an analysis tool in the sense that $(c_n(f))_{n\in\mathbb{N}}$ unveils certain properties of f that were not directly accessible beforehand. Another desirable property would be that by manipulating $(c_n(f))_{n\in\mathbb{N}}$ one can change certain aspects of f while leaving other characteristics of f unchanged.

3. *Computationally fast:* Since applied harmonic analysis is an applied field we are of course also interested in representations that are connected to a computationally fast transform. We will treat this aspect somewhat negligently in this workshop.

It is not quite clear from the discussion on representation systems above, why the name of this research area contains the phrase *"harmonic analysis"*. The reason for this is, that virtually every argument evolves around the Fourier transform or Fourier series. Hence, our first example of a representation system has to be the system of complex exponentials: $(x \mapsto e^{-2\pi inx})_{n\in\mathbb{Z}}$ for the Hilbert space $L^2([0,1])$. Indeed, we have that

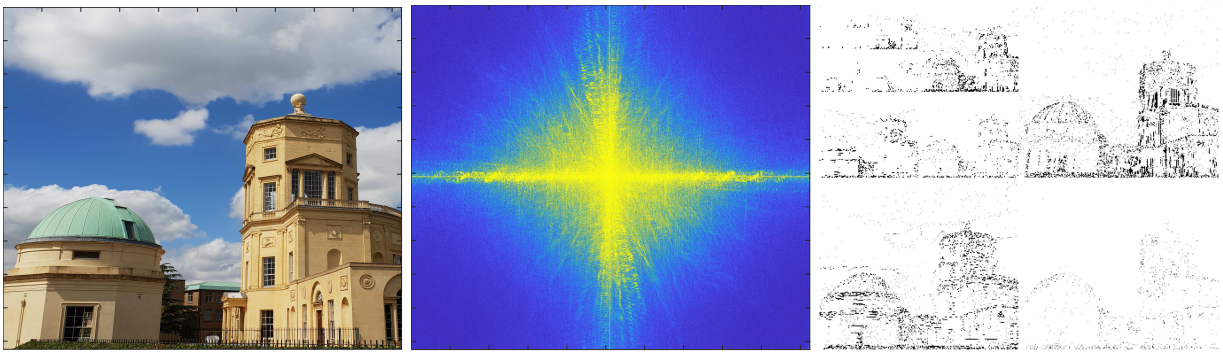

Figure 1.1: **Left:** Pixel representation of a picture of the Radcliffe Observatory; **Center:** Coefficients of the representation with respect to complex exponentials; **Right:** Coefficients corresponding to a wavelet representation. The pixel- and wavelet representations are very interpretable, while the Fourier representation is hard to decipher. The wavelet representation is the only one that can be considered efficient.

1. If $\|f\|_{H^s} \leq 1$ then
$$\left\| f - \sum_{n=-N}^{N} c_n(f) e^{-2\pi i n \cdot} \right\|_{L^2} \lesssim N^{-s+\frac{1}{2}},$$
i.e., if a function is sufficiently smooth, then the representation by the Fourier series is very efficient. However, if f is, for example, discontinuous, then we get very slow approximations. The very slow convergence of the Fourier series for discontinuous functions is called *Gibbs phenomenon*;

2. The sequence $(c_n(f))_{n\in\mathbb{Z}}$ shows which frequencies are present in the signal. On the other hand, the $(c_n(f))_{n\in\mathbb{Z}}$ conceals all local structures of f, see also Figure 1.1. Moreover, the decay of the coefficients contains information about the smoothness of the function. Finally, differentiation of a function is possible by reweighting the Fourier coefficients;

3. In practice the Fourier transform is computed using the fast Fourier transform. This is a well-known highly computationally efficient algorithm.

We will recall some basics of Fourier analysis and then introduce multiple representation systems, including Gabor frames, wavelet systems, and curvelet and shearlet systems. Each of these systems have an application area, where the others fall short.

Although we have acquired a rough understanding of the typical questions of applied harmonic analysis, it might not be clear, why this field of research should be interesting for a researcher mainly interested in partial differential equations. Apart from curiosity there are a couple of arguments to look into this subject, that can hopefully convince a scholar of differential equations.

- **Numerical analysis:** For every efficient way to discretise functions, there is a numerical analyst that uses this method to solve PDEs. I cannot back this statement up, due to the fact that is impossible to check, but it reflects my experience. Let $f \in L^2(\Omega)$, $\Omega \subset \mathbb{R}^2$, $L : H^1(\Omega) \to L^2(\Omega)$ we aim to find $u \in H_0^1(\Omega)$ such that
$$Lu = f. \tag{1.0.2}$$
Assume, that we have a weak formulation
$$a(u, v) := \langle Lu, v \rangle, \text{ for all } v \in H_0^1(\Omega).$$

Let $(\varphi_n)_{n\in\mathbb{N}}$ be any basis for $H_0^1(\Omega)$, then we define $U_N := V_N := \operatorname{span}\{\varphi_n : n \le N\}$. If a is sufficiently nice, say elliptic, then by Cea's Lemma, the solutions $u_N \in U_N$ of

$$a(u_N, v) = \langle f, v\rangle, \text{ for all } v \in V_N,$$

satisfy $\|u_N - u\|_{H^1} \lesssim \inf_{v\in V_N} \|u - v\|_{H^1}$. Therefore, if the basis allows very efficient approximations of the solutions of (1.0.2), then u_N will yield a very good approximation of the solution u already for very small N. If it is associated to a fast transform, than this allows us to solve the discrete problems fast.

Another idea is to choose the spaces V_N, U_N adaptive. We will see later that for wavelet bases, this leads to adaptive PDE solvers, that run in optimal complexity.

- **Helpful toolkit:** The tools that will be presented might also be helpful in theoretical analysis. For example, in the construction of Hairer's *regularity structures* [23], wavelets are used to define a so called reconstruction operator. The short-time Fourier transform has been found to be a valuable tool to understand the mapping properties of *pseudo-differential operators* [18, Chapter 14]. Similarly, wavelets turn out to be an essential tool to establish mapping properties of *Calderon-Zygmund operators*, [32].

- **Real-world applications:** Finally, there are of course plenty of very famous real-world applications of the techniques described in the sequel. The main codecs for *audio, image, and video compression* (mp3, jpeg, H.264/MPEG) are all based on the discrete cosine transform. JPEG2000 is based on the wavelet transform. One of the most important recent results in physics is the detection of a *gravitational wave* generated by the coalescence of two black holes. This was achieved in the Laser Interferometer Gravitational-Wave Observatory in 2015. This achievement is based on a very thorough post-processing of the raw measured data. This requires many tools from signal-processing and in particular, a time-frequency transform using the so-called Wilson bases. In another application in physics, the wavelet transform was used to denoise and deblur the images from the original non-refurbished *Hubble telescope*, which led to its repair in 1993, [25].

We shall start by recalling some essentials on the Fourier transform. We will see that this transform is fundamentally limited by a number of uncertainty principles. Overcoming these limitations has lead to many generalised transforms. Among those are the short-time Fourier transform and Gabor systems, the wavelet transform and the associated systems, and finally directional transforms leading to curvelet and shearlet systems. Most of the remainder is based on the books: [7, 12, 18, 27, 31].

Disclaimer: As applied harmonic analysis is a large and active field and this workshop is limited to a few lectures, the following exposition needs to be reasonably brief. As a result, many theorems in this manuscript will remain unproved, some will be proved only for special cases. Moreover, many results are stated in a simplified version, to make their proofs simple enough for a short lecture. I encourage every participant of this course to find the omitted proofs in one of the references mentioned above, or better yet, prove the results themselves. Moreover, if a result appears to have overly restrictive assumptions, then it might be a good exercise to try to generalise the result as far as possible.

Chapter 2

The basics of Fourier analysis

We start by introducing the essentials of Fourier analysis that are most fundamental for the upcoming discussion. We will omit some proofs and refer to [18] instead.

2.1 The Fourier transform

For $d \in \mathbb{N}$, $f \in L^1(\mathbb{R}^d)$ we define the *Fourier transform* of f by

$$\hat{f}(\xi) := \int_{\mathbb{R}^d} f(x) e^{-2\pi i \langle x, \xi \rangle} dx.$$

If the expressions get to long to put a $\hat{}$ over them, we also write $\mathcal{F}(f)$ for the Fourier transform of a function f. Since

$$|\int_{\mathbb{R}^d} f(x) e^{-2\pi i \langle x, \xi \rangle} dx| \leq \int_{\mathbb{R}^d} |f(x)||e^{-2\pi i \langle x, \xi \rangle}| dx \leq \|f\|_{L^1},$$

we get that $\hat{f} \in L^\infty$ and since $e^{-2\pi i \langle x, \xi_1 \rangle} - e^{-2\pi i \langle x, \xi_2 \rangle} \to 0$ if $\xi_1 - \xi_2 \to 0$ it is not hard to see that $\hat{f}$ is continuous. In other words, the Fourier transform maps from $L^1(\mathbb{R}^d)$ to $L^\infty(\mathbb{R}^d) \cap C(\mathbb{R}^d)$.

2.2 Basic operations

For $t \in \mathbb{R}^d$, we define the *translation operator* $T_t : L^1(\mathbb{R}^d) \to L^1(\mathbb{R}^d)$ by

$$(T_t f)(x) := f(x - t), \text{ for all } x \in \mathbb{R}^d.$$

For $A \in GL(\mathbb{R}^d)$ we define the *dilation operator* by

$$D_A f(x) := \sqrt{|\det(A)|} f(Ax), \text{ for all } x \in \mathbb{R}^d.$$

For $a \in \mathbb{R} \setminus 0$, we set $D_a := D_A$, where $A = \text{diag}(a)$. For $f, g \in L^1$ we define the convolution of f and g by

$$f * g(x) := \int_{\mathbb{R}^d} f(y) g(x - y) dy.$$

The following theorem combines a couple of observations about the basic operations.

Theorem 2.2.1. *Let $f, g \in L^1(\mathbb{R}^d)$. Then,*

- *for $t \in \mathbb{R}^d$: $\mathcal{F}(T_t f)(\xi) = e^{-2\pi i \langle \xi, t \rangle} \hat{f}(\xi)$ and $\mathcal{F}(e^{2\pi i \langle t, \cdot \rangle} f)(\xi) = T_t \hat{f}(\xi)$;*
- *for $A \in GL(\mathbb{R}^d)$: $\mathcal{F}(D_A f) = \frac{1}{\sqrt{|\det(A)|}} \hat{f}(A^{-T} \cdot) = D_{A^{-T}} \hat{f}$;*
- *$\mathcal{F}(f * g) = \hat{f} \cdot \hat{g}$ and $\mathcal{F}(f \cdot g) = \hat{f} * \hat{g}$.*

If $f \in C^k$ and $D^\ell_{x_j} f \in L^1$, for all $\ell = 1, \dots, k$, and $j = 1, \dots, d$, then

$$\mathcal{F}\left(D^k_{x_j} f\right)(\xi) = (2\pi i \xi_j)^k \hat{f}(\xi), \text{ for all } \xi \in \mathbb{R}^d.$$

On the other hand, if $x^\ell_j f \in L^1(\mathbb{R}^d)$ for $\ell = 1, \dots, k$ and $j = 1, \dots, d$, then $\hat{f} \in C^k(\mathbb{R}^d)$ and

$$\mathcal{F}\left(x^k_j f(x)\right) = \left(\frac{i}{2\pi}\right)^k D^k_{\xi_j} \hat{f}(\xi).$$

The theorem above shows the general principle, that smoothness in spatial domain corresponds to decay in frequency domain, and functions with fast decay are transformed to very smooth functions. To understand if the converse principle holds, i.e., if we can identify smoothness or decay properties of f from its Fourier transform $\hat{f}$ we first need to be able to invert the Fourier transform.

2.3 The main theorems

The first step towards an inversion of the Fourier transform and towards a Fourier transform on the Hilbert space $L^2(\mathbb{R}^d)$ is the theorem of Plancherel.

Theorem 2.3.1 (Plancherel)**.** *Let $d \in \mathbb{N}$. If $f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$, then*

$$\|f\|_{L^2(\mathbb{R}^d)} = \left\|\hat{f}\right\|_{L^2(\mathbb{R}^d)}.$$

As a consequence, $\mathcal{F}$ extends to a unitary operator on $L^2(\mathbb{R}^d)$ such that

$$\langle f, g \rangle = \langle \hat{f}, \hat{g} \rangle, \text{ for all } f, g \in L^2(\mathbb{R}^d).$$

The Fourier transform is now obviously invertible and in some cases, we can even give a formula for $\mathcal{F}^{-1}$.

Theorem 2.3.2 (Inversion Formula)**.** *Let $d \in \mathbb{N}$. If $f \in L^1 \cap L^2(\mathbb{R}^d)$ and $\hat{f} \in L^1(\mathbb{R}^d)$, then*

$$f(x) = \int_{\mathbb{R}^d} \hat{f}(\xi) e^{2\pi i \langle x, \xi \rangle} d\xi, \text{ for all } x \in \mathbb{R}^d.$$

Proof. By Plancherel's theorem, $\mathcal{F}$ is unitary on L^2 and thus its inverse is equal to its adjoint. Let $g \in L^1(\mathbb{R}^d)$, then

$$\langle \mathcal{F}(f), g \rangle = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(x) e^{-2\pi i \langle x, \xi \rangle} \overline{g}(\xi) dx d\xi = \int_{\mathbb{R}^d} f(x) \overline{\int_{\mathbb{R}^d} g(\xi) e^{2\pi i \langle x, \xi \rangle} d\xi} dx = \langle f, \mathcal{F}^*(g) \rangle$$

and thus

$$\mathcal{F}^*(g)(x) = \int_{\mathbb{R}^d} g(\xi) e^{2\pi i \langle x, \xi \rangle} d\xi.$$

□

We can now make the correspondence between smoothness of a function and its Fourier transform more precise.

Lemma 2.3.3. *Let $d, n \in \mathbb{N}$ and $f \in L^2(\mathbb{R}^d)$. Then: $D^\alpha(f) \in L^2(\mathbb{R}^d)$ for all multiindices $|\alpha| \leq n$, if and only if $\int_{\mathbb{R}^d} (1 + |\xi|^2)^n |\hat{f}(\xi)|^2 d\xi < \infty$.*

Proof. Let f in $C_c^\infty(\mathbb{R}^d)$. By Theorem 2.2.1 we conclude that

$$\mathcal{F}(D^\alpha(f))(\xi) = (2\pi i \xi)^\alpha \hat{f}(\xi) \text{ almost everywhere.}$$

Hence, by Plancherel's theorem

$$\sum_{|\alpha| \leq n} \|D^\alpha(f)\|_{L^2}^2 = \sum_{|\alpha| \leq n} \int_{\mathbb{R}^d} |(2\pi i \xi)^\alpha|^2 |\hat{f}(\xi)|^2 d\xi.$$

It is not hard to see that

$$\sum_{|\alpha| \leq n} |(2\pi i \xi)^\alpha|^2 \sim (1 + |\xi|^2)^n, \text{ for all } \xi \in \mathbb{R}^d.$$

Thus, for all $f \in C_c^\infty(\mathbb{R}^d)$

$$\sum_{|\alpha| \leq n} \|D^\alpha(f)\|_{L^2}^2 \sim \int_{\mathbb{R}^d} (1 + |\xi|^2)^n |\hat{f}(\xi)|^2 d\xi.$$

The general case follows by the density of $C_c^\infty(\mathbb{R}^d)$ in $L^2(\mathbb{R}^d)$. □

2.4 Two examples

Let $\chi_{[-a/2,a/2]}$ be the characteristic function of $[-a/2, a/2]$. We shall compute the Fourier transform of $\chi_{[-a/2,a/2]}$. However, before we do, we can already observe, that we cannot expect $\mathcal{F}(\chi_{[-a/2,a/2]})$ to decay very quickly. Indeed, by the inversion formula, no discontinuous $L^1(\mathbb{R}^d)$ function can have a Fourier transform that is again in $L^1(\mathbb{R}^d)$. On the other hand, Theorem 2.2.1 shows that $\mathcal{F}(\chi_{[-a/2,a/2]}) \in C^k$ for all $k \in \mathbb{N}$.

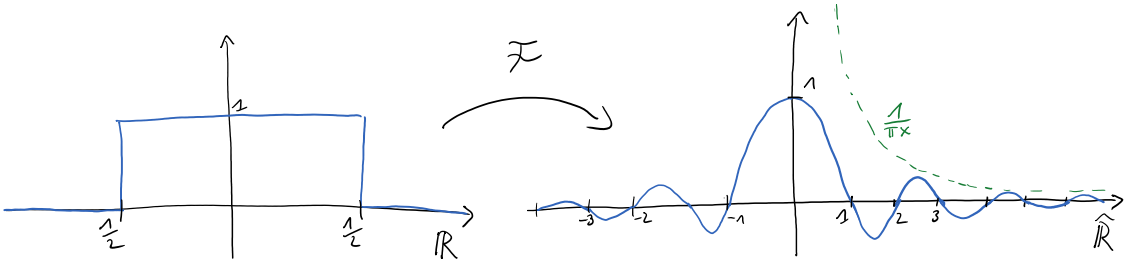

Figure 2.1: **Left:** The box function $\chi_{[-1/2,1/2]}$, **Right:** The cardinal sine function.

We have that

$$\mathcal{F}(\chi_{[-a/2,a/2]})(\xi) = \int_{-\frac{a}{2}}^{\frac{a}{2}} e^{-2\pi i \xi x} dx = \frac{e^{-\pi i \xi a} - e^{\pi i \xi a}}{-2\pi i \xi} = \frac{\sin(\pi a \xi)}{\pi \xi} =: \operatorname{sinc}(\xi).$$

The function sinc is called *cardinal sine function*. The second fairly important function is the *Gaussian function* given by

$$x \mapsto g(x) := e^{-\pi x^2}.$$

This function has the very convenient property that it is a fixed point of the Fourier transform. In other words, we have that $\widehat{g} = g$.

The Gaussian is fairly well localised around 0. But rescaling increases this localisation even more. On the other hand it destroys the localisation of the Fourier transform by basic computation of Theorem 2.2.1. In this simple example, it appears as if localisation in space and frequency domain exclude one another. This connection will be made much more precise in the following section, where we study precisely this trade-off between localisation in space and in frequency.

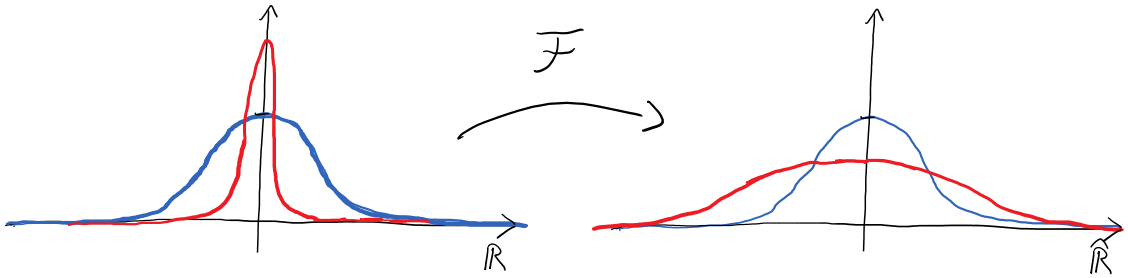

Figure 2.2: **Left:** The Gaussian function g in blue and a rescaled variant $2g(2\cdot)$ in red; **Right:** The Fourier transform $\mathcal{F}(g) = g$ and the Fourier transform $\mathcal{F}(\sqrt{2}g(2\cdot)) = 1/\sqrt{2}g(\cdot/2)$.

2.5 Uncertainty principles

An uncertainty principle identifies a limit to the joint localisation of a signal in spatial- and frequency domain. We start with one of the simplest uncertainty principles: Assume $f \in L^2(\mathbb{R})$ be such that supp f is compact. Then there exists $\Omega > 0$ such that

$$\hat{f}(\xi) = \int_{-\Omega}^{\Omega} f(x) e^{-2\pi i x \xi}, \text{ for all } \xi \in \mathbb{R}.$$

In fact, the expression above also makes sense for $\xi \in \mathbb{C}$ and shows that

$$\xi \to \hat{f}(\xi)$$

is an entire function. Hence, if $\hat{f} \neq 0$, then it cannot vanish on a subset of $\mathbb{R}$ that has an accumulation point. We conclude that supp f and supp $\hat{f}$ compact implies $f = 0$.

The classical uncertainty principle is often named Heisenberg-Pauli-Weyl inequality and makes a statement on the localisation of a function and its Fourier transform around two points. It is also demonstrated that shifted and modulated Gaussians have the best time-frequency localisation.

Theorem 2.5.1 ([18]). *Let $f \in L^2(\mathbb{R})$ and $a, b \in \mathbb{R}$, then*

$$\int_{\mathbb{R}} (x-a)^2 |f(x)|^2 dx \int_{\mathbb{R}} (\xi - b)^2 |\hat{f}(\xi)|^2 dx \geq \frac{1}{16\pi^2} \|f\|_{L^2}^4.$$

Equality holds for shifted and modulated Gaussian functions.

Proof. We shall only proof the result for $a = b = 0$ and $f \in \mathcal{S}(\mathbb{R})$ (Schwartz functions). For $s < t$ we compute by partial integration and the Leibniz rule that

$$\int_s^t xf(x)f'(x)dx = \left(t|f(t)|^2 - s|f(s)|^2\right) - \int_s^t |f(x)|^2 + xf'(x)f(x)dx.$$

Letting $s \to -\infty$ and $t \to \infty$ we get that

$$\int_{\mathbb{R}} |f(x)|^2 dx = -2 \int_{\mathbb{R}} xf'(x)f(x)dx$$

and by the Cauchy-Schwarz inequality, we conclude

$$\int_{\mathbb{R}} |f(x)|^2 dx \leq 2 \left(\int_{\mathbb{R}} |xf(x)|^2 dx\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}} |f'(x)|^2 dx\right)^{\frac{1}{2}} = 4\pi \left(\int_{\mathbb{R}} |xf(x)|^2 dx\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}} |\xi \hat{f}(\xi)|^2 d\xi\right)^{\frac{1}{2}},$$

where the last step follows from Theorem 2.2.1. **Question:** Why do we have equality for the Gaussian? □

Another popular method to prove the uncertainty principle is using the fact that for any two self-adjoint (potentially unbounded) operators A, B in a Hilbert space $\mathcal{H}$ it holds that

$$\|(A-a)f\|\|(B-b)f\| \geq \frac{1}{2}|\langle (AB-BA)f, f\rangle|, \text{ for all } f \in \mathcal{H}.$$

This result can be applied to the multiplication and differentiation operators defined by

$$(Af)(x) := xf(x) \text{ and } (Bf)(x) := \frac{1}{2\pi i} f'(x).$$

The classical uncertainty principle can interpreted thinking of $|f|^2$ as a probability density function and setting $a = \int x|f(x)|^2 dx$, then

$$\left(\int_{\mathbb{R}} (x-a)^2 |f(x)|^2 dx\right)$$

is the standard deviation of f. In other words, the product of the standard deviations of $|f|^2$ and $|\hat{f}|^2$ is lower bounded.

Before we wrap up the introduction on Fourier analysis, we shall mention one more uncertainty principle. This result compares essential supports of spatial- and frequency representations of a function.

We say that a function $f \in L^2(\mathbb{R})$ is ϵ-concentrated on a measurable set $T \subseteq \mathbb{R}$ if

$$\left(\int_{T^c} |f(x)|^2 dx\right)^{\frac{1}{2}} \leq \epsilon \|f\|_{L^2}.$$

If $0 < \epsilon < 1/2$, then most of the energy of f is located in T. If $\epsilon = 0$ then T is the support of f.

Theorem 2.5.2 (Donoho-Stark Uncertainty Principle, [18]). *Suppose $f \in L^2(\mathbb{R})$, $f \neq 0$ is ϵ_T-concentrated on $T \subset \mathbb{R}$ and $\hat{f}$ is ϵ_Ω-concentrated on $\Omega \subseteq \mathbb{R}$. Then*

$$|T||\Omega| \geq (\max\{1 - \epsilon_T - \epsilon_\Omega, 0\})^2.$$

This result now shows really clearly the essence of the uncertainty principles. If the energy in spatial domain is really concentrated on a small set, then it has to be really spread out in frequency and vice versa.

Chapter 3

Short-time Fourier transform and Gabor frames

The uncertainty principle of the Fourier transform demonstrated that we cannot have good localisation of a signal in space and frequency at the same time. Additionally, the example of the cardinal sine function shows that very small changes to a signal, on a very small domain, can have huge effects on the Fourier transform of that signal. This means that as long as we do not know the whole signal, we cannot compute even parts of the Fourier transform and vice versa. At the very least, this is not how the most well-known converter of frequency signals to information works like. The human ear does not require us to listen to a full song before we can understand parts of it.

To overcome this lack of locality, the so-called *short-time Fourier transform* was introduced.

Definition 3.0.1. *Let $g \neq 0$. Then, the* short-time Fourier transform *of a function $f \in L^2(\mathbb{R})$ with window g is defined by*

$$\mathcal{V}_g f(t,\xi) := \int_{\mathbb{R}} f(x)\overline{g(x-t)}e^{-2\pi i x\xi}dx = \langle f, M_\xi T_t g\rangle \text{ for } \xi, t \in \mathbb{R},$$

where M_ξ is the modulation operator*:*

$$M_\xi h(x) := e^{-2\pi i x\xi}h(x), \text{ for all } x \in \mathbb{R}.$$

So instead of taking the Fourier transform of the whole function, we introduce a window function ϕ, localise f in space, by multiplying with $T_t g$, and then take the Fourier transform. If g is a function with fast decay, we can think of $\mathcal{V}_g f(t,\cdot)$ as the representation of all frequencies that occur in a neighborhood of t. In this sense, this representation is very closely related to a musical score.

Similar to the Fourier transform, this transform is an isometry and admits a helpful inversion formula:

Theorem 3.0.2 (Orthogonality Relations and inversion formula, [18])**.** *For $f_1, f_2, g_1, g_2 \in L^2(\mathbb{R})$ we have*

$$\langle V_{g_1} f_1, V_{g_2} f_2\rangle_{L^2(\mathbb{R}^2)} = \langle f_1, f_2\rangle\overline{\langle g_1, g_2\rangle}.$$

For any $\gamma, g \in L^2(\mathbb{R})$ and $\langle g, \gamma\rangle \neq 0$, we have that for all $f \in L^2(\mathbb{R})$:

$$f = \frac{1}{\langle \gamma, g\rangle}\int_{\mathbb{R}}\int_{\mathbb{R}} V_g f(x,\xi) M_\xi T_x \gamma d\xi dx.$$

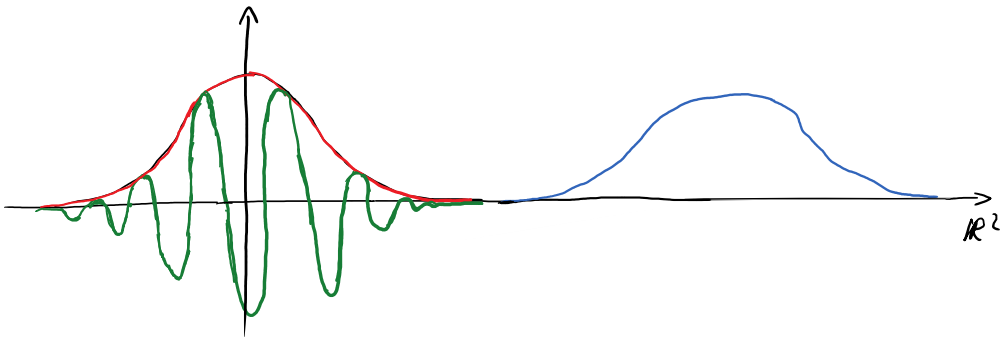

Figure 3.1: A collection of elements from a Gabor system, resulting from translations and modulations of a generating function.

3.1 Gabor systems

We can now introduce our first representation system, apart from the system of complex exponentials from the introduction.

Definition 3.1.1. *Let $g \in L^2(\mathbb{R})$ and let $a, b > 0$. Then the* Gabor system *$G(g, a, b)$ is defined by*

$$G(g, a, b) := \{g_{am,bn} := M_{bn} T_{am} g := e^{-2\pi i b n \cdot} g(\cdot - am) : m, n \in \mathbb{Z}\}.$$

We want to study this system from the point of view that was layed out in the beginning of the lecture. However, it is not clear if $G(\phi, a, b)$ forms a basis, or is even a spanning set. Moreover, even if $G(\phi, a, b)$ were spanning, there does not seem to be a direct method to retrieve the coefficients $(c_{n,m}(f))_{m,n\in\mathbb{Z}}$ of an expansion

$$f = \sum_{m,n\in\mathbb{Z}} c_{m,n}(f) g_{am,bn},$$

let alone show that they contain any information about f.

To study these questions we first need to make a small detour and introduce *frames*.

3.2 Frames

A frame is a generalisation of an orthonormal basis. Recall that by Parseval's identity, every orthonormal basis $(\phi_n)_{n\in\mathbb{N}}$ for a Hilbert space $\mathcal{H}$ satisfies:

$$\|f\|_{\mathcal{H}}^2 = \sum_{n\in\mathbb{N}} |\langle f, \phi_n \rangle_{\mathcal{H}}|^2.$$

We get to the definition of a frame by replacing the equality above, by two inequalities.

Definition 3.2.1. *Let $\mathcal{H}$ be a Hilbert space. A sequence $(\phi_n)_{n\in\mathbb{N}} \subset \mathcal{H}$ is called a* frame*, if there exist $0 < A \leq B$ such that for all $f \in \mathcal{H}$:*

$$A\|f\|_{\mathcal{H}}^2 \leq \sum_{n\in\mathbb{N}} |\langle f, \phi_n \rangle_{\mathcal{H}}|^2 \leq B\|f\|_{\mathcal{H}}^2.$$

If $A = B$ is possible, we call the frame tight*.*

Clearly, every orthonormal basis is a frame. However, the elements of a frame do not have to be orthonormal, or even linearly independent. Nonetheless, a frame needs to span the Hilbert space. Indeed, if $\overline{\text{span}\{\phi_n, n \in \mathbb{N}\}} \neq \mathcal{H}$, then there exists $0 \neq f \in \overline{\text{span}\{\phi_n, n \in \mathbb{N}\}}^{\perp}$. This implies that

$$\sum_{n\in\mathbb{N}} |\langle f, \phi_n\rangle_{\mathcal{H}}|^2 = 0,$$

which contradicts $A > 0$. A couple of standard examples of frames are

- The Mercedes Benz frame:

$$\phi_1 = \begin{pmatrix}1\\0\end{pmatrix}, \quad \phi_2 = \begin{pmatrix}-1/2\\\sqrt{3}/2\end{pmatrix}, \quad \phi_3 = \begin{pmatrix}-1/2\\-\sqrt{3}/2\end{pmatrix},$$

  is a tight frame for $\mathbb{R}^2$ with frame bounds $A = B = 3/2$.

- Let $(\phi_n)_{n\in\mathbb{N}}$ be an orthonormal basis for $\mathcal{H}$, then

$$\phi_1, \phi_1, \phi_2, \phi_2, \ldots,$$

  is a frame for $\mathcal{H}$ with frame bound $A = B = 2$.

- Let $(\phi_n)_{n\in\mathbb{N}}$ be an orthonormal basis for $\mathcal{H}$, then

$$\phi_1, \phi_1, \phi_2, \phi_3, \phi_4 \ldots,$$

  is a frame for $\mathcal{H}$ with frame bound $A = 1$, $B = 2$.

We introduce two operators associated to a frame.

Definition 3.2.2. *Let $(\phi_n)_{n\in\mathbb{N}} \subset \mathcal{H}$ be a frame for $\mathcal{H}$, then the operator*

$$T : \mathcal{H} \to \ell^2(\mathbb{N}), \quad f \mapsto (\langle f, \phi_n\rangle)_{n\in\mathbb{N}}$$

is called analysis operator of $(\phi_n)_{n\in\mathbb{N}}$. *The operator*

$$T^* : \ell^2(\mathbb{N}) \to \mathcal{H}, \quad (c_n)_{n\in\mathbb{N}} \mapsto \sum_{n\in\mathbb{N}} c_n\phi_n$$

is called synthesis operator of $(\phi_n)_{n\in\mathbb{N}}$.

We make two observations: first, T is a bounded operator since

$$\|Tf\|_{\ell^2}^2 = \sum_{n\in\mathbb{N}} |\langle f, \phi_n\rangle|^2 \leq B\|f\|_{\mathcal{H}}^2$$

per definition of a frame. Second, as already suggested by the notation, T^* is the adjoint of T. Indeed, if we ignore summability issues, then for $c \in \ell^2$, $g \in \mathcal{H}$

$$\langle T^*c, g\rangle_{\mathcal{H}} = \sum_{n\in\mathbb{N}} c_n\langle\phi_n, g\rangle_{\mathcal{H}} = \sum_{n\in\mathbb{N}} c_n\overline{\langle g, \phi_n\rangle_{\mathcal{H}}} = \langle c, Tg\rangle_{\ell^2}.$$

Now we have seen, that the analysis operator is continuous, and per construction it is also injective and its inverse is a bounded operator defined on $ran(T)$. Moreover, reconstructing f from $T(f)$ is very simple, by using the concept of a dual frame, its definition first requires us to study the so-called *frame operator*.

Definition 3.2.3. *Let $(\phi_n)_{n\in\mathbb{N}} \subset \mathcal{H}$ be a frame for $\mathcal{H}$. Then,*

$$\mathcal{S} = T^*T : \mathcal{H} \to \mathcal{H}, f \mapsto \sum_{n\in\mathbb{N}} \langle f, \phi_n\rangle \phi_n$$

is called frame operator *of $(\phi_n)_{n\in\mathbb{N}}$.*

We have the following theorem:

Theorem 3.2.4 ([6])**.** *Let $(\phi_n)_{n\in\mathbb{N}} \subset \mathcal{H}$ be a frame with frame bounds A, B. Then, the operator $\mathcal{S}$ is self-adjoint with spectrum $\sigma(\mathcal{S}) \subseteq [A, B]$. In particular, $\mathcal{S}$ possesses a bounded inverse.*

Proof. We have that

$$\langle \mathcal{S}f, f\rangle = \langle Tf, Tf\rangle = \|Tf\|^2_{\ell^2}.$$

Hence

$$A\|f\|^2_{\mathcal{H}} \leq \langle \mathcal{S}f, f\rangle \leq B\|f\|^2_{\mathcal{H}}.$$

So the numerical range of $\mathcal{S}$ is a subset of $[A, B]$ which yields the result. □

Definition 3.2.5. *Let $(\phi_n)_{n\in\mathbb{N}} \subset \mathcal{H}$ be a frame for $\mathcal{H}$ with frame operator $\mathcal{S}$, then we define by $(\widetilde{\phi_n})_{n\in\mathbb{N}} := (\mathcal{S}^{-1}\phi_n)_{n\in\mathbb{N}}$ the* canonical dual frame *of $(\phi_n)_{n\in\mathbb{N}}$*

It is not hard to verify that $(\mathcal{S}^{-1}\phi_n)_{n\in\mathbb{N}}$ is a frame itself. Finally, we get a convenient way of reconstructing a signal from its analysis coefficients.

Theorem 3.2.6 ([6])**.** *Let $(\phi_n)_{n\in\mathbb{N}} \subset \mathcal{H}$ be a frame for $\mathcal{H}$ with frame operator $\mathcal{S}$ then we have*

- *The reconstruction formula: For all $f \in \mathcal{H}$*

$$f = \sum_{n\in\mathbb{N}} \langle f, \phi_n\rangle \widetilde{\phi_n}.$$

- *The decomposition formula: For all $f \in \mathcal{H}$*

$$f = \sum_{n\in\mathbb{N}} \langle f, \widetilde{\phi_n}\rangle \phi_n.$$

Proof. We have that for all $f \in \mathcal{H}$

$$f = \mathcal{S}^{-1}\mathcal{S}f = \mathcal{S}^{-1}\sum_{n\in\mathbb{N}} \langle f, \phi_n\rangle \phi_n = \sum_{n\in\mathbb{N}} \langle f, \phi_n\rangle \mathcal{S}^{-1}\phi_n.$$

This yields the reconstruction formula. From $f = \mathcal{S}\mathcal{S}^{-1}f$, we deduce the decomposition formula by similar means. □

We have now collected all the frame theory, that we will need. However, one comment is in order. In contrast, to the three main objectives of applied harmonic analysis that were posed at the beginning of the lecture, we see, that, if we work with frames, instead of bases, then we need to ask more questions. Indeed, for a frame, there are now two types of efficient representation or sparsity. We say that a signal f is *analysis sparse* with respect to a frame if Tf can be well approximated with only few terms. We say that a signal f is *synthesis sparse* with respect to a frame if there exists a vector c with few non-zero entries such that f is well approximated by T^*c. The same two points of view now appear when we ask about an interpretation and manipulation of a transform.

3.3 Gabor frames

We continue our analysis of Gabor systems, and aim at understanding under which conditions a Gabor system $G(g,a,b)$ forms a frame. We start with very good news. First of all, if $G(g,a,b)$ is a frame, then its dual frame is very easy to compute.

Theorem 3.3.1 ([18]). *Let $g \in L^2(\mathbb{R})$, $a,b>0$. Suppose that $G(g,a,b)$ forms a frame for $L^2(\mathbb{R})$, then the canonical dual frame of $G(g,a,b)$ has the form $(M_{bm}T_{an}(\mathcal{S}^{-1}g))_{m,n\in\mathbb{Z}}$.*

In other words, the canonical dual frame of a Gabor frame is again a Gabor frame and we can compute its window function easily. Moreover, it there are easily verifiable conditions guaranteeing that a system is a Gabor frame. We have the following result describing a large set of Gabor frames known as Painless Nonorthogonal Expansions, [13].

Theorem 3.3.2 ([12]). *Let $0<ab\leq 1$, and let g be such that* supp $g \subset [0,1/b]$. *Then $G(g,a,b)$ forms a frame for $L^2(\mathbb{R})$ if and only if there exist $A,B>0$ such that*

$$A \leq \frac{1}{b}\sum_{m\in\mathbb{Z}} |g(t-am)|^2 \leq B \text{ for a.e. } t\in\mathbb{R}. \tag{3.3.1}$$

We see that if $ab=1$, then (3.3.1) cannot be satisfied if g is continuous. If $ab<1$, then even a smooth function g can satisfy (3.3.1). Finally, if ab and g is as in the theorem, then $G(g,a,b)$ cannot form a frame. There appears to be something special about the threshold $ab\leq 1$. In fact, this upper bound hold even without imposing any assumption on the support of g.

Theorem 3.3.3 ([12]). *Let $g\in L^2(\mathbb{R})$, $a,b>0$:*

- *If $G(g,a,b)$ forms a frame, then $ab\leq 1$*
- *If $G(g,a,b)$ forms an orthonormal basis, then $ab=1$.*

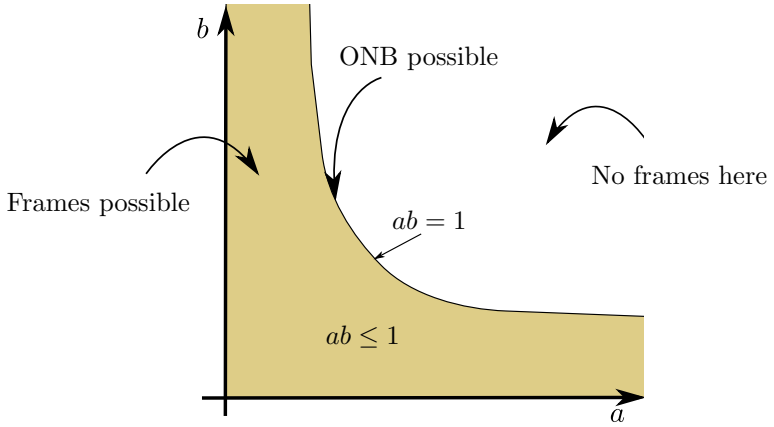

Figure 3.2: Depiction of the situation of Theorem 3.3.3. A Gabor frame only exists if the sampling parameter a,b satisfy $ab\leq 1$. If $ab=1$ then even ONBs are possible.

We shall conclude this discussion on Gabor systems by one of the main motivations for wavelet systems, that will be introduced in the next chapter. In fact, while Theorem 3.3.3 does not rule out Gabor systems that form ONBs, it turns out, that all such systems must have generating windows that have very bad time frequency localisation. This is the content of the famous Balian-Low Theorem:

Theorem 3.3.4 (Balian-Low Theorem, [18]). *Let $g \in L^2(\mathbb{R})$. If $G(g,a,b)$ is an orthonormal basis, then*

$$\left(\int_{\mathbb{R}} t^2|g(t)|^2dt\right)\left(\int_{\mathbb{R}} \xi^2|\widehat{g}(\xi)|^2dt\right) = \infty. \tag{3.3.2}$$

Proof. Let $G(g,a,b)$ be an orthonormal basis. Then $g \neq 0$ and thus (3.3.2) is equivalent to

$$\int_{\mathbb{R}} t^2|g(t)|^2dt = \infty \quad \text{or} \quad \int_{\mathbb{R}} \xi^2|\widehat{g}(\xi)|^2dt = \infty.$$

We assume towards a contradiction that both integrals above are finite. As $g \in L^2$ we conclude with Lemma 2.3.3 that $g \in H^1$. We define two operators

$$\begin{aligned} X : L^2(\mathbb{R}, tdt) \to L^2(\mathbb{R}), \qquad (Xf)(t) &:= tf(t); \\ P : H^1(\mathbb{R}) \to L^2(\mathbb{R}), \qquad (Pf)(t) &:= -2\pi i f'(t). \end{aligned}$$

Per assumption $X(f), P(f) \in L^2$ and $G(g,a,b)$ is an orthonormal basis, hence

$$\langle Xg, Pg\rangle = \sum_{m,n\in\mathbb{Z}} \langle Xg, g_{am,bn}\rangle\langle g_{am,bn}, Pg\rangle. \tag{3.3.3}$$

Let us analyse the expressions individually. We have that

$$\langle Xg, g_{am,bn}\rangle = \langle Xg, M_{am}T_{bn}g\rangle = \langle g, XM_{am}T_{bn}g\rangle.$$

Moreover,

$$XM_{am}T_{bn}g(t) = te^{-2\pi iamt}g(t-bn) = M_{am}T_{bn}Xg(t) + bnM_{am}T_{bn}g(t).$$

By the orthogonality of $M_{am}T_{bn}g(t)$ and g we get that

$$\langle Xg, g_{am,bn}\rangle = \langle g, M_{am}T_{bn}Xg(t)\rangle.$$

Finally, the adjoint of $M_{am}T_{bn}$ can be computed to be $e^{2\pi iambn}M_{-am}T_{-bn}$. This yields that

$$\langle Xg, g_{am,bn}\rangle = e^{2\pi iambn}\langle g_{-am,-bn}, Xg\rangle.$$

By the same argument, we get that

$$\langle g_{am,bn}, Pg\rangle = e^{-2\pi iambn}\langle Pg, g_{-am,-bn}\rangle.$$

This shows with (3.3.3) that

$$\langle Xg, Pg\rangle = \langle Pg, Xg\rangle.$$

We have that for $g \in \mathrm{dom}XP \cap \mathrm{dom}PX$:

$$\langle (XP - PX)g, g\rangle = -2\pi i\langle g, g\rangle.$$

Let us assume that $g \in \mathrm{dom}XP \cap \mathrm{dom}PX$. (The general case can be shown via approximation.) Then we get that

$$0 = \langle Pg, Xg\rangle - \langle Xg, Pg\rangle = \langle (XP - PX)g, g\rangle \neq 0,$$

which is a contradiction. □

Chapter 4

Wavelets

There are two main problems with Gabor systems. First of all, if the window function has a very large support, then the transform is not very well localised. This means that, for example, if a signal is smooth except for one discontinuity, then many coefficients of a Gabor representation of this signal will be large and so the representation is not very sparse. Additionally, the Balian-Low theorem showed that we cannot have nice Gabor orthonormal bases.

An idea, that surprisingly settles both problems is to use a window with flexible size. The key idea is the following. Instead of translating and modulating a generator window, we now use translations and dilations. Thereby, the signal is analysed at different resolutions.

4.1 Continuous wavelet transform

To stay parallel to the previous constructions, we first start with a continuous transform.

Definition 4.1.1. *Let $\psi \in L^2(\mathbb{R})$. The* continuous wavelet transform associated with the wavelet ψ *of a function $f \in L^2(\mathbb{R})$ is defined as*

$$\begin{aligned} W_\psi(f)(a,b) :&= \int_{\mathbb{R}} f(t) a^{-\frac{1}{2}} \psi\left(\frac{t-b}{a}\right) dt \\ &= \langle f, T_b D_{a^{-1}} \psi \rangle \\ &= (f * D_{a^{-1}} \psi^*)(b), \end{aligned}$$

where $\psi^*(t) = \psi(-t)$, and $a, b \in \mathbb{R}$.

Interestingly enough, the continuous wavelet transform is an isometry if the wavelet satisfies an admissibility condition, and it admits an inversion formula.

Theorem 4.1.2 ([12]). *Let $0 \neq \psi \in L^2(\mathbb{R})$ be such that*

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

Then, any $f \in L^2(\mathbb{R})$ satisfies

$$f = \frac{1}{C_\psi} \int_0^\infty \int_{\mathbb{R}} W_\psi(f)(a,b) T_b D_{a^{-1}} \psi db \frac{da}{a^2} \tag{4.1.1}$$

and

$$\|f\|_{L^2}^2 = \frac{1}{C_\psi} \int_0^\infty \int_{\mathbb{R}} |W_\psi(f)(a,b)|^2 db \frac{da}{a^2}.$$

Proof. Recall that $W_\psi(f)(a,b) = (f * D_{a^{-1}}\psi^*)(b)$. We only demonstrate the isotropy property.

$$\begin{aligned}
\int_0^\infty \int_{\mathbb{R}} |W_\psi(f)(a,b)|^2 db &= \int_0^\infty \int_{\mathbb{R}} W_\psi(f)(a,b) \cdot W_\psi(f)(a,b) db \frac{da}{a^2} \\
&= \int_0^\infty \int_{\mathbb{R}} (f * D_{a^{-1}}\psi^*)(b)(f * D_{a^{-1}}\psi^*)(b) db \frac{da}{a^2} \\
&= \int_0^\infty \int_{\mathbb{R}} \hat{f}(\xi)\mathcal{F}(D_{a^{-1}}\psi^*)(\xi)\hat{f}(\xi)\mathcal{F}(D_{a^{-1}}\overline{\psi}^*)(\xi) d\xi \frac{da}{a^2} \\
&= \int_0^\infty \int_{\mathbb{R}} |\hat{f}(\xi)|^2 |\mathcal{F}(D_{a^{-1}}\psi^*))(\xi)|^2 d\xi \frac{da}{a^2} \\
&= \int_{\mathbb{R}} |\hat{f}(\xi)|^2 \int_0^\infty |\mathcal{F}(D_{a^{-1}}\psi^*))(\xi)|^2 \frac{da}{a^2} d\xi.
\end{aligned}$$

We compute that for all $\xi \in \mathbb{R}$

$$\mathcal{F}(D_a\psi^*))(\xi) = D_a\mathcal{F}(\psi^*)(\xi) = \sqrt{a}\overline{\widehat{\psi}(a\xi)}.$$

Hence we get that

$$\int_0^\infty |\mathcal{F}(D_{a^{-1}}\psi^*))(\xi)|^2 \frac{da}{a^2} = \int_0^\infty \frac{|\widehat{\psi}(a\xi)|^2}{a} da = \int_0^\infty \frac{|\widehat{\psi}(a)|^2}{a} da,$$

where the last line holds by transforming $a \to a\xi$ if $\xi \neq 0$ and by the symmetry of $\widehat{\psi}$ (since ψ is real-valued). □

The property

$$\int_0^\infty \frac{|\psi(\xi)|^2}{\xi} d\xi < \infty$$

is often called *Calderon condition* and (4.1.1) is called Calderon's reproducing formula. To guarantee the Calderon condition, we need that $\widehat{\psi}(0) = 0$. This is equivalent to

$$\int_{\mathbb{R}} \psi(x) dx = 0.$$

In other words, an admissible wavelet needs to oscillate.

We have claimed that this wavelet transform overcomes the problems of the short-time Fourier transform. Indeed the following theorem shows that the behaviour of the wavelet transform asymptotically only depends on local smoothness. Variants of the following result exist in various forms in the literature. We present here the simplest estimate imaginable.

Theorem 4.1.3. *Let $L \in N$, $\psi \in L^2(\mathbb{R})$, supp $\psi \subset [-1,1]$, and let ψ have L vanishing moments, i.e, ψ is orthogonal on all polynomials of degree less than L. If $f \in L^2(\mathbb{R})$ is L-times continuously differentiable on a neighborhood of $b \in \mathbb{R}$, then*

$$|W_\psi(f)(a,b)| \lesssim a^{L+\frac{1}{2}} \text{ for } a \to 0.$$

Proof. We have that

$$|W_\psi(f)(a,b)| = |\int_{\mathbb{R}} f(x) T_b D_{a^{-1}} \psi(x) dx| = |\int_{B_a(b)} f(x) T_b D_{a^{-1}} \psi(x) dx|.$$

Using a Taylor expansion of f of order $L-1$ around d we get that

$$|\int_{B_a(b)} f(x) T_b D_{a^{-1}} \psi(x) dx| \lesssim \int_{B_a(b)} |x^* - b|^L |T_b D_{a^{-1}} \psi(x)| dx,$$

for an $x^* \in B_a(b)$. Thus

$$|W_\psi(f)(a,b)| \lesssim a^L \|T_b D_{a^{-1}} \psi\|_{L^1} = a^{L+\frac{1}{2}}.$$

□

Using the inversion formula one can also produce a converse to Theorem 4.1.3. In Figure 4.1 we compute the continuous shearlet transform of a piecewise smooth signal. We observe that the transform decays slowly at translation points associated to singularities of the signal. Additionally, the areas of slow decay become better and better localised for smaller a.

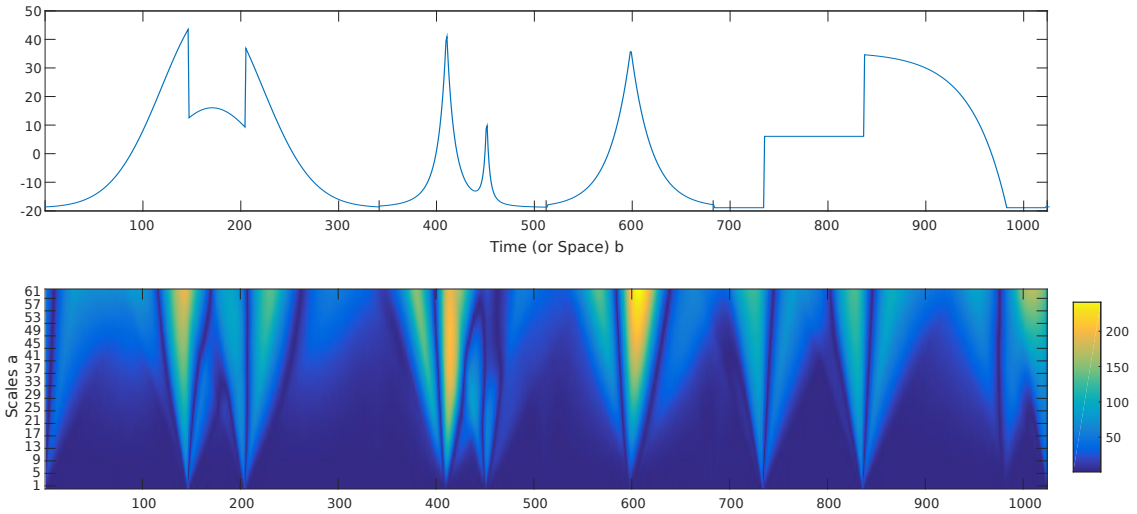

Figure 4.1: **Top:** A signal with multiple singularities of different types. **Bottom:** Continuous wavelet transform of the signal, with b varying along the x axis and varying scale a along the y-axis. The wavelet used in the computation is a symmetric wavelet with four vanishing moments, called symlet.

4.2 Discrete wavelet systems

Yet again, we find ourselves in the position that we constructed a continuous transform with very interesting properties and we would like to transform it into a discrete system.

Definition 4.2.1. *Let $\psi \in L^2(\mathbb{R})$ be a wavelet and $a, b > 0$. Then, the associated* wavelet system *is defined as*

$$\mathcal{W}(\psi, a, b) := \left\{\psi_{j,m} := a^j \psi(a^j \cdot - bm) : j, m \in \mathbb{Z}\right\}.$$

In contrast to Gabor frames, there do exist nice wavelet orthonormal bases. Because of this, we do not want to analyse conditions on the frame properties of wavelets in too much detail. In essence, if ψ is admissible and decays sufficiently fast in time and frequency, then there exist $a, b > 0$ such that $\mathcal{W}(\psi, a, b)$ is a frame. To construct wavelet systems that are orthonormal bases, the celebrated method of multiresolution approximation is usually used.

Definition 4.2.2. *A* multiresolution approximation *is a sequence of closed subspaces $(V_j)_{j\in\mathbb{Z}}$ of $L^2(\mathbb{R})$ such that the following conditions are fulfilled:*

1. *$V_j \subset V_{j+1}$ for all $j \in \mathbb{Z}$;*
2. *$\overline{\bigcup_{j\in\mathbb{Z}} V_j} = L^2(\mathbb{R})$;*
3. *$\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$;*
4. *$f \in V_0$ if and only if $f(2^j \cdot) \in V_j$;*
5. *There is a function $\varphi \in L^2(\mathbb{R})$ such that $\{T_m\varphi : m \in \mathbb{Z}\}$ is an orthonormal basis for V_0. We call φ* scaling function for the MRA.

We can think of V_j as spaces of different resolution. Indeed, as the resolution increases, i.e., $j \to \infty$, we have

$$\|f - P_{V_j} f\|_{L^2} \to 0, \text{ for every } f \in L^2(\mathbb{R}).$$

On the other hand, if $j \to -\infty$, then

$$\|P_{V_j} f\|_{L^2} \to 0, \text{ for every } f \in L^2(\mathbb{R}).$$

A multiresolution approximation will turn out to be a very helpful tool to generate multiscale orthonormal bases for $L^2(\mathbb{R})$. This is done by introducing the *wavelet spaces* W_j, which are defined by

$$W_j := V_{j+1} \ominus V_j.$$

The spaces W_j can now be thought of as containing the details of the signal that were not observable at resultion j but are observable at resolution $j + 1$. Per construction, we have that

$$V_j = W_{j-1} \oplus V_{j-1} = W_{j-1} \oplus W_{j-2} \oplus V_{j-2} = \dots.$$

This demonstrates that

$$L^2(\mathbb{R}) = \overline{\bigoplus_{j\in\mathbb{Z}} W_j} = \overline{\bigoplus_{j\geq j_0} W_j \oplus V_0}.$$

Additionally, it is not hard to see, that if $f \in W_0$ then $f(2^j \cdot) \in W_j$. This shows that, if we find a function ψ such that $\{T_m\psi : m \in \mathbb{Z}\}$ is an orthonormal basis for W_0, then

$$\{T_m D_{2^j} \psi : j, m \in \mathbb{Z}\}$$

and for all $j_0 \in \mathbb{Z}$

$$\{T_m D_{2^j} \psi : j \geq j_0, m \in \mathbb{Z}\} \cup \{T_m D_{2^{j_0}} \phi : m \in \mathbb{Z}\}$$

are orthonormal bases for $L^2(\mathbb{R})$. The problem of producing a wavelet bases is now reduced to finding an MRA and an associated wavelet function. In fact, for any scaling function there exists an associated wavelet function by the following result.

Theorem 4.2.3 ([31]). *Let $\phi \in L^2(\mathbb{R})$ be a scaling function for an MRA $(V_j)_{j\in\mathbb{Z}}$, then there exists a function ψ such that for all $j \in \mathbb{Z}$: $\{\psi_{j,m} : m \in \mathbb{Z}\}$ is an orthonormal basis for the associated wavelet spaces W_j. One possibility to construct ψ is by setting*

$$\widehat{\psi}(\xi) = e^{-2\pi i \frac{\xi}{2}} h\left(\frac{\xi}{2} + \frac{1}{2}\right) \widehat{\phi}\left(\frac{\xi}{2}\right),$$

where h is a 1-periodic filter such that

$$\widehat{\phi}(2\xi) = \frac{1}{\sqrt{2}} h(\xi)\widehat{\phi}(\xi).$$

Instead of proving the result above, we will analyse a standard example of a multiresolution approximation due to Haar [22].

Definition 4.2.4. *Let $\phi := \chi_{[0,1]}$. We call ϕ the Haar scaling function and define for $j \in \mathbb{Z}$*

$$V_j := \overline{\left\{\phi_{j,m} = 2^{-\frac{j}{2}}\phi(2^j \cdot -m) : m \in \mathbb{Z}\right\}}$$

the Haar scaling spaces. *Moreover, we define by $\psi := \chi_{[0,\frac{1}{2})} - \chi_{[\frac{1}{2},1]}$ the* Haar wavelet *and for $j \in \mathbb{Z}$*

$$W_j := \overline{\left\{\psi_{j,m} = 2^{-\frac{j}{2}}\psi(2^j \cdot -m) : m \in \mathbb{Z}\right\}}.$$

The spaces W_j are then called the Haar wavelet spaces.

Of course we want to find out if $(V_j)_{j\in\mathbb{Z}}$ is an MRA ϕ the scaling function, and if $(W_j)_{j\in\mathbb{Z}}$ are the associated wavelet spaces.

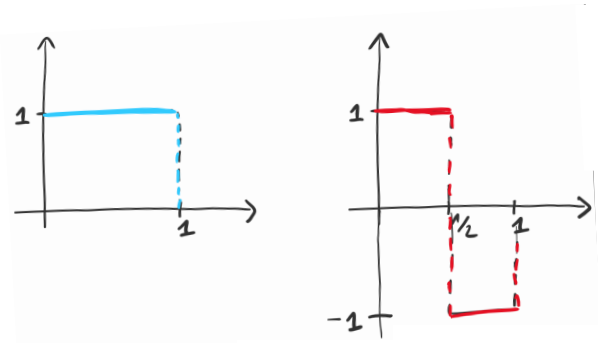

Figure 4.2: **Left:** The Haar scaling function, **Right:** The Haar wavelet. These functions are clearly orthogonal to each other.

Theorem 4.2.5 ([31]). *The Haar scaling spaces form an MRA. Additionally, we have that*

$$W_j \oplus V_j = V_{j+1}.$$

Proof. We check each property of an MRA individually:

We have that $\phi_{j,m} = \chi_{2^{-j}[m,m+1]} = \chi_{2^{-j-1}[2m,2m+1]} + \chi_{2^{-j-1}[2m+1,2m+2]}$. This yields $V_j \subset V_{j+1}$. The space V_j contains all functions that are piecewise constant on intervals with start and endpoints in $2^{-j}\mathbb{Z}$. This implies the second and third property. If $f \in V_0$, then there is a sequence $f_n = \sum_{k=1}^{n} c_k \phi_{0,m}$ converging to f for $n \to \infty$. Clearly, $f_n(2^j \cdot) = \sum_{k=1}^{n} 2^{-\frac{j}{2}} c_k \phi_{j,m} \in V_j$ and $f_n(2^j \cdot)$ converges to $f(2^j \cdot)$. The converse holds with the same argument. Since supp $T_m \phi = [m, m+1]$ we have that $T_m \phi$, $T_n \phi$ are orthogonal if $m \neq n$. Thus the last property follows.

Finally, it is obvious that W_j is orthogonal to V_j since all elements of V_j are constant in $\mathbb{R} \setminus 2^{-j}\mathbb{Z}$ and elements of W_j integrate to 0 over intervals of length 2^{-j} with start points in $2^{-j}\mathbb{Z}$, see also Figure 4.2 for an illustration. How to get

$$V_{j+1} = W_j \oplus V_j$$

is sketched in Figure 4.3. □

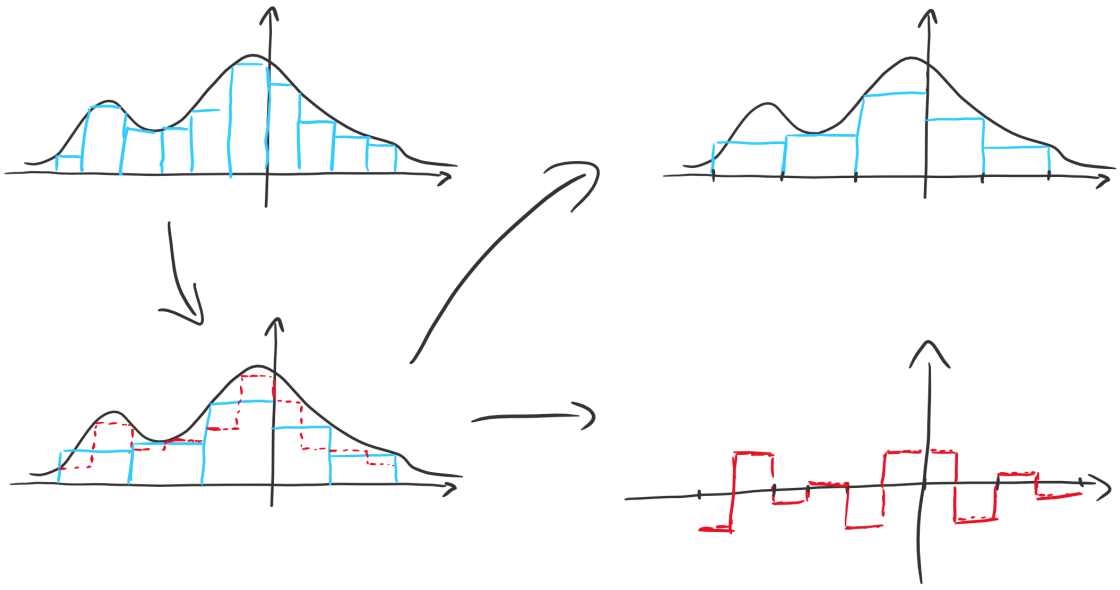

Figure 4.3: Illustration of the projection of a function into V_j and W_j. The top left shows a function approximated in a scaling space. We then see that if two adjacent scaling functions are grouped together, they can be interpreted as a scaling function on a lower scale plus a wavelet update, which is exactly given by the Haar wavelet. This shows the decomposition of V_{j+1} into V_j and W_j.

We have just introduced the first orthogonal wavelet bases, the Haar wavelet basis. If one contemplates what a good wavelet basis would be, then it, unfortunately, becomes apparent quickly, that a wavelet basis is a very bad wavelet basis. Indeed, three properties seem to be particularly important for wavelets, for various reasons that we have seen and will see repeatedly in this manuscript:

- *Spatial decay*, i.e., $\psi(x) \lesssim (1+|x|)^{-P}$ for hopefully very large P. Even better would be supp ψ compact.
- *Regularity/ Fourier decay*, i.e., $\widehat{\psi}(x) \lesssim (1+|x|)^{-M}$ for hopefully very large M. Even better would be supp $\widehat{\psi}$ compact, but by the uncertainty principle this can only be achieved if supp ψ is not compact.
- *Vanishing moments*, i.e., ψ should be orthogonal on all polynomials of order less than $L-1$ for a large L.

In fact, the Haar wavelet has compact support, but is not continuous, i.e., its Fourier transform is not in $L^1(\mathbb{R})$ and it has only one vanishing moment. Nonetheless, there are plenty of nice wavelet constructions.

We finally present a celebrated theorem from [12] which demonstrates that for any given decay, in space, frequency and number of vanishing moments, there exists a wavelet with these characteristics, which yields an orthonormal basis for $L^2(\mathbb{R})$.

Theorem 4.2.6 ([12]). *There are constants $c > 0$ such that for every $r \in \mathbb{N}$ there exists an MRA with scaling function ϕ and associated wavelet ψ such that*

- $\phi, \psi \in C^{\frac{r}{c}}(\mathbb{R})$;
- *ψ has r vanishing moments;*
- $\operatorname{supp} \phi, \operatorname{supp} \psi \subset [0, 2r]$.

4.3 Higher dimensions, bounded domains

The main area of application of wavelets is image processing. Hence, it is clear that we need a two-dimensional variant of the wavelet transform. Moreover, as most images are defined on bounded domains, whereas wavelet systems are defined on $\mathbb{R}$ we need to adapt to this situation as well. Both of these issues are also important if one wants to solve PDEs using a wavelet discretisation. We shall mention how to overcome both of these issues briefly below.

4.3.1 Higher dimensions

It is clear, that for any orthonormal basis $(\phi_n)_{n \in I}$ for a Hilbert space $\mathcal{H}$ and some index set I, the set

$$\{\phi_n \otimes \phi_m : n, m \in I\}$$

is an orthonormal basis for $\mathcal{H} \otimes \mathcal{H}$. This gives a simple construction of a wavelet basis for $L^2(\mathbb{R}^d)$, $d \in \mathbb{N}$, but the resulting system is not a proper multiscale system. In fact, the elements can have very different sizes in each coordinate direction. A different approach is to define a two-dimensional multiresolution approximation $(V_j^2)_{j \in \mathbb{Z}}$ by

$$V_j^2 := V_j \otimes V_j$$

where V_j is a one-dimensional multiresolution approximation. The associated wavelet spaces should again correspond to the differences between one level of resolution to the next. Thus, we set

$$V_{j+1}^2 := V_j^2 + W_j^2.$$

Since

$$V_{j+1}^2 = (V_j + W_j) \otimes (V_j + W_j) = V_j^2 \oplus (V_j \otimes W_j) \oplus (W_j \otimes V_j) \oplus (W_j \otimes W_j),$$

we get that

$$W_j^2 = (V_j \otimes W_j) \oplus (W_j \otimes V_j) \oplus (W_j \otimes W_j).$$

This implies the following theorem.

Theorem 4.3.1 ([31]). *Let $(V_j)_{j\in\mathbb{Z}}$ be an MRA for $L^2(\mathbb{R})$ with scaling function ϕ and wavelet ψ. We define for $(x_1, x_2) \in \mathbb{R}^2$*

$$\begin{aligned}\psi^1(x_1, x_2) &:= \phi(x_1)\psi(x_2)\\ \psi^2(x_1, x_2) &:= \psi(x_1)\phi(x_2)\\ \psi^3(x_1, x_2) &:= \psi(x_1)\psi(x_2).\end{aligned}$$

Then

$$\left\{\psi^k_{j,m} := 2^{-j}\psi^k(2^j x - m) : \quad m \in \mathbb{Z}^2, k = 1, 2, 3\right\}$$

is an orthonormal basis for W_j^2. Moreover,

$$\begin{aligned}&\left\{\psi^k_{j,m} : \quad m \in \mathbb{Z}^2, j \in \mathbb{Z}, k = 1, 2, 3\right\} \text{ and}\\ &\left\{\psi^k_{j,m} : \quad m \in \mathbb{Z}^2, j \geq j_0, k = 1, 2, 3\right\} \cup \left\{\phi_{j_0,m} := 2^{-j_0}\phi(2^{j_0}x - m) : \quad m \in \mathbb{Z}^2\right\}\end{aligned}$$

are ONBs for $L^2(\mathbb{R}^2)$ and any $j_0 \in \mathbb{Z}$.

It goes without saying that one can make a similar construction for any dimension $d \in \mathbb{N}$.

4.3.2 Bounded domains

We shall discuss three approaches to adapt a wavelet basis for $L^2(\mathbb{R})$ to a bounded domain $[0, 1]$.

The first idea that comes to mind is to simply truncate. Let $(\phi_n)_{n\in\mathbb{N}}$ be a wavelet basis, then we define for $n \in \mathbb{N}$:

$$\phi_n^{[0,1]} := \chi_{[0,1]}\phi_n.$$

This approach destroys a couple of properties. In fact, $\phi_n^{[0,1]}$ is not an orthonormal basis for $L^2([0, 1])$. But it is not hard to see, that it is still a tight frame with frame bounds $A = B = 1$. We have that

$$\left\langle f, \psi^{[0,1]}_{j,m}\right\rangle_{L^2([0,1])} = \left\langle \tilde{f}, \psi_{j,m}\right\rangle_{L^2(\mathbb{R})}$$

for an $\tilde{f}$ which equals f on $[0, 1]$ and 0 on $(-\infty, 0) \cup (1, \infty)$. Hence, if f is smooth but $\tilde{f}$ is not, then $\langle f, \psi^{[0,1]}_{j,m}\rangle_{L^2([0,1])}$ behaves as if f was not smooth.

A second approach is to periodise the wavelet elements by setting for all $m \leq 2^j$

$$\psi^{[0,1]}_{j,m} := \sum_{k\in\mathbb{Z}} \psi_{j,m}(\cdot - k).$$

This construction yields an orthonormal basis with the same boundary behaviour as the truncated wavelet system if the periodisation of f is not smooth.

The last approach, which, in fact, overcomes the problem of artificial singularities at the boundaries is to adapt the boundary elements, i.e., the wavelets the support of which intersects $\{0\}$ and $\{1\}$, in a way such that they still have vanishing moments. This is done by constructing an MRA for $L^2([0, 1])$ by setting V_0 to be the span of all shifted scaling functions fully supported in $(0, 1)$ and all polynomials up to a certain order L. The resulting wavelet spaces are then orthogonal on these polynomials, hence have vanishing moments. We shall not provide the details of this construction, and refer to [10]. Some boundary adapted scaling functions described above are depicted in Figure 4.4.

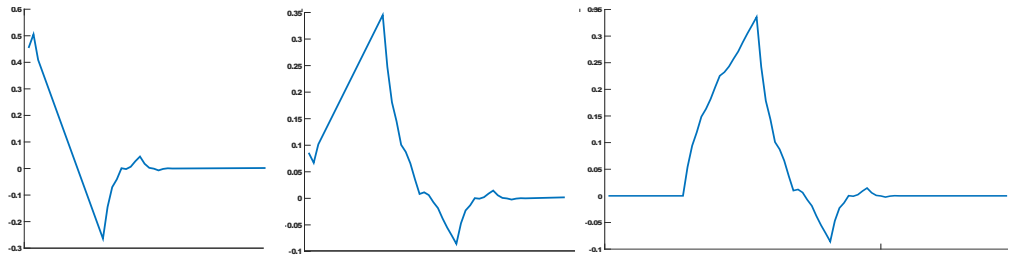

Figure 4.4: Three elements of a boundary adapted wavelet system. The right-most image shows a scaling function which was not adapted to the boundary. In the middle, already minor adaptation is visible. The leftmost scaling function was strongly adapted to make the reproduction of polynomials possible.

4.4 Sobolev spaces and approximation theory

4.4.1 Characterisation of smoothness classes

Consider a wavelet $\psi \in L^2(\mathbb{R})$ such that $\mathcal{W}(\psi, 2, 1)$ is a frame for $L^2(\mathbb{R})$. Then

$$|f|^2_{H^1(\mathbb{R})} = \|f'\|^2_{L^2} \sim \sum_{j,m\in\mathbb{Z}} |\langle f', \psi_{j,m}\rangle|^2 = \sum_{j,m\in\mathbb{Z}} 2^{2j} |\langle f, (\psi')_{j,m}\rangle|^2.$$

In the situation above, we can characterise the Sobolev semi-norm by a weighted ℓ^2 norm of wavelet coefficients with respect to $\mathcal{W}(\psi', 2, 1)$. Another example, is given by considering the Shannon wavelet $\psi := \mathcal{F}^{-1}(\chi_{[-1,-1/2]\cup[1/2,1]})$ with the associated scaling function $\phi := \mathcal{F}^{-1}(\chi_{[-1/2,1/2]})$. We have that

$$\begin{aligned}
\|f\|^2_{H^s(\mathbb{R})} &\sim \int_{\mathbb{R}} (1+|\xi|^2)^s |\hat{f}(\xi)|^2 d\xi \\
&= \int_{-1/2}^{1/2} (1+|\xi|^2)^s |\hat{f}(\xi)|^2 d\xi + \sum_{j=0}^{\infty} \int_{[-2^j,-2^{j-1}]\cup[-2^{j-1},-2^j]} (1+|\xi|^2)^s |\hat{f}(\xi)|^2 d\xi \\
&\sim \int_{-1/2}^{1/2} |\phi(\xi)|^2 |\hat{f}(\xi)|^2 d\xi + \sum_{j=0}^{\infty} \int_{-2^j}^{2^j} 2^{2sj} |\psi(2^{-j}\cdot)|^2 |\hat{f}(\xi)|^2 d\xi \\
&= \sum_{m\in\mathbb{Z}} |\langle \phi_{0,m}, f\rangle|^2 + \sum_{j=0}^{\infty} \sum_{m\in\mathbb{Z}} 2^{2sj} |\langle \psi_{j,m}, f\rangle|^2,
\end{aligned}$$

where the last estimate follows from the fact, that $(2^{-\frac{j}{2}} e^{-2\pi i \xi m/2^j})_{m\in\mathbb{Z}}$ is an orthonormal basis for $L^2([-2^j, 2^j])$ and the Parseval and Plancherel identities.

In fact, a characterisation of norms describing smoothness by wavelets is possible in much more generality. Indeed, we have the following theorem, the proof of which can be found in [9, Theorem 3.7.7]. We denote from now on the Besov spaces

$$B^s_{p,q} := \left\{ f \in \mathcal{S}'(\mathbb{R}) : \left\| \left(\left\| 2^{sj} \mathcal{F}^{-1} \gamma_j \mathcal{F} f \right\|_{L^p} \right)_{j\geq 0} \right\|_{\ell_q} \right\},$$

where $(\gamma_j)_{j\in\mathbb{Z}}$ is a smooth partition of unity function such that supp $\gamma_j \subset [-2^{j+1}, -2^{j-1}] \cup [2^{j-1}, 2^{j+1}]$ if $j \geq 1$ and supp $\gamma_0 \subset [-2, 2]$. This definition is taken from [37], but there exist plenty of alternative definitions of Besov spaces, see e.g. [7].

Theorem 4.4.1 ([7]). *Assume that $\psi, \phi \in L^r$ for some $r \in [1, \infty]$ such that*

$$\{T_m D_{2^j} \psi : j \geq 0, m \in \mathbb{Z}\} \cup \{T_m \phi : m \in \mathbb{Z}\}$$

forms an orthonormal basis. Let $0 < p \leq r$ and let $n \in \mathbb{N}$ be the number of vanishing moments of ψ. Additionally, let s, $q_0 > 0$ be such that $\psi \in B^s_{p,q_0}$. Then we have for all $t > 0$ such that $1/p - 1/r < t < \min\{s, n\}$ that for all $q > 1$ that

$$\|f\|_{B^t_{p,q}} \sim \left(\sum_{j \in \mathbb{Z}} \left(2^{tj} 2^{(\frac{1}{2} - \frac{1}{p})j} \left(\sum_{m \in \mathbb{Z}} |\langle f, \psi_{j,m} \rangle|^p \right)^{1/p} \right)^q \right)^{1/q}. \tag{4.4.1}$$

A similar result holds for higher dimensions $d \in \mathbb{N}$, if all terms of the form $1/p - 1/r$ or $(1/p - 1/2)$ are replaced by $d(1/p - 1/r)$ or $d(1/p - 1/2)$. Moreover, the result can be generalised to bounded domains, and even to wavelets that do not form orthonormal bases, but only so-called bi-orthogonal bases.

Two features of wavelet coefficients lead to high Besov regularity, i.e., a high t in (4.4.1). On the one hand, summability in ℓ^p for a sufficiently small $p > 0$, and, on the other hand, weighted summability in j. These two types of decay are associated to two different types of approximation as we shall see in the following two subsections.

4.4.2 Linear approximation and preconditioning

Setting $p = q = 2$ and taking $t \in \mathbb{N}$ we have $B^t_{2,2}(\mathbb{R}) = H^t(\mathbb{R})$ with equivalent norms and hence we get the reproduction of Sobolev norms. This norm equivalence leads to two interesting conclusions. It enables us to estimate the error of approximating a function f by a sum of wavelets, if we know the smoothness of f and it allows to find suitable preconditioning matrices if we choose to discretise a differential equation by wavelets.

Let ψ, ϕ be such that (4.4.1) is satisfied for $p = q = 2, t \in \mathbb{N}$. Assume $f \in H^t(\mathbb{R})$ with $\|f\|_{H^t} = 1$. We define

$$f_J := \sum_{m \in \mathbb{Z}} \langle f, \phi_m \rangle \phi_m + \sum_{0 \leq j \leq J, m \in \mathbb{Z}} \langle f, \psi_{j,m} \rangle \psi_{j,m}.$$

Then we have that

$$\|f - f_J\|^2_{L^2} = \sum_{j,m \in \mathbb{Z}} |\langle f - f_N, \psi_{j,m} \rangle|^2 + \sum_{m \in \mathbb{Z}} |\langle f - f_N, \phi_m \rangle|^2.$$

Per construction of f_N and by the orthogonality of the system, we get with (4.4.1) that

$$\|f - f_J\|^2_{L^2} = \sum_{j > J, m \in \mathbb{Z}} |\langle f, \psi_{j,m} \rangle|^2 \lesssim 2^{-2Jt} \|f\|^2_{H^t} = 2^{-2Jt}.$$

It is clear that this rate cannot be improved if uniform approximation for all f with $\|f\|_{H^t} \leq 1$ is requested. The statement above can be extended to more general approximation spaces by defining for an MRA $(V_j)_{j \in \mathbb{Z}}$ $t > 0$ and $p, q > 1$

$$A^t_{p,q} := \left\{ f \in L^p : \left(2^{sj} \mathrm{dist}_{L^p}(f, V_j)\right)_{j \geq 0} \in \ell^q \right\}. \tag{4.4.2}$$

We call $A^t_{p,q}$ the linear approximation spaces of $(V_j)_{j \in \mathbb{Z}}$. Then, it turns out that under the assumptions of (4.4.1): $A^t_{p,q} = B^t_{p,q}$ with equivalent norms. See [9] for the details.

Next, we assume that we are given a differential equation

$$Lu = f,$$

where L is a bounded, invertible differential operator from $H^t(\mathbb{R}) \to L^2(\mathbb{R})$. Then, we have a weak formulation

$$a(u,v) := \langle Lu, v\rangle = \langle f, v\rangle \text{ for all } v \in L^2(\mathbb{R}).$$

Assume that $a(\cdot,\cdot)$ is a bounded sesquilinear form such that $a(v,v) \sim \|v\|_{H^t}^2$. Using the analysis and synthesis operators of a wavelet orthonormal basis, we can now rewrite the differential equation as

$$TLT^*c = Tf.$$

We have that (formally) for $c \in \ell^2$

$$\langle TLT^*c, c\rangle = \langle LT^*c, T^*c\rangle \sim \|T^*c\|_{H^t}^2 \sim \left\|\left(2^{jt}c_{j,m}\right)_{j,m}\right\|_{\ell^2}^2 = \left\langle P^{2t}c, c\right\rangle,$$

where P is the diagonal matrix, that maps $(c_{j,m})_{j\geq 0, m\in\mathbb{Z}}$ to $(2^j c_{j,m})_{j\geq 0, m\in\mathbb{Z}}$. In other words, TLT^* is an unbounded operator on ℓ^2 but $P^{-t}TLT^*P^{-t}$ is bounded from above and below. We have transformed the differential equation into a well-conditioned discrete linear problem. Using the approximation results we know that if the solution u is smooth, then solving the truncated linear system, up to a maximum scale J yields a reconstruction of the solution approximating u very well.

We shall see, that this is not the most efficient way to solve the differential equation above, though.

4.4.3 Non-linear approximation

The approximation above is based on linear approximation. This approximation is easy to find, but we will see below, that it is usually far worse than a non-linear approximation. For a function $f \in \mathcal{H}$ and a sequence $(\phi_n)_{n\in\mathbb{N}} \subset \mathcal{H}$, we define the *best N-term approximation error of f with respect to $(\phi_n)_{n\in\mathbb{N}}$* by

$$\sigma_N(f) := \inf_{c\in\ell_2 : \|c\|_0 = N} \left\| f - \sum_{n\in\mathbb{N}} c_n\phi_n \right\|_{\mathcal{H}}.$$

Here $\|c\|_0$ denotes the number of non-zero entries of c.

We consider the following example. Let ψ, ϕ have compact support and be such that (4.4.2) is satisfied for $p = q = 2, t = 1/2$. Let $f = \chi_{[0,1]} \notin H^{\frac{1}{2}}(\mathbb{R})$. Therefore, it is not hard to see with (4.4.2) that

$$\|f - f_J\|_{L^2} \gtrsim 2^{-\frac{J}{2}-\delta},$$

for all $\delta > 0$. To form f_J we need $O(2^J)$ many wavelet elements, if we only consider those which are supported in a compact set K.

On the other hand, for every j there only exist $O(1)$ elements $\psi_{j,m}$ the support of which contains $\{0\}$ or $\{1\}$. Hence, $\langle f, \psi_{j,m}\rangle = 0$ for all but $c > 0$ many $m \in \mathbb{Z}$, for every $j \geq 0$. Moreover, $|\langle f, \psi_{j,m}\rangle| \lesssim 2^{j/2}|\operatorname{supp}\ \psi_{j,m}| = 2^{-j/2}$.

Let Λ_j contain the indices of all nonzero coefficients $(\langle f, \psi_{j,m}\rangle)_{m\in\mathbb{Z}}$ and define for $N = cJ$

$$f_{(N)} = \sum_{j=1}^{J} \sum_{m\in\Lambda_j} \langle f, \psi_{j,m}\rangle \psi_{j,m} + \sum_{m\in\Lambda_0} \langle f, \phi_m\rangle \phi_m.$$

It is straight-forward to compute that

$$\|f - f_{(N)}\|_{L^2}^2 \lesssim 2^{-J} = 2^{-N/c}.$$

As a consequence, $\sigma_N(f) \lesssim 2^{-N/c}$. We see that, the linear approximation method required N elements to achieve an approximation error of $O(N^{\frac{1}{2}})$, while the non-linear approximation rate achieves an exponential error decay.

The reason this example works is, of course, that, as (4.4.1) reveals, f has a much higher smoothness in a Besov scale as in a Sobolev scale. Indeed, whenever a function has significantly higher Besov regularity than Sobolev regularity, then a best N-term approximation will be significantly better than a linear approximation. This is substantiated in the following theorem proved in [7, Theorem 4.2.2].

Theorem 4.4.2 ([7]). *Assume that ψ, ϕ, p and $t = 1/p - 1/2$ are such that (4.4.1) holds. Then, for all $f \in B^t_{p,p}$: $\sigma_N(f) \lesssim N^{-t}$.*

Proof. We invoke Stechkin's Lemma, see e.g. [14]. It states that if we denote by $I_N(c)$ the indices of the N largest coefficients in modulus of a sequence $(c_i)_{i\in I}$, then

$$\left(\sum_{I \setminus I_N(c)} |c_i|^2 \right)^{\frac{1}{2}} \leq N^{-t} \|c\|_{\ell^p}, \tag{4.4.3}$$

for $t = \frac{1}{p} - \frac{1}{2}$.

$f \in B^t_{p,p}$ and $t = 1/p - 1/2$ in (4.4.1) shows that $(\langle f, \psi_{j,m}\rangle)_{j,m} \in \ell^p$. Setting $c = (\langle f, \psi_{j,m}\rangle)_{j,m}$ in (4.4.3) yields the result. □

As Stechkin's lemma admits a converse for $t = 1/p - 1/2 - \delta$ for all $\delta > 0$ one can also obtain a converse of Theorem (4.4.2). Additionally, the theorem holds for higher-dimensional wavelet systems after replacing $t = 1/p - 1/2$ by $t = d(1/p - 1/2)$ and N^{-t} by $N^{-t/d}$, where d is the dimension of the space.

4.4.4 Adaptive solution of operator equations

If the solution of a PDE has higher regularity in a Besov scale than in a Sobolev scale, it could make sense to build the associated Galerkin spaces using only those coefficients that correspond to the best N-term approximation. Of course these coefficients are unknown as they depend on the solution of the PDE. Surprisingly, it is possible to construct an algorithm that choses the Galerkin spaces *adaptively*. Moreover, the number of computational operations of this algorithm to produce an approximation to the solution u of error ϵ is asymptotically equivalent to the smallest number N so that $\sigma_N(u) \leq \epsilon$.

This truely impressive result is due to Cohen, Dahmen, and DeVore [8]. Its proof is based on the approximation results, the preconditioning and the fact that some differential operators are almost diagonal when discretised with respect to wavelet bases. We shall not introduce this property of almost diagonality of differential operators here, but refer to this mysterious property from now on by saying that L is *compressible with respect to a wavelet basis*.

Theorem 4.4.3 ([8]). *Let Ω be a bounded domain, ψ, ϕ generators of an orthonormal wavelet basis, $L : H^t_0(\Omega) \to H^{-t}(\Omega)$ be compressible with respect to that basis, bounded, invertible and elliptic. Finally, assume that $f \in H^{-t}(\Omega)$ and*

$$Lu = f$$

*where u is such that $\sigma_N(u) = O(N^{-s})$. Then there exists an algorithm, depending on (ψ, ϕ, L, s, t), called **SOLVE**, such that $c_\epsilon = \textbf{SOLVE}[\epsilon, L, f]$ and*

- $\|u - T^* c_\epsilon\| = O(\|c_\epsilon\|_0^{-s})$;
- *Only $O(\|c_\epsilon\|_0)$ flops are required to compute c_ϵ.*

Chapter 5

Directional systems

We saw in the previous section that wavelets are very efficient in approximating one-dimensional piece-wise constant functions with a finite number of jumps. It turns out that in higher dimensions, this is not correct anymore. In fact, consider a non-empty set $B \subset (0,1)^2$ with a smooth boundary. Then χ_B is piece-wise constant, and $\widehat{\chi_B} \notin L^1(\mathbb{R}^2)$. One can show by using the projection-slice theorem that there even is a $\theta \in \mathbb{R}^2$ such that

$$(t \mapsto \widehat{\chi_B}(t\theta)) \notin L^1(\mathbb{R}).$$

Therefore, for any partition of unity $(\gamma_j)_{j\in\mathbb{Z}}$ as in the definition of a Besov space, we have that $\|\gamma_j \widehat{\chi}_B\|_\infty \lesssim 2^{-j}$ does not hold for $j \to \infty$. Hence, per definition is follows that

$$\chi_B \notin B^1_{1,1}(\mathbb{R}^2).$$

Thus, invoking a converse of Theorem 4.4.2 with $d = 2$ we get that $\sigma_N(\chi_B) \lesssim N^{-\frac{1}{2}-\epsilon}$ does not hold for any $\epsilon > 0$.

We see that we do not have exponentially fast decay of the N-term approximation error for functions with distributed singularities in higher dimensions. Nonetheless, it could be, that $N^{-\frac{1}{2}}$ was, in fact, the best N-term approximation rate one could hope for, for functions of this sort. This is not the case, as we will demonstrate in the sequel. To make these statements more precise, we first give a more precise definition of a function class of piece-wise smooth functions.

Definition 5.0.1. *Let $\nu > 0$. The class of* cartoon-like functions $\mathcal{E}^2(\mathbb{R}^2, \nu)$ *is defined as the set of functions $f : \mathbb{R}^2 \to \mathbb{R}$ of the form $f = f_0 + \chi_B f_1$. Here, we assume that $B \subset (0,1)^2$ where $\partial B \in C^2$ and the curvature of ∂B is bounded by ν. Moreover, $f_i \in C^2(\mathbb{R}^2)$ with $\|f_i\|_{C^2} \leq 1$ and* supp $f_i \subseteq (0,1)^2$ *for $i = 0, 1$.*

We have the following lower bound on the worst case N-term approximation error of this function class by any representation system.

Theorem 5.0.2 ([28, 16])**.** *Let $\Psi = (\psi_\lambda)_{\lambda\in\Lambda} \subset L^2(\mathbb{R}^2)$. Then, we have that*

$$\sup_{f\in\mathcal{E}^2(\mathbb{R}^2,\nu)} \sigma_N(f, \Psi) \gtrsim N^{-1},$$

where $\sigma_N(f, \Psi)$ denotes the best N-term approximation error of f with respect to Ψ.

Remark 5.0.3.

- *The theorem above only holds under an additional assumption on the type of best N-term approximation. In fact, this lower bound requires an N-term approximation to be constructed under the restriction of polynomial depth search. In other words, for a fixed polynomial p, only elements of Ψ where $|\lambda| \leq p(N)$ are allowed to be used for the N-term approximation.*

- *The lower bound of Theorem 5.0.2 already holds fir the smaller function class of piecewise constant cartoon-like functions, i.e., when $f_0 = 0$ and $f_1 = 1$ in Definition 5.0.1.*
- *It appears to be a bit arbitrary to focus on C^2 regularity in the definition of cartoon-like functions and the optimality result. In fact, generalisations exist to piecewise C^k functions with C^k regular boundaries. The lower bound on the N-term approximation rate is then $N^{-k/2}$, see e.g. [5].*

Considering that there is a considerable gap between the approximation by wavelets and the lower bound of Theorem 5.0.2 we shall be interested in finding an alternative representation system that performs better when curve-like singularities are present.

5.1 Shearlets

To understand how to construct a system that offers optimal approximation of piece-wise constant or smooth functions we observe in Figure 5.1 why the isotropic scaling of wavelets is suboptimal to capture singularities along lower dimensional manifolds.

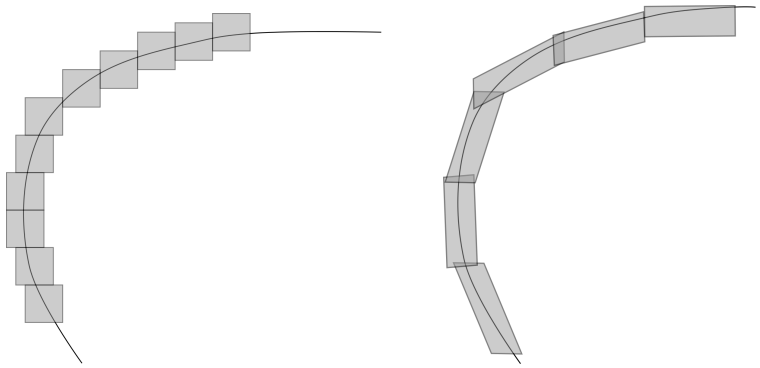

Figure 5.1: **Left:** Isotropically shaped squares overlapping a curve. **Right:** The same curve covered by anisotropically shaped and rotated rectangles.

It appears to be worthwhile to study function systems with different types of localisation and a method to also rotate the elements. The first system of this type was the *curvelet system*, [4]. Which is a generalisation of 2d wavelets, with an anisotropic scaling matrix and rotated elements. It is very similar to the shearlet systems that we shall introduce in the following section and hence, we only make appropriate comments there, as to where the differences lie.

5.1.1 Continuous shearlet transform

Returning to the standard procedure that we have already observed for the short-time Fourier transform and the Gabor systems, or the wavelet transform and the wavelet systems, we first introduce a continuous transform and then demonstrate how this leads to discrete systems.

For $a \in \mathbb{R}^+$, $s \in \mathbb{R}$, we denote the *anisotropic scaling matrix* A_a and the *shearing matrix* S_s by

$$A_a := \begin{pmatrix} a & 0 \\ 0 & a^{\frac{1}{2}} \end{pmatrix}, \text{ and } S_s := \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}.$$

Based on these two matrices, we can now define the shearlet transform.

Definition 5.1.1. *Let $\psi \in L^2(\mathbb{R}^2)$, then we define*

$$\begin{aligned} \mathcal{SH} : L^2(\mathbb{R}^2) &\to L^\infty(\mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^2) \\ f &\mapsto ((a,s,t) \mapsto \langle f, \psi_{a,s,t} \rangle), \end{aligned}$$

where

$$\psi_{a,s,t}(x) := a^{-\frac{3}{4}}\psi(A_a^{-1}S_s(x-t)).$$

We call $\mathcal{SH}$ *the* shearlet transform*.*

The curvelet transform is set up very similarly, but by replacing the shearing matrix S_s by a rotation matrix.

If ψ has compact support, then the anisotropy in the scaling matrix gives rise to shearlet elements $\psi_{a,s,t}$ the support of which obeys the scaling law

$$width = length^2.$$

For $a \to 0$ the elements the width of the elements is essentially of the order of a and the length of the order of $\sqrt{a}$. Now we basically have three questions about the shearlet transform: Is $\mathcal{SH}$ an isometry between $L^2(\mathbb{R}^2)$ and $L^2(\mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^2, \lambda)$ for a measure λ? Do we have an inversion formula? Can we extract any interesting features of f from $\mathcal{SH}(f)$? As in the wavelet case we first make an assumption on the type of generating functions that make the transform into an isometry.

Definition 5.1.2. *Let* $\psi \in L^2(\mathbb{R}^2)$*. We call* ψ *an* admissible shearlet*, if*

$$C_\psi := \int_{\mathbb{R}^+} \frac{|\widehat{\psi}(\xi)|^2}{|\xi_1|^2} d\xi < \infty$$

and $C_\psi \neq 0$*.*

For any admissible shearlet, we have that the shearlet transform is an isometry up to a multiplication by C_ψ.

Theorem 5.1.3 ([19])**.** *Let* ψ *be an admissible shearlet, then for all* $f \in L^2(\mathbb{R}^2)$*:*

$$\|f\|_{L^2}^2 = \frac{1}{C_\psi} \int_{\mathbb{R}^+} \int_{\mathbb{R}} \int_{\mathbb{R}^2} |\mathcal{SH}(f)(a,s,t)|^2 dt ds \frac{da}{a^3}.$$

Proof. We have that

$$\begin{aligned}
\int_{\mathbb{R}^+} \int_{\mathbb{R}} \int_{\mathbb{R}^2} |\mathcal{SH}(f)(a,s,t)|^2 dt ds \frac{da}{a^3} &= \int_{\mathbb{R}^+} \int_{\mathbb{R}} \int_{\mathbb{R}^2} |\langle f, \psi_{a,s,t}\rangle|^2 dt ds \frac{da}{a^3} \\
&= \int_{\mathbb{R}^+} \int_{\mathbb{R}} \int_{\mathbb{R}^2} \left|\left\langle \hat{f}, \widehat{\psi}_{a,s,t} \right\rangle\right|^2 dt ds \frac{da}{a^3} \\
&= \int_{\mathbb{R}^+} \int_{\mathbb{R}} \int_{\mathbb{R}^2} \left| \int_{\mathbb{R}^2} \hat{f}(\xi) a^{\frac{3}{4}} \widehat{\psi}\left(A_a S_s^T \xi\right) e^{-2\pi i \langle \xi, t\rangle} d\xi \right|^2 dt ds \frac{da}{a^3} \\
&= \int_{\mathbb{R}^+} \int_{\mathbb{R}} \int_{\mathbb{R}^2} \left|\hat{f}(\xi)\right|^2 \left| a^{\frac{3}{4}} \widehat{\psi}\left(A_a S_s^T \xi\right)\right|^2 d\xi ds \frac{da}{a^3}.
\end{aligned}$$

where the last line follows from Parsevals theorem. Carelessly using Fubini, leads us to

$$\int_{\mathbb{R}^+} \int_{\mathbb{R}} \int_{\mathbb{R}^2} |\mathcal{SH}(f)(a,s,t)|^2 dt ds \frac{da}{a^3} = \int_{\mathbb{R}^2} |\hat{f}(\xi)|^2 \left(\int_{\mathbb{R}^+} \int_{\mathbb{R}} \left|\widehat{\psi}(A_a S_s^T \xi)\right|^2 ds \frac{da}{a^{\frac{3}{2}}} \right) d\xi.$$

Showing that $\int_{\mathbb{R}^+} \int_{\mathbb{R}} |\widehat{\psi}(A_a S_s^T \xi)|^2 a^{-\frac{3}{2}} ds da = C_\psi$ is an exercise. □

Additionally, we have a reconstruction formula.

Theorem 5.1.4 ([19]). *Let ψ be an admissible shearlet, then for all $f \in L^2(\mathbb{R}^2)$:*

$$f = \frac{1}{C_\psi} \int_{\mathbb{R}^+} \int_{\mathbb{R}} \int_{\mathbb{R}^2} \mathcal{SH}(f)(a,s,t)\psi_{a,s,t} dt ds \frac{da}{a^3}.$$

Finally, we want to understand what kind of information on f we can extract from $\mathcal{SH}(f)$. For the wavelet transform, we could extract the local smoothness of a function from the decay. The Gabor and Fourier transform also allowed analysis of the smoothness (but not locally) by analysing the decay. As shearlets have an additional directional component, the shearlet transform gives an even more precise account about the local smoothness than the wavelet transform. To make this more precise we need to recall the definition of the wavefront set.

Definition 5.1.5. *Let $f \in L^2(\mathbb{R}^2)$. We say that $(x,\lambda) \in \mathbb{R}^2 \times \mathbb{R}$ is a* regular directed point *of f if there exists a neighborhood U_x of x, a smooth function ϕ with* supp $\phi \in U_x$, *$\phi(x) = 1$ on a neighborhood of x, and a neighborhood V_λ of λ such that*

$$\mathcal{F}(\phi f)(\xi) \text{ decays rapidly for } |\xi| \to \infty, \text{ if } \frac{\xi_2}{\xi_1} \in V_\lambda.$$

We denote the set of regular directed points of f by $\mathcal{R}(f)$ and we define the wavefront set *of f by $\mathcal{WF}(f) = (\mathbb{R}^2 \times \mathbb{R}) \setminus \mathcal{R}(f)$.*

We have the following characterisation of the wavefront set.

Theorem 5.1.6 ([19]). *Let ψ be an admissible shearlet, such that $\widehat{\psi} \in C^\infty(\mathbb{R}^2)$, and*

$$|\widehat{\psi}(\xi)| \lesssim \frac{\min\{|\xi_1|^L, 1\}}{(1+|\xi|)^M}, \text{ and } |\psi(x)| \lesssim \frac{1}{(1+|x|)^K} \text{ for all } x, \xi \in \mathbb{R}^2, \text{ and}$$

for all $K, L, M \in \mathbb{N}$ (The implicit constant is allowed to depend on K, M, L). If $f \in L^2(\mathbb{R}^2)$ and (x_0, λ_0) is a regular directed point of f, then there exist neighborhoods of x_0 and λ_0 called U_0 and V_0 such that for all $P \in \mathbb{N}$

$$|\mathcal{SH}(f)(a,s,t)| \lesssim a^P, \text{ for all } (s,t) \in U_0 \times V_0.$$

Additionally, if for $(x_1, \lambda_1) \in \mathbb{R}^2 \times \mathbb{R}$ there exist neighborhoods U_1 and V_1 such that for all $P \in \mathbb{N}$

$$|\mathcal{SH}(f)(a,s,t)| \lesssim a^P, \text{ for all } (s,t) \in U_1 \times V_1,$$

then $(x_1, \lambda_1) \in \mathcal{R}(f)$.

The result above collects a couple of results from [19], in a very weak form. In fact, the argument also works for much less restrictive assumptions on ψ. However, then the concept of wavefront set needs to be replaced by that of N-wavefront set, where one does not require rapid decay of the Fourier transform of the truncated function but polynomial decay of order N.

Sketch of Proof of Theorem 5.1.6. Let (x_0, λ_0) be a regular directional point of f. If supp $\psi = [-1,1]^2$ then for $a < 1$

$$\text{supp } \psi_{a,s,t} = t + S_{-s}\left([-a,a] \times [-\sqrt{a}, \sqrt{a}]\right) \subset B_{(1+|s|)\sqrt{a}}(t).$$

Hence, for any $\phi \in C^\infty$ such that $\phi = 1$ on a neighborhood of t we have that for sufficiently small $a > 0$

$$\langle f, \psi_{a,s,t} \rangle = \langle \phi f, \psi_{a,s,t} \rangle.$$

Applying Plancherel's identity yields

$$\langle f, \psi_{a,s,t} \rangle = \langle \mathcal{F}(\phi f), \widehat{\psi}_{a,s,t} \rangle. \tag{5.1.1}$$

Now we make the bold assumption that supp $\hat{\psi} \subset ([-c_2, -c_1] \cup [c_1, c_2]) \times [-c_2, c_2]$ for constants $c_1, c_2 > 0$. Of course, due to the uncertainty principle, this assumption is never fulfilled, but by the decay assumptions on $\widehat{\psi}$ we see that it is almost satisfied.

Now we compute:

$$\text{supp } \widehat{\psi}_{a,s,t} = \text{supp } \widehat{\psi}_{a,s,0} = \left\{ \xi \in \mathbb{R}^2 : A_a S_{-k}^T \xi \in ([-c_2, -c_1] \cup [c_1, c_2]) \times [-c_2, c_2] \right\}.$$

In other words,

$$\xi \in \text{supp } \widehat{\psi}_{a,s,t} \Leftrightarrow |\xi_1| \in a^{-1}[c_1, c_2], \text{ and } |\xi_2 - k\xi_1| \leq a^{-\frac{1}{2}} c_2.$$

By the reverse triangle inequality we conclude that $||\xi_2| - |k\xi_1|| \leq a^{-\frac{1}{2}} c_2$ and hence

$$\xi \in \text{supp } \widehat{\psi}_{a,s,t} \implies \frac{|\xi_2|}{|\xi_1|} \in \left[\frac{|k\xi_1| - a^{-\frac{1}{2}} c_2}{|\xi_1|}, \frac{|k\xi_1| + a^{-\frac{1}{2}} c_2}{|\xi_1|} \right] \subset k + \sqrt{a} \left[-\frac{c_2}{c_1}, \frac{c_2}{c_1} \right].$$

Hence, per definition of the wavefront set we have that for $\xi \in \text{supp } \widehat{\psi}_{a,s,t}$ that for any $P \in \mathbb{N}$

$$|\mathcal{F}(\phi f)(\xi)| \lesssim (1 + |\xi|)^{-P} \lesssim a^P.$$

Plugging this estimate into (5.1.1) yields the result. In reality, we do not have compact supports but only fast decay of the shearlet and its Fourier transform, but this does not introduce large errors.

The converse of the theorem is more technical and based on the reconstruction formula. □

Figure 5.2 demonstrates the behavior of some parts of the shearlet transform of a characteristic function with smooth boundary curve.

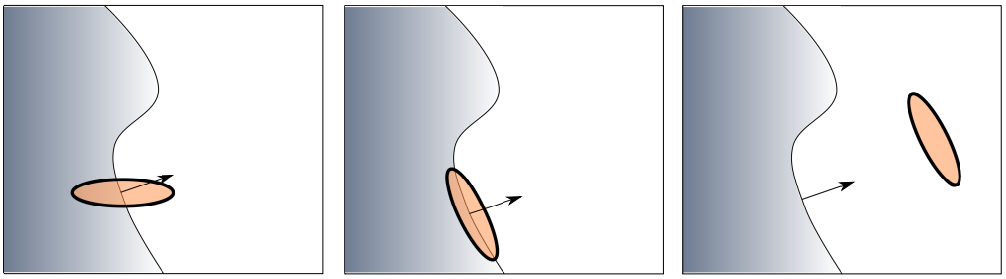

Figure 5.2: Depiction of the different scenarios of a shearlet interacting with a singularity. The wavefront set of a function χ_B where $B \subset \mathbb{R}^2$ with $\partial B \in C^\infty$ is precisely the set of all (x, λ) such that $x \in \partial B$ and $\lambda = \vec{n}_2/\vec{n}_1$, where $\vec{n}$ is a normal at ∂B in x. In the scenario on the left, the shearing parameter is not corresponding to the normal direction, hence the associated shearlet coefficients will decay rapidly for a to 0. The shearlet in the center has a matching shearing parameter and the coefficients will decay slowly with a to 0. In the rightmost image, the shearlet is not intersecting the singularity. Again the decay will be very fast for a to 0.

5.1.2 Discrete shearlet transform

We give the definition of a discrete shearlet system first and then discuss to what extend this is a reasonable discretisation of the continuous transform.

Definition 5.1.7. *Let $\phi, \psi, \widetilde{\psi} \in L^2(\mathbb{R}^2)$ and $c > 0$. Then the* cone-adapted shearlet system *is defined by* $\mathcal{SH}(\phi, \psi, \widetilde{\psi}, c) := \Phi \cup \Psi \cup \widetilde{\Psi}$*, where*

$$\begin{aligned}
\Phi &:= \left\{\phi(\cdot - m) : m \in c\mathbb{Z}^2\right\}, \\
\Psi &:= \left\{\psi_{j,k,m} := 2^{\frac{3j}{4}}\psi(S_k A_{2^j} \cdot - cm) : j \in \mathbb{N}, |k| \leq \left\lceil 2^{\frac{j}{2}} \right\rceil, m \in c\mathbb{Z}^2\right\}, \\
\widetilde{\Psi} &:= \left\{\widetilde{\psi}_{j,k,m} := 2^{\frac{3j}{4}}\widetilde{\psi}(\widetilde{S}_k \widetilde{A}_{2^j} \cdot - m) : j \in \mathbb{N}, |k| \leq \left\lceil 2^{\frac{j}{2}} \right\rceil, m \in c\mathbb{Z}^2\right\}.
\end{aligned}$$

We notice, that we have not directly discretised the continuous transform, but introduced two systems, where the shearing parameter k is now bounded. This is done, because in practice having unbounded shearing can lead to very long elements and an unequal treatment of different directions. Additionally, a low-frequency part called Φ was introduced, which plays a role similar to the space spanned by scaling functions in the construction of a wavelet basis.

Under certain conditions on the sufficiently smoothness of $\phi, \psi, \widetilde{\psi}$ and the number of vanishing moments of $\psi, \widetilde{\psi}$ there exists a $c^* > 0$ such that for all $c \leq c^*$ we have that $\mathcal{SH}(\phi, \psi, \widetilde{\psi}, c)$ forms a frame for $L^2(\mathbb{R}^2)$, [26]. It is an open question if shearlet bases exist and even if there exist tight shearlet frames with compactly supported generators $\phi, \psi, \widetilde{\psi}$.

In any case, we can now establish the N-term approximation rate of a shearlet frame for cartoon-like functions in the following theorem. We shall only sketch the proof and simplify the statement. A detailed proof and theorem statement can be found in [28, Theorem 1.3].

Theorem 5.1.8 ([28])**.** *Let $\nu > 0$. There exist $\phi, \psi, \tilde{\psi} \in L^2(\mathbb{R})$ and $c > 0$ such that the associated shearlet system forms a frame $(\psi_\lambda)_{\lambda\in\Lambda}$ with dual frame $(\widetilde{\psi_\lambda})_{\lambda\in\Lambda}$ and for every $f \in \mathcal{E}^2(\mathbb{R}^2, \nu)$*

$$\|f - f_N\|_{L^2} \leq CN^{-1}\log(N)^{\frac{3}{2}} \text{ for } N \to \infty,$$

where

$$f_N = \sum_{\Lambda_N} \langle f, \psi_\lambda\rangle \widetilde{\psi_\lambda},$$

and Λ_N contains the indices corresponding to N largest coefficients of $(|\langle f, \psi_\lambda\rangle|)_\Lambda$.

Proof. We give a very rough scetch of the proof, which, while very imprecise, still captures the main essence of the argument.

First of all, considering the asymptotic behavior depicted in Figure 5.2 we see that, for sufficiently large j only those coefficients $|\langle f, \psi_\lambda\rangle|$ are large, where the shearlets intersect the singularity curve tangentially. One can estimate, for fixed j all the coefficients k, m, such that $\psi_{j,k,m}$ or $\widetilde{\psi}_{j,k,m}$ are in the geometric position described above by $2^{j/2}$. Additionally, we estimate

$$|\langle f, \psi_{j,k,m}\rangle| \leq 2^{\frac{3}{4}j}|\operatorname{supp}\, \psi_{j,k,m}| \lesssim 2^{-\frac{3}{4}j}.$$

Combining these estimates, we see that

$$\|(\langle f, \psi_\lambda\rangle)_{\lambda\in\Lambda}\|_{\ell^p}^p \lesssim \sum_{j\in\mathbb{N}} 2^{\frac{j}{2}} 2^{-p\frac{3}{4}j}.$$

Hence, as long as $p > 2/3$ we have that $(\langle f, \psi_\lambda\rangle)_{\lambda\in\Lambda} \in \ell^p$ for all $\epsilon > 0$. Since $(\widetilde{\psi_\lambda})_{\lambda\in\Lambda}$ is a frame, its synthesis operator is bounded. We conclude that

$$\|f - f_N\|_{L^2}^2 \lesssim \sum_{\Lambda\setminus\Lambda_N} |\langle f, \psi_\lambda\rangle|^2 \lesssim N^{-2t},$$

for $t = \frac{1}{p} - \frac{1}{2}$ by the Stechkin inequality (4.4.3). Since $p > 2/3$ arbitrary, this yields

$$\|f - f_N\|_{L^2}^2 \lesssim N^{-2+\epsilon}$$

for every $\epsilon > 0$. Turning the ϵ exponent into a log term requires a bit more work. □

5.1.3 Further developements

In addition to wavelets, curvelets, and shearlets there is an immense variety of representation systems with different advantages. Contourlets [15] form a system based on anisotropic scalings but with a filter-based implementation leading to fast implementations. Ridgelets [3] form extremely anisotropic function system where all elements remain of fixed length but decrease in width with increasing scale. Ridgelets have also successfully been used to discretise transport equations [21]. All these systems fall in the concept of α-molecules [20]. This framework describes systems with a certain time-frequency localisation and α-scaling, i.e., elements that obey the scaling law

$$width = length^{\alpha}.$$

Additionally, as mentioned earlier, higher-order regularity of boundary curves allows better approximation rates in principle. The surflet system [5] is based on local Taylor approximations to offer optimal approximation rates for functions with singularity surfaces in C^k for $k \in \mathbb{N}$. Another related approach is to replace the shearing operation with higher-order deformations is the basis of the bendlet and taylorlet transform [30, 17]. Finding a proper way to discretise these transforms and establishing frame properties of the discrete systems is still an open problem.

A representation system that has recently significantly gained in interest is that of neural networks, and in particular deep neural networks. Let $N_0, N_1, \ldots, N_L \in \mathbb{N}$, $\varrho : \mathbb{R} \to \mathbb{R}$, and $W_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$ for $\ell = 1, \ldots, L$. Then the function

$$x \mapsto W_L(\varrho(W_{L-1}\varrho(\ldots W_2(\varrho(W_1(x))))))$$

is called a neural network with L layers, activation function ϱ and architecture $(N_0, N_1, \ldots, N_L)$. The recent increased interest in these functions is based on the fact, that they form the computational architecture for modern machine learning techniques, called deep learning [29]. Neural networks form a very powerful parametrised system. First results showed that the system is universal, in the sense that every continuous function on a compact domain can be approximated arbitrarily well by a neural network [11, 24]. Measuring the approximation fidelity against the number of free parameters in a network gives an analogue to a best N-term approximation. Approximation rates of neural networks have been established for many function classes [1, 33, 34, 35, 38]. Some of these results are also based on approximation rates established for wavelets and shearlets, [2, 36].

Bibliography

[1] A. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory*, 39(3):930–945, 1993.

[2] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen. Optimal approximation with sparsely connected deep neural networks. *preprint arXiv:1705.01714*, 2018.

[3] E. J. Candes. *Ridgelets: Theory and applications*. ProQuest LLC, Ann Arbor, MI, 1998. Thesis (Ph.D.)–Stanford University.

[4] E. J. Candès and D. L. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Comm. Pure Appl. Math.*, 57(2):219–266, 2004.

[5] V. Chandrasekaran, M. Wakin, D. Baron, and R. G. Baraniuk. Compressing piecewise smooth multi-dimensional functions using surflets: Rate-distortion analysis. *Rice University ECE Technical Report*, 2004.

[6] O. Christensen. *An introduction to frames and Riesz bases*. Birkhäuser Boston, Inc., Boston, MA, 2003.

[7] A. Cohen. *Wavelet methods in numerical analysis*. North-Holland, Amsterdam, 2000.

[8] A. Cohen, W. Dahmen, I. Daubechies, and R. DeVore. Tree approximation and optimal encoding. *Appl. Comput. Harmon. Anal.*, 11(2):192–226, 2001.

[9] A. Cohen, W. Dahmen, and R. DeVore. Adaptive wavelet methods for elliptic operator equations: convergence rates. *Math. Comp.*, 70(233):27–75, 2001.

[10] A. Cohen, I. Daubechies, and P. Vial. Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.*, 1(1):54–81, 1993.

[11] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.*, 2(4):303–314, 1989.

[12] I. Daubechies. *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.

[13] I. Daubechies, A. Grossmann, and Y. Meyer. Painless nonorthogonal expansions. *Journal of Mathematical Physics*, 27(5):1271–1283, 1986.

[14] R. A. DeVore. Nonlinear approximation. In *Acta numerica*, pages 51–150. Cambridge Univ. Press, Cambridge, 1998.

[15] M. N. Do and M. Vetterli. Contourlets: a directional multiresolution image representation. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–I. IEEE, 2002.

[16] D. L. Donoho. Sparse components of images and optimal atomic decompositions. *Constr. Approx.*, 17(3):353–382, 2001.

[17] T. Fink. Higher order analysis of the geometry of singularities using the taylorlet transform. *arXiv preprint arXiv:1703.00303*, 2017.

[18] K. Gröchenig. *Foundations of time-frequency analysis*. Springer Science & Business Media, 2013.

[19] P. Grohs. Continuous shearlet frames and resolution of the wavefront set. *Monatsh. Math.*, 164(4):393–426, 2011.

[20] P. Grohs, S. Keiper, G. Kutyniok, and M. Schäfer. Alpha-molecules: Curvelets, shearlets, ridgelets, and beyond. In *Wavelets and Sparsity XV*, pages 885802–1 –885802–11. Proceedings of the SPIE, San Diego, CA, 2013.

[21] P. Grohs and A. Obermeier. Optimal adaptive ridgelet schemes for linear advection equations. *Appl. Comput. Harmon. Anal.*, 2015. in press.

[22] A. Haar. Zur Theorie der orthogonalen Funktionensysteme. *Math. Ann.*, 69(3):331–371, 1910.

[23] M. Hairer. A theory of regularity structures. *Invent. Math.*, 198(2):269–504, 2014.

[24] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[25] S. Jaffard, Y. Meyer, and R. D. Ryan. *Wavelets: tools for science and technology*, volume 69. Siam, 2001.

[26] P. Kittipoom, G. Kutyniok, and W.-Q. Lim. Construction of compactly supported shearlet frames. *Constr. Approx.*, 35(1):21–72, 2012.

[27] G. Kutyniok and D. Labate. *Shearlets: Multiscale analysis for multivariate data*. Springer Science & Business Media, 2012.

[28] G. Kutyniok and W.-Q. Lim. Compactly supported shearlets are optimally sparse. *J. Approx. Theory*, 163(11):1564–1589, 2011.

[29] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[30] C. Lessig, P. Petersen, and M. Schäfer. Bendlets: A second-order shearlet transform with bent elements. *Appl. Comput. Harmon. Anal.*, 2017.

[31] S. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, 2008.

[32] Y. Meyer and R. Coifman. *Wavelets: Calderón-Zygmund and multilinear operators*, volume 48. Cambridge University Press, 2000.

[33] H. Mhaskar. Neural networks for optimal approximation of smooth and analytic functions. *Neural Comput.*, 8(1):164–177, 1996.

[34] H. N. Mhaskar. Approximation properties of a multilayered feedforward artificial neural network. *Adv. Comput. Math.*, 1(1):61–80, Feb 1993.

[35] P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *arXiv:1709.05289*, 2018.

[36] U. Shaham, A. Cloninger, and R. R. Coifman. Provable approximation properties for deep neural networks. *Appl. Comput. Harmon. Anal.*, 2016.

[37] H. Triebel. *Theory of function spaces*. Birkhäuser/Springer Basel AG, Basel, 2010.

[38] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Netw.*, 94:103–114, 2017.